Федеральное государственное учреждение
«Федеральный исследовательский центр «Информатика и управление»
Российской академии наук»

на правах рукописи

ЗАБЕЖАЙЛО МИХАИЛ ИВАНОВИЧ

КОМБИНАТОРНЫЕ СРЕДСТВА ФОРМАЛИЗАЦИИ ЭМПИРИЧЕСКОЙ ИНДУКЦИИ

05.13.17 – теоретические основы информатики

Диссертация на соискание ученой степени
доктора физико-математических наук

Научный консультант:
д.т.н., профессор,
Заслуженный деятель науки РФ
В.К.Финн

Москва – 2016

ОГЛАВЛЕНИЕ

Введение	4
Глава 1 Общая характеристика области интеллектуального анализа данных (ИАД) и место в ней рассматриваемой проблематики восстановления эмпирических зависимостей средствами ДСМ-метода.	16
1.1. Общая характеристика области ДСМ-ИАД.	19
1.2. ИАД и интеллектуальные системы.	21
1.3. Некоторые примеры актуальных задач ИАД	25
1.4. Интеллектуальные системы и системы анализа данных в ИАД	32
1.5. Возможности и перспективы использования интеллектуальных систем в задачах компьютерного анализа данных.	39
Глава 2 ДСМ-ИАД: общее описание.	43
2.1. Задача о восстановлении эмпирических зависимостей из данных: от классов сходства к классам эквивалентности.	43
2.2. Эвристика Ф.Бэкона – Д.С.Милля.	60
2.3. Общая схема ДСМ-рассуждения.	64
Глава 3 Процедурная платформа ДСМ-ИАД.	70
3.1. Представления о сходстве в ДСМ-ИАД.	70
3.1.1. Общие представления о сходстве как формальной конструкции.	70
3.1.2. Примеры формализации сходства для различных типов данных в ДСМ-методе.	74
3.2. Структура средств ДСМ-ИАД: общее описание.	91
3.2.1. Синтаксическая платформа ДСМ-ИАД	91
3.2.2 Квази-Аксиоматические Теории как платформа для представления знаний в интеллектуальном анализе данных.	95
3.2.3. Общие свойства ДСМ-правил правдоподобного вывода.	105
3.3. Решающие предикаты, правила правдоподобного вывода и стратегии ДСМ-рассуждения.	110
3.3.1. Правила правдоподобного вывода ДСМ-метода: работа с данными, представленными в виде множеств.	110
3.3.2. Правила правдоподобного вывода ДСМ-метода: работа с данными, представленными в виде числовых векторов.	121
3.3.3. Правила правдоподобного вывода ДСМ-метода: некоторые дополнительные замечания.	130
Глава 4 Алгоритмы и оценки сложности вычислений в рамках ДСМ-ИАД.	135
4.1. Основные элементы ДСМ-алгоритмики восстановления эмпирических зависимостей из данных. Проблема быстрого роста объемов исчерпывающего ДСМ-перебора вариантов.	135
4.2. Оценки сложности вычислений при работе с булевскими данными.	141
4.3. Некоторые оценки сложности вычислений для случаев обработки других типов данных.	160
Глава 5 Приближенный ДСМ-метод.	174
5.1. Общие представления о методах и приемах сокращения перебора при реализации ДСМ-ИАД.	174
5.2. Псевдо-деревья и их каркасы.	177
5.3. Процедуры направленного перебора в диаграммах взаимной вложимости классов эквивалентности.	196
5.4. Корректность процедурной конструкции приближенного ДСМ-метода.	212

Глава 6 Некоторые функциональные особенности ДСМ-ИАД.	223
6.1. Эффекты немонотонности ДСМ-рассуждений.	223
6.2. О функциональности отношения причинности в ДСМ-методе.	226
6.3. «Наследуемость» эмпирических зависимостей и порождение закономерностей в рамках ДСМ-ИАД.	235
6.4 Некоторые комментарии к проблеме эпистемологических оснований ИАД	239
Заключение	246
Результаты, выносимые на защиту.	246
Перспективы дальнейших разработок.	247
Список рисунков и таблиц	248
Литература	249
Приложения Обзор приложений ДСМ-ИАД.	281
П.1. Общая характеристика приложений ДСМ-ИАД.	281
П.2. Анализ свойств физиологически активных веществ.	285
П.2.1. Фармакологические приложения.	287
П.2.2. Прогноз химической канцерогенности веществ.	290
П.2.3. Мониторинг потенциально опасных химических соединений.	293
П.3. Text-mining: текст-процессинг и задачи компьютерного анализа содержания полнотекстовых документов.	301
П.3.1. Проблема аппроксимации понятий в полнотекстовом информационном поиске.	304
П.3.2. Автоматическая классификация полнотекстовых документов.	312
П.3.3. Тематические карты содержания полнотекстовых документов.	315
П.3.4. Управление знаниями в системах электронного документооборота.	324
П.4. Формализация метода геологических аналогий.	328
П.4.1. Метод геологических аналогий и ДСМ-рассуждения.	339
П.4.2. Пример ДСМ-прогноза нефтегазоносности региона.	331
П.5. Банковский клиринг как задача управления и оптимизации в больших системах.	337
П.5.1. Проблема управления в больших системах комплексного характера.	337
П.5.2. ДСМ-рассуждения в оптимизации банковского клиринга.	341
П.6. Экспертиза сложных технических проектов.	355
П.7. О некоторых комбинаторных свойствах ассоциативных зависимостей	362
П.8 ДСМ-ИАД: как это работает.	371
П.8.1. ДСМ-метод на примере.	371
П.8.2. К архитектуре организации промышленных ДСМ-вычислений.	376

Введение

Диссертационная работа посвящена специальному классу задач восстановления зависимостей из эмпирических данных в процессе обучения на прецедентах (примерах и контрпримерах). Для реконструкции таких зависимостей используется формализованная модель комплекса эвристик, традиционно ассоциируемых с проблемой эмпирической индукции. Предлагается логико-комбинаторный подход, который позволяет сформировать средства контроля корректности (достаточности оснований для принятия) восстанавливаемых зависимостей.

В задачах рассматриваемого типа анализируется частично-определенное отношение **ОБЪЕКТ => СВОЙСТВА**, заданное примерами и контрпримерами. Отношение => характеризует причинно-следственные зависимости в ситуации, когда наличие (или же, наоборот, отсутствие) анализируемых **СВОЙСТВ** определяется направленными влияниями детерминистского характера – наличием (или же, наоборот, отсутствием) в структуре (описании) изучаемого **ОБЪЕКТА** соответствующего структурного фрагмента-носителя («причины»).

Развиваемый в диссертационной работе логико-комбинаторный подход позволяет формировать оценки сложности вычислений и оптимизировать объемы перебора вариантов при поиске решений математических задач, возникающих рамках предложенной техники восстановления зависимостей из эмпирических данных (примеров и контрпримеров - **ОБЪЕКТОВ**, соответственно обладающих или же, наоборот, не обладающих анализируемыми **СВОЙСТВАМИ**).

Восстанавливаемые зависимости используются для прогнозирования **СВОЙСТВ** новых **ОБЪЕКТОВ** (путем экстраполяции этих зависимостей на новые **ОБЪЕКТЫ** и контроля специальных условий достаточности оснований для принятия результатов такой экстраполяции).

Отличительными особенностями развиваемой техники восстановления эмпирических зависимостей являются:

- возможности оперировать, в том числе, и малыми (статистически незначимыми) выборками примеров и контрпримеров (случай нестатистического анализа данных);
- возможности оперировать структурными (т.е. хорошо структурированными) описаниями **ОБЪЕКТОВ** – множествами (признаков), множествами и заданными на них отношениями (графами, цепочками символов конечного алфавита), а также кортежами компонентов подобного типа (возможно, расширяемыми и числовыми значениями соответствующих параметров);
- возможности оперировать в неметрическом (более точно - дометрическом[1]) случае, когда основным объектом исследования оказывается реляционная структура причинных зависимостей в имеющихся эмпирических данных, заданная неявным образом (т.е прецедентами - примерами и контрпримерами) и требующая своего восстановления.

Одним из успешных инструментов решения задач подобного типа является так называемый ДСМ-метод автоматического восстановления зависимостей из эмпирических данных ([*Финн 10, ДСМ АПГ, Финн 11*] и др.)

Однако, весьма чувствительной проблемой здесь оказываются неприемлемо большие для индустриальных приложений ДСМ-метода объемы вычислений уже на ряде выборок сравнительно небольшого размера[2] (например – в некоторых важных для конкретных

[1] Т.е. в ситуации, когда на всей выборке исходных эмпирических данных определить корректным образом ту или иную адекватную характеристикам анализируемого отношения => метрику не представляется возможным. Однако, после восстановления соответствующей реляционной структуры в таких данных - формирования характеризующих эту структуру классов сходства и классов эквивалентности – появляются возможности для корректного задания метрики в каждом из реконструированных классов эквивалентности.

[2] Содержащие от нескольких десятков до полутора-двух сотен **ОБЪЕКТОВ**.

приложений задачах анализа социологических данных, прогнозирования лекарственных свойств физиологически активных химических соединений и др.).

Развиваемая в данной работе математическая техника дискретной оптимизации позволяет сформировать модели и средства целенаправленного управления перебором в решении комбинаторных задач обсуждаемого типа, в том числе – сформировать так называемый приближенный ДСМ-метод (допускающий, помимо прочего, также возможности реализации параллельных вычислений при решении названных задач дискретной оптимизации).

Актуальность темы.

Одним из наиболее известных и весьма детально развитых направлений в области разработки математических моделей и методов *машинного обучения* на прецедентах является проблематика так называемой *Теории статистического обучения*[3] (*statistical learning theory – SLT)* - см., например, [*Вапник 74, Вапник 79, Вапник 84, Vapnik 98, Vapnik 00, Воронцов 04, Воронцов 04-а*] и др. . Ключевой проблемой *Теории статистического обучения* является формирование оценок вероятности ошибки при переносе (экстраполяции) построенных на обучающей выборке зависимостей на новые (не входящие в эту выборку) объекты. Однако в случае малых (статистически не значимых) выборок здесь на сегодняшний день имеется целый ряд еще не решенных вопросов.

В более широком контексте (т.е. в ситуации, когда для порождения зависимостей из данных используются также нестатистические модели и алгоритмы – см., например, [*Michie 94, Mitchell 97, Bishop 06, Langley 11, Mohri 12*] и др.) весьма чувствительны ограничения на область практического применения подобного инструментария связаны с, как правило, экспоненциально быстро растущими оценками сложности вычислений, характерными для возникающих здесь комбинаторных задач.

Достаточно широко распространенными технологиями решения обсуждаемой нами задачи обучения на прецедентах являются также так называемые

- *нейронные сети* (см., например, [*Кохонен 80, Chester 93, Круглов 01, Рутковская 06*] и др.), где при поиске решения используется специальный вариант математических моделей многопараметрической нелинейной оптимизации, и
- *генетические алгоритмы* (см., например, [*Гладков 06, Рутковская 06, Шашкин 10*] и др.), где решение задач оптимизации при выборе наиболее подходящего класса зависимостей ищется моделированием локальных процедур случайного выбора, вариации и\или комбинирования анализируемых параметров по аналогии с механизмами естественного отбора в живой природе.

Однако, в анализе связей вида **ОБЪЕКТ => СВОЙСТВА** при поиске ответа на вопрос «*почему возникает анализируемое явление*[4]?», предлагаемый нейронными сетями и генетическими алгоритмами вариант ответа вида «*такой результат дал нам используемый алгоритм расчета*» вряд ли сможет убедить независимого[5] эксперта в наличии достаточных оснований для принятия полученных соответствующим инструментарием результатов. В ряде прикладных областей возможности выявлять *полный набор* каузальных «влияний», определяющих наличие (или же отсутствие) анализируемого явления, оказывается критически важной. Так, например, при выборе эффективных средств противодействия сбоям в работе сложного технического оборудования (в частности, при решении задач так называемого troubleshooting'а или же при организации противодействия компьютерным

---

[3] В профессиональной литературе по обсуждаемой тематике часто используется также близкий термин - *Теория вычислительного обучения* (*computational learning theory - COLT*).

[4] Почему у **ОБЪЕКТА** *O* наличествуют **СВОЙСТВА** ***P*** ?

[5] Например, того, кто не принимал активного участия в разработке (и настройке) соответствующей нейронной сети или же конкретного вида локальных процедур эволюционной обработки данных для соответствующего генетического алгоритма (и не владеет детальными знаниями об «изюминках» соответствующих математических моделей или их алгоритмики).

атакам[6]), в задачах медицинской диагностики и выбора эффективных средств терапии и др., не учесть тот или иной негативный каузальный фактор в организации противодействия соответствующим угрозам, вообще говоря, значит оставить объект защиты по-прежнему в потенциально уязвимом состоянии.

Наконец, возможности эффективно оперировать с хорошо структурированными данными нечислового характера (графами, цепочками символов конечного алфавита и т.п.) также представляет интерес как в плане развития соответствующих математических моделей и методов, так и в части применения основанного на них инструментария анализа данных при решении прикладных задач, востребованных исследовательским сообществом и индустрией.

Дополнительными фактором, подчеркивающим актуальность рассматриваемой в диссертационной работе тематики можно считать проблематику так называемых Big Data. Объемы данных, требующих здесь эффективного анализа, не оставляют надежд на решение этой задачи «вручную». Время, необходимое для анализа реальных (формируемых в значимых приложениях) данных оказывается критичным параметром (время реакции на те или иные события в поведении объекта управления – так называемое *процессно-реальное время* – становится важнейшим фактором, характеризующим эффективность процессов анализа данных и поддержки принятия решений). Таким образом, в области Big Data использование предлагаемых подходов, математических моделей и методов анализа данных может оказаться полезным при создании компьютерных систем, которые У.Р.Эшби [*Эшби 59*] классифицировал как *усилители интеллектуальных возможностей эксперта*.

С учетом изложенного выше можно констатировать, что актуальной научной проблемой является создание математического аппарата (методов, моделей и алгоритмов), который не только обеспечивал бы

- восстановление зависимостей из эмпирических данных в процессе обучения на прецедентах (примерах и контрпримерах) а также
- эффективный контроль корректности (условий достаточности оснований для принятия) восстанавливаемых зависимостей,

но при этом позволял бы

- оперировать, в том числе, и малыми (статистически незначимыми) выборками прецедентов,
- оперировать структурными (т.е. хорошо структурированными) описаниями анализируемых прецедентов – множествами (признаков), множествами и заданными на них отношениями (графами, цепочками символов конечного алфавита), а также кортежами компонентов подобного типа (возможно, расширяемыми и числовыми значениями соответствующих параметров);
- выявлять исчерпывающий набор каузальных «влияний», определяющих наличие изучаемых причинных зависимостей в имеющихся исходных данных
- оперировать в ситуации, когда основным объектом исследования оказывается, заданная неявным образом прецедентами и требующая своего восстановления реляционная структура причинных зависимостей в имеющихся эмпирических данных.

Целью диссертационной работы является разработка такого математического аппарата (методов, моделей и алгоритмов), а также демонстрация возможностей его эффективного применения в приложениях из различных предметных областей.

Научная новизна.

В процессе разработки предложенных в диссертационной работе математических методов, моделей и алгоритмов

- для идентификации порождаемых эмпирических зависимостей развита оригинальная техника формирования классов эквивалентности на исходно задаваемом множестве

[6] См. выше обсуждение примеров *диагностики и атрибуции целенаправленных активностей*.

прецедентов. Такие классы эквивалентности реконструируются по классам сходства прецедентов, которые формируются на основе отношения структурного сходства описаний прецедентов, формализуемого с помощью соответствующей алгебраической операции;

- для анализа корректности (выполнимости условий достаточности оснований для принятия) порождаемых эмпирических зависимостей а также прогноза их расширения (экстраполируемости) на описания новых прецедентов используются специальные комбинаторные объекты – диаграммы частичного порядка взаимной вложимости построенных классов эквивалентности;
- получен ряд оценок вычислительной сложности, характеризующих переборные задачи, которые возникают при формировании подобных классов эквивалентности и диаграмм их взаимной вложимости на нескольких типах структурных описаний исходно заданных прецедентов;
- развита оригинальная алгоритмическая техника формирования приближенных описаний диаграмм частичного порядка анализируемых классов эквивалентности (так называемая процедурная конструкция *приближенного ДСМ-метода*), которая позволяет организовать целенаправленный управляемый перебора вариантов при порождении эмпирических зависимостей рассматриваемого типа, в том числе:
    a) строить в первую очередь (используя так называемые каркасы псевдо-деревьев замыканий Галуа, формируемых на множестве исходно заданных прецедентов, причем - строить полиномиально быстро) так называемые *полезные* (для прогноза свойств новых объектов, для проверки выполнимости условия каузальной полноты и др.) эмпирические зависимости,

а затем, если потребуется,

    b) достраивать (уже, вообще говоря, экспоненциально сложными вычислениями) множество таких зависимостей до состояния, в котором восстановлены *все* содержащиеся в исходных данных эмпирические зависимости;

- сформулированы и доказаны утверждения, определяющие корректность предложенной процедурной конструкции приближенного ДСМ-метода.

Работоспособность предложенных математических моделей и алгоритмов продемонстрирована на примерах решения конкретных прикладных задач.

Методы исследования.

В диссертационной работе использованы:

- логико-комбинаторные и алгебраические методы дискретной математики;
- методы взаимной сводимости и построения оценок вычислительной сложности трудных переборных задач,
- методы анализа и дискретной оптимизации вычислительных алгоритмов.

Теоретическая значимость.

Разработанные математические модели, методы и алгоритмы позволяют организовать порождение эмпирических зависимостей в том числе на малых выборках сложно структурированных описаний прецедентов, исходно заданных для обучения.

Предложенная в диссертационной работе математическая техника формирования и анализа диаграмм взаимной вложимости классов эквивалентности[7] прецедентов демонстрирует комбинаторную природу хорошо известного в исследованиях по искусственному интеллекту и ассоциируемого с проблемой эмпирической индукции класса эвристик (см., например, [*Милль 00*, *Пойа 70, Пойа 75, Пирс 00, Пирс 05, Финн 10*] и др.).

[7] Реконструируемых по классам сходства, которые, в свою очередь, формируются на основе формализуемого с помощью соответствующей алгебраической операции отношения структурного сходства описаний прецедентов.

Развитая процедурная конструкция целенаправленного построения приближенных описаний таких диаграмм формирует основу для создания специального класса программных систем искусственного интеллекта, ориентированных на использование формализованных моделей правдоподобных рассуждений (построения индуктивных обобщений, рассуждений по аналогии, абдуктивных объяснений и т.п.).

Практическая значимость.

Алгоритмическая техника формирования приближенных описаний диаграмм частичного порядка рассматриваемых классов эквивалентности (так называемая процедурная конструкция приближенного ДСМ-метода) дает возможности оперировать при восстановлении эмпирических зависимостей рассматриваемого класса исходными выборками данных любого (в т.ч. – большого размера[8]).

Предложенная техника формирования структурных описаний диаграмм частичного порядка классов эквивалентности позволяет организовать параллельную обработку данных (в том числе, для промышленных приложений - на базе масштабируемых облачных вычислений).

Предложенная техника восстановления зависимостей успешно применяется при решении ряда прикладных задач интеллектуального анализа данных.

Область исследования согласно Паспорту специальности 05.13.17 – «Теоретические основы информатики»:

- *разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных* (п.5);
- *моделирование формирования эмпирического знания* (п.7);
- *исследование и когнитивное моделирование интеллекта, включая моделирование поведения, моделирование рассуждений различных типов* (п.8);

а также

- *исследование информационных структур, разработка и анализ моделей информационных … структур* (п.2);
- *исследование и разработка средств представления знаний* (п.4)

В соответствии с обозначенными в формуле специальности 05.13.17 – «Теоретические основы информатики» направлениями, область данного исследования характеризуют в том числе и «*…исследования методов преобразования информации в данные и знания*; *… исследование … моделей данных и знаний*, *… методов машинного обучения и обнаружения новых знаний*; *исследования принципов создания и функционирования … программных средств автоматизации указанных процессов*».

Таким образом, выполненное в диссертационной работе исследование комбинаторной структуры соответствующего класса формальных моделей эмпирической индукции соответствует специальности 05.13.17 – «Теоретические основы информатики».

Апробация работы.

Основные положения и результаты работы докладывались и обсуждались на всероссийских и международных конференциях, конгрессах, чтениях и семинарах. В их числе – на Международных конференциях НТИ (ВИНИТИ РАН), Национальных конференциях с международным участием «Искусственный интеллект» (Российская ассоциация искусственного интеллекта), Конференциях и семинарах IEEE (США), Российско-британской конференции «Идеи Д.С. Милля об индукции и логике наук о человеке и обществе в когнитивных исследованиях и системах искусственного интеллекта» (РГГУ), научных семинарах ВИНИТИ РАН, МФТИ, ВМК МГУ, ВЦ РАН, ИПУ РАН и др..

[8] См., в частности, уже упоминавшуюся выше проблему Big Data.

Материалы данной диссертационной работы легли в основу спецкурса «Формальные модели рассуждений в задачах компьютерного восстановления зависимостей из данных», читавшегося автором на протяжении ряда лет студентам старших курсов на Факультете управления и прикладной математики Московского физико-технического института.

Публикации по теме диссертации в изданиях из списка ВАК:

1. Забежайло М.И. Ненаследуемость эмпирического противоречия в ДСМ-методе и немонотонные рассуждения // НТИ[9], Сер.2. - 1984. - N11. - С.14-17.
2. Забежайло М.И К проблеме расширения ДСМ-метода автоматического порождения гипотез на данные с числовыми параметрами // НТИ, Сер.2. - 1992. - N11. - с.11-21
3. Забежайло М.И. К проблеме расширения ДСМ-метода автоматического порождения гипотез на данные с числовыми параметрами II // НТИ. Сер.2. - 1993. - N2. - с.6-16.
4. Забежайло М.И. Формальные модели рассуждений в принятии решений: приложения ДСМ-метода в системах интеллектуального управления и автоматизации научных исследований // НТИ, Сер.2. - 1996. - N5-6. - С.20-32.
5. Забежайло М.И. О комбинаторной природе одной задачи оптимизации // НТИ, Сер.2. - 1996. - N1. - С.19-27.
6. Забежайло М.И. К проблеме формализации метода геологических аналогий. // Известия РАН. Сер. Теория и системы управления. - 1997. - N2.- С.151-164.
7. Забежайло М.И., Емеленко М.Н., Липчинский Е.А., Максин М.В. К пониманию термина “прикладная семиотика” // НТИ, Сер.2 - 1998. - N1. – С. 11-18.
8. Забежайло М.И. К проблеме автоматического понимания полнотекстовых документов в информационном поиске. // Известия РАН. Сер. Теория и системы управления. - 1998. - N5. - С.167-176.
9. Забежайло М.И. Интеллектуальный анализ данных – новое направление развития информационных технологий // НТИ, Сер. 2.- 1998. - №8. - С. 6-17.
10. Забежайло М.И. О функциональности отношения причинности, используемого в ДСМ-рассуждениях // НТИ, Сер.2. - 2013. – N7.- С.1-7.
11. Забежайло М.И. О некоторых возможностях управления перебором в ДСМ-методе // Искусственный интеллект и принятие решений. – 2014. – Часть I: № 1, С.95 -110.
12. Забежайло М.И. О некоторых возможностях управления перебором в ДСМ-методе // Искусственный интеллект и принятие решений. – 2014. – Часть II: № 3, С.3 - 21.
13. Забежайло М.И., Синякова Е.В. К вопросу об «интеллектуальности» интеллектуального анализа данных // НТИ, Сер. 2. - 2014. - № 3. - С. 1-9.
14. Забежайло М.И. Приближенный ДСМ-метода на примерах // НТИ, Сер.2. – 2014. – №10, С. 1-12.
15. Забежайло М.И. К вопросу о достаточности оснований для принятия результатов интеллектуального анализа данных средствами ДСМ-метода // НТИ, Сер.2. – 2015. – №1. – С.1-9.
16. Забежайло М.И. О некоторых оценках сложности вычислений в ДСМ-рассуждениях // Искусственный интеллект и принятие решений. - 2015. – Часть I: №1. – С. 3-17.
17. Забежайло М.И. О некоторых оценках сложности вычислений в ДСМ-рассуждениях // Искусственный интеллект и принятие решений. - 2015. – Часть II: №2. – С. 3-17.

Другие публикации по теме диссертации:

18. Zabezhailo M.I. et al. Reasoning Models for Decision Making: Applications of JSM-Method for Intelligent Control Systems//Architectures for Semiotic Modeling and Situation Analysis in Large Complex Systems. - Proc. of the Workshop of 10th (1995) IEEE Symp. on Intelligent Control (Eds.: J.Albus, A.Meystel, D.Pospelov, T.Reader). - 27 -29 August 1995, Monterey, CA../ AdRem, Inc., 1995. - Pp. 99-108.

[9] НТИ - Научно-Техническая Информация.

19. Zabezhailo M.I. To the Scalable Technology of Automated Document Understanding Based on Quasi-Axiomatic Theories.// Proc. 1997-th International Conference on Intelligent Systems and Semiotics “A Learning Perspective” - ISAS’97 (NIST, Gaithersburg, MD, September 22-25, 1997). - NIST Special Publications. - N918. - 1997. - Pp.100-102.
20. Забежайло М.И., Финн В.К., Козлова С.П., Катамадзе Т.Г., Авидон В.В., Рабинков А.А. Об одном методе автоматического формирования гипотез и его программной реализации // НТИ, Сер.2. - 1982. - N4. - С.20-26.
21. Забежайло М.И., Финн В.К., Авидон В.В., Катамадзе Т.Г., Блинова В.Г., Бодягин Д.А., Рабинков А.А. Об экспериментах с базой данных с неполной информацией посредством ДСМ-метода автоматического порождения гипотез // НТИ, Сер.2. - 1983. - N2. - С.28-32.
22. Забежайло М.И., Ивашко В.Г., Кузнецов С.О., Михеенкова М.А., Хазановский К.П., Аншаков О.М. Алгоритмические и программные средства ДСМ-метода автоматического порождения гипотез. // НТИ, Сер.2. - 1987. - N10. - с.1-13.
23. Забежайло М. И. Новые информационные технологии и системы НТИ // НТИ, Сер. 2. — 1990. — №5. — С. 2—9.
24. Забежайло М.И. Некоторые тенденции в развитии интеллектуальных систем // Программные продукты и системы. - 1990. - N 4. - С.86-94.
25. Забежайло М.И. Интеллектуальные системы и задача восстановления эмпирических зависимостей структурно-числового характера // Итоги науки и техники. Сер." Информатика". Т.15 "Интеллектуальные информационные системы" - М: ВИНИТИ. - 1991. - С. 102-114.
26. Забежайло М.И. Новые информационные технологии в научных исследованиях и технологических разработках // НТИ. Сер. 2.- 1992.- № 6.-С.1-11.
27. Забежайло М.И. Интеллектуальные системы: на пути к новым поколениям. // Новости искусственного интеллекта. N1, 1992. -Сс.8-24.
28. Забежайло М.И. (Zabezhailo M.I.) The extension of the JSM-method: plausible reasoning dealing with numeric data // Новости Искусственного Интеллекта. Специальный выпуск: Международная конференция по искусственному интеллекту "Восток-Запад - 93" (Artificial Intelligence News. Special issue: EWAIC-93, Moscow, 7-9 Sept. 1993). 1993.
29. Забежайло М.И. Формальные модели рассуждений в принятии решений: приложения ДСМ-метода в системах интеллектуального управления и автоматизации научных исследований // НТИ, Сер.2. - 1996. - N5-6. - С.20-32.
30. Забежайло М.И. О комбинаторной природе одной задачи оптимизации // НТИ, Сер.2. - 1996. - N1. - С.19-27.
31. Забежайло М.И., Емеленко М.Н., Липчинский Е.А., Максин М.В. К разработке платформенно-независимой версии программной системы, реализующей ДСМ-метод автоматического порождения гипотез// Труды 3-й Международной конференции “НТИ-’97: Информационные ресурсы, интеграция, технологии” (Москва, 26-28 ноября 1997 г.), М.: ВИНИТИ, Часть 1, С.91-93.
32. Забежайло М.И. Информационный поиск в полнотекстовых документах: некоторые проблемы и перспективы. // Новости искусственного интеллекта. - 1998. - N.1.- С. 24-59
33. Забежайло М.И. Data Mining & Knowledge Discovery in Data Bases: предметная область, задачи, методы и инструменты // Труды 6-ой национальной конференции по искусственному интеллекту с международным участием. - Пущино. - 1998. - Том 1. - С. 592-600.
34. Zabezhailo M.I., Finn V.K., Leybov A.E.,Melnikov N.I., Pankratova E.S. CBR-technology in the chemical safety control // Proc. EWCBR’94 (France, November 7-11, 1994).- 1994.

35. Zabezhailo M.I., Finn V.K. Intelligent Information Systems. - International Forum on Information and Documentation (FID, The Hague, Netherlands). - 1996. - V21. - N2. - Pp.21-31.
36. Zabezhailo M.I. On the Application of the Semiotic Modelling Technique in the Problem of Clearing Payments Control.// Proc. 12th European Conference on Artificial Intelligence (ECAI'96, Budapest, 11-16 August, 1996). - Workshop N30 "Applied Semiotics". Pp.17-21.
37. Zabezhailo M.I., Finn V.K., Gergely T. JSM-method as an Instrument for Semiotic Modelling: Application to Geological Forecasting.// Proc. 12th European Conference on Artificial Intelligence (ECAI'96, Budapest, 11-16 August, 1996). - Workshop N30 "Applied Semiotics". Pp.13-16.

Личный вклад автора.

В диссертационной работе представлены только результаты, которые полученные лично автором: исследование проблематики восстановления зависимостей по прецедентам, формулировки и доказательства утверждений, постановки задач, методы и алгоритмы их решения. Из совместных публикаций в диссертацию включены лишь результаты автора.

Структура и объем работы.

Диссертационная работа состоит из Введения, 6 Глав, Заключения, Списка литературы (525 наименований) и 8 Приложений. Работа содержит 17 рисунков и 19 таблиц. Общий объём работы - 379 страницы (280 стр. – основной текст и 99 стр. - Приложения).

Основные математические результаты работы представлены в **Разделах 2.1** и **3.1**, **Главах 4** и **5**. Содержательные свойства формализуемых эвристик анализируются в **Разделах 2.2**, **3.2** и **6.4**.

Краткое содержание диссертационной работы по Главам.

Глава 1 носит вводный характер и посвящена обсуждению общих характеристик области интеллектуального анализа данных (ИАД) а также места, которое занимает в ней проблематика математических моделей и методов восстановления зависимостей из накапливаемой эмпирической информации. Задача формализации эмпирической индукции (ЭИ) – обучения на прецедентах - рассматривается в первую очередь на примере так называемого ДСМ-метода автоматического порождения эмпирических зависимостей из данных и формулируется в контексте исследований в области искусственного интеллекта – как задача адекватной (позволяющей контролировать доказуемую корректность их использования) формализации ассоциируемого с понятием ЭИ комплекса эвристик (индуктивного обучения, рассуждений по аналогии и абдуктивных объяснений).

В качестве основой области исследований выбран важный (и мало изученный на текущий момент) частный случай общей проблематики обучения на прецедентах – работа с данными нечислового характера. Рассмотрены особенности анализа причинно-следственных связей между описаниями нечисловых объектов (реляционных структур, возможно, дополненных числовыми значениями некоторых существенных параметров) и множествами присущих им свойств. Особо подчеркнута актуальность развития математических моделей и методов, позволяющих получать корректные решения задачи такого типа на малых (статистически незначимых) выборках прецедентов. Обозначены актуальные для современного уровня развития интеллектуального анализа данных проблемы и ограничения.

Приведены примеры актуальных областей приложения рассматриваемых технологий ИАД (задачи компьютерного прогнозирования физиологических активностей химических соединений, автоматизированной обработки больших массивов полнотекстовых документов и др.).

Даны общие представления об архитектуре используемых в анализе данных интеллектуальных компьютерных систем. Обсуждаются возможности и ограничения

наиболее продвинутых математических моделей и техник компьютерного анализа данных. Сформулированы характеристики исследовательской ситуации, в которой адекватным представляется использование именно интеллектуальных систем анализа данных.

Глава 2 посвящена описанию базовых смысловых элементов организации ДСМ-ИАД. Формулируется задача обучения на прецедентах. Дано подробное описание общей для различных типов нечисловых данных (дополнительно характеризуемых также и числовыми параметрами) процедурной схемы, предлагаемой для ее решения. Рассматриваемая схема основана на формализации трех базовых эвристик – индуктивного обучения, выводов по аналогии и абдуктивного объяснения – и характеризуется строгими условиями корректности их применения.

В **Разделе 2.1** предложена схема формализации задачи обучения на прецедентах, выполненной алгебраическими средствами. В основе этой конструкции – анализ структурного сходства описаний прецедентов (примеров, характеризуемых наличием целевых свойств, и контрпримеров, характеризуемых их отсутствием), в котором собственно сходство формализуется как бинарная алгебраическая операция. На базе операции сходства определяется отношение сходства (характеризуемое непустыми результатами вычисления этой операции), с его помощью на множестве описаний исходно заданных прецедентов «одного знака»[10] строятся классы сходства. Затем, путем фиксации каждого конкретного результата вычисления операции сходства, выделяются соответствующие подклассы сформированных классов сходства. Показано (**Утверждение 2.1.1**), что каждый из таких подклассов как подмножество исходно заданного множества описаний прецедентов представляет собою класс эквивалентности. Показано также, что каждый подобный класс эквивалентности может быть порожден специальным оператором замыкания, реализующим один и тот же набор действий при работе с различными типами описаний прецедентов. Таким путем восстанавливается структура покрытия классов сходства классами эквивалентности описаний прецедентов.

Показано, что задача восстановления зависимостей из имеющихся эмпирических данных может быть формализована как задача восстановления соответствующих классов эквивалентности на структурных описаниях прецедентов. (Система возникающих здесь отношений эквивалентности оказывается производной как от выбранного формального уточнения сходства в виде алгебраической операции, так и от конкретной структуры исходно заданной выборки прецедентов). Задача прогнозирования свойств новых, заранее не изученных прецедентов, в таком контексте формализуется как задача формирования классов эквивалентности (отдельно) на множестве прецедентов обоих знаков (примеров и контрпримеров) и последующей проверки вложимости описания нового прецедента хотя бы в один из классов эквивалентности лишь одного знака.

Определяя естественный порядок (по вложимости подмножеств исходно заданного множества прецедентов) на множествах восстанавливаемых из исходной выборки классов эквивалентности каждого знака, можно сформировать диаграммы частичного порядка. Структура и взаимосвязи (в части встречаемости таких результатов вычисления сходства, что сходство примеров некоторого класса эквивалентности может совпадать со сходством контрпримеров некоторого класса эквивалентности противоположного знака) изучаемых диаграмм противоположного знака определяют все основные особенности алгоритмики порождения эмпирических зависимостей в рассматриваемом здесь случае. Каждая из таких диаграмм простыми действиями может быть дополнена до решетки. Однако, нетрудно показать, что в общем случае порождаемые здесь решетки не являются дистрибутивными[11].

В **Разделе 2.2** приведено неформальное описание семантики так называемых правил индуктивного вывода Д.С.Милля (эвристики Ф.Бэкона-Д.С.Милля) - одного из базовых элементов задействованного в обсуждаемом подходе комплекса эвристик, формализуемых

[10] Т.е. отдельно – на примерах и контрпримерах.

[11] Подробно эта ситуация будет рассмотрена позднее – в **Разделе 4.2** (**Глава 4**).

соответствующими логико-математическими средствами. Вместе с рядом других эвристик (в первую очередь - рассуждениями по аналогии в духе Д.Пойа и абдуктивными выводами в стиле Ч.С.Пирса) эта эвристика положена в основу ДСМ-метода автоматического порождения зависимостей из эмпирических данных.

В **Разделе 2.3** представлено общее описание и наивная семантика ДСМ-метода. Приведены ссылки на построенную В.К.Финном, Д.П.Скворцовым и О.М.Аншаковым дедуктивную имитацию процедурной конструкции теоретико-множественной версии ДСМ-метода средствами специального класса многозначных логик. Такая имитация позволяет показать, что для каждого утверждения, которое может быть получено применением (к конкретным исходным данным) используемых в рамках ДСМ-метода правил правдоподобного вывода, в соответствующей (имитирующей ДСМ-рассуждения) дедуктивной теории средствами достоверного вывода может быть также получено аналогичное утверждение, и наоборот. (Таким способом демонстрируется корректность задействованного в рамках ДСМ-метода способа формализации используемых эвристик).

Глава 3 посвящена описанию процедурной структуры ДСМ-метода - формальному уточнению представленных в **Главе 2** базовых элементов развиваемого подхода:

- алгебраической формализации сходства как алгебраической операции при работе с различными типами описаний прецедентов из исходной обучающей выборки,
- детализации функциональных особенностей формализуемого комплекса эвристик (эмпирической индукции в стиле Ф.Бэкона-Д.С.Милля, выводов по аналогии, порождения абдуктивных объяснений в стиле Ч.С.Пирса и др.),
- обсуждению возможностей так называемых квази-аксиоматических теорий (специальных расширений аксиоматических теорий, использующих не только правила достоверного вывода, но также и правила правдоподобного вывода, назначение которых - порождение эмпирических зависимостей из фактов) как средства представления знаний в компьютерных системах ИАД, ориентированных на решение задач эмпирической индукции,
- детальному описанию используемых в ДСМ-ИАД инструментов анализа данных (так называемых решающих предикатов, правил правдоподобного вывода и стратегий формализованных ДСМ-рассуждений).

В **Разделе 3.1** рассмотрены варианты математической формализации понятия сходства на различных типах данных (множествах, графах, цепочках символов конечного алфавита, векторах числовых значений параметров, …), цель которых - порождение (в явной или неявной форме) соответствующего отношения *толерантности*. С помощью операций с числовыми матрицами средствами специального вида – матрицами комонотононого изменения числовых значений параметров в их строках – формализуется отношение сходства на числовых векторах. Определена операция сходства (множеств) монотонных матриц, которая на специальном классе подмножеств таких матриц, порождаемом из исходно заданного множества прецедентов (столбцов числовых значений параметров), формирует нижнюю полурешетку (**Теорема 3.1.2.15** о корректности определения сходства для работы с числовыми значениями параметров).

В **Разделе 3.2.1** представлена система логических языков для описания правил правдоподобного вывода, характеризующих процедурный механизм операций порождения и выявления взаимных зависимостей на множествах классов эквивалентности прецедентов – см. выше **Раздел 2.1**. Семантические особенности используемого (в рамках формируемой формализации) комплекса эвристик (индуктивного обучения, выводов по аналогии и абдуктивного объяснения) обсуждаются в **Разделе 3.2.2**. Общие свойства правил правдоподобного вывода ДСМ-метода (условия применимости, рациональность, инвариантность ядра при работе с различными типами данных и др.) анализируются в **Разделе 3.2.3**. Явный вид решающих предикатов и собственно правил правдоподобного вывода (в том числе – для операций с теоретико-множественными и числовыми данными) представлен к **Разделах 3.3.1-3**.

Центральная роль в работе принадлежит математическим результатам об оценках вычислительной сложности и оптимизации алгоритмики формируемой процедурной конструкции обучения на прецедентах, рассмотренным в **Главах 4** и **5**.

Так в **Главе 4** сформулирован и доказан комплекс утверждений, которые определяют вычислительную сложность основных комбинаторных задач, характеризующих диаграммы вложимости классов эквивалентности прецедентов для рассматриваемых в работе типов данных. Показано (**Теоремы 4.2.10**, **Теорема 4.3.4** и **Следствия 4.3.5-6** а также **Теорема 4.3.11**), что емкость диаграмм обсуждаемого типа растет экспоненциально быстро с ростом объема исходной выборки прецедентов, а сама задача о числе входящих в такие диаграммы элементов (классов эквивалентности описаний прецедентов) принадлежит классу *#PC* – перечислительно полных задач. Показано, что верхняя и нижняя границы рассматриваемых диаграмм как в теоретико-множественном случае, так и при работе с векторами числовых значений параметров, могут быть порождены полиномиально сложными алгоритмами. Получены оценки вычислительной сложности поиска в анализируемых диаграммах элементов, которые удовлетворяют заданным ограничениям на структуру и емкость порождаемых классов эквивалентности (размеры классов, параметры порождающего каждый из них сходства и т.п.). Так в частности, показано (**Теорема 4.2.20** и **Следствие 4.2.31**, **Следствия 4.3.5-7**, **Теорема 4.3.19**), что задача поиска класса эквивалентности, который порождается вточности заданным числом прецедентов из исходной выборки, принадлежит классу *NPC* (*NP*-полных задач).

Глава 5 посвящена описанию разработанного автором диссертационного исследования метода дискретной оптимизации перебора вариантов при восстановлении (по исходно заданной выборке) диаграмм взаимной вложимости классов эквивалентности прецедентов, описываемых множествами признаков (теоретико-множественный случай). Показано, что каждая диаграмма может быть представлена объединением поддиаграмм специального вида - так называемых псевдо-деревьев (в корне каждого из которых находится описание одного из исходно заданных прецедентов, а остальные вершины содержат лишь входящие в корень элементы - образующие). Предложена технология порождения приближенных описаний таких псевдо-деревьев, стартующая с порождения так называемого каркаса – диаграммы взаимной вложимости замыканий всех одноэлементных подмножеств находящегося в корне псевдо-дерева описания прецедента (множества образующих). Показано (**Утверждение 5.2.11**), что каждый каркас порождается процедурой полиномиальной вычислительной сложности. Показано, что при построении каркаса имеется возможность диагностировать (полиномиально сложной вычислительной процедурой) наличие в соответствующем псевдо-дереве в том числе и экспоненциально сложных архитектурных фрагментов. Выделены (**Утверждение 5.2.15**) полиномиально быстро проверяемые необходимые и достаточные условия, при выполнении которых соответствующий фрагмент псевдо-дерева представляет собою полный гиперкуб (на некотором подмножестве образующих). Для случая, когда соответствующий комбинаторно-сложный фрагмент псевдо-дерева составляет лишь некоторую (требующую восстановления) часть такого гиперкуба, разработан метод последовательных приближений (целенаправленного порождения все более сложных описаний архитектурных фрагментов подобного типа), оперирующий

- специальным сжимающим отображением на множестве исходно заданных прецедентов, позволяющим последовательно переходить к целенаправленно выбираемым подзадачам при восстановлении каждого анализируемого псевдо-дерева, а также
- специальный оператор, монотонно расширяющий текущее приближенное описание реконструируемого псевдо-дерева вновь порождаемыми при анализе подзадач (см. выше – сжимающее отображение) архитектурными фрагментами.

Предложенный метод позволяет порождать в первую очередь те фрагменты восстанавливаемых диаграмм, которые полезны для прогнозирования свойств новых

прецедентов. Далее (разумеется за счет все более и более объемных вычислений) может быть порождено и точное описание каждой такой диаграммы. Показано, что этот метод позволяет организовать процесс порождения целевых диаграмм вложимости в режиме *параллельных* вычислений.

В **Разделе 5.4** представлено строгое обоснование корректности разработанного метода последовательных приближений (так называемого *приближенного* ДСМ-метода). Описан (основанный на утверждаемых **Теоремой 5.4.2** результатах о представимости диаграмм вложимости так называемыми каркасами всех допустимых рангов) механизм управления порождением последовательных приближений заданной точности.

Представленные в **Главе 5** результаты имеют центральное значение для всей рассматриваемой диссертационной работы, т.к. здесь сформулирована алгоритмически корректная и строго обоснованная процедурная конструкция, позволяющая (не смотря на ряд имеющихся ограничений в части сложности вычислений – см. выше - Главу 4) оперировать при порождении эмпирических зависимостей из данных и прогнозировании свойств новых объектов выборками прецедентов любого размера. (Фактически эффективный размер выборки определяется возможностями вычислительной установки, на которой может быть организован интеллектуальный анализ данных средствами *приближенного* ДСМ-метода, выполняемого в режиме *параллельных* вычислений).

В **Главе 6** рассмотрены некоторые дополнительные характеристики предложенного инструментария ИАД. Демонстрируются эффекты немонотонности правдоподобного вывода в формализующих ДСМ-ИАД квази-аксиоматических теориях. Доказана (**Утверждение 6.2.2**) функциональность формализованного и анализируемого предложенными средствами отношения причинности. Представлен ряд замечаний о возможности контроля устойчивости порождаемых рассматриваемым способом эмпирических зависимостей при расширении исходной выборки прецедентов.

В **Приложениях** дан обзор применения интеллектуальных компьютерных систем, основанных на предложенном комплексе математических моделей методов и алгоритмов, при решении задач ИАД в ряде практически значимых прикладных предметных областей:

- анализе свойств физиологически активных веществ (**П.2.1 -3**);
- автоматической обработке коллекций полнотекстовых документов (**П.3.1-4**);
- геологическом прогнозировании нефтегазоносности территорий (**П.4.1-2**);
- предконкурсной экспертизе сложных технических проектов (**П.6**) и др. .

Основные математические результаты работы представлены в **Разделах 2.1** и **3.1**, **Главах 4** и **5**. Обсуждению содержательных характеристик формализуемых эвристик (индуктивного обучения, выводов по аналогии и абдуктивного объяснения) посвящены **Разделы 2.2**, **3.2** и **6.4**. Работоспособность развиваемого подхода при решении прикладных задач продемонстрирована в **Приложениях**.

Благодарности.

Автор считает свои приятным долгом выразить глубокую благодарность своему учителю – профессору Виктору Константиновичу ФИННУ, без внимательного и доброго отношения, понимания и поддержки которого эта работа вряд ли увидела бы свет.

Автор признателен члену-корреспонденту РАН Константину Владимировичу РУДАКОВУ за поддержку и внимание к работе, а также профессору Константину Вячеславовичу ВОРОНЦОВУ за обсуждения содержания работы и весьма полезные критические замечания.

Особую благодарность автор считает необходимым выразить профессору Александру Александровичу ГРУШО за внимание к работе, поддержку, подробные обсуждения и профессиональные рекомендации.

Глава 1

Общая характеристика области интеллектуального анализа данных (ИАД) и место в ней рассматриваемой проблематики восстановления эмпирических зависимостей средствами ДСМ-метода

Анализ каузальных (причинно-следственных) зависимостей – одно из традиционных направлений исследований в области искусственного интеллекта. Каузальность (причинно-следственная связь вида ***a***=>***b***) обычно понимается здесь в следующем смысле:

если воспроизвести все условия, характеризующие ***a****,*
то (как «следствие») будет иметь место и ***b****.*

Термин *как «следствие»* взят здесь в кавычки намеренно: так можно привести примеры ситуаций, когда анализируемое отношение => определяется, в частности, на декартовом квадрате множеств формул языка логики предикатов первого порядка и представляет собою отношение логического следования, а выражение ***a***=>***b*** понимается как импликация[1] ***a→b.***

Систематический обзор всех следствий, порождаемых из заданного множества формул, в подобных ситуациях может быть организован средствами так называемой *автоматической дедукции* (см., например, [*ADed, Нильсон 73*] и др.), в основании которой – *метод резолюций* Дж.Робинсона [*Robinson 65, Robinson 65- a, Robinson 67*] и *обратный метод* С.Ю.Маслова [*Маслов 64, Маслов 68*]. Будучи предложенной исследовательскому сообществу в середине 60-х годов двадцатого века, эта техника стала основой для широкого спектра исследований в области разработки и использования при решении прикладных задач так называемых компьютерных систем-prover'ов, предназначенных для порождения зависимостей как вывода следствий и автоматического доказательства теорем (см., например, [*Чень 83*] и др.).

Позднее – в 80-90 годы естественным развитием этого подхода стали так называемые *продукционные экспертные системы* (см., например, [*ES, Попов 87*] и др.) и лежащие в основе большинства из них инструментальные среды на базе той или иной версии языка программирования PROLOG (см., например, [*Пролог, Clocksin 03*] и др.).

Особого внимания при обсуждении этого направления, по-видимому, заслуживает предложенная Г.Плоткиным (см. [*Plotkin 70, Plotkin 71, Plotkin 72*]) математическая техника формирования по заданному множеству формул $\varphi_1,\varphi_2,\ldots,\varphi_n$ исчисления предикатов первого порядка их так называемого *минимального общего обобщения* – новой формулы ψ, характеризуемой двумя свойствами:

- каждая из исходно заданных формул $\varphi_1,\varphi_2,\ldots,\varphi_n$ представляет собою логическое следствие вновь сформированной формулы ψ,
- неверно, что найдется еще одна (отличная от ψ) формула ψ_0 такая, что $\psi \rightarrow \psi_0 \rightarrow \varphi_i$ (для всех i= 1,2, … , n).

Целевая формула ψ порождается по формулам $\varphi_1,\varphi_2,\ldots,\varphi_n$ алгоритмом, подобным базовому для автоматической дедукции алгоритму унификации (с той лишь разницей, что на каждом шаге ищется не порождающая их максимальное общее следствие подстановка-унификатор, а, наоборот, минимальное в смысле отношения следования обобщение для текущей пары формул, из которого каждая из них будет логически следовать). Описанная возможность породить для заданного множества формул $\varphi_1,\varphi_2,\ldots,\varphi_n$ их минимальное общее обобщение ψ существенным образом упрощает процесс дальнейшего обзора всех общих следствий для анализируемого множества формул $\varphi_1,\varphi_2,\ldots,\varphi_n$.

[1] В традиционном для математической логики понимании бинарной связки →.

Однако, не менее хорошо известны и примеры ситуаций, когда причинно-следственная связь вида ***a***=>***b*** задана на декартовом произведении разных по своей природе множеств. Так, в частности, активно развиваемая как в части создания и совершенствования соответствующих математических моделей, так и в части получения различных прикладных результатов, проблематика анализа связи

ХИМИЧЕСКАЯ СТРУКТУРА => ЛЕКАРСТВЕННЫЕ СВОЙСТВА

физиологически активных веществ (см., например, [*QSAR*] и др.) оперирует при анализе целевых свойств изучаемых соединений в том числе и *структурными* описаниями изучаемых объектов (например, молекулярными графами или же их проблемно-ориентированными представлениями в виде систем типовых подграфов специальной структуры; характеризующими данное соединение спектрами излучения или поглощения и др.). При этом анализируемое отношение ***a***=>***b*** может быть интерпретировано следующим естественным образом:

> *лекарственные свойства* ***b*** *имеют структурный носитель* ***a'(b)****, «охватываемый» («содержащийся в …», «покрываемый») описанием структуры соответствующего объекта - химического соединения* ***a****.*

Следует отметить, что в некоторых существенных для этой предметной области[2] случаях собственно химическая структура анализируемого физиологически активного соединения, вообще говоря, не известна и «представляется» лишь как способ получения этого вещества (т.е. как последовательность операций, исполняемых в строго определенном порядке и в соответствии со строго контролируемыми необходимыми условиями – длительностью, процедурами применения сопутствующих ингридиентов и т.п., - по завершению которой порождается объект, *гарантированно* [3] обладающий целевыми свойствами). Таким образом, изучаемая причинно-следственная связь в данном случае характеризуется упорядоченной парой сущностей следующего вида:

<*способ получения физиологически активного средства, лекарственные свойства*>.

Особый статус задач анализа причинно-следственных связей вида ***a***=>***b*** на *неоднородных* парах вида <**ОБЪЕКТ,СВОЙСТВА**> как самостоятельной сферы исследований в области искусственного интеллекта был выделен Дж.Маккарти, обратившим внимание профессионального сообщества на проблему, которая получила название *phenomenal data mining* (см., например, [*McCarthy 00*] и др.). Джон Маккарти показал (в том числе – даже на простейших - ставших в определенной мере традиционными для области data mining - примерах зависимостей вида <*корзина покупок ~ «профиль» покупателя*>) - сколь существенную роль для развития математических моделей, методов и алгоритмов, а также приложений систем искусственного интеллекта играют технологии восстановления зависимостей из фактов (в том числе – методы нечислового машинного обучения на примерах и контрпримерах).

Еще один пример проблематики подобного типа – предложенные В.К.Финном (см., например, и [*Финн 81-а*] др.) задачи так называемого до-определения частично-определенного (заданного небольшими наборами примеров и контрпримеров) бинарного отношения **ОБЪЕКТ=>СВОЙСТВА**. Предпринимаемый здесь интеллектуальный анализ данных фокусируется вокруг поиска «*структурных носителей*» - специфических фрагментов описаний **ОБЪЕКТОВ**, «отвечающих» за наличие у них соответствующих

---

[2] Например, такие ситуации известны при работе с антиканцерогенами – *иммуностимуляторами*. (NB: Подавляющее большинство известных на текущий момент антиканцерогенов – это так называемые *цитостатики* - разрушители клеток, поражающие в той или иной мере вместе с клетками опухоли и здоровые клетки).

[3] Смысл этого термина в данном случае таков: соблюдение *любым* исполняющим эти процедуры *всех* требуемых *условий* приводит его к получению субстанции, *обладающей* требуемыми *свойствами*.

СВОЙСТВ. Разработанные В.К.Финном и его учениками формальные модели, методы а также алгоритмы восстановления из эмпирических данных и анализа зависимостей вида **ОБЪЕКТ=>СВОЙСТВА** получили общее название ДСМ[4]-метод. Эта математическая техника положена в основу специального компьютерно-ориентированного инструментария интеллектуального анализа данных и прогноза свойств новых объектов, который активно и достаточно успешно применяется в различных приложениях (см., например, [*ДСМ-АПГ*] и др.). Однако, в некоторых случаях объем ДСМ-вычислений, необходимых для восстановления всех скрытых в исходных данных эмпирических зависимостей, к сожалению, оказывается неприемлемо большим. В таких ситуациях очевидным образом проявляется потребность в оптимизации ДСМ-перебора вариантов и организации надежного управления необходимыми объемами соответствующих вычислений. Связанный именно с этими проблемами специальный класс комбинаторных задач и алгоритмов будет далее проанализирован с учетом наиболее существенных его характеристик и особенностей.

Представленная задача о до-определении частично-определенного отношения будет рассматриваться нами также и в контексте проблемы формализации *эмпирической индукции*. В самом общем виде эта проблема может быть сформулирована следующим образом:

Дано: (1) объекты (заданные своими описаниями) $\boldsymbol{O}_1, \boldsymbol{O}_2, \ldots, \boldsymbol{O}_n$ каждый из которых соответственно обладает некоторыми свойствами $\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots, \boldsymbol{P}_n$;
(2) новый объект $\boldsymbol{O}_0$ и новое множество свойств $\boldsymbol{P}_0$ (о которых нет информации: обладает ли объект $\boldsymbol{O}_0$ множеством свойств $\boldsymbol{P}_0$).

Требуется: (3) оценить, имеются ли достаточные основания для формирования прогноза относительно наличия (или же, наоборот, отсутствия) у объекта $\boldsymbol{O}_0$ множества свойств $\boldsymbol{P}_0$, и (в случае наличия таких оснований)
(4) сформировать прогноз: обладает ли объект $\boldsymbol{O}_0$ множеством свойств $\boldsymbol{P}_0$.

Вопрос о достаточности оснований для принятия результатов выполняемой процедуры прогнозирования может быть уточнен, например, до формулировки следующего вида:

> *можно ли в рассматриваемом случае сформировать средства контроля надежности порождаемых результатов, сопоставимые, например, с теми средствами, которые обеспечивают надежность результатов достоверного (дедуктивного логического) вывода?*

Используя при формальном уточнении и поиске решения представленной выше задачи эмпирической индукции специальный набор эвристик[5], мы будем стремиться предложить такую их формализацию, которая позволяла бы «довести» заложенные в эти эвристики процедурные идеи до уровня достаточности оснований для принятия получаемых с их помощью результатов. Поиск требуемых условий корректности порождаемых результатов будет распространяться в том числе и на ситуации обработки малых выборок исходных данных (в частности – на ситуации, когда статистически надежные на достаточно емких выборках исходных данных модели и методы вынужденно переходят в статус *эвристики* при обработке малых, статистически незначимых выборок).

Наконец, в процессе анализа отношений **ОБЪЕКТ=>СВОЙСТВА** обсуждаемого вида нас также будут интересовать возможности оперировать с объектами *сложной внутренней*

---

[4] Отражающее взаимосвязи с известными эвристиками («канонами» индуктивного вывода) Ф.Бэкона-Д.С.Милля (см., например, [*Бэкон 35, Милль 00*] и др.).

[5] Понимая при этом термин *эвристический* (см., например, [*Нильсон 73, стр 53*] и др.) как *служащий открытию* (от греческого *heurisko - нахожу*).

структуры (с графами, цепочками символов, кортежами структурных и числовых компонентов).

1.1. Общая характеристика области ДСМ-ИАД.

В содержательном смысле, говоря об *интеллектуальном анализе данных* (ИАД)[6], выполняемом исследователем, обычно имеют ввиду опирающийся на определенную «*гибкость ума»* поиск *инвариантов* (*регулярностей*) в имеющихся данных. При этом упомянутая «*гибкость ума»* ассоциируется с известными способностями человека использовать возможности своего *интеллекта* при решении *трудных задач*, а понятия *инварианты*, *регулярности* и т.п. связывают с поиском *закономерностей* (в данном случае – эмпирических, восстанавливаемых по примерам из накапливаемых данных).

Как правило, интеллектуальный анализ данных отделяют от действий, которые следует выполнять по «жестко заданному» алгоритму, например, от инженерных расчетов. Да, такие расчеты требуют определенного уровня образованности а также известных навыков и квалификации, однако, говорить о демонстрации существенной «гибкости ума», например, при выполнении стандартных вычислений с помощью логарифмической линейки, по-видимому, будет обидной недооценкой интеллектуальных возможностей образованного инженера. «Интеллектуальность» в анализе данных – это все-таки больше, чем «типовые расчеты по формулам»: это, помимо прочего, и подбор соответствующих инструментов экспертизы, и варьирование применения таких инструментов в стремлении объективизировать получаемые результаты, и подбор адекватных средств представления данных об изучаемом явлении, и поиск областей «устойчивости» получаемых результатов при использовании различных инструментов анализа данных, и наконец, анализ и оптимизация получаемых результатов с целью выбрать из них наилучшие, причем имея при этом достаточные основания для такого выбора.

Сегодня практика окружающего нас технологического мира предъявляет обширный набор реальных задач, где возможностей даже самого высококлассного эксперта в области интеллектуального анализа данных уже недостаточно для поиска (в разумные сроки и с разумной глубиной анализа проблемы) решения. Очевидна и бесспорна необходимость усиления интеллектуальных возможностей эксперта за счет эффективного использования современных компьютерных технологий.

В таком контексте необходимость расширения (распространения) методик и технологий выполняемого экспертом интеллектуального анализа данных на компьютерную среду (программное обеспечение, инструменты представления данных и т.п.) приводит нас к задаче формализации соответствующих методик и приемов («инструментов»). Именно эта проблематика – формализация методов и средств интеллектуального анализа данных а также реализация развиваемых подходов и технологий в рамках прикладных программно-технических комплексов – будет находиться в фокусе наших дальнейших обсуждений.

Итак, основным предметом нашего рассмотрения будет описание и детальное изучение возможностей ДСМ-метода автоматического восстановления зависимостей из данных (см., например, [*Финн 10, ДСМ АПГ, Финн 11*] и др.) как *платформы* для организации интеллектуального анализа данных средствами компьютерных технологий. Употребляя термин *платформа*, мы будем иметь ввиду комплекс *методических*, *логических* (в том числе - *дескриптивных* и *аргументационных)*, *алгоритмических* и *инструментальных* средств, позволяющих организовать интеллектуальный анализ данных (в представленном выше смысле этого термина) в рамках соответствующих компьютерных технологий а также прикладных программно-технических систем.

---

[6] Как направления, связанного в том числе и с традиционной проблематикой исследований в области искусственного интеллекта [*AI*].

Переходя собственно к представлению такой платформы, первым шагом очертим набор эвристик, которые станут базой ее формирования. Центральным элементом, который положен в основу развиваемого нами подхода, является известная эвристика Ф.Бэкона-Д.С.Милля (см., например, [*Бэкон 02, Милль 00, Финн 76, Финн 81-а, Финн 83*] и др.) – методологическая схема анализа результатов опытов и формирования индуктивных обобщений, – дополненная схемами абдуктивных рассуждений в стиле Ч.С.Пирса (см., например, [*Пирс 00, Пирс 05*] и др.), а также эвристическими схемами правдоподобных рассуждений по аналогии в стиле Д.Пойа (см. например, [*Пойа 70, Пойа 75*] и др.). Именно эти методические соображения составляют основу развиваемой уже в течение нескольких десятилетий В.К.Финном и его школой логической инфраструктуры ДСМ-метода (см., например, [*Аншаков 86, Аншаков 87, Anshakov 89, Аншаков 93*] и др.).

Возвращаясь к необходимости «воспроизвести» при организации интеллектуального анализа данных компьютерными средствами уже названную нами выше как «системообразующий» признак «гибкость ума», представляется важным обратить особое внимание на следующие три группы факторов (и, соответственно, критически важных характеристик ДСМ-метода), дающих основания говорить о предоставляемых здесь возможностях действительно *интеллектуальной* обработки данных.

Прежде всего – это возможности формировать в рамках формализованного ДСМ-рассуждения (в однородной понятийной и логической среде) *семейства правил* правдоподобного *вывода*, которые можно группировать в те или иные последовательности и тем самым задавать различные *стратегии* интеллектуального анализа данных средствами ДСМ-метода. Подбирая те или иные стратегии при восстановлении эмпирических зависимостей, скрытых в каждой конкретной выборке предложенных для ДСМ-обучения примеров и контрпримеров, можно *моделировать* компьютерными средствами ту самую «*гибкость ума*» ведущего ИАД *исследователя*, которая и отличает этот способ анализа данных от всех других (возможно, не менее важных) технологий восстановления зависимостей, скрытых в изучаемых данных. «Наследуемость» результатов при обработке анализируемых данных разными стратегиями – сигнал о том, что в данном (наследуемом) подмножестве порождаемых эмпирических зависимостей удалось обнаружить *эмпирическую закономерность* (см. подробнее [*Финн 10-а*]).

Не менее существенным элементом обсуждаемой конструкции оказываются возможности выбрать (и при необходимости - варьировать) подходящий вариант языка представления данных об изучаемом явлении. Поясним этот тезис на ставшем, по-видимому» уже классическим примере. При анализе заданных физиологических активностей химических соединений (в частности, при выявлении причинных механизмов, определяющих лекарственные свойства физиологически активных веществ) принято считать, что лекарственные свойства соответствующего химического соединения полностью определяются его структурой. Однако, для формализованного представления этой структуры можно выбрать целый набор, вообще говоря - различных, выразительных средств (формализованных языков и компьютерных инструментов): так анализируемые химические структуры могут быть описаны как

- плоские симметризованные графы,
- раскрашенные пространственные графы (с характерными углами и расстояниями),
- спектры (излучения и\или поглощения)

и др. .

С вычислительной точки зрения (т.е. с учетом сложности вычислений при их обработке) это - три принципиально различных по своим возможностям языка представления данных. Так в третьем случае (см. описания с помощью спектров) мы имеем дело с достаточно простым теоретико-множественным языком, а вот в первом случае – наоборот: ряд принципиально важных задач (например, выделение подграфов определенного вида, вкладывающихся одновременно в несколько исходно заданных для анализа структур, и др.) оказываются уже трудно разрешимыми переборными проблемами. В свою очередь при

переходе ко второму из перечисленных выше языков ситуация с объемами необходимых для проведения ИАД вычислений только усугубляется. Однако, с другой стороны, выразительные возможности каждого из этих языков позволяют анализировать имеющиеся данные с различной степенью детальности описания целевого эффекта – наличия или же, наоборот, отсутствия, достаточных каузальных оснований для прогноза целевых лекарственных свойств вновь предложенных для ИАД химических соединений. В части возможностей варьировать язык представления данных в обсуждаемом примере обратим внимание, в частности, на следующее принципиально важное обстоятельство: выявление эмпирических зависимостей, «устойчиво наследуемых» при переходе от более простых к более сложным средствам представления данных, есть очевидный «сигнал» о том, что мы «нащупали» область существования соответствующей эмпирической закономерности.

Аналогичным образом, при расширении исходно заданной для ДСМ-обучения выборки примеров и контрпримеров новыми объектами эффект наследуемости части зависимостей с исходного на расширенное множество прецедентов (наряду с эффектами устойчивости такого расширения как при использовании различных стратегий ДСМ-рассуждения, так и при варьировании средств представления анализируемых данных) также представляет собою очевидный сигнал о том, что в ходе выполнения ДСМ-ИАД удалось «выявить» область *эмпирической закономерности*.

Наконец, располагая в рамках логической и алгоритмической конструкций ДСМ-метода средствами оценки по «*силе*» и соответствующего *упорядочения* восстанавливаемых эмпирических зависимостей, можно организовать *оптимизацию* порождаемых в ходе ДСМ-обучения и ИАД результатов, выбирая (в рамках *сравнения* зависимостей, формируемых на конкретной исходной выборке *всеми* возможными стратегиями ДСМ-рассуждения) либо оптимальный, либо наилучшие (в смысле введенной оценки качества) варианты эмпирических зависимостей, восстанавливаемых из исходных данных.

Таким образом, роль *инвариантов* (*регулярностей*), обозначенных в самом начале этого обсуждения как одной из характеристических особенностей области интеллектуального анализа данных, в ДСМ-ИАД отводится *эмпирическим закономерностям* (классу эмпирических зависимостей, наследуемых при варьировании средств анализа, варьировании инструментов представления данных а также при расширении исходных данных новыми примерами и контрпримерами). «*Гибкость мысли*» эксперта в области ИАД моделируется в компьютерной системе использованием *классов стратегий* ДСМ-рассуждения, а выбор наиболее подходящего (обеспечивающего *достаточные основания для принятия* полученных *результатов* ДСМ-рассуждения) заключения обеспечивается *оптимизацией*, опирающейся на оценки *качества* формируемых *эмпирических зависимостей*.

1.2. ИАД и интеллектуальные системы.

Переходя к более формализованным представлениям об интеллектуальном анализе данных сразу же условимся, что под ИАД мы будем понимать

решение ***сложных***[7] *задач* ***анализа данных*** *средствами* ***интеллектуальных систем***.

Подробное обсуждение предложенной «формулы» начнем с уточнения понятийных конструкций, использованных в этом определении, а также наиболее существенных взаимосвязей между ними.

Анализ данных (см., например, [*АД*] и др.) мы будем понимать в общепринятом смысле как "*построение и исследование математических методов и вычислительных*

[7] По-видимому, следует напомнить, что (по фактически сложившейся в области искусственного интеллекта традиции) сложными здесь принято считать такие предлагаемые компьютеру задачи, для решения которых специалисту потребовалось бы привлечение его «интеллектуальных способностей».

алгоритмов извлечения знаний из экспериментальных данных". К подробному осуждению характеристик и возможностей современных систем анализа данных[8] мы вернемся чуть позже – в **Разделе 1.4**, а прежде обсудим тот класс задач анализа данных, где применение характерных именно для исследований в области искусственного интеллекта подходов, математических моделей и методов оказывается результативным, актуальным и важным. Именно такими средствами обеспечивается получение здесь интересных и значимых прикладных результатов.

Сложными в рассматриваемом нами контексте представляется уместным считать особый класс *некорректных* вычислительных задач, характерных для исследований в области искусственного интеллекта.

Обычно в современной computer science при обсуждении возможных формализаций понятия *некорректности* в качестве примеров в первую очередь приводят хорошо изученные классы некорректных вычислительных задач, которые связаны с (фактически ставшими классическими) уточнениями понимания некорректности[9] по А.М.Ляпунову [*Боголюбов 83*] или же по А.Н.Тихонову [*Тихонов 69*].

Между тем в области искусственного интеллекта хорошо известна исследовательская ситуация (см. например, [*Попов 87*] и др. - в части практики использования так называемых *экспертных систем*), когда для конкретной решаемой проблемы могут быть предложены *различные средства экспертизы*, причем получаемые этими средствами результаты в тех или иных конкретных случаях могут не давать *однозначного* заключения (т.е. не подлежащего оспариванию - в смысле К.Поппера и А.С.Есенина-Вольпина [*Поппер 83, Есенин-Вольпин 99*] - результата решения рассматриваемой задачи, который принимался бы на *достаточном*[10] основании).

С точки зрения оценки получаемых результатов (их принятия или же, наоборот, непринятия), вообще говоря, допустимая в подобном контексте ситуация, когда, например:
- одним средством экспертизы порождается заключение об *опасности* (токсичности, канцерогенности, мутагенности и т.п.) некоторого конкретного химического соединения,
- а другим - о его *неопасности* (нетоксичности, неканцерогенности и т.п.)

требует от исследователя, как минимум, дополнительных экспертиз (в которых были бы задействованы ранее не использовавшиеся средства анализа данных и поддержки принятия решений). В "полноформатном" исследовании в подобной ситуации представляется целесообразным не только провести детальный анализ корректности (адекватности применения, правильного выполнения всех условий использования и т.п.) всех задействованных средств экспертизы, но и обеспечить сравнительный анализ порождаемых результатов. (Фактически возникает необходимость погружения[11] порождаемых различными средствами экспертизы результатов в общую среду их оценивания, что могло бы дать возможности для формирования соответствующего *отношения предпочтения* и выбора оптимального или же наилучших результатов).

Не менее существенна возможность «*структурирования*» доступных (вовлеченных в процесс ИАД) *средств экспертизы*: на начальной стадии все имеющиеся «инструменты» анализа данных, вообще говоря, «равноправны[12]» (ведь работающий с ними эксперт не имеет аргументов предпочесть один из них другому[13]). Однако, в ходе выполняемого

---

[8] Как, впрочем, и к уточнению понятия ***интеллектуальная система***.

[9] Ставшее общепринятым понимание этого термина предложено Ж.Адамаром: корректно поставленной математической задачей предлагается считать ту, решение которой *существует*, *единственно* и *устойчиво* (см., например, [*Hadamard 02, Тихонов 69*] и др.).

[10] Т.е. *неоспоримом* (в смысле А.С.Есенина-Вольпина) в рамках текущей исследовательской ситуации.

[11] Разумеется, там, где это возможно. (*Влажное* с *длинным* или же *красное* с *громким* сравнивать, по-видимому, бессмысленно).

[12] Не смотря на то, что они могут различаться по используемым моделям представления данных, оценкам вычислительной сложности и т.п..

[13] Иначе имело бы смысл использовать лишь «наиболее подходящие» в текущем случае «инструменты».

процесса ИАД могут выявляться дополнительные обстоятельства существенного характера (например, факты *устойчивости* тех или иных классов зависимостей при варьировании как языка представления данных, так и задействованных средств экспертизы), на основании которых появляются возможности:

- организовать (вообще говоря, лишь частичное) *упорядочение* используемых языков представления данных а также средств экспертизы по уровню «*качества*» порождаемых с их помощью результатов, а также
- уточнить актуальные *области корректности* экспертизы, выполняемой в данном случае (т.е. на имеющихся в данный момент исходных данных) тем или иным конкретным «инструментом» ИАД[14].

Таким образом может быть сформирована система процедур *целенаправленного* повышения надежности (уровня аргументированности, устойчивости, …) эмпирических зависимостей, порождаемых в процессе интеллектуального анализа данных.

Достаточно яркий класс некорректных задач обсуждаемого типа - это задачи о доопределении частично-определенного отношения, на которые еще в конце 70-х годов прошлого столетия обратил внимание исследовательского сообщества В.К.Финн (см., например, работы [*Финн 76, Финн 81*] и др.). Как результат развития этой проблематики далее здесь сформировалось новое научное направление, известное сегодня в связи с разработкой и использованием при решении прикладных исследовательских задач уже упоминавшегося нами ранее ДСМ-метода автоматического восстановления зависимостей из эмпирических данных.

На рубеже столетий задачи этого типа как важный класс актуальных проблем исследований в области искусственного интеллекта выделил также и Дж.Маккарти (см., например, [*McCarthy 00, McCarthy*] и др.).

Джон Маккарти обратил внимание на задачи ***phenomenal*** *data mining*, в которых основным предметом исследования оказываются не собственно отношения в данных (выборках, коллекциях, базах данных и т.п.), а отношения между *данными* и *явлениями*, *имеющим отношение* к этим данным (и не представленными в данных явным образом). Ключевой проблемой здесь оказывается возможность *выводить* (средствами компьютерных программ) из фактов заключения относительно явлений. Таким образом, следуя Дж.Маккарти, представляется важным сфокусировать интерес исследователей на следующих проблемах: что может быть выведено из фактов, какие факты релевантны для формирования таких заключений (о явлениях), наконец, как компьютерная система может формировать заключения подобного типа в хотя бы в ситуациях, когда аналогичные заключения человек легко мог бы выполнить "вручную"?

Исходными данными для вывода заключений о соответствующем явлении вступают факты и отношения между ними. Подобные данные могут быть представлены как в явном виде (как описания соответствующих прецедентов), так и в виде логических выражений в соответствующих базах знаний. Результатом использования средств *phenomenal data mining* оказываются *расширения* уже имеющихся баз данных (как за счет пополнения новой информацией - *доопределения* - ранее лишь частично-определенных отношений, так и за счет формирования новых отношений, уточняющих описание изучаемого явления). Таким образом, речь идет о компьютерном порождении из имеющихся фактов (неявным образом представленных в них) эмпирических зависимостей, характеризующих изучаемое явление. Далее, считает Дж.Маккарти, естественно продвигаться в направлении порождения из данных скрытых зависимостей (т.е. восстановления отношений) между сущностями в изучаемой предметной области, т.к. ее описание в терминах релевантных явлений будет

---

[14] А это в свою очередь позволяет двинуться в сторону еще более детальных уточнений и, возможно, улучшения качества результатов ИАД-экспертизы, используя аналоги предложенного Ю.И.Журавлевым (см., например, работы [*Журавлев 77, Журавлев 77-а, Журавлев 78, Журавлев 80*]) и др.) подхода на базе формирования так называемых корректных алгебр над множеством эвристических алгоритмов.

существенно более компактным, чем в терминах данных[15]. В представленном смысле, по мнению этого бесспорного авторитета в области computer science & artificial intellilgence, формализация рассуждений рассматриваемого типа - одна из трудных проблем в области искусственного интеллекта, а прогресс в ее разрешении пока еще не слишком заметен. Предстоит еще научиться делать многое, оперируя с ограниченными знаниями: необходимо, чтобы продвинутые системы ИАД обладали бы возможностями порождать гипотезы из фактов. По мнению Дж.Маккарти, именно этот путь открывает перспективы эффективного использования математической логики и алгебраической теории баз данных для развития технологий *phenomenal data mining*. Да, факты придется помещать в базы данных, по-видимому, "*вручную*"[16], однако, далее компьютерная система будет иметь возможности *автоматически* порождать зависимости из имеющихся фактов. В качестве примера исследований, развивающих такой подход, Дж.Маккарти ссылается, в частности, на работу [*Lyons 98*], посвященную компьютерным экспериментам в области *phenomenal data mining*.

Говоря об особенностях рассматриваемого нами класса некорректных задач, представляется целесообразным выделить следующие его характеристики:

(1) во многих случаях рассматриваемого типа *одна содержательная* постановка решаемой задачи может иметь несколько, вообще говоря, *различных формальных* постановок-уточнений (и, соответственно, несколько различных инструментов анализа имеющихся данных – средств осуществления ИАД-экспертизы);

(2) *как сравнивать* между собой результаты таких экспертиз? Что такое *наилучший* результат ИАД в этом случае[17];

(3) что такое *достаточность оснований* (в смысле К.Поппера и А.С.Есенина-Вольпина) для принятия результатов при решении рассматриваемого класса задач? Как быть с *множественностью* вариантов формализации исходной содержательной задачи и, соответственно, с множественностью порождаемых здесь решений?

(4) критически важный составной элемент комплекса средств решения задач обсуждаемого типа - это *машинное обучение* на примерах и контрпримерах (как инструмент *порождения* частичных *зависимостей* из исходных фактов, предложенных для обучения);

(5) умение (т.е. наличие средств) работать с малыми выборками (в том числе - сложно структурированных объектов - графов, цепочек символов и т.п.);

(6) проблема *сложности вычислений* (этой области исследований присущи, как правило, "плохие" оценки сложности вычислений. Как следствие, необходимо предложить эффективные средства управления перебором для решения задач в прикладных предметных областях);

(7) проблема экспертизы конкретного объекта (причем *точно* его, а не группы аналогичных объектов *в среднем*);

(8) большое число «дыр» (пропусков, недоопределенных значений) в таблице доопределяемого отношения *data=>phenomena*[18]. В таких ситуациях, как правило, нет оснований говорить о «заслуживающих доверия» оценках частот наблюдаемых событий (и

---

[15] См. актуальность такого подхода по отношению к проблематике *Big Data*.

[16] Как позитивный пример такой работы Дж.Маккарти приводит проект CyC [*Lenat 90*].

[17] Поучительный пример невнимания к этой проблеме - ситуация с ведением ИАД средствами так называемых деревьев решений (см., например, и [*Pearl 87, Kodratoff 92 , Kodratoff 92-a, Quinlan 86, Quinlan 88, Дюк 04*] др.). Здесь результат ИАД, вообще говоря, зависит от того, в каком порядке будут анализироваться имеющиеся данные (прецеденты). В свою очередь, проблема выбора оптимального упорядочения (а с ним - и формирования оптимального дерева решений) оказывается трудной переборной задачей (принадлежит классу *NP*-полных комбинаторных проблем - см. [*Hyafil 76, Murthy 98*] и др.). Наконец, при расширении исходной выборки прецедентов новыми объектами, вообще говоря, придется пересчитывать ранее сформированное оптимальное дерево решений: свойство оптимальности не "наследуется" при расширении исходной выборки и не поддерживается инкрементальными модификациями ранее сформированных древесных структур этого типа.

[18] См. выше - *phenomenal data mining*.

тем более об оценках сходимости частот к вероятностям: выборки заданных для обучения примеров и контрпримеров могут быть еще слишком малы для надежного применения техники статистического анализа данных);

(9) при обработке *больших выборок* исходных данных заключение относительно конкретного нового объекта требуется не «*в среднем* (*по палате*)», а именно для него – этого конкретного объекта. Здесь «в лоб» машинное обучение на базе точных алгоритмов исчерпывающего перебора вариантов на всей выборке на практике не реализуемо из-за проблем со сложностью вычислений, а пре-процессинг статистическими средствами (т.е. отбор совокупности объектов, в *среднем* релевантных анализируемому, для дальнейшего *уточняющего* перебора и поиска решения машинным обучением) еще требуется организовать так, чтобы обеспечить *достаточные основания для принятия* порождаемых результатов;

(10) для рассматриваемого класса задач характерна (и весьма существенна - !) так называемая *немонотонность рассуждений* (эффекты *ненаследуемости* уже порожденных зависимостей при расширении обучающей выборки новыми объектами).

1.3. Некоторые примеры актуальных задач ИАД

Для иллюстрации приведем несколько примеров прикладных предметных областей, где обсуждаемая проблематика (решение некорректных задач рассматриваемого нами типа) представляется весьма актуальной.

(а) **Анализ свойств** (физиологических активностей) **химических соединений**.

Изучается взаимосвязь структуры химического соединения и физиологических активностей (например, лекарственных или же каких-либо иных полезных свойств), которыми данное соединение обладает[19]. Зависимости **СТРУКТУРА=>АКТИВНОСТИ** анализируются обычно по представленной ниже схеме:

Исходные данные.

- Набор соединений[20], обладающих заданными (анализируемыми) свойствами.
- Набор (возможно – не содержащие ни одного объекта) соединений, которые структурно «похожи» на позитивные примеры, однако целевыми свойствами не обладают.
- Набор новых соединений, (предложенных для прогноза) относительно которых требуется сформировать аргументированное заключение о наличии (либо отсутствии) у них целевых свойств.

Требуется.

- Сформировать *аргументированное* заключение о наличии (либо отсутствии) целевых свойств у новых (предложенных для прогноза) соединений.

Изучением и прогнозом свойств по подобной схеме традиционно занимается исследовательское сообщество, ориентированное на *поиск количественных соотношений структура-свойство* (см., например, [*QSAR*] и др.). В фокусе внимания здесь - процедуры построения моделей, позволяющих по структурам химических соединений предсказывать их разнообразные свойства[21]. (В специальной литературе за моделями, позволяющими

---

[19] Или же наоборот (например, в случае анализа негативных свойств – токсичности, канцерогенности, мутагенности и т.п.), не обладает. Причем исходная информация о наличии (или же отсутствии) изучаемых свойств у анализируемых соединения имеет статус *фактов*, зафиксированных в соответствующим образом (например, эти сведения могут быть подучены по результатам соответствующих экспериментов, из литературных источников и т.п.).

[20] Их обычно называют *позитивными* примерами зависимости **СТРУКТУРА => АКТИВНОСТИ**.

[21] Т.е. в основу развиваемого подхода положена известная гипотеза А.М.Бутлерова о химическом строении веществ, утверждавшего, что свойства каждого химического соединения определяются его структурой (см., например, [*Бутлеров 53*] и др.).

прогнозировать количественные[22] характеристики биологической активности, закрепился отдельный англоязычный термин *Quantitative Structure-Activity Relationship* - QSAR. В области приложений объем этого понятия (см. например, [*QSAR*] и др.) часто трактуется расширенно, охватывая любые модели вида **СТРУКТУРА=>СВОЙСТВА**.)

Описание представленного набором фактов частично-определенное (заданное множеством исходных примеров и контрпримеров) отношение **СТРУКТУРА=>СВОЙСТВА** формализуется с помощью достаточно развитого набора математических моделей (см. например, [*QSAR, Голендер 78, Филимонов 06*] и др.). Далее каждая из таких моделей, фактически, используется для решения двух «технических» задач:

- формирования[23] «унифицированного» описания («*интерполяции*») исходно заданного множества *примеров* и *контрпримеров*; а затем
- формирования (там, где это возможно) *экстраполяции* только что построенного «унифицированного» описания на *вновь предложенные* для анализа (т.е. прогнозирования свойств) объекты – химические *соединения*.

С точки зрения обсуждаемой нами проблематики – характеристик и особенностей рассматриваемого класса *некорректных* задач анализа данных – представляется уместным обратить внимание на следующее критически важное обстоятельство:

> при формализации описания структур анализируемых химических соединений можно выбрать по крайней мере *несколько* ***различных*** (т.е. опирающихся на различные физико-химические основания) способов описания имеющихся фактических данных.

В качестве примера приведем (уже упоминавшиеся нами) два несводимых один к другому хорошо известных (и активно используемых в приложениях) варианта проблемно-ориентированного описания структур химических соединений:

- спектры (излучения и\или поглощения);
- структурные графы.

Каждый из названных языков представления данных позволяет получать заслуживающие доверия результаты при решении конкретных прикладных задач. Однако, по своей природе эти подходы принципиальным образом различаются как в части «химии» и «физики» задействованных при их формировании эффектов, так и в плане вычислительной сложности алгоритмов, обеспечивающих обработку представленных ими эмпирических данных. Так с точки зрения алгоритмики анализа данных представления на базе спектров – дескрипторный язык (т.е. – теоретико-множественное описание), для которого ряд типовых вычислительных задач[24] рассматриваемой предметной области имеют приемлемые[25] оценки сложности вычислений. В то же время представления структур как графов – это уже целое семейство языков описания анализируемых данных: можно описывать химические соединения как, например,

- плоские симметризованные графы,
- пространственные симметризованные графы,
- «раскрашенные»[26] пространственные графы с углами и расстояниями.

---

[22] Более корректно здесь следовало бы говорить (см., например, [*Сапегин 92*] и др.) о построении последовательности моделей, где сперва формируется *качественная* (*реляционная*, связывающая основные понятия и отношения – *факторы влияния* и качественные *взаимосвязи* между ними) модель, а уже далее эта *базовая* модель уточняется (там, где это возможно) за счет соответствующих метрических соображений (*метризуемых* параметров и их *количественных* значений).

[23] В том числе – с использованием математических моделей и алгоритмов компьютерного порождения *зависимостей* из *фактов*.

[24] Например, задача выделения общих частей (сходства) в описаниях имеющихся примеров\контрпримеров полиномиально разрешима.

[25] Для эффективного использования этой техники анализа данных в практически значимых приложениях.

[26] Что предполагает наличие проблемно-ориентированных меток у соответствующих вершин и ребер.

При этом, например, выделение множества всех общих подграфов даже для пары плоских симметризованных графов – это уже вычислительно сложная комбинаторная проблема[27], для преодоления влияния которой (в частности - в ряде приложений) приходится использовать те или иные специальные приемы. Так в целом ряде работ (см., например, [*Авидон 74, Голендер 78, Лейбов 91*] и др.) для описания графов химических структур используется специальный проблемно-ориентированный дескрипторный язык – так называемый код ФКСП[28], который сформирован примерно 20 000 подграфов вида

дескрипторный центр 1 --- *углеродная цепочка* --- ***дескрипторный центр*** 2,

а соответствующий каталог дескрипторных центров содержит примерно семь десятков типовых подграфов (фрагментов физиологически активных химических соединений), ориентированных на описания биологических рецепторов. Код ФКСП позволяет представлять анализируемые химические структуры как покрытия соответствующих химических графов множествами [29] подграфов из соответствующего каталога. Таким образом, появляется возможность использования теоретико-множественного языка (вообще говоря – приближенного - описания исходных данных) для работы с плоскими симметризованными графами.

Даже не принимая во внимание весь приведенный в [*QSAR*] список используемых в рассматриваемой нами прикладной области математических моделей и методов и ограничившись лишь только что перечисленным выше набором инструментов, опирающихся на спектральные или же графовые [30] представления исходных данных, можно оказаться в ситуации, когда экспертиза одним инструментом дает результат одного типа (например, утверждает, что анализируемое соединение *будет обладать* целевыми свойствами), в то же время экспертиза инструментом другого вида, вообще говоря, может давать результат противоположного типа (*отрицательный* прогноз по целевым свойствам). Нетривиальной проблемой в такой ситуации оказывается корректное сравнение этих двух формализаций (и полученных в их рамках результатов): какой прогноз *предпочесть*? Каковы *основания* для *доверия* этому прогнозу?

Приведенные аргументы дают очевидные основания говорить о задаче а**нализа свойств** (физиологических активностей) **химических соединений** как об одном из примеров рассматриваемого нами особого класса некорректных вычислительных задач, характеризующих проблематику интеллектуального анализа данных.

(в) **Текст-процессинг** (компьютерный анализ некоторых семантических характеристик отношения ***ТЕКСТ ~ СМЫСЛ***: отнесение текстов к темам, аннотирование, оценка эмоциональной окраски текстов и др.).

Говоря о текст-процессинге в самом общей форме, мы будем иметь ввиду компьютерный анализ взаимосвязей *синтаксических* объектов – текстов на естественном языке – с

---

[27] См., например, [*Гэри 82*] и др. .

[28] Фрагментарный Код Суперпозиции Подструктур.

[29] Если необходимо – нумерованными (т.е. допускающие возможности кратного вхождения того или иного элемента в соответствующее множество).

[30] Мы сознательно избегаем здесь обращений к регрессионным методам анализа связи **СТРУКТУРА => АКТИВНОСТИ**. В случае использования этой техники ситуация оказывается еще более запутанной: ведь в случае порождения регрессионных зависимостей нам приходится иметь дело уже не с *полным* набором *каузальных* факторов, определяющих своим влиянием наличие (или же отсутствие) анализируемого целевого свойства, а лишь с некоторым набором *объяснительных* переменных, используемых для формирования подходящего аналитического приближения наперед заданной точности. Т.е. в этом случае приходится иметь дело, вообще говоря, с некоторым артефактом, причем вопрос о том, насколько *корректно* этот артефакт отражает «*физику*», «*химию*» и «*биологию*» изучаемой связи **СТРУКТУРА => АКТИВНОСТИ,** в каждом конкретном случае должен рассматриваться отдельным образом. (Ведь результаты *регрессионного* анализа возможно, придется сравнивать с результатами экспертизы на основе описания *спектров* или же *графового* представления!).

семантическими конструкциями[31], отражающими (в той или иной мере) содержание этих текстов. В основе такого компьютерного анализа – комплекс подходов и моделей, наиболее известными составными элементами которого, по-видимому, являются

- модель ***СМЫСЛ ~ ТЕКСТ***, предложенная И.А.Мельчуком (см., например, [*Мельчук 95, Мельчук 99, Падучева 05, СТ 12, Большаков 00*] и др.) и развитая далее в существенной степени также Ю.Д.Апресяном и его коллегами (см., например, работы [*Апресян 74, Апресян 88, Апресян 92*] и др.), а также
- набор лингвистических формализмов (порождающие грамматики четырех типов и др.), предложенных Н. Хомским (см., например, [*Chomsky 82, Хомски 72, Хомски 72-а*] и др.).

Одним из оснований этого подхода является идея выделения *семантических примитивов* (слов и словосочетаний), формирующих систему так называемых семантических атомов (сем). Именно через них предполагается выражать любые смыслы (неявным образом) представленные в соответствующих текстах. Для этого предлагается соответствующая модель языка, позволяющая на основе строгих правил структурировать смыслы при переходе от языковых примитивов к сложно структурированным комплексам (построенным с «участием» таких примитивов) – текстам[32]. При этом считается, что информация о смыслах текстов не возникает в процессе функционирования языковой модели, а лишь принимает иную форму, трансформируясь от текстового представления к смысловому (или же наоборот)[33]. В рамках такого подхода критически важная роль выпадает «каноническим описаниям» смысловых примитивов – специализированным словарям[34] – а также правилам формирования отношений между такими примитивами (нормативной информации о связях[35] между словами и группами слов). Правила используемых здесь порождающих грамматик определяют варианты специальной разметки синтаксических отношений между словами (при этом различие в разметке жестко связано с различием в смыслах, которые сопоставляются соответствующим синтаксическим объектам).

Одной из наиболее успешных областей применения этого подхода оказалась проблематика машинного перевода с языка на язык (см., например, [*Апресян* 88, *Апресян* 92] и др.). Формируемые в процессе по-фразного перевода специальные проблемно-ориентированные графы связей между использованными в тексте словами (словарными конструкциями) обеспечивают высокое качество результата, порождаемого по схеме:

***фраза на языке* 1** =>
=> **семантическая сеть** (отражающая смысл этой фразы) =>
=> ***фраза на языке* 2**

Еще одной областью успешного использования рассматриваемых нами технологий текст-процессинга стала проблематика так называемого *text-mining'*а – компьютерного

---

[31] Ряд источников (например - [*Sells 85*]) обобщает такое представление до цепочки лингвистических связей, в которой на входе – набор звуков (символов соответствующего языка), а на выходе – значение (сопоставляемое этому набору символов).

[32] Фактически речь идет о системе правил перехода от текста к смыслу, обеспечивающих интерпретацию («*понимание*») текста. Формируемое такими правилами семантическое представление является размеченным графом («*семантической сетью*»), сопоставленным синтаксическому представлению текста в виде дерева («*дерева зависимостей*»).

[33] См. *трансформации* у Н.Хомского или же *эквативные преобразования* у И.А.Мельчука.

[34] Толково-комбинаторным, синонимическим словарям. (Толково-комбинаторный словарь формируется толкованиями слов а также сопоставленными им моделями управления. Обычно толкование – это запись на формализованном метаязыке; где семантически более сложные элементы объясняются через более простые. Варианты машинных словарей подобного типа, активно используемых в компьютерной текст-процессинге, дают, в частности, в работы [*СО-СО 87, Лахути 97, Белоногов 10*] и др.).

[35] Так называемые машинные грамматики формируются на базе описания выделяемых из словарей общих свойств соответствующих классов семантических примитивов (лексем).

анализа полнотекстовых документов на естественном языке, сфокусированная на идентификацию в тексте (и выделение в отдельную систему) «индексов» - характеризующих его содержание понятий и отношений. (В числе первых публикаций на эту тему можно назвать, например, работы - [*IBM-TM, Gotthard 97*] и др.. Среди наиболее интересных современных работ – публикации о разработке и экспериментальном исследовании функционирования компьютерной системы IBM Watson – см., например, [*IBM Watson, IBM Watson-a, Семенихин 14*] и др.). Дополненная эффективным ростом производительности вычислительных установок эта технология компьютерного *text-mining'*а открыла качественно новые перспективы в ставшей к тому времени уже классической проблематике информационно-поисковых систем[36]. Разветвленные структуры полнотекстовых индексов, автоматически формируемых компьютерными системами *text-mining'*а, стали основой для создания большого количества прикладных систем автоматической обработки больших коллекций текстов: сперва - систем автоматического отнесения (отбора из больших массивов) текстов к заданной теме, затем – уже собственно систем автоматической классификации текстов по темам, а далее – даже, например, автоматического аннотирования или же оценки «эмоциональной окраски» текстовых документов[37].

И вот именно здесь мы уже в области *text-mining'*а вновь сталкиваемся с обсуждаемым нами особым классом некорректных задач. В этих исследованиях и разработках он может быть представлен следующей схемой: имея примеры связи (частично-определенного отношения)

ОБЪЕКТ (полнотекстовый документ) =>
=> ***СВОЙСТВА*** объекта (его принадлежность к заданной *тематической* группе, его заданная *эмоциональная* окрашенность[38] и т.п.),

необходимо оценить наличие (или же отсутствие) анализируемого свойства у нового объекта (возможно, выполнив такую оценку на *большом* массиве предъявляемых для анализа документов). Нетрудно убедиться, что

- для формализации описания объекта (полнотекстового документа) имеются возможности выбрать, вообще говоря, *различные* (несовпадающие, несводимые в точности одна к другой) системы индексов (семантических примитивов); а кроме того
- использовать для формального анализа связи ***ОБЪЕКТ*** => ***СВОЙСТВА*** *разные* математические модели и инструменты[39] анализа данных.

Более того, на основе одной и той же системы исходных (базовых) индексов можно построить несколько различных вариантов описания анализируемых текстов, например, стартуя с описания документа как множества встречающихся в тексте понятий и отношений (из наперед заданного «максимального» набора таких «смысловых дескрипторов»), а далее переходя к учету (в выполняемом анализе и порождении прогнозов по свойствам) возможных отношений между такими дескрипторами (а также их смысловыми группами). Таким образом, и в области текст-процессинга возникает уже знакомая нам проблематика перехода от одной содержательной постановки исследовательской задачи к множеству *различных* ее формальных уточнений, а вместе с нею – и к уже представленному нами ранее

---

[36] Эта область information & computer science имеет обширную историю и библиографию, начиная со второй половины пятидесятых годов 20-го века (см., например, [*Ланкастер 72, Маннинг 11*] и др.).

[37] Примеры таких работ нетрудно найти, например, в материалах регулярных конференций ***COLING*** (*International Conference on Computational Linguistics*), ***EACL*** (*Conference of the European Chapter of the Association for Computational Linguistics*), ***EMNLP*** (*International Conference on Empirical Methods in Natural Language Processing*) и др..

[38] Например, оцениваемая как *ирония* или же *сарказм* (см., в частности, [*Rosso 14, Reyes 14*] и др.)

[39] Например, *регрессионный анализ* vs. дискретные методы *машинного обучения* на примерах и контрпримерах.

комплексу вопросов о выборе наилучшего решения, оценке достаточности основания для его принятия и т.п. .

В практических приложениях этот комплекс вопросов обычно отодвинут в сторону в силу сугубо прагматических соображений: при работе с большими массивами текстовых документов пользователь ожидает от компьютерной системы в первую очередь сокращения перебора при поиске релевантных запросу документов. При этом, обычно, параметры поиска стараются настроить так, чтобы максимальным образом обеспечить полноту поиска (стараясь поддерживать при этом показатели точности на приемлемом уровне, – ведь далее отбирать релевантные документы из сформированной компьютерной системой выборки приходится «вручную»). Разумеется, успех (или же, наоборот, неуспех) здесь определяется прежде всего *искусством* пользователя в части *подбора* подходящих параметров поиска. Подобная деятельность, по-видимому, вполне обоснованно может быть отнесена к области, которую Ю.Д Апресян назвал *экспериментальной лингвистикой*, возникшей на базе использования компьютеров: эксперимент здесь «оказывается главным, а его результаты напрямую зависят от того, какие структуры выбраны для описания языка и какая последовательность преобразований рекомендована теорией» [*Большаков 00*]. Однако, с точки зрения уважения к «поиску истины» следует признать, что удовлетворительных ответов на сформулированные нами выше (при характеристике рассматриваемого здесь специального класса некорректных задач) вопросы представленные технологии текст-процессинга (в их текущем состоянии развития) пока что, к сожалению, не дают.

(c) **Диагностика и атрибуция целенаправленных активностей** (диагностика не являющихся флюктуациями сбоев в работе сложных технических комплексов; атрибуция исторических артефактов; идентификация фактов компьютерных атак и их атрибуция).

В самом общем виде задача атрибуции может быть представлена следующей схемой:
- имеется набор объектов, о которых известно, что каждый из них сопоставлен некоторому определенному факту (в частности, набор документов, каждому из которых сопоставлены определенные даты[40] и авторы[41], или же, например, набор сбоев в работе компьютерных систем и факта идентификации каждого такого сбоя как целенаправленного вредоносного воздействия на соответствующую систему[42], наконец, это могут быть, например, образцы почерка, каждому из которых сопоставлен конкретный персонаж - автор[43]);
- предъявлен также новый объект (подобного представленным для «обучения» вида), для которого требуется сформулировать аргументированное суждение о его сопоставимости этому же факту (например, суждение о том, можно ли сформировать заключение о датировке и авторстве конкретного нового исторического документа, можно ли идентифицировать новый сбой в работе компьютерной системы как атаку, наконец, можно ли идентифицировать автора соответствующего рукописного текста?).

В такой схеме, как и выше – в примерах (а) и (в) – приходится иметь дело с частично определенным (заданным набором примеров и, возможно, контрпримеров) отношением

ОБЪЕКТ => его **АТРИБУЦИЯ** ,

---

[40] Периоду его создания.

[41] В данном случае речь идет о задаче атрибуции (исторических) документов. Один из наиболее известных примеров работ этого типа – работы А.А.Зализняка и В.Л.Янина палеографической по датировке берестяных грамот [*Зализняк 00*]. (В дополнение представляется также интересным отметить работу [*Гусакова 04*], где аналогичные результаты воспроизведены средствами интеллектуального анализа данных, причем - без использования не всегда абсолютно бесспорных заключений относительно частот появления признаков в том или ином анализируемом документе, формируемых в рамках палеографического метода).

[42] В случае идентификации прецедентов (фактов) компьютерных атак.

[43] В случае идентификации авторов рукописных текстов по образцам почерка.

а основной предмет интереса состоит в до-определении этого отношения[44] в новой «точке» (или же заданном множестве «точек») – на новой объекте (или же заданном множестве таких объектов). Важнейшей отличительной особенность этой двух-шаговой процедуры

интерполяция («обучение») на заданных прецедентах + экстраполяция на новый прецедент

оказывается то обстоятельство, что на «входе» мы, как правило, располагаем лишь содержательным (неформализованным - !) описанием каждого из заданных для «обучения» прецедентов (фактов), и выбор соответствующего формализованного описания может оказаться существенно *неоднозначным*. Поучительный пример ситуаций подобного типа – проблематика диагностики сбоев в работе сложных технических систем, ориентированная на каузальный анализ возникновения соответствующих инцидентов и организацию противодействия влиянию соответствующих риск-факторов с целью обеспечения бесперебойного функционирования соответствующего технического комплекса. Тематика эта обсуждается уже несколько десятилетий (см. например, [*Davis 83*] и др.), однако, что же в итоге современные computer science и ИТ-индустрия могут предложить в качестве надежной и эффективной технологии, например, для диагностики нештатного функционирования и отладки крупных компьютерных сетей? К сожалению, - очень немногое [45]. Так, например, сегодня в отсутствие общепризнанных моделей угроз [46] штатному функционированию компьютерной сети (как, впрочем, и нормальной работе более широкого класса сложных технических объектов комплексной архитектуры) в приложениях можно обнаружить целый спектр соответствующих моделей (формализованных описаний «природы» изучаемых явлений – технических сбоев), сформированных на их основе языков представления информации о накапливаемых прецедентах и, разумеется, инструментальных средств анализа данных и поддержки принятия решений. Как следствие, вообще говоря, уместен вопрос: ка сравнивать между собою результаты подобных разнородных экспертиз? Что можно считать достаточным основанием для принятия результатов экспертизы, полученных тем или иным инструментом анализа данных?

Особую деликатность обсуждаемая нами проблема атрибуции соответствующих активностей имеет в области *идентификации* фактов компьютерных *атак* и формулирования *аргументированных* заключений относительно их источника. История вывода из строя обширных комплексов оборудования Бушерской атомной станции (Иран)[47], официальное обращение руководства Эстонии в штаб-квартиру НАТО с предложение атаковать (в соответствии 5 и 6-й статьями Договора об образовании НАТО) Российскую Федерацию в ответ на известные проблемы, зафиксированные в ИТ-инфраструктуре Правительства Эстонии в период переноса Мемориала памяти советским солдатам, павшим при освобождении г.Таллина от фашистских войск, наконец, принятое в сентябре 2014 года на саммите НАТО

---

[44] Единство заданных для обучения объектов (их сходство) в данном случае выражается в том, что характеризующая это единство информация здесь распределена по всей совокупности объектов, обладающих этим свойством (носитель этого единства – *целостное* множество характеристик, встречаемость элементов которого в отдельных объектах и обуславливает их «принадлежность» к такому единству [*Гусакова 04*]).

[45] См., например, [*SDN&NFV-2013, SDN&OF-2014, SIGCOM- 2014*] и др.

[46] Одним из наиболее распространенных подходов здесь, по факту, является так называемый *Root Cause Analysis* (см., например, [*RCA*] и др.) Однако, при детальном изучении выясняется, что эта технология не только «грешит» излишним вниманием к субъективным оценкам анализируемой ситуации экспертами, но и не предоставляет возможностей, например, для объективной идентификации *комбинаций* влияний факторов, каждый из которых самостоятельно не является причиной соответствующего сбоя, однако определенные варианты их совместного воздействия уже опасны и ведут к нарушению нормальной работы оборудования. (Так, например, при анализе сбоев в крупных системах хранения данных установлено, что определенный уровень *влажности*, *запыленности* и нештатных *вибраций* каждый сам по себе не ведут к отказам техники, однако, некоторые конкретные их комбинации оказывают фатальное влияние на нормальную работу соответствующего оборудования).

[47] См. например, [*Sanger 12, Sanger 13*] и др. .

в Уэльсе (Великобритания) решение о распространении принципа коллективной обороны стран-участниц северо-атлантического альянса и на кибератаки – вот лишь некоторые доводы, подчеркивающие критически важный характер подобной проблематики.

Говоря о технической стороне проблемы необходимо в первую очередь уделить внимание трем задачам:

- идентификации собственно самого факта компьютерной атаки[48],
- идентификации (атрибуции) источника вредоносного воздействия[49],
- идентификации каузальной природы реализуемого в рамках этой атаки вредоносного воздействия с целью выделения эффективных факторов противодействия и формирования адекватного «ответа» (решения по противодействию соответствующим факторам влияния).

Ключевым элементом в организации необходимого здесь анализа данных и выработки мер противодействия оказывается выбор средств формализации описания соответствующих прецедентов (идентифицированных фактов осуществления компьютерных атак) – того или иного языка (или даже – системы языков) описания анализируемых явлений. Особую техническую остроту проблеме придает необходимость в целом ряде случаев оперировать здесь малыми (т.е. статистически незначимыми) выборками сложно структурированных объектов (комплексных описаний поведения соответствующих технических систем, собранными в так называемые технологические базы данных коллекциями *log*'ов, *trace*'ов и т.п.). С формальной точки зрения – это, к сожалению, не только ситуация так называемых Big Data: здесь в обилии накапливаемых технологических данных еще следует углядеть «*регулярности*», имеющие *отношение*[50] к решаемой содержательной задаче, убедиться в «полноте» описания *предполагаемой* «*физики эффекта*» средствами используемого языка представления данных, научиться сравнивать результаты экспертизы, сформированные различными по своей природе инструментами компьютерного анализа данных.

Хочется надеяться, что развиваемые в рамках интеллектуального анализа данных подходы, модели и методы помогут продвижения к новым интересным результатам и в этой области приложений с характерными для нее вариантами некорректных компьютерных задач обсуждаемого нами типа.

1.4. Интеллектуальные системы и системы анализа данных в ИАД

Теперь перейдем к уточнению используемого нами понятия ***интеллектуальная система.*** Следуя предложенному В.К.Финном (см., например, [*Финн 04*] и др.) пониманию этого термина, будем считать, что архитектура *интеллектуальной системы* формируется тремя основными функциональными блоками:

- *Решателем* задач
- встроенным *Хранилищем Информации*, и
- подсистемой *интерактивного взаимодействия* с пользователем.

В свою очередь,

Решатель задач объединяет подсистемы

---

[48] Ведь известно, что власти Ирана до появления соответствующей информации о вирусе Stuxnet в международных профессиональных источниках выход из строя примерно 20% центрифуг, используемых в технологическом процессе обогащения урана квалифицировали не как результат целенаправленного вредоносного воздействия, а как технические сбои, обусловленные недостаточным уровнем качества оборудования и ошибками в действиях эксплуатационного персонала.

[49] В случае компьютерной атаки, например, на критическую инфраструктуру того или государства (ядерные электростанции, дамбы, систему стратегических коммуникаций и т.п.) вопрос об идентификации источника такой атаки (не является ли им так называемая «третья сила» - та или иная террористическая группировка и т.п.) в определенных случаях может оказаться вопросом о жизни и смерти всего человечества.

[50] В некоторых случаях (как это было, например, в ситуации с Бушерской АЭС) это удается сделать лишь объединив имеющиеся в наличии данные «технического» анализа с информацией из так называемых «традиционных» источников (в том числе - разведки).

- поддержки формализованных рассуждений - *Рассуждатель*,
- исполнения вычислений (числовых расчетов) – *Вычислитель,*
- формирования и управления стратегиями анализа данных – комбинациями используемых средств экспертизы, релевантными целям ИАД, реализуемого средствами интеллектуальной системы, - *Синтезатор* стратегий ИАД.

В свою очередь, *Рассуждатель* – это компьютерная подсистема поддержки формализованных рассуждений, обеспечивающая синтез различных познавательных процедур – «эмпирической индукции, аналогии и дедукции; способности к выдвижению гипотез; … способности отвечать на вопрос «почему», которая может быть реализована посредством абдуктивного объяснения начального состояния базы фактов; способности к обучению на основе позитивных и негативных примеров; способности к фальсификации полученных результатов, и, наконец, способности к адаптации в соответствии с изменением множества фактов и знаний» ([*Финн 04*]).

Встроенное *Хранилище Информации* объединяет

- *Базу Фактов* - подсистему накопления информации о примерах и контрпримерах, используемых *интеллектуальной системой* для восстановления эмпирических зависимостей из данных, и
- *Базу Знаний*, накапливающую релевантные цели ИАД сведения[51] в виде метаданных, зависимостей, гипотез и т.п. и формализованных в декларативной, процедурной или же концептуальной (понятийной) форме.

Подсистема поддержки *интерактивного взаимодействия* с пользователем отвечает за функции интерфейса (в широком понимании этого термина):

- визуализацию исходных данных и порождаемых результатов,
- поддержку интерактивного диалога с пользователем (выполняющим ИАД экспертом),
- сервисы обучения навыкам работы с системой и подсказки вариантов действия в случае потребности (тех или иных «трудных» для пользователя ситуациях).

Следует отметить, что в предлагаемом выше смысле понятия *интеллектуальная система* и *система искусственного интеллекта* (традиционно используемое в области *Artificial Intelligence* - AI), вообще говоря, не являются синонимами. Исторически сложилось[52] так, что к классу *систем искусственного интеллекта* принято относить *любые* компьютерные системы, *релевантные* проблематике исследований в области AI[53]. Однако, с учетом современных представлений об *интеллектуальности* компьютерных систем[54] такой взгляд на проблему вряд ли можно считать безупречным. Полезные примеры здесь можно найти, в частности, в проблематике ***компьютерного анализа данных*** – при решении задач восстановления (автоматического формирования средствами компьютерных систем) зависимостей из накапливаемых эмпирических данных.

Так в этой активно развиваемой уже более 20 лет[55] области сегодня используется достаточно разнообразное множество математических моделей и методов (а вместе с ними – и соответствующих алгоритмических конструкций). Современные технологии *компьютерного анализа данных* (как составной элемент систем анализа данных – в том

---

[51] Для уточнений этого термина можно воспользоваться, например, предложенным в работе [*Финн 04*] достаточно детально формализованном понятием *система знаний для интеллектуальной компьютерной системы*.

[52] См., например, [*C&T, Nilsson 71, Simon 91*] и др.

[53] Одним из аргументов в пользу такого понимания термина *система искусственного интеллекта* может служить, например, известный вариант ответа на вопрос о том, что такое *политическая экономия*, который был предложен Д.С.Миллем (см., например, [*Милль 07*] и др.).

[54] См., например, уже цитированную работу [*Финн 04*], а также сборник [*НИИ 04*], работу [*Забежайло 14*] и др. .

[55] См., например, [*PA & DM, Баргесян 09*] и др..

числе и *интеллектуального*) развиваются прежде всего вокруг следующих четырех классов функциональных характеристик[56]:

- возможностей уже на старте процесса ИАД убедиться в его потенциальной результативности[57] (в частности, в наличии, например, возможностей сформировать доказательство «теоремы существования» для формируемого решения);
- возможностей управлять сложностью вычислений (в частности, возможностей в ситуации доказуемо сложного комбинаторного перебора задействовать средства оптимизации вычислительной сложности алгоритмов формирования финального решения – например: эффективного выделения полиномиально разрешимых частных подклассов соответствующей общей комбинаторной проблемы, сокращения перебора за счет порождения приближенных решений и т.п.);
- возможностей вести каузальный анализ (т.е. выделять множество всех факторов, влиянием которых фактически и определяется «вид» соответствующих эмпирических зависимостей, подбирать эффективные средства противодействия негативному влиянию соответствующей части таких факторов и т.п.);
- возможностей формировать заключение ИАД-экспертизы *в точности* для выбранного единичного объекта (пациента, сбоя в нормальной работе оборудования и т.п.) а не в «*среднем* по группе аналогичных (однотипных)» объектов.

В множестве известных на сегодняшний день математических «техник» и созданных на их основе программных систем анализа данных представляется целесообразным в первую очередь обратить внимание на следующие пять направлений:

(1) Системы восстановления статистических зависимостей из числовых данных [58], базирующиеся на предложенном В.Н.Вапником и А.Я.Червоненкисом комплексе математических моделей и методов так называемой *структурной минимизации эмпирического риска* (см., например, работы [*Вапник 74, Вапник 79, Вапник 84, Vapnik 98, Vapnik 00*] и др.). В процессе разработки так называемого *метода обобщенного портрета* для распознавания образов здесь было найдено условие равномерной сходимости частот к вероятностям событий, обобщенное далее до условия равномерной сходимости средних к математическим ожиданиям по семейству случайных величин. В рамках этого подхода была сформирован математический инструментарий, позволяющий оценивать «емкость» и прогностические возможности заданных соответствующими параметрами семейств алгоритмов распознавания[59]. Была показана возможность не только обосновать *сходимость* класса методов обучения, базирующихся на *структурной минимизации эмпирического риска*, но и получить оценки *скорости* такой сходимости (например, для соответствующих систем кусочно-линейных решающих правил распознавания, обеспечивающих *минимизацию ошибок* на предложенных для обучения исходных данных).

(2) Математическую технику так называемых *корректных алгебр* над множеством некорректных (эвристических) алгоритмов, предложенную Ю.И.Журавлевым (см., например, работы [*Журавлев 77, Журавлев 77-а, Журавлев 78, Журавлев 80*]) и успешно развиваемую его школой[60]. Здесь в ситуации, когда для анализа данных сперва имеется лишь набором, вообще говоря, некорректных (эвристических) алгоритмов, обеспечивающих каждый корректное решение задачи распознавания лишь в некоторой ограниченной области значений входных параметров, имеется возможность определить стандартные алгебраические операции над таким

---

[56] И, соответственно, требований к реализующему эти технологии инструментарию ИАД.

[57] «*Гарантировать*» такую результативность.

[58] Классический вариант общего подхода к проблеме статистического анализа данных можно найти, например, в работах [*Tukey 77, Кендал 76*] и др. .

[59] См., в частности, понятие *размерности Вапника-Червоненкиса* (*VC-dimension*).

[60] См., например, [*Журавлев 06*] и др. .

алгоритмами[61] и, порождая алгебраические замыкания соответствующей степени, формировать «запас» дополнительных алгоритмов, который позволял бы выделить базис в соответствующем пространстве матриц (представляющих различные алгоритмы вычисления оценок), а далее использовать его для построения корректного алгоритма (алгоритма, обеспечивающего корректные результаты распознавания на всей области значений входных данных). Разложение такого алгоритма по базису (разумеется, в случае успешного формирования такого базиса в соответствующем пространстве представляющих используемые алгоритмы матриц числовых оценок) дает возможности его «конструирования» из «кусков» имеющихся (а также, возможно, дополнительно порождаемых при построении упомянутых алгебраических замыканий) некорректных алгоритмов. Таким образом, в этом подходе возможность формирования базиса в соответствующем пространстве матриц гарантирует существование корректного решения соответствующей задачи анализа данных.

(3) Техника нейронных сетей (см., например, [*Кохонен 80, Chester 93, Круглов 01, Рутковская 06*] и др.), реализующая специальный вариант математических моделей многопараметрической нелинейной оптимизации. Здесь в процессе настройки (обучения сети на прецедентах) подбираются подходящие коэффициенты связи между формирующими сеть нейронами так, чтобы сеть могла воспроизводить зависимости, которые имеют место на предложенных для ее обучения данных. Базовая гипотеза, на которой основано использование нейронных сетей в решении задач анализа данных выглядит следующим образом: «хорошо настроенная (обученная)» сеть позволяет получить «правильный» результат и при подаче ей на вход новых – не присутствовавших в исходной обучающей выборке[62] - данных. Считается, что «адекватная» настройка нейронной сети позволяет ей получать разумные результаты и для неполных, или же (частично) искаженных входных данных.

(4) Генетические алгоритмы (см., например, [*Гладков 06, Рутковская 06, Шашкин 10*] и др.), где решение задач оптимизации ищется моделированием локальных процедур случайного выбора, вариации и\или комбинирования анализируемых параметров по аналогии с механизмами естественного отбора в живой природе. В ходе длительного процесса последовательных итераций используемого набора локальных процедур отслеживаются оценки «полезности» порождаемых промежуточных результатов. Делается это с помощью специальной fitness function – функции *пригодности*, *приспособленности*, удачный подбор которой, как и удачный отбор части промежуточных результатов для применения теперь уже к ним используемых локальных процедур эволюционной обработки данных на следующих «шагах» компьютерных «мутаций», определяют успех представляемой техники анализа данных в каждом конкретном случае ее применения.

(5) Математические модели, методы и алгоритмы машинного обучения (см., например, [*Michie 94, Mitchell 97, Bishop 06, Langley 11, Mohri 12*] и др.) в случае использования для решения задач компьютерного анализа данных интересные в первую очередь в части выявления эмпирических зависимостей и закономерностей, скрытых в эмпирических данных (т.е. как модели и алгоритмы, позволяющие реализовать обучение на основе прецедентов). Возможности использования машинного обучения на примерах и контрпримерах как инструмента порождения зависимостей из фактов – критически важная характеристика современных систем анализа данных. Именно здесь можно найти целый ряд интересных работ, позволяющих говорить о

[61] Как операции над числовыми матрицами, формируемых на заданном множестве объектов используемыми алгоритмами вычисления оценок.

[62] Но «*подобных*» тем или иным элементам использованной обучающей выборки.

привнесении специфических черт «интеллектуальности[63]» в область компьютерного анализа данных.

В рамках каждого из только что названных направлений получено множество интересных результатов, однако, к сожалению, ни один из этих подходов не свободен от достаточно «чувствительных» ограничений, существенным образом сужающих сферу его применения до области действительно корректных результатов. Чтобы убедиться в справедливости такой оценки, вернемся к (уже сформулированным нами выше) четырем классам критически важных характеристик, определяющих «точки роста (развития)» современных ИАД-технологий.

Начнем с возможностей вести каузальный анализ изучаемых средствами ИАД явлений – с возможностей выделять в описаниях имеющихся прецедентов факторы влияния и их определенные комбинации, определяющие наличие (или же, наоборот, отсутствие) целевого эффекта - анализируемого явления. Фактически, речь здесь идет о возможности получить ответ на вопрос «*почему возникает анализируемое явление*?» При этом вариант ответа вида «*такой результат дал нам используемый алгоритм расчета*» в случае, например, работы с нейронными сетями или же генетическими алгоритмами вряд ли сможет убедить независимого[64] эксперта в наличии достаточных оснований для принятия полученных соответствующим инструментарием результатов.

Более тонкая ситуация может возникнуть при использовании статистических методов восстановления зависимостей из данных. Здесь, важно иметь ввиду того, что используемая математическая техника ориентирована сперва на интерполяцию имеющихся «табличных» данных [65] вновь порождаемой аналитической зависимостью (например, регрессией определенного вида), а затем – на экстраполяцию этой аналитической зависимости на новый объект (новый факт, представленный новыми значениями соответствующих переменных, для которого и будет построен «прогноз» значения изучаемой функции). Заметим, что в ряде случаев эта математическая техника ограничивается подбором некоторого определенного множества – «*объяснительных*» - переменных и связывающей их аналитической зависимости, позволяющей с заданной степенью точности интерполировать имеющиеся табличные данные (факты). При этом, вообще говоря, не требуется, чтобы задействованные *объяснительные* переменные отражали бы (прямо или косвенно) действительно актуальные *каузальные* факторы и при этом полностью представляли бы *полный набор* каузальных «влияний», определяющих наличие (или же отсутствие) анализируемого явления. В ряде прикладных областей это обстоятельство оказывается критически важным. Так, например, при выборе эффективных средств противодействия сбоям в работе сложного технического оборудования (в частности, при решении задач так называемого troubleshooting'а или же при организации противодействия компьютерным атакам[66]), в задачах медицинской диагностики и выбора эффективных средств терапии и др., не учесть тот или иной негативный каузальный фактор в организации противодействия соответствующим угрозам, вообще говоря, значит оставить объект защиты по-прежнему в потенциально уязвимом состоянии.

Рассуждения «в среднем» по группе аналогичных объектов, характерные для статистических методов анализа данных (в том числе – машинного обучения), также несут

---

[63] См. уже обсуждавшиеся выше архитектурные особенности и характеристики интеллектуальных систем (в предложенном В.К.Финном смысле).

[64] Например, того, кто не принимал активного участия в разработке (и настройке) соответствующей нейронной сети или же конкретного вида локальных процедур эволюционной обработки данных для соответствующего генетического алгоритма (и не владеет детальными знаниями об «изюминках» соответствующих математических моделей или их алгоритмики).

[65] Можно, например, считать, что исходно заданная выборка прецедентов – это числовая функция, заданная таблицей значений на значениях соответствующих переменных, и аналитический вид этой функции и требуется «восстановить» в рамках зафиксированной точности приближения.

[66] См. выше обсуждение примеров *диагностики и атрибуции целенаправленных активностей*.

в себе ряд принципиальных проблем в части корректности, интерпретируемости и применимости порождаемых заключений. Так, например, диагноз «*пациент Ж. на 70% здоров*» вряд ли может служить надежным основанием для принятия решения, в частности, о допуске «*пациента Ж.*» к тем или иным видам работ (получение разрешения на исполнение которых требует соблюдения соответствующего ряда требований и ограничений).

Серьезным препятствием для успешного применения ряда достаточно эффективных математических техник ИАД в приложениях оказываются характеристики вычислительной сложности соответствующих моделей и алгоритмов. Так, например, при поиске корректного алгоритма в моделях класса (2) - см. выше – рост степеней алгебраического замыкания, позволяющего сформировать базис в соответствующем пространстве алгоритмов, представляет собою отдельную проблему, преодоление которой требует специальных усилий в части оптимизации вычислительной сложности требуемых здесь расчетов. Еще более тяжелой оказывается ситуация с нестатистическими методами и моделями машинного обучения. Здесь в подавляющем большинстве случаев фиксируются экспоненциально быстро растущие оценки сложности вычислений (отражающие принадлежность соответствующих переборных задач к идентифицированным классам трудных комбинаторных проблем). Известные попытки бороться с проблемами сложности вычислений теми или иными систематическими методами своим успехом, как правило, обязаны либо возможностям «угадать» тот или иной полиномиально разрешимый подкласс соответствующей общей комбинаторной проблемы, либо критическим образом зависят от удачного выбора начальных условий для соответствующей локальной оптимизации[67] - соответствующей fitness функции и правил отбора подмножества результатов как новых объектов для следующего шага компьютерных «мутаций». Однако, методы *целенаправленного* и *управляемого* формирования тех или иных классов приближенных решений как инструмент систематического противодействия проблемам сложности вычислений в этой области еще ждут своей разработки.

Разумеется, выбор интересных моделей и техник компьютерного анализа данных вовсе не ограничен представленными выше пятью направлениями. Этот список легко пополнить за счет, например:

- различных вариантов техники так называемых деревьев решений (см. например, работы [*Breiman 84, Quinlan 86, Quinlan 88, Pearl 87, Kodratoff 92, Kodratoff 92-a*] и др.);
- комплекса процедур так называемого *Root Cause Analysis* (см., например, комментарии в [*RCA, ДИ*] и др.);
- весьма популярных на рубеже 20-21 веков моделей *Case-Based Reasoning* (CBR) – см., например, работы [*Schank 89, Barletta 91, CBR 94, CBR Express, CBR-PROC*] и др.

или же

- моделей и технологий так называемого *Автоматизированного Открытия* (*Automated Discovery & Deriving Laws Through Analysis of Processes and Equations*) - см., в частности, [*Zytkow 86, Langley 89, Zytkow 96*]

и др[68]. .

---

[67] Так Л.О.Шашкин (см. [*Шашкин 10*] и др.) продемонстрировал интересный эффект успешного порождения полезного подкласса эмпирических зависимостей в ДСМ-методе средствами особым образом настроенного семейства генетических алгоритмов. Детальный анализ этого эмпирического явления (эффекта сокращения необходимого комбинаторного перебора) показал, что успех здесь обусловлен удачным выбором стартовых (для выполняемых далее эволюционных процедур) объектов и fitness функции. Так, в частности, удалось показать, что реализуемый генетическими алгоритмами этого семейства процесс соответствует навигации вверх-вниз по отдельным цепочкам частичного порядка в диаграммах всех простых ДСМ-сходств.

[68] Например, своими математическими моделями и программными системами, позволившими решить ряд важных прикладных задач, известна группа Новосибирских ученых во главе с Н.Г.Загоруйко (см. [*Загоруйко 99*] и др.).

Однако, и здесь (причем - в еще большей степени) сказываются уже обсуждавшиеся нами ограничения. Так, например, для деревьев решений весьма чувствительным слабым местом оказывается зависимость порождаемого результата от последовательности выбора переменных при формировании результирующего дерева. (Фактически речь идет об эффекте *неассоциативности* базового отношения, используемого при построении дерева, - отношения «сходства» переменных). При соответствующем расширении исходных данных здесь, в общем случае, нет «наследуемости» структуры порожденного дерева. Таким образом возникают ситуации, когда не только требуются новые полноформатные вычисления для порождения соответствующего результата, но этот новый результат, вообще говоря, может существенным образом изменить представления о *каузальной природе* анализируемого явления.

Представления о каузальности в рамках так называемого *Root Cause Analysis* также порождают ряд вопросов, без удовлетворительного ответа на которые эта техника вряд ли может претендовать на полноту предлагаемого ею решения соответствующих задач анализа данных. Так, например, каузальные факторы в рамках RCA отбираются по принципу их решающего индивидуального влияния на изучаемый эффект: «*пусть некоторый фиксированный фактор идентифицирован как наличествующий при присутствии изучаемого эффекта*; *если удаление этого фактора из числа наблюдаемых (действующих) влечет исчезновение соответствующего эффекта, то это - каузальный фактор, требующий анализа в рамках RCA*». Естественным образом возникает вопрос: а как быть с комбинациями факторов, влияние каждого из которых в отдельности не влечет фатальных последствий, однако некоторые из их комбинаций порождают соответствующий негативный эффект [69] ? Не менее загадочным образом здесь решается проблема формирования карт взаимодействия таких факторов (см., например, в [*ДИ*] и др.): как быть в ситуации, когда два различных эксперта формируют две различные *Диаграммы Исикавы*? Существуют ли общепринятые (оцениваемые экспертами как надежные, заслуживающие доверия) инструменты автоматического формирования *Диаграмм Исикавы* из фактов (накапливаемых эмпирических данных)?

Привлекавшая внимание пользователей своими возможностями оперировать в том числе и нечисловыми зависимостями техника *Case-Based Reasoning* в приложениях достаточно быстро столкнулась с проблемой однородности понятия *аналогичные прецеденты* на всей области анализа данных и принятия решений. Неявным образом принятая здесь гипотеза об *универсальности* используемой *метрики* (метрической формализации понятия «близости» тех или иных прецедентов) оказалась неработающей в целом ряде ситуаций, а задача о формировании разбиения изучаемого пространства прецедентов на области однородности метрик так и не получила надежного решения (как в части методологии, так и в различных попытках обеспечить потребности соответствующих приложений).

Наконец, достаточно эффектные примеры «угадывания» (из данных) явного вида числовых зависимостей хорошо известных физических законов [70] , демонстрируемые программными системами *Automated Discovery,* оказалось непросто распространить и на обработку новых эмпирических данных. Выбор (формирование) *критерия достаточности оснований* для надежного *принятия* порождаемых *результатов* оказался здесь не только сложной *технической* (процедурной), но и не менее сложной *методологической* проблемой.

---

[69] Например, определенные уровни влажности, запыленности и вибрации каждый в отдельности не вызывают сбоев в дисковых накопителях (компьютерных системах хранения данных), тем не менее их одновременное «присутствие» фиксируется как причина выхода подобного оборудования из строя.

[70] Т.е. в ситуациях, когда явный вид целевой зависимости (например, *Закона тяготения* и т.п.) хорошо известен.

1.5. Возможности и перспективы использования интеллектуальных систем в задачах компьютерного анализа данных

Какие же дополнительные функциональные возможности способны привнести в компьютерный анализ данных системы интеллектуального анализа данных? В первую очередь – это:

- вариативность используемых средств представления данных (об анализируемых фактах, зависимостях и явлениях);
- вариативность задействованных средств экспертизы;
- выбор средств оценки порождаемых результатов (ориентированной на выбор наилучшей экспертизы) и погружение предпринимаемых «вариаций» (см. выше – о выборе средств представления данных и экспертизы) в единую среду оценивания порождаемых заключений и последующей оптимизации выбора финальных результатов экспертизы;
- оценки устойчивости порождаемых зависимостей при расширении исходной выборки данных новыми примерами и контрпримерами.

В части основных функциональных возможностей структурных блоков архитектуру систем ИАД характеризуют:

- условия применимости выбранных процедурных конструкций, определяющие область корректного использования соответствующего ИАД-инструментария;
- средства поиска (восстановления из имеющихся эмпирических данных) зависимостей и закономерностей;
- средства (корректного, аргументированного) распространения таких зависимостей и закономерностей на новые объекты, и
- анализ достаточности оснований для принятия результатов ИАД-экспертизы.

Особая роль в этом процессе интеллектуального анализа данных отводится[71] машинному обучения как собственно инструменту порождения зависимостей из фактов.

Приступая к организации ориентированного на реальные приложения поиска зависимостей в данных, хотелось бы определиться, - как действовать в ситуации, когда:

- запас имеющихся средств экспертизы (которые можно было бы варьировать для оценки устойчивости получаемых результатов) еще недостаточно велик[72];
- требуется иметь доказательство «теоремы существования» для искомого решения;
- необходимо оперировать нечисловыми данными (или же комбинациями вида *числовые + нечисловые* компоненты в описании анализируемых данных);
- необходимо получать (пусть – хотя бы предварительные[73]) результаты и на (пока еще) малых – т.е. статистически незначимых[74]) выборках прецедентов. Это может быть актуально, в частности, в ситуациях, которые «предшествуют» формированию (выполнению) условий, определяющих надежность и достоверность результатов, порождаемых соответствующей математической техникой. Фактически речь идет о таких ситуациях, когда в силу малых размеров исходных выборок выполнимость условий корректности применения к ним соответствующих статистических методов анализа данных не может быть гарантирована, и надежность такого инструментария оказывается «ослабленной» с уровня *достоверного* (на выборках достаточной

---

[71] См. уже цитированные выше замечания Дж.Маккарти ([*McCarthy 00, McCarthy*] и др.) о природе *phenomenal data mining*.

[72] Например, имеется ли возможность использовать процедурные конструкции, аналогичные корректным алгебрам Ю.И.Журавлева для пополнения текущего «запаса» средств ИАД-экспертизы?

[73] И предполагающие дальнейшие уточнения при пополнении исходной выборки данных новыми элементами.

[74] В ряде приложений переход к анализу статистически значимых выборок существенным образом затруднен (например, неудачные запуски космических ракет; испытания новой авиационной техники; синтез и изучение антиканцерогенов-биостимуляторов, не являющихся цитостатиками; атрибуция некоторых видов кибератак и т.п.).

«емкости») *вывода* до «статуса» *эвристики* (которая в некоторых случаях, как и прежде, способна приводить к не вызывающему сомнений результату). Тем не менее, даже в таких случаях хотелось бы располагать возможностями для корректного восстановления эмпирических зависимостей из подобных коллекций данных;

- необходимо порождать характеризующие «*в точности*» *конкретный* изучаемый объект (т.е. именно его, а не описывающие ситуацию «*в среднем*» по *группе подобных* объектов) заключения о скрытых в имеющихся данных каузальных зависимостях, управляя при этом *сложностью* вычислений (в том числе – при использовании алгоритмов исчерпывающего комбинаторного анализа вариантов);
- требуется обладать возможностями вести детальный *каузальный* анализ факторов, определяющих изучаемое явление?

Как самостоятельная *платформа* для организации интеллектуального анализа данных ДСМ-метод предлагает свои варианты положительного ответа по каждой из перечисленных выше позиций. В основании ДСМ-метода – *математические модели* двух типов, каждый из которых используется для решения соответствующего класса задач.

(1) Ядро процедурной конструкции ДСМ-метода - это базовый набор эвристик, объединяющий[75]

- *эмпирическую индукцию*,
- рассуждения по *аналогии*,
- *абдукцию* (представляющую собою рассуждение от утверждения - *наблюдаемого «факта»* - к *объясняющей* его *гипотезе*, дедуктивным следствием которой и является это исходное утверждение)
- процедуры *выдвижения гипотез*,
- эмпирическое объяснение (способность *отвечать на вопрос «почему»,* реализуемую процедурами формирования *абдуктивного объяснения* начального состояния базы фактов, предложенной в качестве исходных данных для проводимого интеллектуального анализа),
- процедуры *фальсификации* результатов, полученных в процессе *обучения* на *позитивных* и *негативных примерах*
- и др. .

Формализация используемых здесь эвристик выполнена средствами специального класса *многозначных логик*[76] (см., например, [*МЛиП, Финн 83, Аншаков 86, Аншаков 87, Финн 88, Финн 91, Финн 10*] и др.). Корректность предложенной формализации демонстрируется построением *дедуктивной имитации* реализуемых в рамках ДСМ-метода правдоподобных выводов средствами специальной дедуктивной теории (см., например, [*Финн 10, Аншаков 93*] и др.). Таким образом, показано, что все эмпирические зависимости, которые могут быть порождены на заданных исходных данных применением правил правдоподобного вывода ДСМ-метода, также могут быть получены в результате дедуктивного вывода в соответствующей (*имитирующей*[77]) дедуктивной теории, и наоборот: каждый из результатов вывода (*достоверного вывода - !*), порождаемый в этой теории, может быть также получен в процедурной конструкции ДСМ-метода применением к исходным данным тех или иных ДСМ-правил правдоподобного вывода.

---

[75] Развернутые комментарии процедурных особенностей используемых эвристик а также «технологии» их «воплощения» в правила правдоподобного вывода ДСМ-метода будет даны далее – в **Главе 2**.

[76] Основные результаты в части разработки и использования логико-математических моделей этого типа при формализации эвристик, положенных в основу ДСМ-метода, получены В.К.Финном, Д.П.Скворцовым и О.М.Аншаковым.

[77] Имитирующей используемую процедурную конструкцию ДСМ-метода.

(2) Еще один класс математических моделей возникает здесь при анализе актуальной разрешимости трудных переборных задач, которые формулируются при построении соответствующих алгоритмических конструкций в процессе формализации эвристик обсуждаемого (см. выше) типа. Такой анализ актуальной разрешимости предполагает

- построение доказуемых оценок сложности вычислений, характеризующих основные переборные (комбинаторные) задачи рассматриваемой предметной области (в том числе – для различных типов данных, обрабатываемых в рамках ДСМ-ИАД);
- разработку методов дискретной оптимизации комбинаторного перебора, характерного для задач восстановления эмпирических зависимостей из данных средствами ДСМ-ИАД,
- разработку специальной проблемно-ориентированной алгоритмической техники реализации приближенных вычислений, позволяющей организовать управляемый целенаправленный анализ вариантов при порождении эмпирических зависимостей с учетом специальных дополнительных условий (при прогнозировании свойств конкретных новых объектов, проверке выполнимости условий достаточности оснований для принятия результатов правдоподобного вывода, при порождении зависимостей с теми или иными структурными характеристиками[78] и т.п.).

Суммируя возможности и потребности ДСМ-метода в его текущем состоянии, отметим:

- построена дедуктивная имитация процедурной конструкции реализуемого ДСМ-методом правдоподобного вывода на примерах и контрпримерах (таким образом доказана корректность формализации используемых эвристик);
- применение ДСМ-метода как основы инструментальных программных систем интеллектуального анализа данных позволило получить решения для ряда важных прикладных задач;

однако,

- в ряде приложений применение ДСМ-инструментария анализа данных затруднено неприемлемо высоким объемом вычислений, необходимых для получения финального результата.

Таким образом,

- актуальным и важным (как для развития «технических» возможностей ДСМ-метода, так и для расширения сферы его успешных приложений) оказывается обеспечить детальный анализ *актуальной разрешимости* выявленных в рассматриваемой области математических задач и найти возможности надежно управлять объемами ДСМ-вычислений в практически значимых приложениях.

Подробному обсуждению этой проблематики, корректной постановке соответствующих математических задач а также разработке эффективных алгоритмов их решения и будут посвящены последующие части данной работы.

В заключение предпринятого нами в этой вводной **Главе** обсуждения специального класса некорректных задач, характерных для проблематики интеллектуального анализа данных, сформулируем несколько дополнительных замечаний относительно перспектив а также возможных путей дальнейшего развития области исследований *Искусственный Интеллект (ИИ)\ Artificial Intelligence (AI).*

Обсуждая особенности Интеллектуального Анализа Данных, представляется уместным регулярно обращаться к вопросу: а где же здесь, собственно, «*интеллект*», ведь

[78] Например, содержащими те или иные структурные компоненты, порожденные с «участием» тех или иных примеров\контрпримеров из исходных данных и т.п. .

в последние годы термин *интеллектуальный* становится все более и более часто встречающимся в связи с приложениями постоянно обновляющихся информационных технологий.

Обилие в нашем сегодняшнем технологическом окружении устройств, *именуемых*[79] *интеллектуальными,* – интеллектуальных датчиков, интеллектуальных интерфейсов, интеллектуальных вычислителей, интеллектуальных сетевых middle-box'ов и т.п. – вынуждает задуматься о корректности использования термина «интеллектуальный»: со времен публикации известного теста Тьюринга в области computer science *интеллектуальные* решения (*интеллектуальные* системы) предполагают по крайней мере компьютерное моделирование тех или иных интеллектуальных функций специалиста (т.е. не просто *расчеты*, а в той или иной мере определенную «*игру ума*»).

В таком контексте некорректное использование наименования *интеллектуальный* представляется, вообще говоря, ненужным: смена названия вряд ли сможет повлиять, например, на функциональные возможности, на удобство эксплуатации или же другие реально «ощутимые» пользователем характеристики соответствующих систем, решений и оборудования.

Достаточно долгий период времени область *ИИ\AI* развивалась в рамках взгляда на нее (*vision*), полностью определявшегося знаменитым *Тестом Тьюринга* [*Turing 50*]. Однако, почти в это же время У.Р.Эшби [*Эшби 59*] ввел в рассмотрение исследовательского сообщества наряду с понятием интеллектуальной системы–«*заменителя»* эксперта также и понятие системы-*усилителя* его интеллектуальных возможностей. Разумеется, в части *ИИ\AI*-позиционирования систем-*усилителей* интеллектуальных возможностей эксперта возникает существенная проблема адекватности Теста Тьюринга как своего рода *индикатора интеллектуальности* соответствующих разработок.

В стремлении преодолеть эту проблему представляется естественным перспективы развития области *ИИ\AI* как самостоятельного исследовательского направления ассоциировать, в частности, и с проблематикой *формализации эвристик* (как очевидных проявлений интеллектуальности человеческого разума).

Продвижение вперед в этом направлении открывает возможности иметь дело как с новыми интересными математическими задачами (в частности, задачами формирования корректных формализаций конкретных эвристик логико-математическим средствами, задачами оценки вычислительной сложности реализующих это формальные модели алгоритмических конструкций, развитием методов управления сложностью необходимых здесь вычислений и др.), так и с распространением развиваемых моделей и компьютерных систем ИАД в сферу (в том числе - масштабных) индустриальных приложений.

Актуальные примеры подобных приложений уже сегодня демонстрируют области так называемых Big Data: это могут быть модели и технологии создания искусственных объектов с заданными свойствами (в частности – физиологически активных веществ и т.п.), диагностика медицинских и сложных технических объектов (включая troubleshooting и оперативное формирование рекомендаций по оптимизации лечебных мероприятий), задачи текст-процессинга на больших коллекциях документов и многое другое.

[79] При ближайшем рассмотрении – всего лишь, фактически, за использование в их конструктивном решении микропроцессоров и\или некоторого набора программируемых функций.

Глава 2

ДСМ-ИАД: общее описание.

2.1. Задача о восстановлении эмпирических зависимостей из данных: от классов сходства к классам эквивалентности.

Интеллектуальный анализ данных (ИАД – см., например, [*ИАД, PA&DM, Забежайло 13*] и др.) сегодня позиционируется как одно из быстро развивающихся направлений не только *theoretical computer science*, но и *applications*, востребованное в том числе и для решения трудных практически значимых задач в целом ряде прикладных предметных областей.

Несложно убедиться, что достаточно обширный класс задач ИАД, ориентированных на восстановление эмпирических зависимостей, неявным образом представленных в тех или иных коллекциях экспериментальных данных, в самом общем виде может быть описан следующей схемой:

Даны:

1) множество (примеров) *объектов*, обладающих целевыми *свойствами*, и множество[1] (контрпримеров) объектов, не обладающих целевыми свойствами,

а также

2) *новый объект* (или же некоторое явным образом представленное множество таких объектов).

Требуется:

оценить *наличие* (или отсутствие) *целевых свойств* у нового *объекта* (новых объектов из заданного множества), т.е.

- дать соответствующий *прогноз* (о наличии целевых свойств) и
- предъявить *основания* (неоспоримые *аргументы*), позволяющие *принять* этот прогноз.

Формализованные поставки задач этого вида могут использовать (т.е. опираться на) те или иные математические техники – методы и модели статистического анализа данных, машинного обучения, нейронных сетей, распознавания образов или автоматической классификации и др.[2]. Имеется обширная литература, представляющая особенности формирования и применения в приложениях как подобных математических моделей, так и реализующих их инструментальных программных систем[3].

Однако, критически важным аспектом, определяющим возможности использования каждого такого формализованного уточнения обсуждаемой общей задачи, остается вопрос о существовании корректного[4] решения порождаемого соответствующим методом (формализованным подходом). Фактически, речь здесь идет о существовании надежных «*источников доверия*» к получаемым соответствующим инструментарием результатам (другими словами – о возможности предъявить *достаточные основания* для принятия получаемых результатов). Наличие (в конкретном случае – при решении той или иной прикладной задачи) таких *достаточных оснований* – весомый аргумент в пользу доверительного отношения к получаемым применением соответствующего инструментария результатам. (Отсутствие возможностей предъявить подобные *достаточные основания* – дополнительный повод удостовериться в обоснованности использования в данном конкретном случае именно этого формального инструментария ИАД).

---

[1] В исходном состоянии, - возможно, пустое.

[2] См., например, ссылки к работе [*Забежайло 13*] и др..

[3] См., например, [*PA&DM, Забежайло 13*] и др..

[4] Причем – ***доказуемо*** корректного!

Классический пример элегантного решения (математическими средствами - !) вопроса о существовании *достаточных оснований* для принятия результатов компьютерного анализа данных был дан Ю.И.Журавлевым для обширного класса задач распознавания образов (см., например, работы [*Журавлев 77*, *Журавлев 77-а, Журавлев 78*] и др.). Предложенные здесь корректные алгебры над множеством, вообще говоря, некорректных эвристических алгоритмов, использующих технику вычисления оценок при решении задач распознавания образов, обеспечивают доказательство существования (при выполнении определенных условий) корректного[5] алгоритма распознавания и предлагают соответствующие инструменты[6] для его восстановления по исходно заданному набору, вообще говоря, некорректных (т.е. вычисляющих корректные результаты лишь на некоторых - характерных для каждого из них - подмножествах возможных исходных данных) алгоритмов.

Следуя мотивации, только что представленной выше на примере использования алгебраических соображений при формировании достаточных оснований для принятия результатов ИАД в соответствующем классе задач распознавания образов, представляется естественным попытаться сформулировать подобную (в части мотивации[7]) конструкцию и для исходной рассматриваемой нами задачи (оценки наличия у вновь предъявляемого объекта целевых свойств, характерных для уже заданного множества объектов-примеров).

Итак, наша цель – не только научиться

- использовать имеющуюся (накапливаемую, доступную, …) информацию об уже идентифицированных экземплярах (примерах) объектов, обладающих целевыми свойствами, причем - вместе с информацией об уже идентифицированных экземплярах (контрпримерах) объектов, не обладающих целевыми свойствами, для идентификации новых (ранее не классифицированных) экземпляров предъявляемых для идентификации объектов,

но и развить «навыки», позволяющие

- формировать соответствующие (желательно - быстро[8]) проверяемые условия, выполнимость которых обеспечивала бы наличие *достаточных оснований* для *принятия результатов* идентификации (используемыми конкретными инструментальными средствами) вновь предъявленных объектов.

Идея (целевого) подхода к решению поставленной задачи:

1. Каждый представленный в исходных данных объект может быть охарактеризован набором признаков (например, определенным набором тех или иных последовательностей 0 и 1 в его «теле»).
2. Все множество таких признаков может быть собрано в (пополняемый и расширяемый) каталог – «алфавит» («атомарных») признаков. Каждый уже идентифицированный ранее объект может быть «закодирован» (представлен) определенным набором (множеством) таких признаков.
3. На множестве идентификационных признаков могут быть заданы отношения (т.е. описания исходно заданных объектов могут стать не множествами признаков, а, например, графами, цепочками и т.п.). Описания изучаемых объектов могут быть расширены также числовыми значениями тех или иных параметров (*множества признаков + отношения + числовые характеристики*).
4. В случае пустоты исходно заданного множества контрпримеров, к имеющимся коллекциям объектов (примерам изучаемых объектов) путем дополнительного

[5] Формирующего математически корректное решение предъявленной прикладной задачи.

[6] Корректные алгебры (см., например, работы [*Журавлев 77*, *Журавлев 77-а, Журавлев 78*] и др.).

[7] Т.е. в части поиска и формирования *проверяемых условий*, позволяющих отвечать на вопрос о наличии (или же об отсутствии) в каждом конкретном случае *достаточных оснований* для того, чтобы *доверять* получаемым в ходе выполняемых ИАД-вычислений *результатам*.

[8] Т.е. средствами полиномиально сложных вычислений.

анализа[9] изучаемой предметной области могут быть добавлены также наборы контрпримеров – комплексов признаков из исходного «алфавита», которые тем не менее на являются объектами, обладающими целевыми признаками.

5. На множествах примеров и контрпримеров может быть определена (алгебраическая) операция сходства *, удовлетворяющая трем условиям:

   - a*a = a
   - a*в = в*a
   - (a*в)*c = a*(в*c) = a*в*c

   Например: операция * определяемая как пересечение множеств, или же как выделение множества всех максимальных общих подграфов двух графов, множества всех максимальных общих подцепочек двух цепочек и т.п. .

   Два объекта (примера или контрпримера) будем считать сходными в том и только в том случае, когда операция сходства дает на них не пустой результат. Таким образом можно определять и отношения **R*** сходства на соответствующих объектах.

Операциональная формализация сходства предоставляет возможность не только судить о сходстве (или же наоборот, несходстве) пар объектов, но *выделять* структурный *«носитель»* этого сходства.

6. С помощью такой операции * на примерах и контрпримерах строятся классы сходства: для каждого заданного объекта исходной коллекции примеров (и, соответственно, контрпримеров) это - множество всех известных объектов, которые сходны[10] с ним. Т.е. класс сходства для заданного элемента **a** есть множество всех сходных с ним элементов **b** исходного множества примеров (или, соответственно, контрпримеров):

$$\mathbf{T}(a) = \{b \mid a\mathbf{R}^{*}b\}.$$

7. Диагностика (идентификация) новых объектов производится с помощью порождаемых таким образом классов сходства: возможность отнесения нового – диагностируемого - объекта к какому-либо из уже реконструированных (по примерам[11]) классов сходства позволяет утверждать, что это – еще один объект, аналогичный (см. класс сходства) ранее уже наблюдавшимся (в исходной обучающей выборке примеров и контрпримеров).

8. Идея техники отнесения нового объекта к классу выглядит следующим образом:

   - класс сходства есть объединение некоторых классов эквивалентности[12] [*Гусакова 87*]. Действительно, фиксируя при вычислении операции сходства **a*b = v** значение результата **v**, и рассматривая класс сходства по таким образом модифицированному отношению $\mathbf{R^*_v}$, получаем отношение эквивалентности $\boldsymbol{E}(\mathbf{R^*_v})$[13]. Таким образом, исходное множество примеров (и контрпримеров) может быть представлено сперва в виде покрытия его порождаемыми описанным образом классами сходства, а далее (за счет реконструкции классов эквивалентности, покрывающих каждый класс сходства) – покрытия его соответствующими классами эквивалентности;

---

[9] Проводимого с привлечением экспертов по изучаемой предметной области.

[10] Т.е. операция сходства в каждом таком случае дает непустой (ненулевой) результат.

[11] Или же контрпримерам.

[12] Порождаемых всеми возможными в данном случае версиями соответствующего отношения эквивалентности.

[13] Транзитивность тривиальным образом следует из вложимости **v** в каждый из объектов соответствующего класса сходства (трансформируемого в класс эквивалентности).

- итак, по построенному классу сходства можно восстановить (с учетом имеющегося контекста – т.е. множества примеров и контрпримеров, задействованных в обучении) порождающие его классы эквивалентности;
- при этом при восстановлении классов эквивалентности каждый класс сходства, объединением которых он является, будет характеризоваться непустыми (ненулевыми) результатами вычисления уже заданной операции сходства. Таким образом каждый восстанавливаемый здесь класс эквивалентности будет определяться теми примерами из исходной выборки, результат вычисления операции сходства для которых будет одним и тем же[14] (например, в случае простого сходства на множествах – всех примеров, содержащих общее подмножество образующих); следовательно
- каждый класс эквивалентности в таком случае – орбита (множество примеров-родителей) для соответствующей ДСМ-гипотезы. (При этом, разумеется, один и тот же пример из исходной ДСМ-выборки может попадать в несколько пересекающихся, но не совпадающих классов эквивалентности[15]);
- помимо этого, классы сходства и соответствующие классы эквивалентности, порождаемые на исходном множестве примеров и контрпримеров (как и в случае рассмотрения собственно ДСМ-рассуждений – см. далее **Главы 2** и **3**) могут характеризоваться выполнимостью дополнительных логических условий – в частности: запрета на контрпримеры, условия единственности причины и т.п.;
- здесь важно заметить, что в ДСМ-рассуждении используются одновременно классы сходства\эквивалентности, порожденные как на позитивных, так и на негативных примерах (контрпримерах). Критерий достаточности оснований для принятия результатов ДСМ-рассуждения допускает принятие лишь тех результатов правдоподобного вывода одного знака, которые не фальсифицируемы результатами правдоподобного вывода другого знака. Таким образом, в дополнение к традиционной для распознавания образов технике порождения решающих правил добавляются дополнительные условия нефальсифицируемости порождаемых результатов в смысле К.Поппера. (При этом позитивные и негативные классы эквивалентности при их порождении в определенном смысле равноправны: результаты одного знака используются при фальсификации результатов другого знака и наоборот. На достаточном основании принимаются лишь непротиворечивые – нефальсифицируемые – результаты, а процедура взаимной фальсификации требует исчерпывающего просмотра зависимостей – классов эквивалентности, - порождаемых на заданном множестве примеров и контрпримеров);
- каждый набор классов эквивалентности можно «сузить» до фактор-множества. В данном случае каждый класс эквивалентности и его представитель в фактор множестве будет задаваться соответствующим результатом вычисления операции сходства (ДСМ-гипотезой), а фактор множество – множеством непротиворечивых ДСМ-гипотез;
- проверкой «аналогии» между элементами фактор-множества и новым – диагностируемым – объектом исполняется идентификация последнего как

[14] Именно этим фактом дополнительно к рефлексивности (идемпотентности соответствующей операции сходства) и симметричности обеспечивается также и транзитивность соответствующего (уже используемой операции сходства) отношения сходства (суженного представленным дополнительным условием до эквивалентности).

[15] Для, разумеется, различных – см. различные ДСМ-гипотезы – конкретных вариантов отношений эквивалентности, участвующих в формировании соответствующих классов сходства.

нового объекта, соотносимого с подмножеством (классом сходства как объединением соответствующих классов эквивалентности) исходно заданных для обучения примеров как *подобный* им;

- при этом необходимо, чтобы вновь идентифицируемый объект попадал лишь в классы эквивалентности одного знака. В противном случае при идентификации порождается противоречие (что в силу формулировки ППВ-II в ДСМ-методе порождает противоречивое доопределение по правилу аналогии);
- наконец, идентификация в порожденных на исходно заданном для обучения множестве примеров классах эквивалентности одноэлементных – т.е. вырожденных - классов (что в рамках ДСМ-рассуждения сигнализирует о невыполнении *Условия Каузальной Полноты* – см., например, [*Финн 10, ДСМ АПГ*] и др. – и также о необходимости пополнять исходную обучающую выборку новыми примерами, позволяющими обеспечить выполнение названного Условия), указывает на необходимость расширения исходной выборки примеров такими новыми примерами, которые были бы сходны с примерами, образующими вырожденные классы эквивалентности[16].

9. Поддерживая (формируя, сопровождая и пополняя) базы описаний анализируемых объектов (например, *сигнатур вирусов при анализе вредоносного программного обеспечения* и т.п.) можно оперативно диагностировать новые (потенциально – наделенные искомыми свойствами) объекты, проверяя (в том числе – в режиме глубокого «распараллеливания») сходство каждого такого объекта с уже накопленными в соответствующих базах признаками (а также их «существенными» - см. обучение на примерах и контрпримерах - *комбинациями* и т.п.). При этом собственно проверка такого сходства реализуется как проверка попадания[17] нового (идентифицируемого в настоящий момент) объекта последовательно в каждый из порожденных на текущий момент (невырожденных - !) классов эквивалентности.

10. Стабильность системы классов сходства[18] (и вместе с ними – соответствующих классов эквивалентности) при расширении исходной обучающей выборки примеров и контрпримеров новыми объектами может рассматриваться как сигнал о том, что накопленную выборку можно оценить как *насыщенную* в части порождения (определяемого соответствующей системой классов эквивалентности) множества эмпирических зависимостей (в т.ч. – тенденций и закономерностей – см., например, [*Финн 10-а*]).

* * *

Нашей основной целью будет продемонстрировать, как представленная выше схема рассуждений о возможностях аргументированного отнесения нового объекта к множеству уже изученных (имеющих достаточно информативное[19] описание их «структуры» и обладающих целевыми свойствами) может специфицироваться (уточняться). Причем такие уточнения должны быть возможны как в части формирования и комбинирования (т.е. «сборки» в те или иные *стратегии* рассуждения) разного вида однородных[20] процедур

---

[16] Что (см. описание структуры ДСМ-рассуждения, например, в [*Финн 10, ДСМ АПГ*] и др.) предоставляет возможности для целенаправленного расширения исходной выборки примеров для обучения.

[17] Причем – непротиворечивого (см. выше) попадания.

[18] См. также замечание о свойствах соответствующих пространств толерантности в работе [*Гусакова 87*].

[19] Здесь представляется уместным напомнить о требованиях к дескриптивным возможностям используемого языка представления исходных данных (см., например, [*Финн 10, Финн 10-а*] и др.): задействованный способ описания данных должен иметь соответствующие дескриптивные возможности для адекватного представления (т.е. записи в виде формальных выражений используемого языка) восстанавливаемых из данных зависимостей.

[20] Выстроенных в единой «логике» рассуждения.

интеллектуального анализа данных, так и в части взаимно согласованной[21] обработки предлагаемыми процедурами ИАД различных типов представлений данных. Спектр доступных здесь вариантов использования дескриптивных возможностей для поиска зависимостей, в неявном виде содержащихся в накапливаемых эмпирических данных, может варьироваться от достаточно «грубых[22]», но «обсчитываемых» процедурами невысокой вычислительной сложности, до весьма детальных[23], однако, как правило, требующих экспоненциально сложных вычислений.

Обратимся к более детальному анализу только что представленных соображений алгебраического характера, а также их взаимосвязей с базовыми компонентами ДСМ-метода автоматического порождения зависимостей из эмпирических данных.

Прежде чем двигаться вперед сформулируем некоторое предположение (своеобразный «*постулат значения*»), которому отводится критически важная роль во всех наших дальнейших построениях: мы будем считать, что у каждого свойства (или же множества свойств) рассматриваемых нами объектов в описаниях последних существует некоторый фрагмент, отвечающий за наличие этого свойства (множества свойств). Другими словами, вне зависимости от выбора конкретного языка представления анализируемых объектов (их представления, например, множествами признаков[24] или же множествами значений признаков с отношениями на признаках – графами, цепочками – и т.п.) в случае наличия у объекта анализируемых свойств в его описании может быть выделен определенный подобъект (подмножество, подграф, подцепочка и т.п.), представляющий собою «*структурный носитель*» соответствующего множества свойств.

I. Начнем с простейшего случая:

Дано: множество **Ω** объектов (примеров) ***O*** , каждый из которых обладает некоторым свойством ***P***. Каждый элемент ***O*** из **Ω** характеризуется некоторым (структурным) описанием[25]. Также задан (соответствующим описанием) новый объект $\boldsymbol{O_0}$, о наличии (либо отсутствии) у него анализируемого свойства ***P*** нет информации.

Требуется: сформировать аргументированное суждение относительно наличия (или же отсутствия) у $\boldsymbol{O_0}$ свойства ***P***.

При формировании такого прогноза и поиске соответствующих аргументов представляется уместным воспользоваться следующими простейшими соображениями:

1) *если все объекты **O** из **Ω** обладают свойством **P**, то естественно предположить, что в их структурных описаниях имеется некоторая «общая часть»* **V**, *которая (в случае единственности причины для **P**) «вкладывается» во все примеры из* **Ω**.
2) В свою очередь в «*многопричинной*» ситуации *в* **Ω** *должны содержаться подмножества объектов (примеров) из множества* $\boldsymbol{\Omega_i}$, *где каждый элемент для каждого такого* $\boldsymbol{\Omega_i}$ *будет иметь в виде общего «фрагмента» - «носителя свойства **P**» - соответствующий подобъект (подмножество образующих, подграф, подцепочку и т.п.)* $\mathbf{V_i}$.

Начнем наши построения с уточнения простейшего случая и будем рассматривать ситуацию, когда в нашем распоряжении имеется множество **Ω** объектов O_1, O_2, … , O_n ,

---

[21] В данном случае подразумевается, что различные (по своим выразительным возможностям) формализованные описания одних и тех же эмпирических данных обрабатываются однотипными по «логике» рассуждения процедурами ИАД.

[22] Например, в виде булевских векторов значений характерных для рассматриваемой ситуации признаков.

[23] Например (в случае работы с биохимическими данными), в виде пространственных графов (рассматриваемых физиологически активных соединений), дополненных числовыми значениями релевантных физико-химических параметров (доз, концентраций, температур и т.п.).

[24] Булевскими векторами значений этих признаков.

[25] Например, множеством значений признаков – множеством образующих (букв) некоторого алфавита **U**, некоторым графом, цепочкой символов, числовым вектором и т.п.

каждый из которых обладает заданным свойством *P*. В исходном состоянии нам не важно, как именно представлены объекты O_i своими формализованными описаниями: необходимо лишь, чтобы в этих описаниях можно было бы выделять общие части[26]. Другими словами, необходимо, чтобы на множестве объектов (включая также и общие составные части $c_i \subseteq O_j$ всех пар, троек и т.д. объектов $O_j \in \mathbf{\Omega}$) была бы определена операция сходства $\otimes$ со следующими свойствами:

(i) $c \otimes c = c$
(ii) $c_r \otimes c_s = c_s \otimes c_r$
(iii) $(c_r \otimes c_s) \otimes c_t = c_r \otimes (c_s \otimes c_t) = c_r \otimes c_s \otimes c_t$.

Пусть $\mathbf{\Omega} = \{O_1, O_2, \dots, O_m\}$. Построим по $\mathbf{\Omega}$ и $\otimes$ множество *Dom*($\mathbf{\Omega}$) следующим способом:

(a) каждый из принадлежащих множеству $\mathbf{\Omega}$ объектов по определению принадлежит множеству *Dom*($\mathbf{\Omega}$), т.е.

$$\mathbf{\Omega} = \{O_1, O_2, \dots, O_m\} \subset Dom\ (\mathbf{\Omega}).$$

(b) если A и B принадлежат множеству *Dom*($\mathbf{\Omega}$), а результат A⊗B применения к ним операции сходства $\otimes$ не является пустым подобъектом $\ulcorner\varnothing\urcorner$, то этот результат A⊗B также помещается в множество *Dom*($\mathbf{\Omega}$) , т.е.

$$\{[A \in Dom(\mathbf{\Omega})] \,\&\, [B \in Dom\ (\mathbf{\Omega})] \,\&\, [(A \otimes B) \neq \ulcorner \varnothing \urcorner]\} \rightarrow [\ (A \otimes B) \in Dom\ (\mathbf{\Omega})\ ],$$

(c) других элементов в множестве *Dom* ($\mathbf{\Omega}$) нет.

Теперь определим отношение сходства $\mathbf{R}^{\otimes}$, выполнив это следующим образом:
пара объектов $< O_{i1}, O_{i2}>$ из $\mathbf{\Omega}$ принадлежит отношению $\mathbf{R}^{\otimes}$

$$O_{i1}\mathbf{R}^{\otimes}O_{i2}$$

тогда и только тогда, когда

$$O_{i1} \otimes O_{i2} \neq \ulcorner \varnothing \urcorner.$$

Классы сходства (толерантности) на $\mathbf{\Omega}$ определяются стандартным способом:
для каждого объекта O_{i1} из $\mathbf{\Omega}$ класс $\mathbf{T}(O_{i1})$ сходных с ним объектов O_{i2} из исходного множества $\mathbf{\Omega}$ задается условием

$$\mathbf{T}(O_{i1}) = \{O_{i2} |\ O_{i1}\mathbf{R}^{\otimes}O_{i2}\}.$$

Далее, фиксируя каждый конкретный результат $V=V_0$ вычисления операции $\otimes$ сходства на элементах множества *Dom* ($\mathbf{\Omega}$):

$$A \otimes B = V_0,$$

будем выделять соответствующие подклассы $\mathbf{E}_{V0}$ сформированных классов сходства $\mathbf{T}$:

$$\mathbf{E}_{V0} = \{< O_{i1}, O_{i2} > |\ O_{i1} \otimes O_{i2} \otimes V_0 = V_0\}.$$

Таким образом, пара $<\mathbf{\Omega}, \otimes>$ определяет пространство толерантности в смысле [*Гусакова 87, Шрейдер 71, Шрейдер 71-а*] и др., т.к. всякому непустому[27] результату применения операции $\otimes$ к каждой конкретной паре объектов $< O_i, O_j >$ можно сопоставить точно такую же пару объектов, принадлежащих (сопряженному с операцией $\otimes$) бинарному

[26] При этом представляется естественным ради сокращения перебора рассматривать лишь соответствующие максимальные общие части анализируемых объектов (примеров).

[27] Т.е. дополнив множество Ω некоторым специальным примером $O_{\varnothing}$, который по определению будем называть *пустым* объектом, можно говорить о пустоте (отсутствии общих частей у сравниваемых объектов, выражающемся в равенстве $O_{\varnothing}$ результата вычисления операции $\otimes$ на этих объектах) и непустоте соответствующего сходства.

отношению $\mathbf{R}^{\otimes}$, которое, как несложно убедиться, по условиям (i)-(iii) оказывается отношением толерантности (сходства).

Итак, для каждого конкретного объекта O_r классом сходства $\mathbf{T}_{\mathbf{\Omega}}(O_r)$ по определению будем считать множество всех тех примеров из $\mathbf{\Omega}$, которые сходны с этим O_r в смысле определенного нами отношения $\mathbf{R}^{\otimes}$. Если при этом зафиксировать конкретное значение V_0 - результат применения операции $\otimes$ к соответствующим элементам $\mathbf{T}_{\mathbf{\Omega}}(O_r)$ - и рассматривать лишь все содержащие максимальную общую часть V_0 объекты, то порождаемое таким способом отношение $\mathbf{E}^{\otimes}$ оказывается отношением эквивалентности. Т.е. имеет место простое

2.1.1. Утверждение

Пусть заданы: множество $\mathbf{\Omega}$, удовлетворяющая ограничениям (i)-(iii) операция $\otimes$, некоторый O_r из $\mathbf{\Omega}$ а также некоторый подобъект V, характеризующий сходство объекта O_r по крайней мере с одним из элементов класса сходства $\mathbf{T}_{\mathbf{\Omega}}(O_r)$. Множество всех таких элементов O_s из $\mathbf{T}_{\mathbf{\Omega}}(O_r)$, что

$$O_r \otimes O_s \otimes V = V$$

представляет собою класс эквивалентности $\mathbf{E}_{\mathbf{\Omega},v}(O_r)$.

Доказательство.

Рефлексивность и симметричность отношения $\mathbf{E}_v^{\otimes}$ (порождаемого фиксацией результата V применения операции сходства $\otimes$) определяется соответствующими свойствами отношения $\mathbf{R}^{\otimes}$. Транзитивность вытекает из следующего простого рассуждения: будучи общей частью одновременно пар $<O_r,O_s>$ и $<O_s,O_t>$, заданный подобъект V оказывается также общей частью примеров O_r и O_t .

□

В представленной ситуации для каждого порождаемого из множества примеров $\mathbf{\Omega}$ применением операции $\otimes$ сходства V_j можно подобрать максимальное по вложению подмножество примеров из $\mathbf{\Omega}$, таких, что все они образуют класс эквивалентности по отношению $\mathbf{E}_v^{\otimes} = \mathbf{E}^{\otimes}(V_j)$. Таким образом, нетрудно убедиться, что для представленного выше **Утверждения 2.1.1** имеет место также

2.1.2. Следствие

Пусть заданы множество примеров $\mathbf{\Omega}$ и отношение $\mathbf{R}^{\otimes}$. Соответствующее пространство толерантности $<\mathbf{\Omega},\mathbf{R}^{\otimes}>$ может быть представлено в виде объединения классов эквивалентности, формируемых по всем возможным V порождаемыми на $<\mathbf{\Omega},\mathbf{R}^{\otimes}>$ отношениями эквивалентности $\mathbf{E}_v^{\otimes}$.

□

Итак, чтобы иметь основания утверждать[28], что новый (предложенный для прогноза свойств) объект O_0 также имеет свойство P, в обсуждаемой схеме рассуждений следует убедиться, что данный O_0 «попадает» (т.е. может быть помещен как новый элемент) хотя бы в один из классов эквивалентности $\mathbf{E}_{\mathbf{\Omega},v}(O_r)$, сформированных на исходном множестве примеров $\mathbf{\Omega}$ всеми возможными в данном случае отношениями эквивалентности $\mathbf{E}_v^{\otimes}$.

Процедурно этот процесс можно было бы реализовать следующим образом:

- перечислим все классы эквивалентности, покрывающие соответствующие классы сходства пространства толерантности $<\mathbf{\Omega},\mathbf{R}^{\otimes}>$,
- для каждого такого класса эквивалентности (номер l) выделим характеризующую его «общую часть» V_l входящих в него примеров из $\mathbf{\Omega}$;
- проверим, найдется ли среди всех таких V_l по крайней мере одно, вкладывающееся в O_0 .

[28] По-видимому, более аккуратно было бы говорить об *отсутствии аргументов*, позволяющих оспорить такое заключение.

(При этом для «экономной» организации такого процесса «диагностики» свойства P у O_0 достаточно ограничиться, например, порождением лишь минимальных по взаимному вложению общих частей вида V_l). Тем не менее при формировании (переборе) множества всех возможных (для заданного $\mathbf{\Omega}$) максимальных общих частей вида V_{l0} необходимо:

(1) выбрать из $\mathbf{\Omega}$ некоторое множества примеров $O_{i1}, O_{i2}, \dots, O_{ik}$,
(2) убедиться, что сходство объектов $O_{i1} \otimes O_{i2} \otimes \dots \otimes O_{ik} = V_{l0}$ не является пустым подобъектом $O_{\varnothing}$;
(3) проверить, что все элементы в множестве $\{ O_{i1}, O_{i2}, \dots, O_{ik} \}$ попарно различны;
(4) убедиться, что в множество $\{ O_{i1}, O_{i2}, \dots, O_{ik} \}$ включены все примеры из исходного множества $\mathbf{\Omega}$, содержащие V_{l0} как подобъект, и наконец, что
(5) в множестве $\{ O_{i1}, O_{i2}, \dots, O_{ik} \}$ имеется не менее двух примеров (т.е. выделяемое сходство V_{l0} не является тривиальным): $k \geq 2$.

Рассмотрим четыре переборные задачи, характеризующие комбинаторные свойства предложенной конструкции.

2.1.3. Определение

1. Задачу, где

Дано: множество примеров $\mathbf{\Omega}$ (для описания «внутренней» структуры элементов которого использован некоторый формализованный язык представления данных), операция сходства $\otimes$ и натуральное число k.

Требуется: установить, имеется ли в покрытии множества $\mathbf{\Omega}$ классами эквивалентности, сформированным по всем возможным отношениям $\mathbf{E}_v^{\otimes}$, такой класс $\mathbf{E}_{\mathbf{\Omega},V_0}$, который содержит ровно k различных элементов из $\mathbf{\Omega}$,

мы далее будем называть задачей ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА РОВНО*** *k*.

2. Задачу, где

Дано: множество примеров $\mathbf{\Omega}$ (для описания «внутренней» структуры элементов которого использован некоторый формализованный язык представления данных), операция сходства $\otimes$ и натуральное число k.

Требуется: установить, имеется ли в покрытии множества $\mathbf{\Omega}$ классами эквивалентности, сформированным по всем возможным отношениям $\mathbf{E}_v^{\otimes}$, такой класс $\mathbf{E}_{\mathbf{\Omega},V_0}$, который содержит не менее k различных элементов из $\mathbf{\Omega}$,

мы далее будем называть задачей ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА НЕ МЕНЕЕ*** *k*.

3. Задачу, где

Дано: множество примеров $\mathbf{\Omega}$ (для описания «внутренней» структуры элементов которого использован некоторый формализованный язык представления данных), операция сходства $\otimes$ и натуральное число k.

Требуется: установить, имеется ли в покрытии множества $\mathbf{\Omega}$ классами эквивалентности, сформированным по всем возможным отношениям $\mathbf{E}_v^{\otimes}$, такой класс $\mathbf{E}_{\mathbf{\Omega},V_0}$, который содержит не более k различных элементов из $\mathbf{\Omega}$,

мы далее будем называть задачей ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА НЕ БОЛЕЕ*** *k*.

4. Задачу, где

Дано: множество примеров $\mathbf{\Omega}$ (для описания «внутренней» структуры элементов которого использован некоторый формализованный язык представления данных), операция сходства $\otimes$.

Требуется: установить, сколько элементов (различных классов эквивалентности) имеется в покрытии множества $\mathbf{\Omega}$ классами эквивалентности, сформированным по всем возможным отношениям $\mathbf{E}_v^{\otimes}$,

мы далее будем называть задачей ***ЧИСЛО КЛАССОВ ЭКВИВАЛЕНТНОСТИ***.

□

Ниже (в **Главе 4**) мы подробно проанализируем характер сложности этих комбинаторных задач (в т.ч. – доказуемую трудно-разрешимость **Задач 1** и **4** для некоторых полезных типов данных). Здесь лишь отметим, что уже в простейшем случае – при рассмотрении объектов O_i в множестве **Ω** как непустых подмножеств букв некоторого конечного алфавита U, – имеют место следующие три утверждения:

2.1.4. Утверждение

Задача ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА РОВНО*** *k* принадлежит классу ***NPC*** (***NP***-полных задач)

2.1.5. Утверждение

Задача ***ЧИСЛО КЛАССОВ ЭКВИВАЛЕНТНОСТИ*** принадлежит классу #***PC*** (перечислительно полных задач)

2.1.6. Утверждение

Задачи ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА НЕ МЕНЕЕ*** *k* и ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА НЕ БОЛЕЕ*** *k* разрешимы полиномиально быстро.

□

II. Теперь обратимся к более сложному случаю: к ситуации, когда объекты (примеры) из исходно заданного множества **Ω** могут обладать наборами свойств, т.е. к ситуации, когда **A –** это уже не одноэлементное множество свойств $\boldsymbol{P} = \{P_1, P_2, \dots, P_m\}$, и каждому объекту O_i из Ω сопоставлено свое (непустое) подмножество $\boldsymbol{P}_i$ свойств из ***P***.

В предложенном контексте сформулированные выше условия (1) – (5) нам придется дополнить естественным требованием об отборе (в каждый из классов эквивалентности, формируемых нами для «диагностирования» вновь предложенных на прогноз объектов) тех и только тех объектов из исходного множества Ω, которые имеют одни и те же (общие - !) свойства из ***P***. Таким образом, нам придется рассматривать пары $< O_i, \boldsymbol{P}_i >$,,причем такие, что для каждого удовлетворяющего условиям (1) – (5) набора объектов $O_{i1}, O_{i2}, \dots, O_{ik}$ множество их общих свойств $\boldsymbol{P}_V$ не пусто, т.е.

(6) для каждого удовлетворяющего условиям (1) – (5) множества примеров $\{O_{i1}, O_{i2}, \dots, O_{ik}\}$, общая часть которых V не является пустым подобъектом $O_\varnothing$; множество их общих свойств $\boldsymbol{P}_V$ удовлетворяет условию $\boldsymbol{P}_V \neq \varnothing$.

При этом в каждом таком случае пара $<V, \boldsymbol{P}_V>$ будет представлять V как *причину* наличия у объектов из множества$\{O_{i1}, O_{i2}, \dots, O_{ik}\}$ (т.е. «*структурный носитель*[29]») множества свойств $\boldsymbol{P}_V$.

Теперь отнесение нового (предложенного для прогноза свойств) объекта O_0 к классу эквивалентности $\mathbf{E}_{\Omega,V}$ будет означать, что свойства этого объекта полностью определяет множество $\boldsymbol{P}_V$ из соответствующей пары $<V, \boldsymbol{P}_V>$. (Разумеется, для попадания O_0 в класс $\mathbf{E}_{\Omega,V}$ требуется, чтобы O_0 содержал этот непустой подобъект V).

Воспользуемся для описания связи *объект O обладает свойствами* ***P*** частично определенным отношением $=>_1$, а для описания связи *подобъект* V *есть причина* (*структурный носитель*) *множества свойств* $\boldsymbol{P}_V$ - частично определенным $=>_2$. Далее, используем предложенные в [*Финн 10* и др.] логические средства (язык описания решающих предикатов и правил правдоподобного вывода ДСМ-метода), затем выберем для описания примеров из **Ω** их представления в виде множеств букв некоторого алфавита U, а в качестве операции сходства ⊗ используемо операцию ∩ пересечения множеств, после чего запишем с их помощью представленные выше условия (1) – (2). Нетрудно убедиться, что это будет[30] формула следующего вида:

---

[29] Разумеется, с точностью до по-элементного вида исходного множества примеров Ω.

[30] С точностью до индексов у J-функций, характеризующих в ДСМ-предикатах типы истинностных значений и номер текущего шага ДСМ-рассуждения, а также отбора в соответствующий класс эквивалентности из

$$\boldsymbol{M}^{+}_{a}(V,W) = \exists k\ \underline{\boldsymbol{M}^{+}_{a}}(V,W,k) =$$

$$= \exists k\ \exists O_1 ... \exists O_k\ \exists \boldsymbol{P}_1 ... \exists \boldsymbol{P}_k \Big([\ \mathbf{J}_{<1,r>}(O_1 ==>_1 \boldsymbol{P}_1)\ \&$$

$$\&\ \forall \boldsymbol{P}\ (\ \mathbf{J}_{<1,r>}(O_1 ==>_1 \boldsymbol{P}) \rightarrow (\boldsymbol{P} \subseteq \boldsymbol{P}_1))\ ]\ \& \ ...$$

$$... \ \&\ [\ \mathbf{J}_{<1,r>}(O_k ==>_1 \boldsymbol{P}_k)\ \&$$

$$\&\ \forall \boldsymbol{P}\ (\ \mathbf{J}_{<1,r>}(O_k ==>_1 \boldsymbol{P}) \rightarrow (\boldsymbol{P} \subseteq \boldsymbol{P}_k))\ ]\ \&$$

$$\&\ [\ (\bigcap_{i=1}^{k} O_i = V)\ \&\ (V \neq \varnothing)\ ]\ \&$$

$$\&\ [\ \forall i \forall j\ (\ [(i \neq j) \& (1 \leq i,j \leq k)] \rightarrow (O_i \neq O_j)\ )\ ]\ \&$$

$$\&\ \forall X\ \forall Y\ \Big[\ [\mathbf{J}_{<1,r>}(X ==>_1 Y)\ \&$$

$$\&\ \forall \boldsymbol{P}\ (\ \mathbf{J}_{<1,r>}(X ==>_1 \boldsymbol{P}) \rightarrow (\boldsymbol{P} \subseteq Y)\ )\ \&\ (V \subset X)] \rightarrow$$

$$\rightarrow [\ (W \neq \varnothing)\ \&\ (W \subseteq Y)\ \&\ ((X = O_1) \vee ... \vee (X = O_k))\ ]\ \Big]\ \&$$

$$\&\ (k \geq 2)\ \Big)$$

Другими словами, спецификация (уточнение) предложенной выше схемы диагностики свойств новых объектов путем отнесения их (если это возможно) к порождаемым на исходно заданной для обучения выборке примеров, реализованная формальным уточнением соответствующих

- языка представления данных (в данном случае тип данных – множества признаков) и
- формального представления операции сходства на используемом типе представления данных как операции пересечения множеств,

привела нас к формализованной записи одной из базовых конструкций ДСМ-метода[31] - так называемого предиката простого сходства (см. подробнее [*Милль 00, Финн 10*] и др.). Зафиксируем этот факт и двинемся дальше.

III. Перейдем к анализу ситуации, когда исходное множество Ω сформировано как *примерами* (т.е. объектами, которые обладают подлежащими изучению свойствами), так и *контрпримерами* (т.е. объектами, которые, – как известно из каких-либо «внешних» по отношению к выполняемому ИАД источников, – не обладают подлежащими изучению свойствами, однако имеют «подобную» *примерам* структуру[32]): $\boldsymbol{\Omega} = \boldsymbol{\Omega}^{+} \cup \boldsymbol{\Omega}^{-}$, однако, $\boldsymbol{\Omega}^{+} \cap \boldsymbol{\Omega}^{-} = \varnothing$. В такой ситуации особого обсуждения заслуживают следующие два случая:

a) в множествах $\mathbf{E}_{\Omega+}$ и $\mathbf{E}_{\Omega-}$ классов эквивалентности, построенных соответственно на $\boldsymbol{\Omega}^{+}$ и $\boldsymbol{\Omega}^{-}$, найдутся соответственно два класса $\mathbf{E}_{\boldsymbol{\Omega}+,\mathbf{V}+}$ и $\mathbf{E}_{\boldsymbol{\Omega}-,\mathbf{V}-}$, для которых $V^{+} = V^{-}$ при одном и том же множестве свойств $\boldsymbol{P}_V$;

b) предлагаемый для прогноза свойств новый объект O_0 «попадает» одновременно как в некоторый «позитивный» класс эквивалентности $\mathbf{E}_{\boldsymbol{\Omega}+,\mathbf{V}+}$, так и в некоторый негативный класс эквивалентности $\mathbf{E}_{\boldsymbol{\Omega}-,\mathbf{V}-}$.

В первом из выделенных случаев представляется естественным считать, что с точки зрения решаемой задачи о прогнозе свойств новых объектов оба названных класса эквивалентности $\mathbf{E}_{\boldsymbol{\Omega}+,\mathbf{V}+}$ и $\mathbf{E}_{\boldsymbol{\Omega}-,\mathbf{V}-}$ являются носителями некоторой структурной «патологии[33]»

каждого множества всех примеров с одинаковым описанием структуры C_i того из них, который имеет максимальное по вложению множество свойств **A**.

[31] Заметим, сформированной из совершенно других понятийных оснований – из уточнений логическими средствами известных индуктивных канонов рассуждения Д.С.Милля (см., например, [*Милль 00*] и др.)

[32] В их структурах встречаются такие же, как и в структурах *примеров* фрагменты (подобъекты).

[33] Например, в части своих выразительных (дескриптивных) возможностей выбранный язык представления данных может оказаться в конкретном анализируемом случае подобного типа на столько «бедным», что его

(ведь соответствующий структурный фрагмент $V = V^+ = V^-$ оказывается одновременно и *причиной* и *антипричиной*[34] соответствующего множества свойств $\boldsymbol{P}_V$) и, следовательно, должны быть исключены из представленной выше процедуры прогнозирования.

Во втором случае от прогноза свойств нового объекта O_0 отнесением его к классам $\mathbf{E}_{\Omega^+,V^+}$ и $\mathbf{E}_{\Omega^-,V^-}$ также следует отказаться (либо признать такой прогноз «противоречивым»[35], ведь достаточными основаниями отнести O_0 только к расширению множества примеров $\boldsymbol{\Omega}^+$ или же только к расширению множества контрпримеров Ω^- мы в данной ситуации не располагаем).

IV. Теперь перейдем к анализу ситуации, когда предложенный для прогноза свойств новый объект O_0 соотносим сразу с несколькими порождаемыми на исходном множестве примеров $\boldsymbol{\Omega}$ классами эквивалентности *одного знака*. Здесь, если (в случае работы исключительно с объектами знака α[36]) выделены все классы эквивалентности $\mathbf{E}^1_{\Omega_\alpha V_{1,\alpha}}$, $\mathbf{E}^2_{\Omega_\alpha V_{2,\alpha}}$, … $\mathbf{E}^k_{\Omega_\alpha V_{k,\alpha}}$, соотносимые с заданным O_0 (т.е. все соответствующие $V_{1,\alpha}$, $V_{2,\alpha}$, … , $V_{k,\alpha}$ содержатся как фрагменты в O_0), то множество свойств $\boldsymbol{P}_0$ этого анализируемого объекта O_0 представляется естественным сформировать как объединение множеств свойств $\boldsymbol{P}_{V_{1,\alpha}}$, $\boldsymbol{P}_{V_{2,\alpha}}$, , $\boldsymbol{P}_{V_{k,\alpha}}$, каузально обусловленных их структурными носителями $V_{1,\alpha}$, $V_{2,\alpha}$, … , $V_{k,\alpha}$:

$$\boldsymbol{P}_0 = \boldsymbol{P}_{V_{1,\alpha}} \cup \boldsymbol{P}_{V_{2,\alpha}} \cup \ \cup \boldsymbol{P}_{V_{k,\alpha}}$$

При этом (как мы уже договорились выше в разделе **III**), если в O_0 окажутся вкладываемыми «структурные носители» разных знаков (т.е. как причины, так и антипричины наличия соответствующих свойств), то представляется естественным либо вообще отказаться от прогноза свойств нового объекта O_0 на имеющейся сходной выборке примеров $\boldsymbol{\Omega}$, либо объявить результаты такого прогноза вновь установленным *эмпирическим противоречием*.

Не сложно убедиться, что

- представленный выше в разделе **III**a) случай в рамках стандартной конструкции ДСМ-метода (см., например, [*Финн 10*] и др.) воспроизводится в структуре ДСМ-правил правдоподобного вывода первого рода (так называемых ППВ-I), где непротиворечивость порождаемой эмпирической зависимости обусловлена выполнимостью используемого решающего предиката лишь одного знака. В противном случае – это либо порождение эмпирического противоречия, либо отказ от прогноза (сохранение недоопределенности);
- представленные выше в разделах **III**b) и **IV** случаи воспроизводятся в структуре ДСМ-ППВ второго рода (ППВ-II).

Зафиксируем и эти два факта, после чего продолжим двигаться дальше.

V. Обратимся к анализу ситуации, когда исходное множество $\boldsymbol{\Omega}$ содержит как *примеры*, так и *контрпримеры*: $\boldsymbol{\Omega} = \boldsymbol{\Omega}^+ \cup \boldsymbol{\Omega}^-$. Попробуем найти возможности выделять «патологии» представленных выше в разделах **III**a) и **III**b) типов без исчерпывающего формирования из исходного множества примеров $\boldsymbol{\Omega}$ всех возможных классов эквивалентности рассматриваемого вида.

Прежде всего, расширим множество $\mathbf{E}_\Omega$ всех порождаемых из каждого $\boldsymbol{\Omega}$ классов эквивалентности ровно $|\boldsymbol{\Omega}|$ новыми (одноэлементными) классами $\mathbf{E}_{\{O1\}}$, $\mathbf{E}_{\{O2\}}$, …, $\mathbf{E}_{\{On\}}$, (где $n = |\boldsymbol{\Omega}|$ – число элементов в $\boldsymbol{\Omega}$), считая по определению, что для каждого O_i из $\boldsymbol{\Omega}$ множества $\{O_i\}$ есть *вырожденный* (одноэлементный) класс эквивалентности. Тогда условие **III**a) может быть упрощено до вида:

средствами *структурные различия* (по крайней мере некоторых) *причин* и *антипричин* могут оказаться здесь попросту *не представимыми*.

[34] Т.е. причиной отсутствия.

[35] Разумеется, понимая это противоречие не как *логическое*, а как новый э*мпирический факт*.

[36] Где α есть либо +, либо -.

(7) ни один из примеров не должен содержать в себе ни один из порождающих классы эквивалентности на контрпримерах «структурных носителей» анализируемых свойств. Симметричное условие должно связывать также контрпримеры с порождаемыми на множестве примеров «структурными носителями» анализируемых свойств.

Легко видеть, что на языке ДСМ-ППВ-I это будет (см., например, [*Финн 10*] и др.) условие так называемого *запрета на контрпримеры*.

Наконец, подобными только что приведенным выше для случая *запрета на контрпримеры* вариациями «взаимодействия» позитивных и негативных классов эквивалентности несложно воспроизвести «логику» формирования процедур ИАД, являющихся аналогами и для остальных входящих в состав ДСМ-метода решающих предикатов для ППВ-I. При этом, как и ранее, *операционная структура* таких процедур (т.е. своего рода *базовая логика* исполнения ИАД) является общей для различных типов представления анализируемых данных а также для различных уточнений операции сходства ⊗. В части же *достаточности оснований* для принятия порождаемых в ходе исполняемого ИАД результатов – прогнозов свойств новых объектов – базовым элементом используемой конструкции оказывается во всех рассматриваемых случаях *непротиворечивое* (т.е. не порождающее «конфликтов» при «*позиционировании*» конкретного объекта *относительно* релевантных ему *классов эквивалентности*) отнесение соответствующих объектов к классам эквивалентности, покрывающим исходное множество примеров Ω (а с ним - и порождаемые соответствующими отношениями сходства $\mathbf{R}^{\otimes}$ пространства толерантности $<\boldsymbol{\Omega}, \mathbf{R}^{\otimes}>$).

VI. Итак, вопрос о наличии (или же, наоборот, отсутствии) у нового объекта $O\tau$ из множества $\boldsymbol{\Omega}^{\tau}$ искомых свойств $\boldsymbol{P}$ может быть сведен к проверке «попадания» этого объекта в соответствующий класс эквивалентности, восстановленный на пространстве толерантности $<\boldsymbol{\Omega}, \mathbf{R}^{\otimes}>$. (При этом «попадание» в соответствующий класс эквивалентности означает, что сходство нового объекта $O\tau$ со всеми входящими в этот класс эквивалентности объектами есть формирующее этот класс v, а прогнозируемые свойства $\boldsymbol{P}\tau$ этого нового объекта $O\tau$ есть общая часть свойств $\boldsymbol{P}_i$ всех примеров из этого же класса).

Разумеется, для корректного прогнозирования свойств нового объекта $O\tau$ необходимо убедиться, что он попадает в классы эквивалентности лишь одного знака - (+)[37] или (-)[38]. Попадание в классы эквивалентности разных знаков квалифицируется как *эмпирическое противоречие* (противоречивый прогноз). Непопадание ни в один из соответствующих классов эквивалентности означает, что у нас по-прежнему нет данных, чтобы сформировать обоснованный прогноз свойств нового объекта.

Таким образом, для формирования непротиворечивого прогноза требуется (по исходным данным – множествам $\boldsymbol{\Omega}^{+}$ и $\boldsymbol{\Omega}^{-}$):

- сформировать все классы эквивалентности на множестве примеров и на множестве контрпримеров, после чего
- проверить, попадает ли новый объект хотя бы в один класс эквивалентности одного знака, не попадая при этом ни в один из классов эквивалентности противоположного знака.

Заметим, что каждый из классов эквивалентности соответствующего знака характеризуется парой сущностей

$$< \mathrm{v}, \{O^{\mathbf{v}}_{\mathrm{i}1}, O^{\mathbf{v}}_{\mathrm{i}2}, \ldots, O^{\mathbf{v}}_{\mathrm{im(v)}}\} > ,$$

где v – формирующее класс эквивалентности конкретное сходство, а

$$\boldsymbol{O}(\mathrm{v}) = \{O^{\mathbf{v}}_{\mathrm{i}1}, O^{\mathbf{v}}_{\mathrm{i}2}, \ldots, O^{\mathbf{v}}_{\mathrm{im(v)}}\}$$

[37] Примеры.
[38] Контрпримеры.

есть множество *всех тех* и *только тех* примеров из $\mathbf{\Omega}$, которые и образуют именно этот класс сходства. Определив на подмножествах множества примеров[39] $\mathbf{\Omega}$ отношение частичного порядка $\subseteq$ как вложимость соответствующих подмножеств примеров, получим частично упорядоченное множество

$$< \boldsymbol{Eq}(\mathbf{\Omega},\otimes), \subseteq > ,$$

где $\boldsymbol{Eq}(\mathbf{\Omega},\otimes)$ - есть множество всех классов эквивалентности, порождаемых на множестве примеров $\mathbf{\Omega}$ с помощью операции сходства $\otimes$ (и связанного с ней отношения $\mathbf{R}^{\otimes}$), а частичный порядок здесь задается взаимной вложимостью $\subseteq$ соответствующих подмножеств примеров из $\mathbf{\Omega}$, формирующих эти классы эквивалентности.

Таким образом, в рассматриваемой ситуации приходится иметь дело с *особым классом сущностей (**инвариантов**):*

– парами вида < v,$\boldsymbol{O}$(v) >, характеризующими все те и только те примеры из исходного множества $\mathbf{\Omega}$, исчерпывающим (т.е. максимальным по вложению) сходством которых и является соответствующее v;

- диаграммами взаимной вложимости таких сущностей (порождаемыми естественным порядком на соответствующих подмножествах исходного множества примеров $\mathbf{\Omega}$).

Рассмотрим особенности этих характерных для рассматриваемой нами области анализа данных сущностей (*инвариантов*) более подробно.

VII. Пусть имеется множество примеров[40] $\mathbf{\Omega}$, образованное ровно n объектами:

$$\mathbf{\Omega} = \{O_1,O_2,...,O_n\}.$$

Будем считать, что каждый из таких объектов представлен некоторым «*структурным описанием*»: *множеством признаков* (булевским вектором заданной длины), *графом* (плоским и симметризованным, пространственным, …), *цепочкой символов* конечного алфавита и т.п. . Будем также считать, что в нашем распоряжении имеется конструктивная процедура идентификации подобъектов *O'* в «структурных описаниях» рассматриваемых нами объектов. (Так, например, для объектов, заданных множествами признаков, это будут непустые подмножества таких признаков; для объектов, представленных графами, - множества подграфов; для объектов, представленных цепочками символов, - подцепочки таких символов и т.п.).

Тот факт, что «подструктура» *O'* представляет собою подобъект[41] какого-либо из рассматриваемых нами объектов O_i, мы будем представлять выражением специального вида:

$$O' \sqsubseteq O_i \ .$$

Условимся, что для пары множеств - объектов $\{O_{j1},O_{j2},...,O_{js}\}$ из исходного $\mathbf{\Omega}$ и их подобъектов $\{O'_{i1},O'_{i2},...,O'_{ir}\}$ – выражение

$$\{O'_{i1},O'_{i2},...,O'_{ir}\} \subseteq^* \{O_{j1},O_{j2},...,O_{js}\}$$

означает (т.е. представляет собою сокращенную запись того), что для каждого *k* из множества значений {1,2, … ,r} и каждого *l* из множества значений {1,2, … ,s} имеет место «вложимость»

$$O'_{ik} \sqsubseteq O_{jl} \ .$$

Для заданных множества примеров $\mathbf{\Omega} = \{O_1,O_2,...,O_n\}$ и способа выделения «подструктур» *O'* в описаниях («структурах») объектов их $\mathbf{\Omega}$ будем рассматривать пары множеств вида:

---

[39] На контрпримерах такая же конструкция формируется абсолютно аналогичным образом.

[40] Для контрпримеров может быть сформирована абсолютно аналогичная конструкция.

[41] Возможно, совпадающий с самим соответствующим объектом.

$$<\{O'_{i1},O'_{i2},...,O'_{ir}\},\{O_{j1},O_{j2},...,O_{js}\}>$$

таких, что одновременно выполнены следующие три условия:

(1) $\{O'_{i1},O'_{i2},...,O'_{ir}\} \subseteq^* \{O_{j1},O_{j2},...,O_{js}\}$

(2) $\forall O'_{i0}\ (\ [\{O'_{i0}\} \subseteq^* \{O_{j1},O_{j2},...,O_{js}\}] \rightarrow [(O'_{i0}= O'_{i1}) \vee (O'_{i0}= O'_{i2}) \vee \ldots \vee (O'_{i0}= O'_{ir})]$

(3) $\forall O_{j0}\ (\ [\{O'_{i1},O'_{i2},...,O'_{ir}\} \subseteq^* \{O_{j0}\}\ ] \rightarrow [(O_{j0}= O_{j1}) \vee (O_{j0}= O_{j2}) \vee \ldots \vee (O_{j0}= O_{js})]\ )$

т.е. $\{O'_{i1},O'_{i2},...,O'_{ir}\}$ представляет собою максимальное по вложению множество общих подструктур для объектов из соответствующего множества $\{O_{j1},O_{j2},...,O_{js}\}$; кроме того – это множество объектов $\{O_{j1},O_{j2},...,O_{js}\}$ содержит *все те* и *только те объекты* из исходного множества $\mathbf{\Omega} = \{O_1,O_2,...,O_n\}$, которые в качестве подструктур одновременно содержат *все те* и *только те подобъекты*, которые собраны в соответствующем множестве $\{O'_{i1},O'_{i2},...,O'_{ir}\}$.

Таким образом, рассматриваемые здесь пары вида

$$<\boldsymbol{O'},\boldsymbol{O}>,$$

где $\boldsymbol{O'} = \{O'_{i1},O'_{i2},...,O'_{ir}\}$, а $\boldsymbol{O} = \{O_{j1},O_{j2},...,O_{js}\}$ и представляют рассмотренные нами выше классы эквивалентности, восстанавливаемые при анализе классов сходства, формируемых на исходном множестве примеров $\mathbf{\Omega}$ и отношении сходства $\mathbf{R}^{\otimes}$. Чтобы убедиться в этом, условимся считать, что задействованная операция сходства $\otimes$ и порождает нам соответствующий механизм выделения общих подобъектов O' для каждой пары объектов O_{j1} и O_{j2} из исходного множества $\mathbf{\Omega}$. (При этом учтем, что в ряде случаев – например, при выделении сходства графов или же цепочек символов – результатом выделения общих подструктур оказывается соответствующее *множество* таких подобъектов – максимальных общих подграфов, максимальных по вложению общих подцепочек и т.п.. В следствие этого корректное определение необходимой нам операции сходства следует также дать на *множествах* анализируемых примеров, в реально необходимых нам вычислениях оперируя с *одноэлементными множествами* таких объектов: в частности - сходством двух одноэлементных множеств графов в этом случае будет множество всех их максимальных общих подграфов, сходством двух одноэлементных множеств цепочек символов конечного алфавита будет множество их максимальных по вложению общих подцепочек и т.п.).

Несложно убедиться, что при заданных $\mathbf{\Omega}$ и $\otimes$ между компонентами каждой пары $<\boldsymbol{O'},\boldsymbol{O}>$, имеет место взаимно-однозначное соответствие. Другими словами, справедливо

2.1.7. Утверждение

Пусть заданы множество примеров $\mathbf{\Omega}$ и операция сходства $\otimes$. Для каждой пары вида $<\boldsymbol{O'},\boldsymbol{O}>$, удовлетворяющей только что сформулированным выше условиям **(1)** – **(3)**, каждому множеству вида $\boldsymbol{O'}$ соответствует единственное множество вида $\boldsymbol{O}$ и наоборот[42].

Доказательство.

==>. Предположив противное и представив, что одному $\boldsymbol{O'}$ можно сопоставить два различных множества $\boldsymbol{O_1}$ и $\boldsymbol{O_2}$, следует объединить их в общее множество примеров $\boldsymbol{O_{12}}$. Нетрудно убедиться, что условия (1) - (3) окажутся выполнены и для новой пары $<\boldsymbol{O'},\boldsymbol{O_{12}}>$. Однако, в этом случае окажется невозможна одновременная выполнимость условия **(3)** для каждой из пар $<\boldsymbol{O'},\boldsymbol{O_1}>$, $<\boldsymbol{O'},\boldsymbol{O_2}>$ и $<\boldsymbol{O'},\boldsymbol{O_{12}}>$.

<==. Предположив по аналогии с предыдущим случаем, что одному множеству $\boldsymbol{O}$ сопоставлены два различных множества $\boldsymbol{O'_1}$ и $\boldsymbol{O'_2}$, объединим эти два множества в новое

---

[42] Далее в **Главах 3** и **4** при работе с данными в виде множеств признаков эти соображения будут дооформлены до соответствующих утверждений относительно распространенных в некоторых разделах математики эффектов *двойственности* .

множество общих подструктур $\boldsymbol{O'_{12}}$ и придем к противоречию (на невыполнимости условия **(2)** одновременно для всех трех пар $<\boldsymbol{O'_1},\boldsymbol{O}>$, $<\boldsymbol{O'_2},\boldsymbol{O}>$ и $<\boldsymbol{O'_{12}},\boldsymbol{O}>$).

□

Достаточно очевидный вариант процедуры порождения рассматриваемых классов эквивалентности может быть сформирован следующим образом:

если O' – общий фрагмент описаний нескольких прецедентов O_{i1}, O_{i2} , … , O_{ik} , т.е.

$$O' = O_{i1} \otimes O_{i2} \otimes \ldots \otimes O_{ik} \ ,$$

то формируемый на основе этого O' класс эквивалентности $\boldsymbol{Eq}_{<O'>}(\boldsymbol{\Omega},\mathbf{R}^{\otimes})$ может быть порожден последовательной проверкой вложимости всех оставшихся (т.е. не входящих в множество $\boldsymbol{O} = \{O_{i1},O_{i2}, \ldots ,O_{ik}\}$) прецедентов из $\boldsymbol{\Omega}$ в порождаемый этим множеством $\boldsymbol{O}$ фрагмент класса $\boldsymbol{Eq}_{<O'>}(\boldsymbol{\Omega},\mathbf{R}^{\otimes})$, т.е. проверкой выполнимости соотношения

$$O' = O_{ij} \otimes O'$$

для всех оставшихся за пределами множества $\{O_{i1}, O_{i2} , \ldots , O_{ik} \}$ объектов O_{ij} из $\boldsymbol{\Omega}$.

Таким способом на подмножествах вида $\boldsymbol{O} = \{ O_{i1}, O_{i2} , \ldots , O_{ik} \}$ множества прецедентов $\boldsymbol{\Omega}$ может быть определен выполняющий все предложенные проверки и (в определенных случаях) формирующий соответствующее расширение этого множества $\{O_{i1}, O_{i2} , \ldots , O_{ik}\}$ оператор $\boldsymbol{Closure}^{\otimes}$:

$$\boldsymbol{Closure}^{\otimes}: 2^{\Omega} \rightarrow 2^{\Omega}$$

такой, что для любых $\boldsymbol{O}$, $\boldsymbol{O_i}$ и $\boldsymbol{O_j}$ из 2^{Ω}:

(i) $\boldsymbol{O} \subseteq \boldsymbol{Closure}^{\otimes}(\boldsymbol{O})$

(ii) $\boldsymbol{O_i} \subseteq \boldsymbol{O_j}$ влечет $\boldsymbol{Closure}^{\otimes}(\boldsymbol{O_i}) \subseteq \boldsymbol{Closure}^{\otimes}(\boldsymbol{O_j})$

(iii) $\boldsymbol{Closure}^{\otimes}(\boldsymbol{Closure}^{\otimes}(\boldsymbol{O})) = \boldsymbol{Closure}^{\otimes}(\boldsymbol{O})$.

Выполнимость условий (i)-(iii), в свою очередь, означает, что $\boldsymbol{Closure}^{\otimes}$ - это оператор *замыкания* на подмножествах множества прецедентов. *Неподвижными точками* такого замыкания и будут формируемые классы эквивалентности.

Утверждение 2.1.7 предоставляет нам расширенные возможности для организации упорядочений в множестве $\boldsymbol{Eq}_{<O',O>}(\boldsymbol{\Omega},\mathbf{R}^{\otimes})$ всех пар вида $<\boldsymbol{O'},\boldsymbol{O}>$, порождаемых на заданном пространстве толерантности $<\boldsymbol{\Omega},\mathbf{R}^{\otimes}>$. Действительно, имеются по крайней мере следующие два варианта действий:

a) Будем в множестве $\boldsymbol{Eq}_{<O',O>}(\boldsymbol{\Omega},\mathbf{R}^{\otimes})$ представлять каждую пару вида $<\boldsymbol{O'},\boldsymbol{O}>$ ее вторым компонентом $\boldsymbol{O}$, и на множестве таких (характеризующих соответствующие классы эквивалентности) подмножеств примеров из исходного множества $\boldsymbol{\Omega}$ определим отношение частичного порядка $\subseteq$ в соответствии с их взаимной вложимостью. Таким образом,

$$< \boldsymbol{Eq}_{<O',O>}(\boldsymbol{\Omega},\mathbf{R}^{\otimes}) \big|_{O} , \subseteq >$$

будет представлять собою частично-упорядоченное множество, а его диаграмма сможет дать нам дополнительную информацию в части оптимизации поиска именно таких классов эквивалентности, которые релевантны прогнозированию свойств заданного нового объекта $\boldsymbol{O}\tau$ (см. выше п.**VII** в текущем **Разделе**).

Однако имеется и еще одна возможность:

b) Будем в множестве $\boldsymbol{Eq}_{<O',O>}(\boldsymbol{\Omega},\mathbf{R}^{\otimes})$ представлять каждую пару вида $<\boldsymbol{O'},\boldsymbol{O}>$ ее первым компонентом $\boldsymbol{O'}$, и на множествах таких (однозначно характеризующих

соответствующие классы эквивалентности) подструктур – выделенных[43] фрагментов «структурных» описаний исходных объектов из $\mathbf{\Omega}$, задействовав ранее уже представленные (см. выше в текущем п.**VII**) «инструменты» $\sqsubseteq$ и $\subseteq^*$, можно определить[44] также отношение частичного порядка $\preccurlyeq$, отражающее взаимную вложимость соответствующих множеств фрагментов «структурных» описаний. Таким путем может быть задано еще одно описание выделенного нами для изучения структурных особенностей $\boldsymbol{Eq}_{<O',O>}(\mathbf{\Omega},\mathbf{R}^{\otimes})$ частично-упорядоченного множества

$$< \boldsymbol{Eq}_{<O',O>}(\mathbf{\Omega},\mathbf{R}^{\otimes})\,|_{O'}\,, \preccurlyeq >,$$

диаграмма взаимной вложимости (упорядочения) элементов которого - также удобный инструмент поиска ответа на вопрос о прогнозе свойств заданного своим «структурным» описанием нового объекта $\boldsymbol{O}\tau$ (см. выше - п.**VII**).

Далее (в **Главах 3** и **4**) мы подробно обсудим способы и особенности формального уточнения операций сходства для работы с различными типами структурных описаний исходно заданных объектов из множества $\mathbf{\Omega}$. Особое внимание будет уделено изучению комбинаторных свойств некоторых специальных подмножеств множества $\boldsymbol{Eq}_{<O',O>}(\mathbf{\Omega},\mathbf{R}^{\otimes})$ – вычислительной сложности формирования его границ а также удовлетворяющих определенным условиям на структуру и размеры определяемых парами вида $<\boldsymbol{O'},\boldsymbol{O}>$ классов эквивалентности. Затем в **Главе 5** будут подробно рассмотрены алгоритмы формирования диаграмм вложимости

$$< \boldsymbol{Eq}_{<O',O>}(\mathbf{\Omega},\mathbf{R}^{\otimes})\,|_{O}\,, \subseteq > \quad \text{и} \quad < \boldsymbol{Eq}_{<O',O>}(\mathbf{\Omega},\mathbf{R}^{\otimes})\,|_{O'}\,, \preccurlyeq >,$$

а также предложен ряд приемов оптимизации перебора при формировании подобных комбинаторно сложных объектов.

Итак, нами продемонстрирована достаточно общая схема *контроля корректности* осуществления ИАД, формализованного средствами ДСМ-метода. Представленная схема предлагает собственное *корректное* решение вопроса о существовании *достаточных оснований* для принятия результатов выполняемого ИАД. Принципиально важные характерные особенности данной схемы – это:

1) возможности при восстановлении эмпирических зависимостей, содержащихся в неявном виде в накапливаемых экспериментальных данных, использовать *семейство однородных* (выстроенных в *единой «логике»* анализа данных) *вычислительных процедур*;
2) возможности *варьировать* в ходе осуществляемого ИАД (т.е. в ходе применения предложенного семейства однородных вычислительных процедур) «*выразительные возможности» средств представления* изучаемых *данных* (в том числе - возможность «перемещаться» при выборе подходящих дескриптивных средств используемого языка представления данных от простейших – быстро обрабатываемых полиномиально сложными алгоритмами, до весьма детальных, – однако, требующих исчерпывающего и, как правило, экспоненциально сложного перебора вариантов),
3) при этом *выделять* и *прослеживать* (например, на последовательно формируемых расширениях новыми объектами исходной обучающей выборки примеров) соответствующие *инварианты* – те (*устойчиво сохраняющиеся* – т.е. порождаемые средствами ДСМ-решателя задач) классы эквивалентности и сопоставленные им *эмпирические зависимости*, которые (в силу их устойчивого «поведения»[45]) можно рассматривать как обнаруженные в имеющихся данных *эмпирические закономерности* (см. [*Финн 10, Финн 10-а*] и др.).

---

[43] Как «*носители*» особой *характеристики* – конкретного сходства, порождающего соответствующий класс эквивалентности.

[44] Для каждого типа обрабатываемых структурных описаний (множеств, графов, цепочек символов, …) – свое.

[45] Т.е. в силу сохранения («*стабильности*») выявленной «структуры» классов эквивалентности при расширении исходной выборки примеров (и контрпримеров) новыми объектами.

2.2 Эвристика Бэкона-Милля.

Переходя к описанию интересующего нас конкретного варианта КАТ, обратимся, прежде всего, к более детальному представлению соответствующего механизма реализации правдоподобных рассуждений, в основу которого положена *эвристика Бэкона -Милля* - совокупность правил и методологических принципов, способных обеспечить причинный анализ накапливаемых экспериментальных данных, порождение и (в определенной степени[1]) верификацию характеризующих эти данные зависимостей.

Ключевыми идеями предложенной **Ф.Бэконом** конструкции (см, [*Bacon 60, Бэкон 35*] и др., а также соответствующий фрагмент **Раздела 1.2.4**) являются специфический механизм организации *эмпирической структурной индукции* и учет в процессе порождения результатов как *примеров*, так и *контрпримеров* изучаемого явления. **Д.С.Милль** [*Mill 74, Милль 00*], развив идеи **Бэкона** об использовании "*таблиц причин*", предложил для поиска таковых (причин) использовать сформулированные им *методы* ("*каноны*") *сходства*, *различия*, *остатков* и *сопутствующих изменений*.

Согласно **Миллю** (см., например, [*Милль 00*: Книга III "Индукция", Глава VIII "Четыре метода опытного исследования", стр. 310-323] и др.) основное внимание должно быть уделено базирующимся на анализе "сходства" и "разницы" [*Милль 00*: стр.310]

- методу сходства (или взаимного согласия - см. ниже **Замечание 2.2.1**),
- методу различия,

основанному на их взаимодействии

- соединенному методу сходства и различия,
- методу остатков, и
- методу сопутствующих изменений.

Четыре предложенных им "*метода опытного исследования*" (вместе с объединением первых двух из них в соединенный метод сходства и различия) **Милль** формулирует в виде следующих пяти ***Правил*** ("*канонов*"):

Правило 1. (Метод сходства[2]).

> "Если два или более случая подлежащего исследованию явления имеют общим лишь одно обстоятельство, то это обстоятельство, в котором только и согласуются все эти случаи, есть причина (или следствие) данного явления" ([*Милль 00*: стр.313]).

Правило 2. (Метод различия).

> "Если случай, в котором исследуемое явление наступает, и случай, в котором оно не наступает, сходны во всех обстоятельствах, кроме одного, встречающегося лишь в первом случае, то это обстоятельство, в котором только и разнятся эти два случая, есть следствие, или причина, или необходимая часть причины явления" ([*Милль 00*: стр.314]).

Правило 3. (Соединенный метод сходства и различия).

> "Если два или более случая возникновения явления имеют общим лишь одно обстоятельство, и два или более случая невозникновения того же явления имеют общим только отсутствие того же самого обстоятельства, то это обстоятельство, в котором только и разняться оба ряда случаев, есть или следствие, или причина, или необходимая часть причины изучаемого явления." ([*Милль 00*: стр.317-318]).

---

[1] Детально мы уточним это обстоятельство ниже.

[2] См. также **Замечание 2.2.1**.

Правило 4. (Метод остатков).

> "Если из явления вычесть ту его часть, которая как известно из прежних индукций, есть следствие некоторых определенных предыдущих, то остаток данного явления может быть следствием остальных предыдущих" ([*Милль 00*: стр.319]).

Правило 5. (Метод сопутствующих изменений).

> "Всякое явление, изменяющееся определенным образом всякий раз, когда некоторым определенным образом изменяется другое явление, есть либо причина, либо следствие этого явления, либо соединено с ним какой-либо причинной связью" ([*Милль 00*: стр.319]).

□

2.2.1. Замечание

Формально говоря, миллевский термин *method of agreement* следовало бы в те времена - речь идет о второй половине 19 века - переводить на русский язык как *метод* (*взаимного*) *согласия* (или *согласования*). Тем не менее, некоторыми первыми переводчиками **Милля** на русский язык был выбран термин *метод сходства*, удивительным образом предвосхитивший сегодняшнее понимание миллевских идей, базирующееся на алгебраической теории сходства (см., например, [*Гусакова 87, Финн 88*, *Финн 91*] и др.).

В качестве контрпримера, т.е. в противоположность к избранному нами толкованию (а в те годы можно было найти сразу несколько достаточно популярных вариантов изложения или перевода работ **Милля** на русский язык - см. помимо [*Милль 00*] также [*Милль 64, Mill 66, Рождественский 67, Бунге 68*] и др.), приведем вариант перевода, где используется термин *метод согласования*:

> "Основное правило согласования заключается в следующем: "*Если двум или большему числу случаев исследуемого явления природы обще одно лишь обстоятельство, то обстоятельство, в котором все случаи согласуются, есть причина, производящая данное явление, его agens*"." [*Бунге 68*: стр.5[3]].

Формально строгое отношение к толкованию термина method of agreement сегодня может опираться, например, на следующие два обстоятельства.

1. Третье из предложенных в англо-русском словаре **В.К.Мюллера** [*Мюллер 87*: стр.28] семи толкований глагола agree представляет словарная статья:

   соответствовать, гармонировать,
   быть сходным, быть по душе.

2. Девятое из предложенных **COLLINS COBUILD English Language Dictionary** [*COCO 87*, p.30] девяти толкований глагола agree представляет словарная статья

   > **9**. If two stories, accounts, totals, or other things **agree**, they are the same as each other or are so similar that both can be considered as being true or correct[4].

Таким образом, мы имеем все основания по-прежнему следовать точке зрения, сформулированной еще редактором издания [*Милль 00*] - приват-доцентом Императорского Московского Университета **В.Н.Ивановским** - в его Примечании к толкованию термина *method of agreement*:

---

[3] Курсивом выделен фрагмент, где **Н.Бунге** цитирует работу [*Mill 68*].

[4] "Если две истории, расчета, итоговых суммы или иные вещи **agree**, то они совпадают друг с другом или же столь сходны, что могут одновременно рассматриваться как истинные или корректные". (Пер. автора).

> “Иногда *agreement* переводят как *согласие, совпадение*. Проф. **Минто** чрезвычайно удачно называет этот метод методом *единственного сходства*, обращая этим внимание на тот “процесс исключения, посредством которого причинная связь сводится к какой-нибудь одной черте явления... Доказательство опирается здесь не просто на сходство, а на сходство в одном обстоятельстве, при различии во всех прочих” <Проф. ***В.Минто***, *Логика*, 2-е изд., стр.368-417>. (См. [*Милль 00*: Прим. ред. стр. 313]).

2.2.2. Замечание

Весьма существенным (с нашей точки зрения) элементом формулировки ***Правила 1*** (см. также приведенное нами выше редакторское Примечание **В.Н.Ивановского**) является указание на “*исчерпывающий*” характер того одного обстоятельства, “в котором только и согласуются” все анализируемые случаи явления: в описание этого обстоятельства должны быть включены все те (и только те) компоненты, которые и характеризуют взаимную согласованность анализируемых прецедентов. Обозначенное здесь понимание искомого результата как своего рода “*неподвижной точки*” некоторого отображения (ср.: нельзя ограничиться лишь *фрагментом* описания искомого “одного обстоятельства”; аналогично, не следует ограничиваться и анализом лишь *некоторых* прецедентов, необходимо принять во внимание все релевантные исследуемому явлению прецеденты) будет играть важнейшую роль в представленной нами ниже логико-математической формализации миллевских ***Правил***.

2.2.3. Замечание

Легко видеть, что, если ***Правила 1-4*** отражают качественный характер взаимосвязи *причина-следствие* (для описания которого вполне достаточно уточнения “структурных” характеристик изучаемых сущностей - “явлений, обстоятельств и их частей”), то ***Правило 5*** (Метод сопутствующих изменений) предполагает использование более мощных выразительных средств, ориентированных на учет “количественных изменений причины [*Милль 00*: стр.323])“. Эта специфика Метода сопутствующих изменений также будет особым образом учтена нами в представленной ниже логико-математической формализации миллевских “канонов”.

Среди множества попыток осуществить математическую формализацию идей ***Бэкона-Милля*** отметим прежде всего работу ***Г. фон Вригта*** [*von Wright 60*], в которой досконально исследованы возможности так называемой *логики условий*. Автор использовал три понятия для реконструкции индуктивных рассуждений: понятие *необходимого условия*, понятие *достаточного условия*, а также понятие *необходимого и достаточного условия*. Кроме того им рассматривается процедура *вероятностной индуктивной элиминации* возможных причинных связей с помощью наблюдаемых событий. ***Г. фон Вригт*** показал, что с помощью индуктивной элиминации можно получить *достоверный вывод*: когда все возможные достаточные условия элиминированы, то одно оставшееся условие будет конъюнкцией всех возможных обуславливающих (т.е. каким-то образом связанных с интересующим нас свойством) свойств. При поиске необходимого условия оставшееся после элиминации условие будет дизъюнкцией всех обуславливающих свойств. Следует отметить, что точная реализация подхода ***Г. фон Вригта*** затруднена значительным объемом перебора при выделении искомых условий и обобщений.

Другая серия работ по формализации индуктивных методов ***Бэкона-Милля*** выполнена польскими логиками. ***Е.Лось***, ставя в [*Los 77*] вопрос о формализации канонов ***Милля***, предлагает специальный формально корректный язык, на основе которого и предпринимается попытка решить поставленную задачу. Далее предлагаются правила вывода, позволяющие по имеющим место в момент времени ***T*** явлениям $z_1,\ldots,z_n$ и имеющим место в момент времени $T+t$ явлениям $z_{n+1},\ldots,z_{n+m}$ делать заключения о наступлении в момент времени $T+t$ некоторых явлений $z_{n+m+1},\ldots,z_{n+m+r}$, определенным образом связанных с явлениями $z_1,\ldots,z_{n+m}$. Далее предлагается аксиоматизация соответствующих логик, определяемых средствами построенного языка.

В работе [*Czerwinski 60*], рефератом которой является работа [*Магинскас 63*], также делается попытка сформулировать миллевские каноны в языке, близком к языку самого ***Милля***, но более точном и строгом с логической точки зрения. Предлагаемая формализация вновь основывается на понятии "*время*", вводимом при помощи счетно-бесконечного упорядоченного множества элементарных временных промежутков - "секунд", а также понимании "времени" t как конечной непустой последовательности следующих одна за одной "секунд". Умозаключения по методам ***Милля*** рассматриваются как *достоверные умозаключения*, опирающиеся на некоторые общие предположения (помимо посылок, описывающих экспериментальные данные, зависящие от конкретных предметов и конкретных времен). Эти общие предположения должны утверждать некоторого рода *детерминизм* и *отсутствие влияний* на явление, причину которого мы ищем, со стороны явлений, не принадлежащих некоторому классу (таким образом, постулируется требование *относительной изолированности* изучаемых систем). Далее рассматриваются различные уточнения (осуществляемые в рамках выбранного подхода) понятия причины: так называемые *причина явления*, *причина события* и *основная причина явления*. К сожалению, авторы рассматриваемых работ не ставят перед собой задачу эффективной реализации предлагаемого подхода в виде каких-либо *алгоритмических* конструкций.

Наконец, практически исчерпывающий обзор[5] литературы, посвященной основанным на использовании различных *вероятностных* конструкций *индуктивным методам*, можно найти в известной работе ***Г.Кайберга*** [*Кайберг 78*].

[5] Он охватывает около 2000 работ.

2.3 ДСМ-метод: общее описание и наивная семантика.

Предпринятая **В.К.Финном** а также его учениками и коллегами (см., например, [*Финн 76, Финн 81-а, Финн 83, Забежайло 82-83, САФИД 85, САФИД 88*] и др.) попытка формализации миллевских "канонов" средствами многозначных логик привела к созданию так называемого *ДСМ-метода автоматического порождения зависимостей из эмпирических данных* (см., например, обзор публикаций по ДСМ-методу [*Кузнецов 91*] и др.). Среди положенных в основу этой формализации идей важнейшая роль отводится:

- миллевским представлениям о "*внутренней структуре*" *правил правдоподобного вывода*;
- юмовским представлениям об *эпистемологии* (т.е. механизмах *оценивания* результатов правдоподобного вывода);
- возможностям использования *многозначных логик* (в частности, представлениям о необходимости разделения соответствующих *внутренних* и *внешних языков* при формализации правдоподобных рассуждений, конкретизирующим известные идеи **Д.А.Бочвара** [*Бочвар 38*] и т.п.);
- возможностям использовать в реализации правдоподобных рассуждений бэконовские идеи об одновременном анализе *позитивных* и *негативных прецедентов* изучаемого явления;
- представления о *верификации* порождаемых в процессе правдоподобного вывода результатов путем *обзора всех возможных фальсификаторов* для них, а также представления о возможностях принятия порождаемых результатов в случае отсутствия для них (подходящих по "силе") фальсификаторов, порождаемых из имеющегося эмпирического материала (ср. с обсуждавшимися ранее идеями **К.Поппера** и **А.С.Есенина-Вольпина**);
- возможностям учесть при порождении *индуктивных обобщений* бэконовские представления о роли *позитивных* и *негативных "влияний"* на анализируемый результат, а также (подобные описанным **Д.Пойа**) представления о структуре правдоподобных выводов по *аналогии*;
- представления (учитывающие соображения **Ф.Бэкона** и организованные в стиле **К.Поппера**) о *достаточности оснований* для принятия результатов правдоподобного вывода; и наконец,
- представления о роли формальных уточнений понятия *сходство* (и, в частности, формализации сходства как алгебраической операции) для построения адекватных (машинно-ориентированных) формализаций правдоподобных выводов в стиле **Бэкона-Милля**.

При этом оказалось, что полученная формальная конструкция (ее детальное описание - чуть ниже), помимо прочего,

- предлагает (реализует) конструктивный вариант *абдукции* в стиле **Ч.С.Пирса** (т.е. представляет собой конструктивный[1] механизм вывода *к гипотезе*); и
- реализует оригинальный вариант *немонотонных рассуждений*, расширяющий маккартиевские представления об эффектах немонотонности выводов в интеллектуальных системах.

С неформальной точки зрения ДСМ-метод характеризуется следующей последовательностью действий (а также условиями их реализуемости):

[1] Т.е. в явном виде порождающий искомый результат - гипотезу.

0. Изучаются причины наличия (или отсутствия) заданных свойств у объектов, наделенных “внутренней структурой”. При этом предполагается, что выполняются следующие *условия* (*применимости ДСМ-метода*):
    - анализируемая предметная область характеризуется хорошо структурированными данными и плохо формализованными знаниями;
    - в рассматриваемой предметной области существуют “причинные зависимости”, характеризующие (возможно, в неявной форме) наличие и\или отсутствие у изучаемых объектов избранных для анализа свойств;
    - имеется возможность подобрать в анализируемой предметной области как некоторое количество примеров, когда рассматриваемые объекты обладают изучаемыми свойствами, так и некоторое количество примеров, когда рассматриваемые объекты не обладают таковыми свойствами.
1. Рассматриваются два множества примеров явлений (объектов, характеризуемых наличием или отсутствием анализируемых свойств):

    - те, где наличие или отсутствие нужных свойств заранее установлено (мы будем называть это множество *обучающей выборкой*, образованной множествами *примеров* и *контрпримеров*), и

    - те, где наличие или отсутствие нужных свойств предстоит установить[2] в ходе ДСМ-рассуждения (т.е. рассуждения, осуществляемого с использованием представленных ниже правил правдоподобного вывода в рамках ДСМ-метода).
2. В основу ДСМ-рассуждения положен анализ сходства (элементов обучающей выборки), формализованного как алгебраическая операция.
3. Формулируются правила правдоподобного вывода (в рамках ДСМ-метода автоматического порождения зависимостей из данных), составляющие два множества

    - правила *порождения* (в ходе анализа сходства на элементах обучающей выборки) *эмпирических зависимостей*, характеризующих *причины* наличия (или отсутствия) у элементов обучающей выборки анализируемых свойств (так называемые *правила первого рода* - ППВ-I), и

    - правила *переноса* (там, где это возможно) найденных с помощью ППВ-I зависимостей на новые примеры явлений, наличие (или отсутствие) искомых свойств у которых и требуется установить в ходе ДСМ-рассуждения (так называемые *правила второго рода* - ППВ-II).
4. Определяется порядок применения ППВ-I и ППВ-II - стратегии ДСМ-рассуждения (т.е. конкретные последовательности ДСМ-правил правдоподобного вывода, используемые в ходе ДСМ-рассуждения). На множестве ДСМ-ППВ задается отношение *частичного порядка*, позволяющее естественным образом упорядочить “*по силе*” элементы множества *гипотез*, порождаемых теми или иными ДСМ-ППВ.
5. Сформулирован *критерий достаточности оснований* для принятия результатов ДСМ-рассуждения: требуется, чтобы порождаемое ДСМ-анализом обучающей выборки множество гипотез непротиворечивым образом “объясняло” каждый элемент обучающей выборки[3]. В случае невыполнения этого условия предполагается *управляемое расширение* обучающей выборки примерами, которые сходны с “необъясненными”, в стремлении уже на расширенной выборке получить правдоподобный вывод на достаточном основании.
6. Стандартная методика реализации правдоподобного вывода в рамках ДСМ-метода состоит в следующем:

    - анализом обучающей выборки (т.е. применением к ее элементам ППВ-I) порождаются эмпирические зависимости, которые затем

---

[2] Если это возможно. (Позже мы продемонстрируем примеры ситуаций, где обсуждаемый прогноз средствами ДСМ-метода реализовать не удается).

[3] Т.е. каждый элемент обучающей выборки мог быть получен в результате непротиворечивого ДСМ-рассуждения, стартующего с этой обучающей выборки.

- используются для прогнозирования целевых свойств у тех исходно заданных примеров, свойства которых и предстоит установить в рамках ДСМ-рассуждения. Затем объекты, свойства которых спрогнозированы[4] этой процедурой, добавляются[5] к обучающей выборке, и
- процесс порождения зависимостей (теперь уже на новой, расширенной выборке) итерируется до тех пор, пока свойства всех изначально заданных для прогноза примеров не будут установлены, либо до тех пор, пока в процессе порождения гипотез не перестанут выводиться новые гипотезы. (Последнее состояние мы будем называть состоянием *стабилизации* ДСМ-вывода).

7. Строится дедуктивная имитация процесса правдоподобного вывода в рамках ДСМ-метода (т.е. "погружение" конкретного ДСМ-рассуждения в специальную аксиоматическую теорию), позволяющая проводить дедуктивный анализ совместимости Базы знаний (БЗ) соответствующей интеллектуальной системы и результатов правдоподобного вывода (т.е. гипотез, получаемых применением правил правдоподобного вывода ДСМ-метода). Таким образом задаются дополнительные средства управления принятием результатов в соответствующей КАТ, основанные на анализе выводимости соответствующих утверждений (средствами достоверного вывода) из аксиом БЗ интеллектуальной системы.

Обратимся к описанию *наивной*[6] *семантики* ДСМ-метода. Пусть заданы два конечных множества:

- $\boldsymbol{C} = \{C_1, C_2, ..., C_n\}$ - объектов, характеризуемых наличием определенной "внутренней структуры" таким образом, что есть возможность выделять "подобъекты" в объектах из множества $\boldsymbol{C}$ (при этом мы будем считать что множество всех возможных подобъектов для объектов из $\boldsymbol{C}$ есть некоторое множеств $\boldsymbol{C'}$), и
- $\boldsymbol{A} = \{A_1, A_2, ..., A_m\}$ - анализируемых свойств у объектов из множества $\boldsymbol{C}$.

На декартовом произведении $\boldsymbol{C}$x$\boldsymbol{A}$ задается частично определенное отношение $==>_1$ *обладать свойством*, интерпретируемое как *объект* C_i *обладает свойствами* A_j и образованное парами $<C_i,A_j>$ из множеств следующих трех типов

- $\boldsymbol{\Omega}^+$ - множество пар $<C_i,A_j>$ (их мы будем также называть *позитивными примерами* или просто *примерами*), таких что объект C_i действительно обладает свойствами A_j;
- $\boldsymbol{\Omega}^-$ - множество пар $<C_i,A_j>$ (их мы будем также называть *негативными примерами* или *контрпримерами*), таких что объект C_i не обладает свойствами A_j;
- $\boldsymbol{\Omega}^\tau$ - множество пар $<C_i,A_j>$ (их мы будем также называть *недоопределенными примерами*), таких что о наличии у объекта C_i свойств A_j нет сведений (эти знания и предполагается породить в ходе ДСМ-рассуждения).

В свою очередь на декартовом произведении $\boldsymbol{C'}$x$\boldsymbol{A}$ задается частично определенное отношение $==>_2$ *быть причиной*, интерпретируемое как *подобъект* C'_i *является причиной наличия\отсутствия свойств* A_j и образованное парами $<C'_i,A_j>$ из множеств следующих четырех типов

---

[4] В процессе прогноза возможна ситуация, когда соответствующая гипотеза о свойствах "нового" объекта может быть порождена как применением ППВ, утверждающих наличие искомых свойств (так называемый *позитивный* прогноз), так и применением ППВ, утверждающих отсутствие искомых свойств (*негативный* прогноз). В этом случае результат применения ДСМ-ППВ объявляется *противоречивым*. Более точно: в случае *равносильности* (в смысле частичного порядка на соответствующих ППВ) позитивного и негативного прогнозов результат объявляется противоречивым. В противном случае предпочтение отдается "знаку", который имеет более сильная (в смысле названного упорядочения) гипотеза.

[5] За исключением результатов противоречивого прогноза.

[6] Использование этого термина отражает полуформальный (т.е. требующий более детального уточнения) характер ряда используемых в ней конструкций.

- ω^+ - множество пар $<C'_i,A_j>$ (их мы будем также называть *позитивными гипотезами первого рода*), таких что подобъект C'_i действительно обладает свойствами A_j (т.е. C'_i считается "причиной" присутствия для A_j);
- ω^- - множество пар $<C'_i,A_j>$ (их мы будем также называть *негативными гипотезами первого рода*), таких что объект C'_i не обладает свойствами A_j (т.е. C'_i считается "причиной" отсутствия для A_j или, другими словами, "антипричиной, контрпричиной" для присутствия A_j);
- ω^0 - множество пар $<C'_i,A_j>$ (их мы будем также называть *противоречивыми*[7] *гипотезами первого рода*), таких что информация о наличии\отсутствии у подобъекта C'_i свойств A_j противоречива;
- ω^τ - множество пар $<C'_i,A_j>$ (их мы будем также называть *недоопределенными гипотезами первого рода*), таких что о наличии у подобъекта C'_i свойств A_j нет сведений (эти знания и предполагается породить в ходе ДСМ-рассуждения).

Описание наивной семантики ДСМ-рассуждения завершает следующая процедурная схема:

- стартуя с элементов множеств Ω^+ и Ω^- , применением правил правдоподобного вывода первого рода (ППВ-I), мы будем стремиться породить все возможные элементы множеств ω^+, ω^- и ω^0; после чего
- использовать порожденные гипотезы из множеств ω^+ и ω^- для доопределения (там, где это возможно) применением правил правдоподобного вывода второго рода (ППВ-II) элементов множества Ω^τ, затем
- пополнить вновь доопределенными[8] элементами множества Ω^τ соответствующие множества примеров и контрпримеров, после чего итерировать процесс ДСМ-вывода до ситуации, когда соответствующее множество $\Omega^\tau = \varnothing$, либо же до наступления состояния стабилизации (см. выше).

(Таким образом, речь идет о максимально возможном доопределении использованием правил правдоподобного вывода в рамках ДСМ-метода - т.е. в процессе ДСМ-рассуждения - частично определенных отношений $==>_1$ и $==>_2$.)

Задавая те или иные конкретные спецификации понятия *внутренняя структура* объектов из множества $\boldsymbol{C}$ (и, соответственно, *структура подобъектов* из множества $\boldsymbol{C'}$), мы будем получать те или иные варианты наивной семантики для соответствующей формализации ДСМ-метода (т.е., другими словами, для формализации обработки средствами ДСМ-метода тех или иных конкретных *типов данных*).

В качестве примера рассмотрим более подробно такую спецификацию в случае обработки данных, представленных множествами. (Мы будем говорить в этом случае, что *тип* обрабатываемых *данных* - *множества*[9]).

Пусть $\mathbf{U^{(1)}} = \{c_1, c_2, \dots, c_n\}$ универсум образующих[10] для объектов, а $\Omega_C = \{C_1, C_2, \dots, C_m\} \subseteq 2^{\mathbf{U^{(1)}}}\backslash\varnothing$ - множество объектов (т.е. непустых множеств образующих из $\mathbf{U^{(1)}}$), построенных над универсумом $\mathbf{U}$[11]. Пусть $\mathbf{U^{(2)}} = \{a_1, a_2, \dots, a_n\}$ - множество образующих для свойств объектов, а

---

[7] Точный смысл этого термина будет определен чуть позже уточнением следующей идеи: противоречивыми оказываются те гипотезы первого рода, которые могут быть порождены одновременно применением и *позитивных*, и *негативных* ППВ-I.

[8] При этом результаты противоречивого доопределения (т.е. те, которые получены применением равных по силе ППВ-II, утверждающих *наличие* искомых свойств, с одной стороны, и их *отсутствие* - с другой) помещаются в множество Ω^0 и исключаются из соответствующего пополнения множеств примеров и контрпримеров.

[9] Ниже речь пойдет также об обработке не только собственно множеств, но и нумерованных множеств (мультимножеств), кортежей (в том числе - кортежей мультимножеств), а также числовых данных в виде точек и интервалов на действительной оси.

[10] При этом, не теряя общности, мы будем считать, что в $\mathbf{U}$ нет одинаковых элементов.

[11] Т.е. множество объектов, построенных из образующих $a_1, a_2, \dots, a_n$.

$$\mathbf{\Omega_A} = \{A_1, A_2, \dots, A_m\} \subseteq 2^{\mathbf{U^{(2)}}}\backslash\varnothing$$

есть множество свойств объектов из $\mathbf{\Omega_C}$. Пусть, наконец,

$$\mathbf{==>_1} \subseteq \{2^{\mathbf{U^{(1)}}}\backslash\varnothing\} \times \{2^{\mathbf{U^{(2)}}}\backslash\varnothing\}$$

есть определенное на декартовом произведении множеств $\{2^{\mathbf{U^{(1)}}}\backslash\varnothing\}$ и $\{2^{\mathbf{U^{(2)}}}\backslash\varnothing\}$ частично определенное отношение "*обладать свойством*", такое, что выполнимость выражения

$$\mathbf{J_\nu}\ (C ==>_1 A),$$

где

$$\mathbf{J_\nu}\ \varphi = \begin{cases} 1, & \text{если формула } \varphi \text{ принимает истинностное значение } [\varphi] = \nu\ ; \\ 0\ , & \text{в противном случае;} \end{cases}$$

мы будем интерпретировать как

- объект C обладает свойствами A в случае $\nu = 1$,
- объект C не обладает свойствами A в случае $\nu = 0$,
- не известно, обладает ли объект C свойствами A, в случае $\nu = \tau$.

Пусть $\mathbf{\Omega}$ есть множество пар <C,A>. Будем рассматривать случай, когда

$$\mathbf{\Omega} = \mathbf{\Omega^+} \cup \mathbf{\Omega^-} \cup \mathbf{\Omega^\tau}\ ,$$

где

$$\mathbf{\Omega^+} \cap \mathbf{\Omega^-} = \mathbf{\Omega^-} \cap \mathbf{\Omega^\tau} = \mathbf{\Omega^+} \cap \mathbf{\Omega^\tau} = \varnothing\ ,$$

а множества $\mathbf{\Omega^i}$ ($i \in \{+,-,\tau\}$) будут интерпретироваться нами как

- множество позитивных примеров

$$\mathbf{\Omega^+} = \{<C,A>|\ \mathbf{J_1}(C==>_1 A)\}$$

- множество негативных примеров

$$\mathbf{\Omega^-} = \{<C,A>|\ \mathbf{J_0}(C==>_1 A)\}$$

и

- множество недоопределенных примеров

$$\mathbf{\Omega^\tau} = \{<C,A>|\ \mathbf{J_\tau}(C==>_1 A)\} \subseteq (\{2^{\mathbf{U^{(1)}}}\backslash\varnothing\} \times \{2^{\mathbf{U^{(2)}}}\backslash\varnothing\}) \setminus (\mathbf{\Omega^+} \cup \mathbf{\Omega^-}).$$

"Технологическая цепочка" правдоподобных рассуждений в рамках ДСМ-метода определяется последовательным применением правил правдоподобного вывода в рамках следующей схемы:

- использованием правил первого рода (см. подробнее ниже **Главу 3**), т.е. анализом сходства объектов и свойств на множествах $\mathbf{\Omega^+}$ - позитивных примеров и $\mathbf{\Omega^-}$ - негативных примеров порождаются, соответственно, множества $\boldsymbol{Hp}(\mathbf{U^{(1)},U^{(2)},\Omega^+,\Omega^-})$ - позитивных[12] гипотез (т.е. утверждений вида

  $$\mathbf{J_1}\ (C' ==>_2 A),$$

  представляющих гипотезы о причинах наличия анализируемых свойств в рассматриваемых примерах из множества $\mathbf{\Omega^+}$) и $\boldsymbol{Hp}(\mathbf{U^{(1)},U^{(2)},\Omega^-,\Omega^+})$ - негативных гипотез (т.е. утверждений вида

  $$\mathbf{J_{-1}}\ (C' ==>_2 A),$$

[12] Упоминание обоих множеств $\mathbf{\Omega^+}$ и $\mathbf{\Omega^-}$ в выражении $\boldsymbol{Hp}(\mathbf{U^{(1)},U^{(2)},\Omega^+,\Omega^-})$ означает, что порождение гипотез осуществляется на множестве объектов $\mathbf{\Omega^\alpha}$ с учетом (как, например, при реализации условия *Запрета на Контрпримеры* - см. подробнее ниже **Главу 3**) структуры и свойств объектов также и в множестве $\mathbf{\Omega^{-\alpha}}$.

представляющих гипотезы о причинах отсутствия анализируемых свойств в рассматриваемых примерах из множества $\boldsymbol{\Omega^-}$);

- противоречивые гипотезы первого рода помещаются в множество $\boldsymbol{Hp^{(0)}(U^{(1)},U^{(2)},\Omega)}$;
- использованием правил второго рода (см. подробнее ниже **Главу 3**), т.е. анализом ДСМ-аналогий между порожденными непротиворечивыми гипотезами и вновь анализируемыми объектами из множества $\boldsymbol{\Omega^\tau}$, порождаются утверждения о наличии или отсутствии анализируемых свойств у объектов из множества примеров $\boldsymbol{\Omega^\tau}$. Таким образом осуществляется *доопределение* исходно недоопределенных примеров из множества $\boldsymbol{\Omega^\tau}$;
- результаты противоречивого доопределения помещаются в множество $\boldsymbol{\Omega^0}$;
- затем вновь доопределенные (непротиворечивым образом - !) объекты присоединяются к имеющимся множествам $\boldsymbol{\Omega^+}$ - позитивных примеров (для объектов, которым спрогнозировано наличие анализируемых свойств) и $\boldsymbol{\Omega^-}$ - негативных примеров (для объектов, которым спрогнозировано отсутствие анализируемых свойств). Далее процесс анализа сходств и доопределения все еще недоопределенных объектов итерируется до состояния исчерпанности множества $\boldsymbol{\Omega^\tau}$, либо до состояния стабилизации.

Эта “технологическая цепочка” задействована при обработке любых адекватных применению ДСМ-метода типов данных - множеств, мультимножеств, кортежей, числовых векторов и др.. Различия при переходе от обработки одного типа данных к обработке другого имеют место только, так сказать, на “микро уровне” и состоят лишь в “перенастройке” средств представления объектов и их свойств в соответствующих правилах правдоподобного вывода - ППВ (вид которых, в свою очередь, - на “макро уровне” - остается одним и тем же для всех этих типов данных).

Глава 3

Процедурная платформа ДСМ-ИАД.

3.1 Представления о сходстве в ДСМ-методе.

Приведенные выше неформальное описание *эвристики* **Бэкона-Милля** и общая структура ДСМ-рассуждения дают достаточно веские основания для того, чтобы обратиться к формальному уточнению сходства как базовой компоненты формализованной конструкции ДСМ-метода.

Целью такого уточнения является переход от *интуитивного* понимания сходства у **Милля** к его корректной *математической формализации*[1], причем выполненной таким средствами, которые позволяли бы осуществить эффективную алгоритмическую реализацию процедур анализа сходства в рамках тех или иных компьютерных систем.

3.1.1. Общие представления о сходстве как формальной конструкции.

Исторические корни исследования природы сходства в науке (см., например, [*Воронин 91*] и др.), пожалуй, ярче всего проявляются в науках о жизни, и в частности, в биологии. Значительное число теоретических построений здесь в той или иной мере связаны с анализом сходства объектов исследования через их соотнесение с соответствующей классификационной схемой. Как правило, последовательность действий в этом случае такова:

- собирается коллекция (по возможности - *полная*[2]) изучаемых объектов эмпирического мира;
- для описания собранной коллекции формируется некоторое множество признаков, различные совокупности которых информативны в опознании и взаимном различении элементов этой коллекции;
- на множестве всех отобранных признаков задается некоторая структура (обычно - это иерархия подмножеств признаков, либо иерархия подмножеств признаков и специальных отношений на них) - классификационная схема;
- сходными (в рамках описываемой конструкции) провозглашаются объекты, соответствующие “близким” (например, в значительной мере совпадающим) фрагментам построенной классификационной схемы.

Математическая формализация представленной последовательности действий, цель которой - порождение (в явной или неявной форме) соответствующего отношения *толерантности* (т.е. рефлексивного и симметричного отношения - см., например, [*Якубович 68*, *Шрейдер 71-а*, *Цаленко 85*] и др.), осуществляется, как правило, средствами одного из трех нижеследующих формализмов:

- математической теории меры (в частности, вероятностного характера - см., например, [*Воронин 91*] и цитируемую там библиографию),
- метрического анализа близости (см., например, метрическую формализацию сходства в задачах распознавания образов [*РКП 89*]), и

---

[1] Здесь, если следовать уже достаточно длительное время распространенной в работах по Искусственному Интеллекту терминологии о моделировании интеллектуальных способностей человека, можно вести речь о переходе от так называемых “*правополушарных*” – характерных (локализованных) для правого полушария мозга у человека-правши (см., например, [*Спрингер 83*] и др.) интуитивных, плохо формализованных - построений к “*левополушарным*” (точно определенным, алгоритмически “прозрачным” и, следовательно, корректно переносимым в те или иные алгоритмические модели) конструкциям, ориентированным на использование в разработке интеллектуальных компьютерных систем.

[2] Разумеется, в том смысле, в котором это практически возможно.

- уточнения сходства как алгебраической операции, удовлетворяющей соответствующим аксиомам (см., например, работу [*Гусакова 87*] и приведенный в ее третьем параграфе краткий обзор операционально задаваемых сходств).

В первом из перечисленных случаев задействована идея *близости* в смысле *меры*, технически описываемая средствами расширения соответствующей булевой алгебры (т.е. множества, замкнутого относительно обычным образом определяемых операций объединения, пересечения и дополнения, имеющего к тому же наибольший и наименьший по вложению элементы) с помощью определенной на ее элементах специальной аддитивной функции - меры, удовлетворяющей следующим аксиомам:

(1) $\mu(A)\geq 0$ для любого элемента A рассматриваемой булевой алгебры ***B***;
(2) $\mu(A)=0$ означает, что A есть наименьший элемент в ***B***;
(3) $(A\cap B = \varnothing) \supset \mu(A\cup B) = \mu(A) + \mu(B)$ для любых A и B из ***B*** [3].

Сходными в данной формализации объявляются объекты, *мера сходства* которых "*достаточно велика*" (в частности, объекты A и B считаются сходными, если мера их объединения $\mu(A\cup B)$ *мало отличается* от мер $\mu(A)$ и $\mu(B)$).

Во втором из представленных выше случаев в основу формализации кладется определение понятия *расстояния* с помощью специальной функции - *метрики*, удовлетворяющей следующим аксиомам:

(1) $\rho(x,y)\geq 0$ и $\rho(x,x)=0$ для всех x и y;
(2) $\rho(x,y) = \rho(y,x)$ для всех x и y;
(3) $\rho(x,z) \leq \rho(x,y) + \rho(y,z)$ для всех x и y[4].

Сходными в данной ситуации объявляются объекты, которые "*достаточно*" *близки*, т.е. *расстояние* между которыми не превышает заданного порога. (Следует заметить, что часть решений в рамках первой модели формализации - а именно, те случаи, где мера μ задается на элементах соответствующего *метрического пространства*, - очевидным образом может быть переформулирована также и средствами второй модели формализации. В свою очередь, нетривиальный пример "ослабления" выразительных средств второй модели[5] до, фактически, перехода в "область действия" первой модели дан в работе [*Кочетков 89*].

Одним из популярных вариантов развития второй модели является формализм так называемых *коэффициентов ассоциации* (иногда их также называют *мерами сходства*) *по A.Tversky* (см., например, [*Tversky 77, Bradshaw 97*] и др.). Фактически, здесь речь идет об использовании при сравнении "сходства" двух булевских векторов (одинаковой длины) следующих четырех параметров:

- a - число единиц в первом из векторов, не встречающихся во втором;
- b - число единиц во втором из векторов, не встречающихся в первом;
- c - число единиц, входящих одновременно в оба вектора;
- d - число нулей, входящих одновременно в оба вектора.

Далее из этих коэффициентов строятся алгебраические выражения того или иного конкретного вида, с помощью которых и дается та или иная "интегральная оценка" меры "близости" анализируемых булевских векторов.

Так, например, величина

$$H = \frac{d+c}{a+b+c+d}$$

---

[3] Так называемая аксиома аддитивности меры.

[4] Так называемая аксиома треугольника.

[5] Получаемый отказом от "метрического" понимания близости и переходом из метрических пространств к топологическому определению понятия близости с использованием понятия *система окрестностей*.

характеризует меру близости в смысле метрики Хэмминга (см., например, [*Биркгоф 76*, *Peterson 61*] и др.), а величина

$$T = \frac{c}{a+b+c}$$

представляет собой так называемую меру сходства в смысле Tanimoto (см., например, [*Bradshow 97, Tanimoto 87*] и др.).

Третий из представленных выше случаев развивает процедурный подход, в рамках которого сходство задается с помощью алгебраической операции, позволяющей, помимо прочего, предъявить и "носитель" сходства - объект того же сорта, что и те, к которым применяется операция сходства (результатом которой оно, в свою очередь, и является). При этом операция сходства * должна удовлетворять аксиомам идемпотентности, коммутативности и ассоциативности[6], например, (в случае бинарного сходства):

(1) a*a = a для всех a;
(2) a*b = b*a для всех a и b;
(3) a*b*c = (a*b)*c = a*(b*c) для всех a, b и c.

Здесь (см., например, [*Гусакова 87*] и др.) для моделирования сходства используется техника так называемых *карт*, т.е. упорядоченных троек

$$\boldsymbol{Cd} = \langle \mathbf{X}, \mathbf{Y}, \mathfrak{R} \rangle,$$

где

X - называется множеством *объектов,*
Y - множеством *свойств*, а
$\mathfrak{R}$ - есть отношение на **X**x**Y**, характеризующее наличие у каждого из объектов в множестве **X** тех или иных свойств из множества **Y**.

С отношением $\mathfrak{R}$ всегда связана пара отображений $\boldsymbol{s}_{\mathfrak{R}}, \boldsymbol{g}_{\mathfrak{R}}$, являющаяся *соответствием Галуа* ([*Гусакова 87, Кон 68*] и др.), ассоциированным с этим отношением:

$$\forall \mathbf{A} \in \boldsymbol{P}(\mathbf{X})\ \boldsymbol{s}_{\mathfrak{R}}(\mathbf{A}) = \{y \in \mathbf{Y} \text{ таких, что для } \forall x \in \mathbf{A} \text{ имеет место } x \mathfrak{R} y\}$$

$$\forall \mathbf{B} \in \boldsymbol{P}(\mathbf{Y})\ \boldsymbol{g}_{\mathfrak{R}}(\mathbf{B}) = \{x \in \mathbf{X} \text{ таких, что для } \forall y \in \mathbf{B} \text{ имеет место } x \mathfrak{R} y\}$$

где $\boldsymbol{P}(\mathbf{A})$ - множество всех подмножеств множества **A** ($\mathbf{A} \in \{\mathbf{X}, \mathbf{Y}\}$.

Простейшие свойства отображений $\boldsymbol{s}_{\mathfrak{R}}$ и $\boldsymbol{g}_{\mathfrak{R}}$ определяются соотношениями:

$$\mathbf{A}_1 \subseteq \mathbf{A}_2 \rightarrow \boldsymbol{s}_{\mathfrak{R}}(\mathbf{A}_1) \supseteq \boldsymbol{s}_{\mathfrak{R}}(\mathbf{A}_2) \qquad \text{(i1)}$$

$$\mathbf{B}_1 \subseteq \mathbf{B}_2 \rightarrow \boldsymbol{g}_{\mathfrak{R}}(\mathbf{B}_1) \supseteq \boldsymbol{g}_{\mathfrak{R}}(\mathbf{B}_2) \qquad \text{(i2)}$$

$$\boldsymbol{s}_{\mathfrak{R}}(\mathbf{A}_1 \cup \mathbf{A}_2 \cup \ldots \cup \mathbf{A}_n) = \boldsymbol{s}_{\mathfrak{R}}(\mathbf{A}_1) \cap \boldsymbol{s}_{\mathfrak{R}}(\mathbf{A}_2) \cap \ldots \cap \boldsymbol{s}_{\mathfrak{R}}(\mathbf{A}_n) \qquad \text{(i3)}$$

$$\boldsymbol{g}_{\mathfrak{R}}(\mathbf{B}_1 \cup \mathbf{B}_2 \cup \ldots \cup \mathbf{B}_n) = \boldsymbol{g}_{\mathfrak{R}}(\mathbf{B}_1) \cap \boldsymbol{g}_{\mathfrak{R}}(\mathbf{B}_2) \cap \ldots \cap \boldsymbol{g}_{\mathfrak{R}}(\mathbf{B}_n) \qquad \text{(i4)}$$

$$\boldsymbol{s}_{\mathfrak{R}}(\boldsymbol{g}_{\mathfrak{R}}(\boldsymbol{s}_{\mathfrak{R}}(\mathbf{A}))) = \boldsymbol{s}_{\mathfrak{R}}(\mathbf{A}) \qquad \text{(i5)}$$

$$\boldsymbol{g}_{\mathfrak{R}}(\boldsymbol{s}_{\mathfrak{R}}(\boldsymbol{g}_{\mathfrak{R}}(\mathbf{B}))) = \boldsymbol{g}_{\mathfrak{R}}(\mathbf{B}) \qquad \text{(i6)}$$

С учетом понятия *карты*, два типа сходства задают

3.1.1.1. Определение (См.[*Гусакова 87*])

---

[6] Ассоциативность (вместе с коммутативностью) обеспечивает независимость конечного результата - сходства - от порядка (т.е. конкретной последовательности по-парных сравнений) его вычисления.

Объекты $O_1, O_2,..., O_n$ локально сходны, если существует хотя бы одна карта $\boldsymbol{Cd} = \langle \mathbf{X},\mathbf{Y},\mathfrak{R}\rangle$, такая, что

(1) $\{O_1, O_2,..., O_n\} \subseteq \mathbf{X}$ и

(2) $s_{\mathfrak{R}}(O_1, O_2,..., O_n) =^7 s_{\mathfrak{R}}(O_1) \cap s_{\mathfrak{R}}(O_2) \cap ... \cap s_{\mathfrak{R}}(O_n) \neq \varnothing$.

o

и

3.1.1.2. Определение (См.[*Гусакова 87*])

Объекты $O_1, O_2,..., O_n$ глобально сходны, если существует хотя бы одна карта $\boldsymbol{Cd} = \langle \mathbf{X},\mathbf{Y},\mathfrak{R}\rangle$, такая, что

(1) $\{O_1, O_2,..., O_n\} \subseteq \mathbf{X}$, и

(2) найдется $\mathbf{B} \subseteq \mathbf{Y}$, для которого $g_{\mathfrak{R}}(\mathbf{B}) = \{O_1, O_2,..., O_n\}$.

o

Таким образом, локально сходны (в смысле соответствующей *карты*) любые два (или более) объекта, имеющие непустое подмножество общих свойств. Глобальное сходство характеризует (ср. с обсуждением свойства "*исчерпываемости*" в **Замечании 3.2.2.** из **Раздела 3.2**) все те и только те объекты, локальное сходство которых представляется некоторым непустым множеством свойств (и вточности им).

Подробный анализ алгебраических свойств пространств толерантности - т.е. формальных объектов, задаваемых упорядоченными парами вида

$$\langle \boldsymbol{M}, \tau \rangle,$$

где

$\boldsymbol{M}$ - множество объектов, а

τ - бинарное отношение на декартовом квадрате $\boldsymbol{M}$x$\boldsymbol{M}$, удовлетворяющее аксиомам рефлексивности и симметричности, т.е. такое, что

(1) $\forall \mathbf{X} \in \boldsymbol{M}$ имеет место $\mathbf{X}\ \tau\ \mathbf{X}$;

(2) $\forall \mathbf{X},\mathbf{Y} \in \boldsymbol{M}$ имеет место $(\mathbf{X}\ \tau\ \mathbf{Y}) \rightarrow (\mathbf{Y}\ \tau\ \mathbf{X})$,

можно найти в работах [*Гусакова 87*] и др.. Среди представленных здесь результатов особо выделим утверждение о невыразимости (в общем случае) глобального сходства средствами локального:

3.1.1.3. Утверждение (См.[*Гусакова 87*] и др.)

Существуют глобальные сходства, не представимые **n**-местным локальным сходством ни при каких $\mathbf{n} \geq 2$.

o

3.1.1.4. Замечание

Сформулированное утверждение характеризует эффект, в определенной мере аналогичный известному эффекту невыразимости транзитивного замыкания бинарного отношения (а также ряда других понятий, конструируемых с использованием понятия *неподвижной точки* некоторого соответствующего отображения) средствами логики предикатов первого порядка.

o

---

[7] См. выше свойства (i3,i4) соответствий Галуа.

Теперь, возвращаясь к формализации эвристики **Бэкона-Милля** и уточнению предложенного выше (см. **Разделы 2.2** и **2.3**) описания ДСМ-метода, мы (в рамках наивной семантики последнего) будем говорить, что

3.1.1.5. Определение

Пусть задано некоторое множество **Ω** объектов ***O***, характеризуемых свойствами $\boldsymbol{P} = \{P_1, P_2, \dots, P_m\}$ и каждому объекту O_i из **Ω** сопоставлено свое (непустое) подмножество $\boldsymbol{P}_i$ свойств из ***P***. Объекты

O_1, обладающий свойствами $\boldsymbol{P}_1$,
O_2, обладающий свойствами $\boldsymbol{P}_2$,
.....
O_n, обладающий свойствами $\boldsymbol{P}_n$,

локально сходны, если

(1) все $\boldsymbol{P}_1, \boldsymbol{P}_2, \dots, \boldsymbol{P}_n$ имеют непустую общую часть ***P'***, а
(2) все $O_1, O_2, \dots, O_n$ имеют непустой общий подобъект ***O'***.

o

3.1.1.6. Определение

Пусть задано некоторое множество **Ω** объектов ***O***, характеризуемых свойствами ***P***. Объекты

O_1, обладающий свойствами $\boldsymbol{P}_1$,
O_2, обладающий свойствами $\boldsymbol{P}_2$,
.....
O_n, обладающий свойствами $\boldsymbol{P}_n$,

глобально сходны, если (в дополнение к условиям 1-2 из **Определения 3.1.1.5**) верно, что

(3) в множество $\{\langle O_1,\boldsymbol{P}_1\rangle, \langle O_2,\boldsymbol{P}_2\rangle, \dots, \langle O_n,\boldsymbol{P}_n\rangle\}$ включены те и только те элементы из **Ω**, для которых множества свойств $\boldsymbol{P}_1, \boldsymbol{P}_2, \dots, \boldsymbol{P}_n$ имеют вточности общую часть ***P'*** (причем непустую -!), а объекты $O_1, O_2, \dots, O_n$ - вточности общий подобъект ***O'*** (непустой-!).

o

Теперь мы продемонстрируем на примерах реализацию представленной схемы формализации сходства в ДСМ-методе.

3.1.2. Примеры формализации сходства для различных типов данных в ДСМ-методе.

Предметом детального рассмотрения для нас будут две - в определенном смысле "базовые" структуры данных - множества и числовые векторы (конечной размерности). С их помощью мы также предложим формализации сходства для нумерованных[8] множеств, конечных кортежей нумерованных множеств и векторов (конечной размерности) числовых интервалов.

Пусть $\mathbf{U^{(1)}} = \{c_1, c_2, \dots, c_{n1}\}$ - некоторое множество образующих, а $\boldsymbol{\Omega}_C = \{O_1, O_2, \dots, O_n\} \subseteq 2^{\mathbf{U^{(1)}}} \setminus \varnothing$ - множество объектов (непустых множеств образующих), построенных над универсумом $\mathbf{U^{(1)}}$[9]. Пусть также $\mathbf{U^{(2)}} = \{a_1, a_2, \dots, a_{n2}\}$ - множество образующих для свойств

[8] Т.е. множеств, допускающих кратное вхождение элементов. Часто для их именования используется также термин *мультимножества.*

[9] Т.е. множество объектов, построенных из образующих $c_1, c_2, \dots, c_n$.

объектов, а $\mathbf{\Omega_A} = \{\boldsymbol{P_1}, \boldsymbol{P_2},..., \boldsymbol{P_n}\} \subseteq 2^{\mathbf{U(2)}}\backslash\varnothing$ - множество свойств (множеств «атомарных» свойств из $\boldsymbol{P} = \{P_1,P_2 , \dots , P_m\}$ объектов O_i из $\mathbf{\Omega_C}$.

Следуя предложенной нами в **Разделе 3.1.1** схеме, операцию * выделения сходства на множествах объектов и их свойств определим с помощью обычной булевской операции пересечения множеств - ∩.

Таким образом, следуя **Определению 3.1.1.5**, дадим

3.1.2.1. Определение

Пусть задано некоторое множество $\mathbf{\Omega}$ объектов $\boldsymbol{O}$, характеризуемых свойствами $\boldsymbol{P}$. Объекты

O_1, обладающий свойствами $\boldsymbol{P_1}$,
O_2, обладающий свойствами $\boldsymbol{P_2}$,
.....
O_n, обладающий свойствами $\boldsymbol{P_n}$,

локально сходны, если

(1) $O_1 \cap O_2 \cap ... \cap O_n = O' \neq \varnothing$ и
(2) $\boldsymbol{P_1} \cap \boldsymbol{P_2} \cap ... \cap \boldsymbol{P_n} = \boldsymbol{P'} \neq \varnothing$.

o

3.1.2.2. Определение

Пусть задано некоторое множество $\mathbf{\Omega}$ объектов $\boldsymbol{O}$, характеризуемых свойствами $\boldsymbol{P}$. Объекты

O_1, обладающий свойствами $\boldsymbol{P_1}$,
O_2, обладающий свойствами $\boldsymbol{P_2}$,
.....
O_n, обладающий свойствами $\boldsymbol{P_n}$,

глобально сходны, если (в дополнение к условиям 1-2 из **Определения 3.1.2.1)** верно, что

(3) в множество $\{<O_1,\boldsymbol{P_1}>,<O_2,\boldsymbol{P_2}>,...,<O_n,\boldsymbol{P_n}>\}$ включены те и только те элементы из $\mathbf{\Omega}$, для которых свойства $\boldsymbol{P_1}, \boldsymbol{P_2},..., \boldsymbol{P_n}$ имеют вточности общую часть $\boldsymbol{P'}$ (непустую-!), а объекты $O_1,O_2,...,O_n$ - вточности общий подобъект O' (непустое подмножество - !).

o

Стандартные свойства обычной операции пересечения множеств ∩ определяют, что справедлива

3.1.2.3. Лемма

Пусть $\boldsymbol{B}$ - непустое конечное множество. Алгебра $\wp = <2^{\boldsymbol{B}}, \cap>$ есть нижняя полурешетка.

o

Как это уже делалось ранее (см. **Раздел 2.1** из **Главы 2**), сформируем по $\mathbf{\Omega}$ и ⊗ (представляемой в данном случае как операция ∩ пересечения множеств) множество $Dom(\mathbf{\Omega})$:

(a) каждый из принадлежащих множеству $\mathbf{\Omega}$ объектов по определению принадлежит множеству $Dom(\mathbf{\Omega})$, т.е.

$$\mathbf{\Omega} = \{O_1, O_2, \dots , O_m\} \subset Dom\ (\mathbf{\Omega}).$$

(b) если A и B принадлежат множеству $Dom(\mathbf{\Omega})$, а результат A∩B применения к ним операции сходства ∩ не является пустым множеством ∅ , то этот результат A∩B также помещается в множество $Dom(\mathbf{\Omega})$, т.е.

$$\{[A \in Dom(\mathbf{\Omega})] \ \&\ [B \in Dom\ (\mathbf{\Omega})] \ \&\ [(A \cap B) \neq \varnothing\ ]\} \rightarrow [\ (A \cap B) \in Dom\ (\mathbf{\Omega})\ ],$$

(c) других элементов в множестве $Dom\ (\mathbf{\Omega})$ нет.

Далее, определив на множестве **Ω** пар вида <O_i,P_i> бинарную операцию $*$ выделения сходства как *по-компонентное* (отдельно – на объектах O_i и на множествах свойств P_i) *пересечение*, и сформировав множество *Dom*(**Ω**) как расширение множества **Ω** всеми *по-парными*, *по-троечными* и *т.д.* сходствами элементов из **Ω,** мы получим

3.1.2.4. Следствие

Алгебра $\wp$ = < *Dom* (**Ω**), $*$> есть нижняя полурешетка.

o

При обработке мультимножеств (нумерованных множеств) объекты будут задаваться множествами пар вида <c_i, q_i>, где c_i -как и ранее выбираются из универсума $\mathbf{U^{(1)}}$, а каждое q_i - неотрицательное целое число - количество (кратность) вхождений образующей c_i в соответствующий объект. При этом при вычислении сходства между объектами для одновременно входящих в них образующих результирующая кратность вхождения есть минимум из кратностей вхождения этой образующей в каждый из сравниваемых объектов. (А при не вхождении образующей в какой-либо из сравниваемых объектов - это просто 0).

Далее не составляет труда, определив по-компонентно операцию сходства как взятие пересечения множеств на теоретико-множественных компонентах и взятие минимума на числовом компоненте (при вычислении кратности вхождений), построить соответствующую алгебру $\wp_{\mathbf{num}}$. В силу идемпотентности, коммутативности и ассоциативности операции взятия минимума из пары чисел, и с учетом **Следствия 3.1.2.4**, имеет место

3.1.2.5. Следствие

Алгебра $\wp_{\mathbf{num}}$ есть нижняя полурешетка.

o

При обработке кортежей нумерованных множеств та же самая идея формализации может быть реализована на каждом из компонентов кортежа. Сходство кортежей как таковых будет считаться непустым в том и только в том случае, когда по каждому из компонентов соответствующее сходство мультимножеств не пусто.

В основе конструкции, обеспечивающей обработку ДСМ-методом данных числового характера, лежит конкретизирующая миллевский *Метод сопутствующих изменений* (см. выше **Раздел 2.2**) идея "*монотонности изменения значений существенных параметров*" на "*области неизменности механизма осуществления обуславливаемых ими свойств*". Этот содержательный принцип допускает уточнение, например, в виде предположения об *односвязности* (в соответствующем пространстве значений параметров) области постоянства механизма, "управляющего" осуществлением исследуемых свойств. (Удобным демонстрационным примером здесь могут послужить представления о постоянстве и "неизменности" - в пределах определенных областей - механизма осуществления биологических активностей в задаче анализа связи **СТРУКТУРА-АКТИВНОСТЬ** химических препаратов в ситуации, когда учитываются и числовые параметры - дозировки, концентрации и т.п.).

Идея ДСМ-ориентированной формализации сходства на числовых объектах (см., например, [*Забежайло 92, Zabezhailo 93*] и др.) состоит, грубо говоря, в следующем:

- в качестве (наделенных исследуемыми свойствами) объектов рассматриваются соответствующие векторы численных значений параметров;
- локально сходными объявляются каждые два объекта, для которых в рассматриваемом множестве параметров найдется непустое подмножество таких, что их значения от одного из выбранных объектов к другому (по всему множеству включенных в рассмотрение объектов) *изменяются монотонно*;
- глобально сходными будут считаться все объекты, попадающие в соответствующий "*участок монотонности* изменения значений параметров".

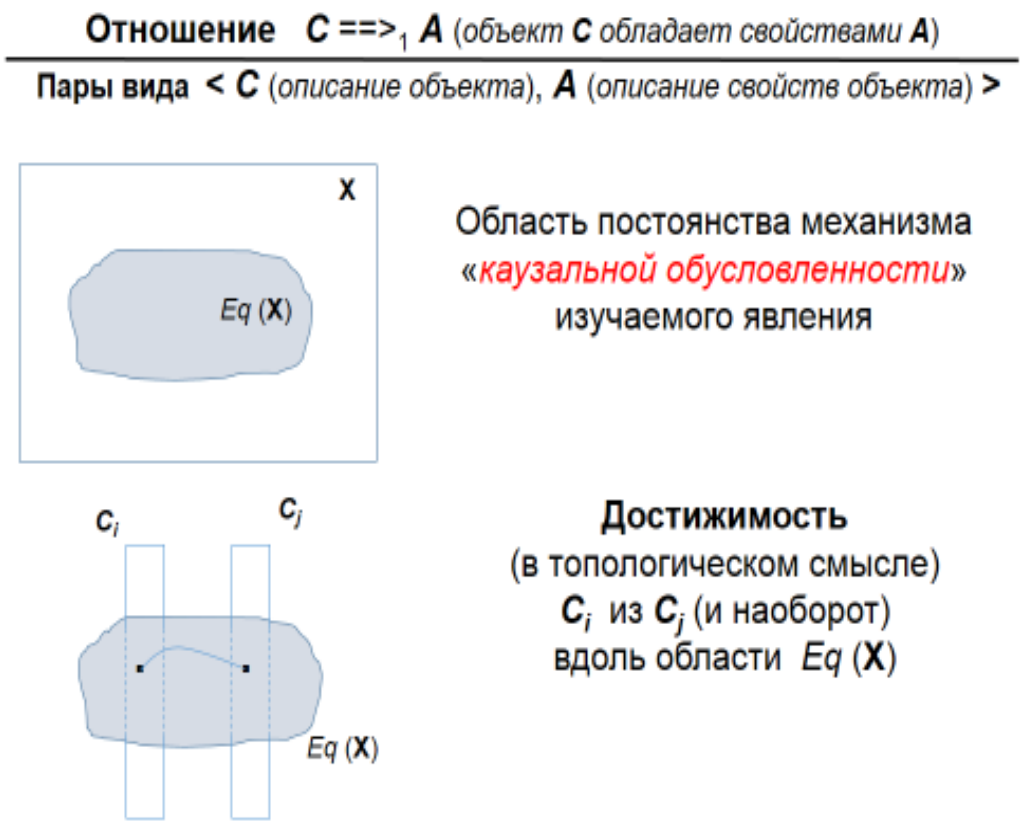


Рис. 3.1.2.1. Уточнение понятия *сопутствующие изменения*.

При этом представляется естественным предположить, что наличие изучаемых свойств у заданных для анализа объектов обусловлено существованием областей «каузальной обусловленности» (другими словами – постоянства «физического механизма» реализации) изучаемого явления (см. Рис. 3.1.2.1).

В подобной ситуации[10] область постоянство такого механизма должна характеризоваться

- набором релевантных параметров (допускающих объективную фиксацию их значений теми или иными измерительными приборами) а также
- интервалами релевантных изучаемому явлению числовых значений каждого из таких (релевантных) параметров.

В свою очередь, переносимость отражающих выделяемые «каузальные обусловленности» зависимостей на новые объекты представляется естественным характеризовать попаданием числовых значений соответствующих (т.е. релевантных области постоянства «физического механизма» реализации изучаемого явления) параметров в релевантные (см. выше) числовые интервалы таких значений. При этом попадание нового объекта в области (постоянства

[10] Можно, в качестве иллюстрации, ориентироваться здесь на следующую схему. Часто в экспериментальной лаборатории реальная ситуация при работе с плохо формализованными предметными областями имеет две основные характеристические особенности:

- в распоряжении исследователя есть набор "измерительных приборов", каждый из которых (будучи использованным) порождает соответствующий "образ" изучаемого объекта (явления);
- отсутствует априорная информация о "релевантных" и "нерелевантных" параметрах (фиксируемых соответствующими приборами) и об областях изменения (релевантных) параметров.

Вполне естественной здесь представляется необходимость предварительной обработки экспериментальных данных, позволяющей учесть систему "**влияний**" на изучаемый объект (явление) и оформить такую систему влияний в виде качественной модели явлений в изучаемой предметной области. При этом естественно рассматривать каждый из x_{ij}, представляющий собою соответствующее значение параметра R_j на объекте C_i, как "образ" изучаемого объекта, получаемый в результате эксперимента на j-том "измерительном приборе").

физического механизма реализации изучаемого явления) *разных знаков* – и в область наличия изучаемых свойств у анализируемых объектов и, одновременно, в область отсутствия таких свойств – следует (как и, например, в теоретико-множественном случае) считать случаями противоречивого[11] прогноза свойств этого нового объекта (см. Рис. 3.1.2.2).

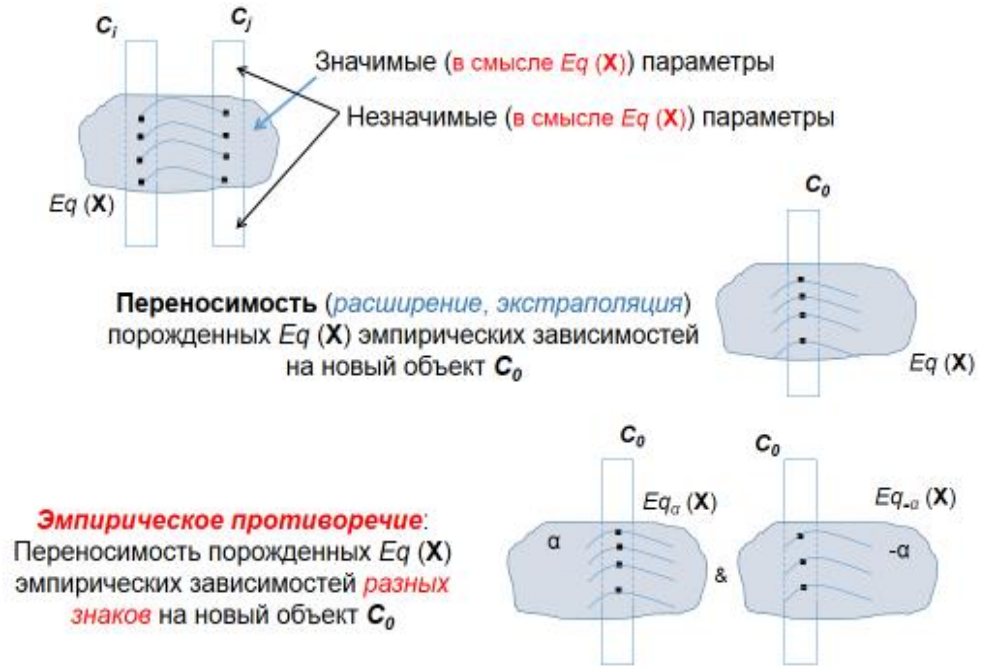


Рис. 3.1.2.2. *Сопутствующие изменения* и свойства новых объектов.

		C_i
	R_1	x_{1i}
	R_2	x_{2i}
	:	:
	:	:
Объекты		
	:	:
	:	:
	R_n	x_{ni}

(1) каждый объект O_i – вектор числовых значений
$< x_{1i}, x_{2i}, \ldots, x_{ni} >$,
каждое x_{ji} из которых есть результат измерения числовых характеристик объекта O_i прибором Dev_j,

(2) наличие (или же отсутствие) у объекта O_i свойств P отражает отношение $==>_1$:
$O_i ==>_1 P$

Рис. 3.1.2.3. Отношение **ОБЪЕКТ**$=>_1$**СВОЙСТВА** для числовых данных

Итак, пусть задано некоторое множество $\mathbf{R}=\{R_1,R_2,...,R_m\}$ параметров, принимающих числовые значения. Множество объектов **C**, характеризуемых значениями параметров из **R**, зададим следующим образом: будем говорить, что вектор числовых значений $<x_{i1},x_{i2},...,x_{im}>$, где каждый из x_{ij} есть значение параметра R_j на объекте C_i, представляет собой описание каждого объекта C_i. из **C**.

Далее (не теряя общности формируемого описания сходства на числовых векторах) ограничимся рассмотрением случая, когда каждый из рассматриваемых объектов O_i либо обладает, либо не обладает интересующим нас набором свойств P.

[11] Имеется ввиду ситуация эмпирического противоречия – см. подробнее **Разделы 2.3** и **3.3**..

Матрицей-универсумом знака *a* (полностью описывающей множество примеров знака *a*) будем называть матрицу $\mathbf{X}^a = \mathbf{C}^a$ x $\mathbf{R} = \{C^a_1, C^a_2, ..., C^a_{m^a}\}$ x $\{R_1, R_2, ..., R_n\}$, где $a \in \{+,-\}$. Здесь каждый из столбцов C_i описывает каждый отдельный эксперимент (т.е. *пример*, когда объект O_i *для* (*a* = +) *обладает и для* (*a* = -) *не обладает свойствами* ***P***), а каждая из строк R_j на текущем столбце C_i представляет числовой результат (т.е. значение соответствующего замеряемого прибором $\mathbf{Dev}_j$ параметра) именно в этом эксперименте с номером i. Итак, $\mathbf{X}^+$ как и ранее (в теоретико-множественном случае) – это матрица *примеров*, а $\mathbf{X}^-$ – это матрица *контрпримеров* (описаний экспериментов, в которых изучаемые свойства A не зафиксированы).

Будем (по определению) считать, что рассматриваемый нами «мир» явлений и эмпирических зависимостей характеризуется следующим

Постулатом Значения: наличие у рассматриваемых объектов изучаемых нами свойств обусловлено (связано с) наличием в описаниях этих объектов соответствующих «зон постоянства» механизма причинной обусловленности («каузальности[12]») для этих свойств.

Реконструкцией подобных «зон постоянства» механизма «каузальности» изучаемых свойств мы и займемся далее, используя ранее уже рассмотренную схему действий:

- на множестве примеров (и контрпримеров) определяется отношение сходства. Далее сходство формализуется как алгебраическая операция;
- с помощью заданного отношения сходства на множестве примеров (по стандартной схеме) формируются классы сходства, а затем
- по имеющимся классам сходства восстанавливаются формирующие их классы эквивалентности.

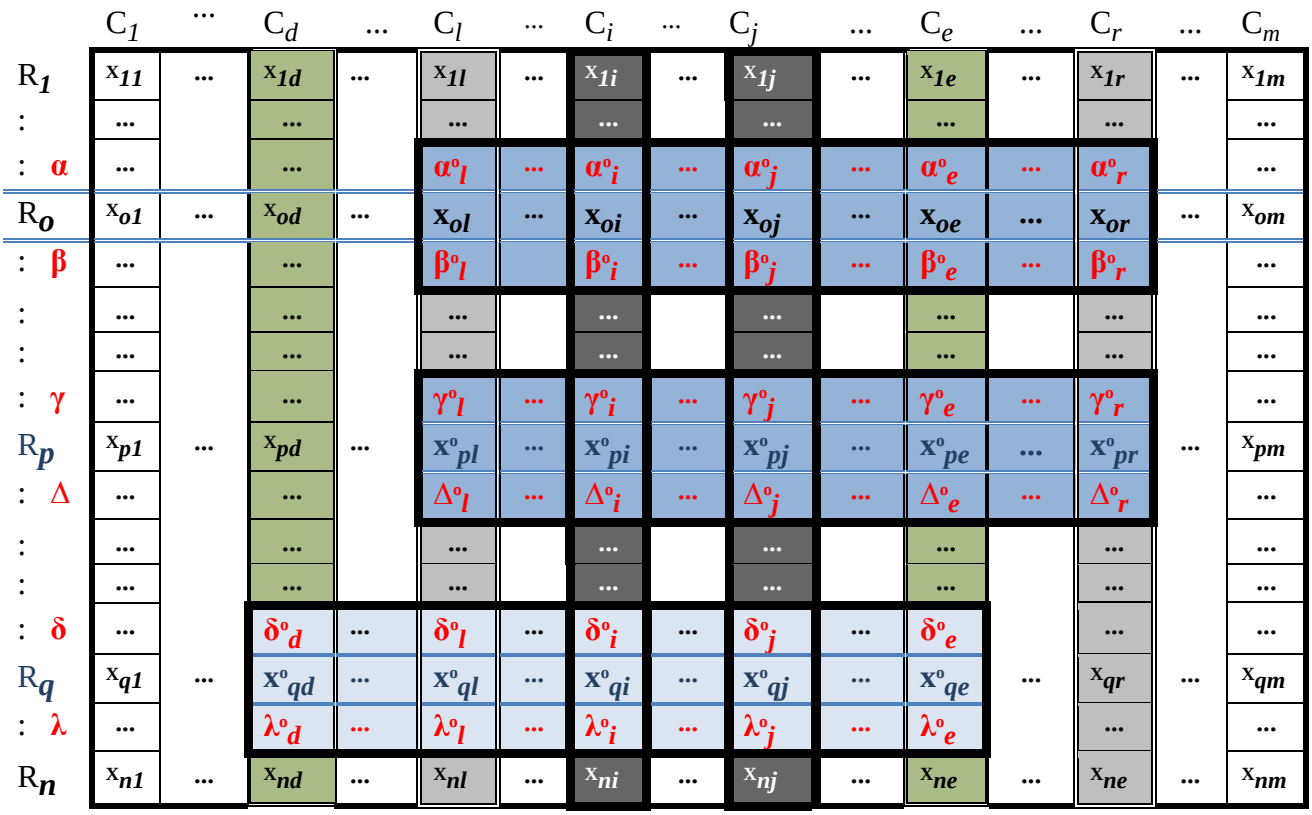


Рис. 3.1.2.4. Матрица-универсум **X**, монотонные подматрицы и м-башни в **X**.

[12] Т.е. как и в ранее рассмотренных случаях речь идет о мире направленных позитивных и негативных детерминистских влияний.

На процедурном уровне при работе объектами, которые представлены числовыми векторами (значений соответствующих параметров в каждом отдельном эксперименте), эта схема выглядит следующим образом:

1. Формируем матрицу-универсум **X** (см., например, Рис. 3.1.2.4).
2. Выбираем в универсуме **X** текущую ось монотонности - строку матрицы-универсума. Упорядочение элементов[13] этой строки $R_{\boldsymbol{o}}$ порождает соответствующее упорядочение столбцов матрицы-универсума. (Далее нижеследующие процедуры мы повторим для каждой из строк матицы **X**, выбираемой в качестве оси монотонности).
3. Для каждой отличной от $R_{\boldsymbol{o}}$ строки $R_{\boldsymbol{\mu}}$ выделяем участки (т.е. соответствующие множества столбцов матрицы-универсума **X**) ко-монотонности с выбранной осью монотонности $R_{\boldsymbol{o}}$.
4. Для каждого такого участка ко-монотонности $R_{\boldsymbol{o}}$ и $R_{\boldsymbol{\mu}}$ выделяем соответствующее (т.е. содержащее те и только те столбцы из **X**, которые формируют рассматриваемый участок ко-монотонности $R_{\boldsymbol{o}}$ и $R_{\boldsymbol{\mu}}$) множество столбцов $\mathbf{C}(C_l, C_r) = \{C_l, C_{l+1}, \ldots, C_r\}$. (Будем говорить о нем как о множество столбцов соответствующей подматрицы – так называемой м-башни $\mathbf{Tow}_X(C_l, C_r)$ - в универсуме **X**). Подматрицу универсума **X**, имеющую вид
$$\mathbf{C}(C_l, C_r) \times \{R_{\boldsymbol{o}}, R_{\boldsymbol{\mu}}\}$$
будем называть *основанием* м-башни $\mathbf{Tow}_X(C_l, C_r)$.
5. Финальная структура м-башни $\mathbf{Tow}_X(C_l, C_r)$ формируется путем последовательного пополнения множества строк $\{R_{\boldsymbol{o}}, R_{\boldsymbol{\mu}}\}$ всеми теми строками $R_{\boldsymbol{p}}$, вдоль которых соответствующие числовые значения $\mathbf{x}_{\boldsymbol{pl}}, \mathbf{x}_{\boldsymbol{pl+1}}, \ldots, \mathbf{x}_{\boldsymbol{pr}}$ на уже выделенном множестве столбцов $\mathbf{C}(C_l, C_r)$ также изменяются ко-монотонным образом (все одновременно не возрастают, или же все одновременно не убывают). Таким образом формируется множество $\mathbf{R}^{\boldsymbol{\mu}}_{\boldsymbol{o}}(C_l, C_r)$ тех и только тех строк матрицы-универсума **X**, которые на столбцах из множества $\mathbf{C}(C_l, C_r)$ изменяются ко-монотонно выбранной оси монотонности $R_{\boldsymbol{o}}$. Итак:
$$\mathbf{Tow}_X(C_l, C_r) = \mathbf{C}(C_l, C_r) \times \mathbf{R}^{\boldsymbol{\mu}}_{\boldsymbol{o}}(C_l, C_r) .$$
Замечание: представляется естественным рассматривать в качестве м-башен лишь те соответствующим образом сформированные подматрицы универсума **X**, которые образованы не менее чем тремя столбцами из исходного множества $\{C_1, C_2, \ldots, C_m\}$.
6. М-матрицей в универсуме **X** будем называть любую (в том числе – двух-столбцовую) подматрицу какой-либо из м-башен в **X**, сформированную на том же (что и используемая м-башня) множестве срок универсума **X**.
7. Будем считать, что два заданных столбца C_i и C_j универсума **X** (представляющие два объекта из характеризующего анализируемое отношение **ОБЪЕКТ** $=>_1$**СВОЙСТВА** множества примеров) сходны (*в контексте X*), если и только если в универсуме **X** существует некоторая м-матрица $\mathbf{V}(C_i, C_j)$, крайними (соответственно слева и справа) столбцами которой и являются эти два столбца C_i и C_j.

У используемой нами для формализации сходства на векторах числовых значений параметров процедурной конструкции имеется проблемное место, которое требует отдельного комментария. Речь идет о ситуации, когда в некоторых строках исходной матрицы-универсума **X** появляется большое количество повторяющихся числовых значений. Обсудим этот эффект более подробно на примере повторно встречающегося в нескольких строках матрицы **X** значительного количества экземпляров 0 и 1. (Назовем их 0-1-участками).

---

[13] Например, по не убыванию (или же, наоборот, не возрастанию).

I. Возможная (*комбинаторная*) «*патология*» при работе в такой ситуации – экспоненциально быстро растущий объем множества возможных упорядочений столбцов в содержащих существенные 0-1-участки строках при выборе их в качестве оси монотонности (ведь число требующих учета с такой ситуации перестановок определяется факториалами от длин соответствующих участков 0 и 1).

II. С учетом *семантики* используемой в данном случае процедурной схемы формализации *сходства* (т.е. с учетом исходно принятого предположения, что механизм «причинной обусловленности» анализируемых свойств объектов характеризуется некоторой областью значений – интервалами изменения – релевантных параметров) представляется уместным заключить, что остающиеся неизменными (на интервале изменения других релевантных параметров) значения соответствующих параметров указывают на *нерелевантность* этих параметров (разумеется - на соответствующих интервалах остающихся неизменными значений) изучаемым свойствам объектов-прецедентов.

III. Указанная нерелевантность может, в частности, сигнализировать о неверно выбранной калибровке (*точности*) *измерения* значений соответствующих параметров. Таким образом, представляется естественным провести необходимые измерения с более высокой точностью (добившись *дифференциации* соответствующих числовых значений, либо получив еще более весомые аргументы в пользу нерелевантности этих параметров в соответствующих интервалах значений), после чего вернуться к исходной процедурной конструкции – выделению областей ко-монотонности изменения числовых значений параметров в модифицированной[14] матрице-универсуме.

IV. Принимая во внимание используемые представления о механизме причинной обусловленности анализируемых свойств, область нерелевантности того или иного параметра можно рассматривать как *область* «*безразличия*» к возможным упорядочениям столбцов матрицы-универсума, которые задаются значениями других параметров при выборе каждого из них на роль актуальной оси монотонности. Таким образом, обсуждаемые 0-1 участки значений «безразличного» параметра можно переупорядочить каждый раз в соответствии с частичным порядком на *актуальной оси монотонности* (на числовых значениях действительно релевантного анализируемым свойствам параметра), не нарушая при этом ни содержательных представлений о семантике исследуемого отношения причинности, ни используемых процедур формализации сходства.

Теперь воспользуемся только что представленным содержательным описанием формируемой процедурной конструкции и перейдем к ее формальному описанию. Пусть

$$\mathbf{X} = \mathbf{C} \times \mathbf{R} = \{C_1, C_2, ..., C_m\} \times \{R_1, R_2, ..., R_n\},$$

есть исходно заданный универсум. Будем рассматривать ее подматрицы - матрицы V вида

$$V = \mathbf{C}(V) \times \mathbf{R}(V),$$

где множество столбцов $\mathbf{C}(V)$ матрицы V есть непустое подмножество множества $\mathbf{C}$ всех столбцов матрицы-универсума $\mathbf{X}$, а множество строк $\mathbf{R}(V)$ матрицы V есть содержащее не менее двух строк подмножество множества $\mathbf{R}$ всех строк матрицы-универсума $\mathbf{X}$.

3.1.2.6. Определение

Матрицу $V = \mathbf{C}(V) \times \mathbf{R}(V)$, где

$$\mathbf{C}(V) \subseteq \mathbf{C}, \text{ а } \mathbf{R}(V) \subseteq \mathbf{R},$$

и кроме того:

а) в $\mathbf{R}(V)$ найдется строка R_o, такая, что перенумеровав столбцы в множестве $\mathbf{C}(V)$ в соответствии с упорядочением элементов в строке R_o, мы получим, что в каждой из оставшихся строк множества $\mathbf{R}(V)$ значения x_{ij} изменяются монотонно вдоль множества столбцов $\mathbf{C}(V)$;

[14] Содержащей более детально измеренные числовые значения релевантных свойствам параметров.

б) если x_{ol} представляет собой минимальное, а x_{or} - максимальное значение параметра R_o на множестве столбцов **C**(V), то неверно, что в матрице **X** найдется столбец $C_o \notin$ **C**(V), такой, что

$$x_{ol} \le x_{oo} \le x_{or};$$

в) **R**(V) есть максимальное по вложению множество строк, удовлетворяющее сформулированным выше условиям а) и б) при фиксированном множестве столбцов **C**(V),

будем называть *монотонной матрицей* (*м-матрицей*) в **X**, а R_o - осью монотонности м-матрицы V.

o

Рассмотрим алгоритм *CONCAT* склейки м-матриц, перерабатывающий любую пару м-матриц V_1,V_2 из **X** в множество (возможно - пустое) "проходящих" (т.е. вкладывающихся на соответствующих множествах столбцов универсума **X** и в V_1 и в V_2 одновременно) м-матриц (порожденных в контексте **X**!).

3.1.2.7. Определение

Алгоритм *CONCAT* задается следующим описанием:

Шаг 1. В множестве осей монотонности каждой из матриц V_1 и V_2 отметим общую ось R_j. (Если это невозможно, или все такие оси уже отмечены, переходим к Шагу 5).

Шаг 2. Построим множество **C**(V_1,V_2) всех столбцов матрицы **X** , таких, что значения x_{ij} в строке R_j на столбцах из **C**(V_1,V_2) лежат между значениями x_{ij} в строке R_j на столбцах матриц V_1 и V_2.

Шаг 3. Построим множество $\mathbf{R}^j(V_1,V_2)$ - максимальное по вложению множество строк матрицы **X**, таких, что

$$\mathbf{R}^j(V_1,V_2) \subseteq \mathbf{R}(V_1) \cap \mathbf{R}(V_2);$$

кроме того, значения x_{ij} в строках из $\mathbf{R}^j(V_1,V_2)$ изменяются ко-монотонно (см. определение м-матрицы) в смысле оси монотонности R_j вдоль столбцов из множества

$$\mathbf{C}^j(V_1,V_2) = \mathbf{C}(V_1,V_2)\text{, где } \mathbf{C}(V_1) \cup \mathbf{C}(V_2) \subseteq \mathbf{C}(V_1,V_2).$$

Если в множестве $\mathbf{R}^j(V_1,V_2)$ окажется всего одна строка (это будет R_j (V_1,V_2)) переходим к Шагу 1.

Шаг 4. Матрицу $\mathbf{V}^j_{1,2} = \mathbf{R}^j(V_1,V_2)$ x $\mathbf{C}^j(V_1,V_2)$ поместим в результирующее множество $V_{1,2}$ и перейдем к Шагу 1.

Шаг 5. Удалим из $V_{1,2}$ повторяющиеся элементы.

Шаг 6. СТОП.

o

3.1.2.8. Пример

Пусть матрицы V_1 и V_2 - те, что представлены ниже:

V_1:

	C_1	C_2	C_3
R_1	1	2	3
R_2	1	2	4
R_3	11	12	13

,

V_2:

	C_4	C_5	C_6
	4	5	6
	3	5	6
	14	15	16

Полагая, что матрица-универсум есть

$$\mathbf{X} = \{R_1,R_2,R_3\} \text{ x } \{C_1,C_2,...,C_6\},$$

имеем: *CONCAT*(V_1,V_2) - одноэлементное множество м-матриц $\{V_{1,2}\}$, где $V_{1,2}$ - матрица

	C_1	C_2	C_3	C_4	C_5	C_6
R_1	1	2	3	4	5	6
R_3	11	12	13	14	15	16

o

3.1.2.9. Определение

Будем говорить, что примеры C_{i1} и C_{i2} локально сходны относительно матрицы **X**, если среди порождаемых в универсуме **X** м-матриц найдется V = **R**(V) x **C**(V) такая, что

$$\mathbf{C}(V) = \{C_{i1}, C_{i1}+1., \ldots, C_{i2}\},$$

а порядок перечисления столбцов в множестве **C**(V) определяется упорядочением числовых значений от столбца C_{i1} к столбцу C_{i2} на R_o - оси монотонности матрицы V.

o

Множество всех м-матриц в универсуме **X** обозначим через $\mu(\mathbf{X})$. На множестве множеств м-матриц в **X** с помощью оператора *CONCAT* может быть определена бинарная операция Π, применяя которую к описаниям примеров (столбцам матриц $\mathbf{X}^+$ и $\mathbf{X}^-$) как к одноэлементным множествам (одностолбцовых) м-матриц, можно найти:

а) максимальные (т.е. нерасширяемые) по числу столбцов м-матрицы - *м-башни* в универсуме **X**, а вместе с каждой из них:

б) явный вид множества параметров (имена строк соответствующих м-матриц), определяющих наличие (в случае анализа положительных примеров) или отсутствие (в случае анализа отрицательных примеров) изучаемых свойств;

в) интервалы допустимых изменений каждого из параметров, оказывающих влияние на наличие (или отсутствие) необходимых свойств (диапазоны изменения числовых значений в каждой из строк соответствующей м-башни).

Таким образом, мы получаем возможность выделять как "сопутствующим образом влияющие" на поведение изучаемых объектов параметры, так и области их адекватных (в смысле наличия или отсутствия исследуемых свойств) изменений.

Итак, представленный алгоритм ***CONCAT*** перерабатывает подаваемую ему на вход пару <V_1,V_2> элементов $\mu(\mathbf{X})$ в множество V_{12}.

3.1.2.10. Лемма

Каждый элемент множества V_{12} - монотонная матрица из $\mu(\mathbf{X})$.

Доказательство.

Для каждого элемента в множестве V_{12} его односвязность обеспечена действиями, выполняемыми на Шаге 2, а ко-монотонность значений в строках - действиями, выполняемыми на Шаге 3 в определении алгоритма ***CONCAT***.

o

Алгоритм CONCAT фактически определяет бинарный оператор $\cap_\mu$ склейки м-матриц, действующий при заданном "универсуме" **X** из $\mu(\mathbf{X})$x$\mu(\mathbf{X})$ в множество всех подмножеств $2^{\mu(X)}$. Ради наглядности изложения будем обозначать результат применения оператора $\cap_\mu$ к матрицам V_1 и V_2 посредством $V_1 \cap_\mu V_2$.

3.1.2.11. Лемма

Оператор $\cap_\mu$ удовлетворяет свойствам идемпотентности и коммутативности, т.е. для любых V_1 и V_2 имеют место:

(1) $V_1 \cap_\mu V_1 = V_1$

(2) $V_1 \cap_\mu V_2 = V_2 \cap_\mu V_1$

Доказательство.
Идемпотентность - очевидное свойство определения алгоритма ***CONCAT***. Коммутативность - очевидное следствие коммутативности теоретико-множественных операций $\cap$ и $\cup$, используемых при порождении всех $\mathbf{R}^j(V_1,V_2)$ и $\mathbf{C}^j(V_1,V_2)$ в определении алгоритма ***CONCAT***.

o

3.1.2.12. Лемма

Пусть $V_1 \cap_\mu V_2 = V_{12}$, а $\mathbf{R}(V_1)$ и $\mathbf{R}(V_2)$ - множества строк соответствующих м-матриц V_1 и V_2. Тогда

$$|V_{12}| \le \min(|\mathbf{R}(V_1)|,|\mathbf{R}(V_2)|).$$

Доказательство.
Легко видеть, что число различных вариантов склейки определяется числом общих осей монотонности склеиваемых матриц.

o

Явный вид ситуации, когда на Шаге 3 алгоритма склейки реализуется строгое включение

$$\mathbf{R}^j(V_1,V_2) \subset \mathbf{R}(V_1)\cap\mathbf{R}(V_2),$$

(но при этом $\mathbf{R}^j(V_1,V_2) \ne \mathbf{R}(V_1)\cap\mathbf{R}(V_2)$),

представляет

3.1.2.13. Пример

Пусть матрицы V_1 и V_2 - те, что представлены ниже.

V_1:

	C_1	C_2	C_3
R_1	a_1	a_2	a_3
R_2	1	2	4
R_3	11	12	13

,

V_2:

C_4	C_5	C_6
a_4	a_5	a_6
3	5	6
14	15	16

При этом: $a_1 < a_2 < a_3 < a_4 < a_5 < a_6$ и R_3 - ось монотонности.

Легко видеть, что ситуация с R_2:

1	2	*	4		
		3	*	5	6

определяет при $\mathbf{R}(V_1) = \mathbf{R}(V_2) = \{R_1,R_2,R_3\}$ лишь

$$\mathbf{R}(V_1,V_2) = \{R_1,R_3\}.$$

Таким образом, $V_1 \cap_\mu V_2 = V_{12}$ есть множество, содержащее единственную матрицу

	C_1	C_2	C_3	C_4	C_5	C_6
R_1	a_1	a_2	a_3	a_4	a_5	a_6
R_3	11	12	13	14	15	16

o

Определим на множествах монотонных матриц из $\mu(\mathbf{X})$ бинарную операцию $\boxplus$ *локального относительного суперсходства*, действующую из $2^{\mu(X)}$ x $2^{\mu(X)}$ в $2^{\mu(X)}$.

Пусть $\mathbf{A} = \{A_1,...,A_2\}$ и $\mathbf{B} = \{B_1,...,B_2\}$ - произвольные множества монотонных матриц из $2^{\mu(X)}$:

$$\mathbf{A} \in 2^{\mu(X)} \text{ и } \mathbf{B} \in 2^{\mu(X)}.$$

3.1.2.14. Определение

Определим $\mathbf{D} = \mathbf{A} \boxplus \mathbf{B}$ как "покомпонентное соединение" множеств матриц **A** и **B**:

$$\mathbf{D} = \mathbf{A} \pitchfork \mathbf{B} =$$
$$= \{A_1 \cap_\mu B_1, ..., A_1 \cap_\mu B_2, ..., A_2 \cap_\mu B_1, ..., A_2 \cap_\mu B_2\}$$

o

Считая, что множеству $2^{\mu(X)}$, по определению, принадлежит каждый столбец матрицы **X**, определим подмножество ***Dom***(**X**) множества $2^{\mu(X)}$ следующим образом:

1) каждый столбец матрицы **X** как одноэлементное множество монотонных матриц входит в ***Dom***(**X**);
2) если $\mathbf{A} \in \boldsymbol{Dom}(\mathbf{X})$ и $\mathbf{B} \in \boldsymbol{Dom}(\mathbf{X})$, то $\mathbf{A} \pitchfork \mathbf{B} = \mathbf{D} \in \boldsymbol{Dom}(\mathbf{X})$.
3) других элементов в ***Dom***(**X**) нет.

Свойства операции $\pitchfork$ определяет

3.1.2.15. Теорема

Алгебра $\wp = <\boldsymbol{Dom}(\mathbf{X}), \pitchfork>$ есть нижняя полурешетка.

Доказательство.

1. Коммутативность.

Имея $\mathbf{A} = \{A_1, ..., A_2\}$ и $\mathbf{B} = \{B_1, ..., B_2\}$, вычислим

$$\mathbf{A} \pitchfork \mathbf{B} =$$
$$= \{A_1, ..., A_2\} \pitchfork \{B_1, ..., B_2\} =$$
$$= \{A_1 \cap_\mu B_1, ..., A_1 \cap_\mu B_2, ..., A_2 \cap_\mu B_1, ..., A_2 \cap_\mu B_2\} = ^{15}$$
$$= \{B_1 \cap_\mu A_1, ..., B_1 \cap_\mu A_2, ..., B_2 \cap_\mu A_1, ..., B_2 \cap_\mu A_2\} =$$
$$= \mathbf{B} \pitchfork \mathbf{A}.$$

2. Ассоциативность.

Рассмотрим $\cap_\mu|_i$ - проекцию оператора $\cap_\mu$ на строку R_i исходной матрицы-"универсума" **X**, задаваемую следующим образом:

> $\cap_\mu|_i$ склеивает (порождая в виде результата не совпадающую с $\varnothing$ м-матрицу) только по оси монотонности R_i лишь такие м-матрицы V_1 и V_2, для которых R_i - общая ось монотонности.

Легко видеть, что $\cap_\mu|_i$ - ассоциативная операция: последовательность вычисления $(V_1 \cap_\mu|_i V_2) \cap_\mu|_i V_3$ дает (при склейке вдоль одной и той же оси R_i для всех трех матриц V_1,V_2 и V_3, благодаря ассоциативности теоретико-множественных операций $\cap$ и $\cup$, тот же результат, что и последовательность вычисления $V_1 \cap_\mu|_i (V_2 \cap_\mu|_i V_3)$. В свою очередь, ассоциативность $\pitchfork$ естественным образом следует из соотношения

$$(V_1 \cap_\mu V_2) \quad = \quad \bigcup_{\text{по всем } R_i \in \mathbf{R}(\mathbf{X})} (V_1 \cap_\mu|_i V_2)^{16}$$

и ассоциативности теоретико-множественной операции $\cup$.

3. Идемпотентность.

Докажем необходимое индукцией по сложности $\mathbf{A} \in \boldsymbol{Dom}(\mathbf{X})$ (т.е. учтем, что каждый элемент из ***Dom***(**X**) порожден посредством операции $\pitchfork$ из "столбцов" матрицы **X**).

1) Если $\mathbf{A} = \{A\}$ - одноэлементное множество матриц и A - матрица, порождаемая одним из столбцов матрицы **X**, то, очевидно, что из идемпотентности оператора $\cap_\mu$ (см. **Лемму 3.1.2.11**) имеем:

---

[15] В силу коммутативности оператора $\cap_\mu$.

[16] Т.е. объединением по всем возможным осям монотонности.

$$\{A\} \amalg \{A\} = \{A\} .$$

2) По определению ***Dom***(**X**), для каждого $\mathbf{B} \in$ ***Dom***(**X**) найдутся $\mathbf{B_1}$ и $\mathbf{B_2}$ из ***Dom***(**X**), такие, что $\mathbf{B_1} \amalg \mathbf{B_2}$. Будем считать, что

$$\mathbf{B_1} \amalg \mathbf{B_1} = \mathbf{B_1} \text{ и } \mathbf{B_2} \amalg \mathbf{B_2} = \mathbf{B_2} ,$$

тогда имеем:

$$\mathbf{B} \amalg \mathbf{B} = (\mathbf{B_1} \amalg \mathbf{B_2}) \amalg (\mathbf{B_1} \amalg \mathbf{B_2}) =$$
$$=^{17} \mathbf{B_1} \amalg \mathbf{B_2} \amalg \mathbf{B_1} \amalg \mathbf{B_2} =$$
$$=^{18} \mathbf{B_1} \amalg \mathbf{B_1} \amalg \mathbf{B_2} \amalg \mathbf{B_2} =$$
$$=^{19} (\mathbf{B_1} \amalg \mathbf{B_1}) \amalg (\mathbf{B_2} \amalg \mathbf{B_2}) =$$
$$= \mathbf{B_1} \amalg \mathbf{B_2} = \mathbf{B} .$$

o

3.1.2.16. Замечание

Свойства коммутативности и ассоциативности операция $\amalg$ сохраняет на всем множестве $2^{\mu(X)}$. Нераспространяемость свойства идемпотентности $\amalg$ с ***Dom***(**X**) на все множество $2^{\mu(X)}$ иллюстрирует

3.1.2.17. Пример

Пусть $\mathbf{A} = \{\mathbf{A_1},\mathbf{A_2}\}$, где $\mathbf{A_1}$ есть матрица

	C_1	C_2	C_3
R_1	1	2	3
R_2	11	12	13

и $\mathbf{A_2}$ есть матрица

	C_4	C_5	C_6
R_1	4	5	6
R_2	14	15	16

Учитывая $\mathbf{A_3}$:

	C_1	C_2	C_3	C_4	C_5	C_6
R_1	1	2	3	4	5	6
R_2	11	12	13	14	15	16

легко видеть, что

$$\mathbf{A} \amalg \mathbf{A} = \{\mathbf{A_1},\mathbf{A_2},\mathbf{A_3}\} \neq \mathbf{A}.$$

o

Теперь, располагая необходимым формальным инструментарием, мы имеем возможность для сходства на векторах действительных чисел (в том числе - и в случае, когда на совокупность свойств ***P*** анализируемых объектов *O* не накладывается никаких предварительных ограничений) дать

3.1.2.18. Определение

Пусть задано некоторое подмножество $\mathbf{\Omega}$ объектов *O* и множество **C** - представляющих эти объекты числовых векторов одинаковой размерности из некоторого множества числовых векторов $\mathbf{\Omega_X}$ (столбцов некоторой матрицы **X**), характеризуемых свойствами ***P***. Объекты

---

[17] По ассоциативности $\amalg$.

[18] По коммутативности $\amalg$.

[19] По исходному предположению.

O_1, обладающий свойствами P_1,
O_2, обладающий свойствами P_2,
.....
O_n, обладающий свойствами P_n,

локально сходны, если

(1) все P_1, P_2, ... , P_n имеют непустую общую часть P ', а

(2) среди порождаемых в **X** м-матриц найдется непустая м-матрица **V** = **R**(V) x **C**(V) такая, что множество ее столбцов **C**(V) = {C_1,C_2,...,C_n} перечисляет все объекты из множества { O_1, O_2, ... , O_n}.

o

3.1.2.19. Определение

Пусть задано некоторое подмножество **Ω** объектов *O* и множество **C** - представляющих эти объекты числовых векторов одинаковой размерности из некоторого множества числовых векторов Ω_X (столбцов некоторой матрицы **X**), характеризуемых свойствами ***P***. Объекты

O_1, обладающий свойствами P_1,
O_2, обладающий свойствами P_2,
.....
O_n, обладающий свойствами P_n,

глобально сходны, если (в дополнение к условиям 1-2 из **Определения 3.1.2.18**) верно, что

(3) м-матрица **V** = **R**(V) x **C**(V) такая, что **C**(V)={C_1,C_2,...,C_n} - м-башня в **X** (т.е. максимальная – нерасширяемая - по числу столбцов м-матрица в матрице **X**).

o

Будем обозначать задаваемое (в случае n=2) **Определением 3.1.2.18** бинарное отношение как LRA_2 (C_1,C_2) и называть его *локальным относительным сходством* (числовых векторов обсуждаемого вида).

3.1.2.20. Лемма

Отношение LRA_2 есть отношение толерантности, т.е. оно рефлексивно и симметрично.

Доказательство.

Рефлексивность следует из определения множества ***Dom***(**X**). Симметричность - из коммутативности операции $\pitchfork$ (как, впрочем, и из самого определения отношения LRA_2).

o

Пусть как и ранее Ω_X некоторое множества числовых векторов - объектов - **C** (столбцов некоторой матрицы **X**), характеризуемых свойствами **A**. Зафиксируем в матрице **X** строку R_0, и пару (различных) столбцов <C_1,C_2>. Рассмотрим параметрическое бинарное отношение

$$P_LRA_2\ (C_i,C_j,R_0,C_1,C_2) \subseteq \Omega_X \times \Omega_X,$$

порождаемое всеми теми парами столбцов из множества Ω_X, для которых при фиксированных C_1,C_2 и R_0 (в качестве оси монотонности) в множестве **Tow**(**X**) м-башен, порождаемых из матрицы **X**, найдется м-башня

$$\mathbf{V}_{R0,C1,C2} = \mathbf{R}(V) \times \mathbf{C}(V),$$

такая, что

a) множество ее столбцов

$$\mathbf{C}(V) = \mathbf{C}(V)=\{C_1,...,C_2\}$$

содержит C_1 и C_2 в качестве максимального и минимального (в смысле соответствующего упорядочения вдоль оси монотонности R_0) элементов,

b) столбцы C_i,C_j связаны отношением LRA_2 так, что (в смысле **Определения 3.1.2.18**) м-башня $\mathbf{V}_{R0,C1,C2}$ и есть та самая м-матрица

$$\mathbf{V} = \mathbf{R}(V) \times \mathbf{C}(V)$$

такая, что $C_i, C_j \in \mathbf{C}(V)=\{C_1,...,C_2\}$[20].

Имеет место

3.1.2.21. Лемма

Отношение $\boldsymbol{P_LRA_2}$ $(C_i, C_j, R_0, C_1, C_2)$ есть[21] отношение эквивалентности.

Доказательство.

Отношение $\boldsymbol{P_LRA_2}$ задано с помощью элементов множества **Tow**(**X**). Таким образом, $\boldsymbol{P_LRA_2}$ есть подмножество $\boldsymbol{LRA_2}$, что в силу **Леммы 3.1.2.20** дает рефлексивность и симметричность отношения $\boldsymbol{P_LRA_2}$.

Далее, если имеют место

$$\boldsymbol{P_LRA_2}\ (C_3, C_4, R_0, C_1, C_2) \text{ и}$$

$$\boldsymbol{P_LRA_2}\ (C_4, C_5, R_0, C_1, C_2)$$

то в **Tow**(**X**) есть элемент **m**(**X**) такой, что

$$C_3, C_4, C_5 \in \mathbf{C}(V) = \{C_1,...,C_2\}.$$

Но тогда из

$$C_3, C_5 \in \mathbf{C}(V)=\{C_1,...,C_2\}$$

следует, что имеет место

$$\boldsymbol{P_LRA_2}\ (C_3, C_5, R_0, C_1, C_2),$$

т.е. отношение $\boldsymbol{P_LRA_2}$ - транзитивно (при фиксированных R_0, C_1 и C_2).

o

Пусть $\boldsymbol{P_LRA^*_2}$ - транзитивное замыкание отношения $\boldsymbol{P_LRA_2}$. В ситуации, когда C_p, и C_q пробегают (при фиксированных значениях параметров R_0, C_1 и C_2) все множество столбцов **C**(**X**) матрицы **X**, имеет (с учетом **Леммы 3.1.2.21**) место

3.1.2.22. Следствие

Всякий раз, когда для произвольных **r** и **s** пара столбцов $<C_r, C_s>$ из множества **C**(**X**) принадлежит транзитивному замыканию $\boldsymbol{P_LRA^*_2}(C_r, C_s, R_0, C_p, C_q)$ отношения $\boldsymbol{P_LRA_2}(C_r, C_s, R_0, C_p, C_q)$, одновременно справедливо, что

$$<C_r, C_s> \in \boldsymbol{LRA_2}(C_r, C_s),$$

т.е. объекты C_r и C_s локально сходны в смысле **Определения 3.1.2.18**.

o

Таким образом каждой м-башне из **Tow**(**X**) может быть сопоставлен некоторый класс эквивалентности по отношению $\boldsymbol{P_LRA_2}$ на множестве столбцов **C**(**X**) матрицы-"универсума" **X**, порождаемый локальным сходством[22] "крайних" ("первого" и "последнего") столбцов рассматриваемой м-башни.

Более того, в отличие от случая обработки данных, представленных множествами (где, как показано, например, в работе [*Гусакова 87*], глобальное сходство не представимо локальным), при работе с числовыми векторами имеет место

3.1.2.23. Теорема

Пусть **V** - произвольная м-башня в матрице **X**, **R**(**X**) -множество строк матрицы **X**, а **C**(**V**) = $\{C_{i1}, ..., C_{i2}\}$ - множество образующих м-башню **V** столбцов матрицы **X**. В множестве столбцов **C**(**V**) найдутся C_{j1} и C_{j2}, такие, что

$$\mathbf{V} \in \mathbf{[(R(X)} \times \{C_{j1}\}) \pitchfork \mathbf{(R(X)} \times \{C_{j2}\})\mathbf{]}.$$

---

[20] Более точно - максимальный элемент на множестве соответствующих м-матриц, определяемых осью монотонности R_0 и "накрывающих" одновременно столбцы C_i, и C_j.

[21] При фиксированных **X**, R_0, C_1, и C_2.

[22] Более точно - одним из локальных сходств (каждое из которых порождается упорядочением, связанным с соответствующей осью монотонности).

Доказательство.

Для каждой пары столбцов B_{s1}, B_{s2} из **C**(**X**) и строки R_r из **R**(**X**) рассмотрим отношение ***P_LRA2***(C_{j1},C_{j2},B_{s1},B_{s2},R_r) как соответствующее множество $\mathbf{H}_{s1,s2,r}$ пар вида <C_{j1},C_{j2}>. Пусть **H**(**X**) - множество всех непустых $\mathbf{H}_{s1,s2,r}$. Если **V**∈**Tow**(**X**), то в **H**(**X**) найдется $\mathbf{H}_{s10,s20,r0}$, такое, что

$$\mathbf{V} = \mathbf{R}(\mathbf{V})x\{B_{s10},...,B_{s20}\},$$

при этом R_{ro} есть ось монотонности **V,** и B_{s10} с B_{s20} определяют min и max значения $X_{r0,j}$ в строке R_{r0} на столбцах из множества $\{B_{s10},...,B_{s20}\}$. Учитывая **Следствие 3.1.2.22**, имеем

$$<B_{s10},B_{s20}> \in \boldsymbol{P_LRA^*_2}(C_{j1},C_{j2},B_{s10},B_{s20},R_{r0}),$$

т.е. в смысле $\mathbf{V}\in\mathbf{Tow}(\mathbf{X})\subseteq\boldsymbol{\mu}(\mathbf{X})$ имеет место и $\boldsymbol{LRA_2}(B_{s10},B_{s20})$.

При этом **V** есть результат склейки м-матриц **R**(**X**)x$\{B_{s10}\}$ и **R**(**X**)x$\{B_{s20}\}$ лишь по оси R_{r0}, т.е.

$$\mathbf{V} \in \mathbf{[}(\mathbf{R}(\mathbf{X}) \text{ x } \{B_{s10}\}) \pitchfork (\mathbf{R}(\mathbf{X}) \text{ x } \{B_{s20}\})\mathbf{]}.$$

o

Таким образом, согласно **Теореме 3.1.2.23** имеется возможность по заданным столбцам $C_{j1},...,C_{j2}$ и самой матрице-"универсуму" **X** "процедурно" указать (путем анализа упорядочений на $\{C_{j1}, ... ,C_{j2}\}$ - в случае, когда они сходны в смысле представленного **Определением 3.1.2.19** глобального сходства - ту пару столбцов C_{j10} и C_{j20}, локальное сходство которых содержит в качестве элемента (из соответствующего множества м-матриц) порождаемую множеством столбцов $\{C_{j1}, ... ,C_{j2}\}$ м-башню. Т.е. если при работе с данными, которые представлены множествами, локального сходства, выражающего заданное глобальное (т.е. пары объектов, результат применения формализующей локальное сходство операции к которым совпадает с интересующим нас глобальным сходством), вообще говоря, может и не существовать (см., в частности, ниже **Пример 3.1.2.24**), то при обработке (представленными средствами -!) числовых векторов такая пара объектов может быть найдена путем достаточно простых вычислений.

3.1.2.24. Пример

Пусть на множестве образующих $\{a_1,a_2,a_3,a_4\}$ заданы три объекта - O_1 = $\{a_2,a_3,a_4\}$, O_2 = $\{a_1,a_3,a_4\}$ и O_3 = $\{a_1,a_2,a_4\}$, - каждый из которых обладает набором свойств *P*. Нетрудно убедиться, что глобальное сходство всех трех названных объектов - пара <a_4,*P*> - не представимо локальным сходством ни одной из трех пар предложенных объектов.

o

С другой стороны, располагая в качестве исходных данных лишь матрицей-"универсумом" **X**, мы не имеем возможности средствами языка логики предикатов первого порядка описать (как и в случае обработки данных типа “множества”) чисто синтаксическим путем глобальное сходство. **Теорема 3.1.2.23** позволяет по имеющемуся глобальному сходству эффективно найти соответствующее ему локальное. В то же время она мало что дает[23] - в попытке описать по заданному локальному сходству порождаемые им (в смысле, возможно, различных - непараллельных! - осей монотонности) глобальные сходства. Действительно, имея исходную матрицу-"универсум" X, мы достаточно легко порождаем множество всех пар локально сходных столбцов из **X**; кроме того, известно, что среди элементов множества локальных сходств находятся (как элементы соответствующих множеств м-матриц) и все глобальные сходства (точнее - соответствующие им м-башни). Мы также знаем, что каждое глобальное сходство (м-башня) есть максимальный (по вложению) элемент соответствующего множества м-матриц, однако до порождения в явном виде диаграмм вложения м-матриц для заданной матрицы-"универсума" **X**, мы не можем указать (в названном выше

[23] Ср. с эффектом невыразимости синтаксическими средствами языка логики предикатов первого порядка транзитивного замыкания отношения - см., например, [*Гусакова 87, Финн 88*]).

смысле), какие из всех возможных локальных сходств окажутся максимальными элементами рассматриваемых диаграмм вложения.

В заключение обратим внимание на одно обстоятельство, характеризующее корректность предложенной формализации локального и глобального сходств при работе с векторами действительных чисел. Говоря во вступительной части настоящего раздела о содержательном описании формализуемой ситуации, мы желали рассматривать как сходные те и только те исследуемые объекты, свойства которых обусловлены одним и тем же "физическим механизмом". Одной из основных наших целей было стремление выделять области постоянства такого механизма. Возможность связать в построенной нами формализации с каждым глобальным сходством отношение эквивалентности $\boldsymbol{P_LRA_2}$ позволяет говорить о каждом множестве глобально сходных объектов как о классе эквивалентности (в смысле соответствующего отношения $\boldsymbol{P_LRA_2}$). Именно этот класс эквивалентности и дает нам информацию (явный вид множества эквивалентных объектов и явный вид множества существенных для этой эквивалентности параметров, представленных строками в матрице **X**) об областях постоянства механизмов осуществления ("физических механизмов реализации") исследуемых нами свойств на предъявленном множестве входных данных - матрице **X** числовых характеристик изучаемых объектов.

Наконец, при работе с векторами числовых интервалов только что представленный механизм манипуляции с числовыми векторами используется для работы с векторами пар действительных числе вида $<\rho_1,\rho_2>$, где первая из компонент пары - левая, а вторая - правая границы соответствующего интервала. Частичный порядок на парах такого вида определяется двумя линейными порядками на их компонентах - границах интервалов: в случае согласованности обоих покомпонентных упорядочений задается соответствующее упорядочение и на самих парах. При несогласованности упорядочений на компонентах соответствующие пары (числовые интервалы) объявляются несравнимыми (т.е. могут быть упорядочены произвольным образом[24]). Допустимы также и иные варианты определения частичного порядка (например, упорядочение интервалов по упорядочению их середин и др.). Далее мы имеем возможность воспроизвести конструкцию, рассмотренную нами выше при анализе точечных числовых данных.

В заключение заметим, что каждый из представленных нами типов данных может быть задействован при формализации описания как “структуры” изучаемых объектов, так и присущих им свойств. Таким образом, для обработки данных в “комплексных” случаях (например, объекты - мультимножества, а свойства - числовые интервалы и т.п.) имеется возможность рассматривать декартовы произведения полурешеток соответствующего типа (задающих “поведение” компонент пары <**ОБЪЕКТ,СВОЙСТВА**>), для корректного определения сходства у которых потребуется лишь назначить “ведущую” компоненту соответствующего декартового произведения. Именно по этой - “ведущей” компоненте будет проверяться так называемое условие “исчерпываемости” (см., например, п.3 в **Определении 3.1.2.2**) в определении глобального сходства. (При этом по второй компоненте будет необходимо лишь наличие непустого локального сходства)[25].

---

[24] Другими словами, здесь потребуется отдельно анализировать все возможные упорядочения каждого такого множества “неразличимых” - в смысле определенного частичного порядка - числовых интервалов.

[25] Нетрудно убедиться, что существуют ситуации, когда невозможно, чтобы найденные по каждой из компонент пары <ОБЪЕКТ,СВОЙСТВА> структуры порождали одновременно и локальные и глобальные сходства. (Например, в названном конкретном случае: исчерпывающее пересечение множеств по первой компоненте по второй порождает лишь фрагмент м-башни, и т.п.).

3.2. Структура средств ДСМ-ИАД: общее описание.

Переходя к детальному обсуждению задействованных в рамках ДСМ-метода средств интеллектуального анализа данных, основное внимание представляется целесообразным уделить общему описанию логического синтаксиса, специфическим особенностям используемых средств представления знаний а также базовой понятийной схемы характеризующих ДСМ-метод формальных моделей рассуждений.

3.2.1. Синтаксическая платформа ДСМ-ИАД

Следуя «каноническим» описаниям [*Финн 88, Anshakov 89, Anshakov 89, Аншаков 93, Финн 10*], приведем описание иерархии логических языков, формирующих синтаксическую платформу ДСМ-ИАД:

- $\boldsymbol{L^0_i}$ - язык так называемых *фактоподобных*[1] формул (используемый как средство формального представления отношений наивной семантики - см. **Раздел 2.2**);
- $\boldsymbol{L_i}$ - внутренний (в смысле Д.А.Бочвара – см. [*Бочвар 38*]) язык задания *фактоподобных* формул;
- $\boldsymbol{L_e}$ - внешний (в смысле Д.А.Бочвара) язык теории (характеризующей представление *фактов*);
- $\boldsymbol{L'_e}$ - внешний языка для формулирования *правил правдоподобного вывода*;
- $\boldsymbol{L''_e}$ - внешний язык для формулирования стратегий правдоподобного рассуждения - последовательностей применения правил правдоподобного и достоверного вывода.

Выразительные возможности названных языков естественным образом расширяются в рамках обсуждаемой иерархии в соответствии с условиями

$$\boldsymbol{L^0_i \subset L_i \subset L_e \subset L'_e \subset L''_e}.$$

Общую для всех обрабатываемых ДСМ-методом типов данных схему спецификации обсуждаемой иерархии логических языков мы подробно разберем на примере работы с данными, представленными множествами. В остальных случаях (при обработке мультимножеств, числовых векторов и др.) все ее основные компоненты имеют аналогичный вид (разумеется, с учетом специфики в формализации сходства анализируемых объектов - см. **Раздел 3.1** данной **Главы**).

Итак, при работе с типом данных - “множества”, язык $\boldsymbol{L^0_i}$ определяется следующими условиями:

- заданы универсум объектов $\mathbf{U^{(1)}} = \{o_1,o_2,...,o_n\}$ и универсум свойств $\mathbf{U^{(2)}} = \{p_1,p_2,...,p_m\}$;
- индивидные константы сортов 1 и 2 определяются как элементы множеств всех подмножеств, соответственно, универсумов $\mathbf{U^{(1)}}$ и $\mathbf{U^{(2)}}$:

$$O \in 2^{\mathbf{U(1)}}, P \in 2^{\mathbf{U(2)}};$$

- заданы двуместные предикатные символы $==>_1$ и $==>_2$, первый из которых соответствует (частично определенному) отношению “обладать свойствами” (см. **Раздел 2.3** предыдущей **Главы 2**), а второй - (частично определенному) отношению “быть причиной (наличия или отсутствия) свойств”;
- заданы логические связки & и $\vee$;

[1] Термином “*фактоподобные*” объединяются два класса формул: *факты* (т.е. выражения, отражающие исходное состояние обучающей выборки - см. **Раздел 2.3**) и *полуфакты* (т.е. выражения, синтаксически устроенные точно также, как и факты, но полученные уже в ходе *правдоподобных рассуждений*).

- если t_1 и t_2 - константы, соответственно, сортов 1 и 2, то $t_1 ==>_1 t_2$ и $t_1 ==>_2 t_2$ - формулы языка $\boldsymbol{L^0_i}$;
- если $\boldsymbol{\Phi_1}$ и $\boldsymbol{\Phi_2}$ - формулы языка $\boldsymbol{L^0_i}$, то $\boldsymbol{\Phi_1 \& \Phi_2}$ и $\boldsymbol{\Phi_1 \vee \Phi_2}$ - также формулы языка $\boldsymbol{L^0_i}$.

В свою очередь язык $\boldsymbol{L_i}$ (как расширение языка $\boldsymbol{L^0_i}$) определяется заданием (в добавление к определению языка $\boldsymbol{L^0_i}$):

- переменных X,... и Y,... сортов, соответственно, 1 (для представления заданных множествами объектов и подобъектов) и 2 (для представления заданных множествами свойств);
- кванторов $\forall$ и $\exists$ для переменных каждого из сортов;

а также

- определением терма сорта i (i=1,2) как переменной сорта i или индивидной константы того же сорта;
- определением терма языка $\boldsymbol{L_i}$ как терма сорта 1 или терма сорта 2;
- определением формулы языка $\boldsymbol{L_i}$ по стандартной схеме:
  - если t_1 и t_2 - термы, соответственно, сортов 1 и 2, то $t_1 ==>_1 t_2$ и $t_1 ==>_2 t_2$ - формулы языка $\boldsymbol{L_i}$;
  - если $\boldsymbol{\Phi_1}$ и $\boldsymbol{\Phi_2}$ - формулы языка $\boldsymbol{L_i}$, то $\boldsymbol{\Phi_1 \& \Phi_2}$ и $\boldsymbol{\Phi_1 \vee \Phi_2}$ - также формулы языка $\boldsymbol{L_i}$;
  - если $\boldsymbol{\Phi_1}(X)$ и $\boldsymbol{\Phi_2}(Y)$ - формулы языка $\boldsymbol{L_i}$, в которые переменные X и Y входят свободно, то $\forall X \boldsymbol{\Phi_1}(X)$, $\exists X \boldsymbol{\Phi_1}(X)$, $\forall Y \boldsymbol{\Phi_2}(Y)$ и $\exists Y \boldsymbol{\Phi_2}(Y)$ - также формулы языка $\boldsymbol{L_i}$.

Внешний язык $\boldsymbol{L_e}$ предполагает приписывание истинностных оценок формулам внутреннего языка (характеризующего *факты* и *полуфакты* – гипотезы (зависимости), порождаемые в ходе ДСМ-рассуждения). При этом условие

$$\boldsymbol{L_i \subset L_e}$$

предполагает, что в $\boldsymbol{L_e}$ выразительные средства языка $\boldsymbol{L_i}$ пополнены следующими возможностями:

- заданы переменные сорта 3 - i,j,... - для натуральных чисел;
- заданы также индивидные константы сорта 3: 1,2,... ;
- задано множество одноместных логических связок[2] $\mathbf{J_{\underline{\nu}}}$, таких, что

$$\mathbf{J_\nu}\,\varphi = \begin{cases} 1 \text{ (логическая истина), если формула } \varphi \text{ принимает истинностное значение } \boldsymbol{\nu}[\varphi] = \underline{\nu}; \\ 0 \text{ (логическая ложь), в противном случае;} \end{cases}$$

при этом:

- множество истинностных значений $\mathbf{V_e}$ = {*логическая истина*[3], *логическая ложь*[4]} $\cup$ {<1,n>,<0,n>,<-1,n>,<τ,n> | где n=1,2,3,...} $\cup$ {<1,ω>,<0,ω>,<-1,ω>,< τ,ω>};
- индекс n внутри пар, характеризующих каждое конкретное истинностное значение вида <i,n> (где i $\in$ {1,-1,0,τ}) будет использован для кодирования числа шагов правдоподобного вывода (числа применений правил вывода - см. **Раздел 2.3**);
- "типы" значений - 1,-1,0 и τ - будут использоваться для представления внутренних (в смысле языка $\boldsymbol{L_i}$) истинностных значений , характеризующих, соответственно,

---

[2] Эти связки известны также как *функции Слупецкого* или же *характеристические функции Россера-Тюркетта*.

[3] Далее везде, где это не оговорено особо, мы будем использовать для обозначения этого истинностного значения символ ***t***.

[4] Далее везде, где это не оговорено особо, мы будем использовать для обозначения этого истинностного значения символ ***f***.

фактическую истину, *фактическую ложь*, *фактическое* (или *эмпирическое*) *противоречие*[5] и *фактическую недоопределенность*[6];

- тип значений ω будет использован для описания предельного случая - сколь угодно большого числа шагов ДСМ-рассуждения;

- бинарные функциональные символы ∩, ∪ и \, которые мы сопоставим обычным теоретико-множественным операциям[7];
- предикатные символы = и ≤ , которые будут задействованы, соответственно, для термов сортов 1 и 2, и для натуральных чисел (переменных и констант сорта 3).

Термы языка L_e определяются обычным образом: термом этого языка является любой терм сортов 1,2 и 3.

Квази-формулой[8] языка L_e является

- каждое выражение вида $t_1 ==>_1 t_2$ или $t_1 ==>_2 t_2$, где t_1 и t_2 - термы сортов, соответственно, 1 и 2; кроме того
- если Φ_1 и Φ_2 - квази-формулы языка L_e , то $\Phi_1 \& \Phi_2$ и $\Phi_1 \vee \Phi_2$ - также квази-формулы языка L_e; и наконец
- если $\Phi_1(X)$ и $\Phi_2(Y)$ - квази-формулы языка L_e, в которые переменные X и Y входят свободно, то $\forall X\Phi_1(X)$, $\exists X\Phi_1(X)$, $\forall Y\Phi_2(Y)$ и $\exists Y\Phi_2(Y)$ - также квази-формулы языка L_e.

Перейдем к определению формул языка L_e:

- если t_1 и t_2 - термы сорта i (где i=1,2), то $t_1=t_2$ формулы языка L_e;
- если t_1 и t_2 - термы сорта 3, то $t_1 \le t_2$ - формула языка L_e;
- если Φ - квази-формула или формула языка L_e, то $J_{\underline{v}}\Phi$ - формула языка L_e;
- если Φ_1 и Φ_2 - формулы языка L_e , то $\Phi_1 \& \Phi_2$ и $\Phi_1 \vee \Phi_2$ - также формулы языка L_e; и наконец
- если $\Phi_1(X)$ и $\Phi_2(Y)$ - формулы языка L_e, в которые переменные X и Y входят свободно, то $\forall X\Phi_1(X)$, $\exists X\Phi_1(X)$, $\forall Y\Phi_2(Y)$ и $\exists Y\Phi_2(Y)$ - также формулы языка L_e.

Легко видеть, что для каждой формулы Φ языка L_e может быть определено ее отрицание $\neg\Phi$ (в булевском смысле), представимое выражением

$$\neg\Phi = J_{\underline{t}}\Phi.$$

Язык L'_e , представляющий собой специальное расширение языка логики предикатов первого порядка и предназначенный для формулирования правил правдоподобного вывода, определяется следующими условиями:

- L'_e содержит в себе выразительные возможности языка L_e, т.е.

$$L_e \subset L'_e$$

- термы языка L_e есть термы языка L'_e;
- если $t_1, t_2, \ldots, t_k$ - термы все сорта 1 или все сорта 2, а k - переменная сорта 3, то $t_1 \cap t_2 \cap \ldots \cap t_k$ и $t_1 \cup t_2 \cup \ldots \cup t_k$ - термы[9] языка L'_e;
- если t_1 и t_2 - термы сорта i (где i=1,2), то $t_1=t_2$ формулы языка L'_e;
- если t_1 и t_2 - термы сорта 3, то $t_1 \le t_2$ - формула языка L'_e;

---

[5] Подробнее процедурную семантику этого типа истинностных значений мы представим ниже, в **Разделах 3.2.2** и **3.2.3**.

[6] Напомним, что одна из наших целей - изучать свойства *частично-определенных* отношений, представленных предикатными символами $==>_1$ и $==>_2$.

[7] При работе с другими типами данных здесь, соответственно, должны быть задействованы те символы операций, с помощью которых задается (см. **Раздел 3.1**) сходство изучаемых объектов.

[8] Приставка *квази-* подчеркивает, что это выражение языка L_e, которому еще не приписана истинностная оценка. Квази-формулы будут также именоваться внутренними формулами противоположность внешним - формулам языка L_e.

[9] При обработке других типов данных здесь оказываются задействованными соответствующие операции (см. **Раздел 3.1**).

- если t_1 и t_2 - термы, соответственно, сорта 1 и 2, то $t_1 ==>_1 t_2$ и $t_1 ==>_2 t_2$ - формулы языка $\boldsymbol{L'_e}$;
- если $t, t_1, t_2, ..., t_k$ - термы все сорта 1 или все сорта 2, а k - переменная сорта 3, то $t=t_1 \vee t=t_2 \vee ... \vee t=t_k$ - формула языка $\boldsymbol{L'_e}$;
- если $\boldsymbol{\Phi_1}$ и $\boldsymbol{\Phi_2}$ - формулы языка $\boldsymbol{L'_e}$, то $\boldsymbol{\Phi_1 \& \Phi_2}$ и $\boldsymbol{\Phi_1 \vee \Phi_2}$ - также формулы языка $\boldsymbol{L'_e}$;
- если $\boldsymbol{\Phi}$ - формула языка $\boldsymbol{L'_e}$, то (полагая, что $\underline{\nu} \in \mathbf{V_e}$ -заданному множеству истинностных значений) $\mathbf{J_{\underline{\nu}}\Phi}$ - формула языка $\boldsymbol{L'_e}$;
- если $\boldsymbol{\Phi_1}(X)$ и $\boldsymbol{\Phi_2}(Y)$ - формулы языка $\boldsymbol{L'_e}$, в которые переменные X и Y входят свободно, то $\forall X \boldsymbol{\Phi_1}(X)$, $\exists X \boldsymbol{\Phi_1}(X)$, $\forall Y \boldsymbol{\Phi_2}(Y)$ и $\exists Y \boldsymbol{\Phi_2}(Y)$ - также формулы языка $\boldsymbol{L'_e}$;
- если $t^{(1)}_1, t^{(1)}_2, ..., t^{(1)}_k$ - термы сорта 1, $t^{(2)}_1, t^{(2)}_2, ..., t^{(2)}_k$ - термы сорта 2, а k - переменная сорта 3, то

$$t^{(1)}_1 ==>_i t^{(2)}_1 \vee t^{(1)}_2 ==>_i t^{(2)}_2 \vee ... \vee t^{(1)}_k ==>_i t^{(2)}_k \quad \text{и}$$

$$t^{(1)}_1 ==>_i t^{(2)}_1 \,\&\, t^{(1)}_2 ==>_i t^{(2)}_2 \,\&\, ... \,\&\, t^{(1)}_k ==>_i t^{(2)}_k$$

(для i =1,2) - формулы языка $\boldsymbol{L'_e}$;

- если $\boldsymbol{\Phi}$ - формула языка $\boldsymbol{L'_e}$, содержащая свободные вхождения n - переменной сорта 3 в выражениях вида $\mathbf{J_{\underline{<\alpha,n>}}\Phi}$ для $\alpha \in \{1,-1,0,\tau\}$, то $\forall n \mathbf{J_{\underline{<\alpha,n>}}\Phi}$ и $\exists n \mathbf{J_{\underline{<\alpha,n>}}\Phi}$ - формулы языка $\boldsymbol{L'_e}$;
- если $\boldsymbol{\Phi}$ - формула языка $\boldsymbol{L'_e}$, в которую n - переменная сорта 3 - входит свободно, то $\forall n \boldsymbol{\Phi}$ и $\exists n \boldsymbol{\Phi}$ - формулы языка $\boldsymbol{L'_e}$;
- если $\boldsymbol{\Phi}$ - формула языка $\boldsymbol{L'_e}$, в которую переменные $X_1, X_2, ..., X_k$ (все сорта 1 или все сорта 2) входят свободно, а k - переменная сорта 3, то $\mathbf{Q}_1 X_1 \mathbf{Q}_2 X_2 ... \mathbf{Q}_k X_k \boldsymbol{\Phi}$ для $\mathbf{Q}_i \in \{\exists, \forall\}$ и i=1,2,...,k - также формулы языка $\boldsymbol{L'_e}$.

Фундаментальной особенностью языка $\boldsymbol{L'_e}$ является возможность выразить его средствами конструкции, основанные на поиске *неподвижной точки* соответствующего оператора. Так, например, в [*Финн 88, Гусакова 87, Anshakov 89* и др.] демонстрируется, что средствами языка $\boldsymbol{L'_e}$ выразимо транзитивное замыкание бинарного отношения (не представимое обычными средствами логики предикатов первого порядка). Отмеченная особенность языка $\boldsymbol{L'_e}$ связана с использованием формул переменной длины (формул с троеточием[10]), где соответствующий числовой параметр k играет роль *параметра эмпирической индукции* при учете примеров, послуживших основанием правдоподобного вывода.

Необходимые уточнения свойств языка $\boldsymbol{L''_e}$ [11], используемые при обработке средствами ДСМ-ИАД различных типов данных (см. варианты определения операции сходства для соответствующих типов данных в **Разделе 3.1**) формируются единообразно, однако, всякий раз при этом выполняется естественное условие вида

$$\boldsymbol{L'_e} \subset \boldsymbol{L''_e}.$$

[10] Исчерпывающее описание формальной природы выражений такого типа (основанное на использовании так называемых кванторов по кортежам переменной длины) предложено, например, в работе [*Скворцов 83*].
[11] Подробнее о стратегиях ДСМ-рассуждения мы поговорим ниже (см. **Раздел 3.3**).

3.2.2. Квази-Аксиоматические Теории как платформа для представления знаний в интеллектуальном анализе данных.

Целью работ по искусственному интеллекту всегда (прямо или косвенно) было создание программно-технических комплексов – систем искусственного интеллекта - выполняющих такие действия, для исполнения которых человеку требуется привлечение возможностей его интеллекта. Одним из подобного сорта действий всегда считалась способность вести рассуждения[1]. Рассуждение как средство получения нового знания - объект главного нашего интереса в интеллектуальном анализе данных. Специфика использования рассуждений в таких условиях (причем как в ситуации, когда рассуждения проводит человек, так и, в равной степени, когда их осуществляет интеллектуальная система) - в реальной (или потенциальной) неполноте информации об объекте исследования.

С методологической точки зрения рассуждение - основа для весьма заманчивого варианта представления знаний в компьютерных системах. Выразив (по крайней мере) часть описывающих конкретную предметную область (ПО) фактов в виде тех или иных зависимостей, можно получить более компактное описание анализируемой ПО[2]. Кроме того, задействовав (в арсенале решающих программных инструментов компьютерной системы) формализованные средства рассуждения, можно организовать систематический обзор всех корректных следствий из имеющихся фактов и зависимостей, порожденных как в ходе достоверного (т.е. дедукции), так и правдоподобного (например, порождения гипотез с использованием аналогий, абдуктивных рассуждений и т.п.) вывода.

Получение нового знания как проблема науки, научного знания и логики научного открытия - вот критически важный элемент во всех формализованных или неформализованных представлениях о рассуждении. До середины 19 века логика научного познания (да и логика вообще) понималась, преимущественно, как теория силлогизмов[3]. Аристотель (см., например, [*Аристотель 81*]) считал предметом науки то общее, что постигалось умом (т.е. в ходе умозаключений). При этом научное знание понималось Аристотелем как результат вывода из общего закона (так называемой *большой посылки*) и частных обстоятельств (так называемой *малой посылки*), полученный с помощью применения *фигуры*[4] *вывода* одного из трех типов

- непосредственного вывода;
- силлогизма, или
- сорита[5].

При этом непосредственный вывод - частный случай силлогизма, в котором отсутствует малая посылка, а заключение имеет те же термы, что и большая посылка. В свою очередь, сорит (или так называемый большой силлогизм) представляет собой последовательность силлогизмов, где заключение предыдущего силлогизма служит одной из посылок последующего. Таким образом, проблема науки у Аристотеля - это проблема дедукции, т.к. научное знание здесь - это запись научных данных в форме силлогизмов. Однако в этой схеме остается непонятным, что такое процесс образования новых понятий. (По Аристотелю каждое понятие есть результат *обобщения*, но процесс обобщений не совпадает с процессом образования силлогизмов. Более того, в науке часто новые понятия получаются

---

[1] Здесь мы понимаем рассуждение неформально, т.е. в широком смысле. Чуть позже мы приведем также формальное уточнение этого понятия.

[2] Именно на это обстоятельство и обратил внимание сообщества исследователей, работающих в области искусственного интеллекта, Дж.Маккарти (см., в частности, [*McCarthy 00, McCarthy*] и др.).

[3] В "*Критике чистого разума*" (1781 год) Кант пишет (см., например, [*Кант 66*]), что логика была уже законченной наукой еще почти 2 тысячелетия назад, во времена Аристотеля.

[4] Т.е. (в современной терминологии) правила вывода - формализованной схемы перехода от посылок к заключению.

[5] От *sorites* (греч.) - кучеобразный (ср.: soros - куча).

вовсе не путем обобщения[6]). Таким образом, проблема порождения нового знания (в том числе - проблема порождения зависимостей) как *ПРОБЛЕМА НОВОГО* оказывается не получившим адекватного развития элементом теории Аристотеля: в силлогизм ввести что-либо новое не возможно, - это аналитический вывод.

Среди усилий, предпринимавшихся с целью преодолеть ограниченный характер аристотелевой теории познания, особого внимания заслуживают работы Ф.Бэкона (см., например, [*Бэкон 35*]) и Д.С.Милля (см., например [*Милль 00*] и др.), посвященные проблеме *метода* научного исследования. Развитие логики как инструментария научного познания к середине 19 века значительным образом продвинули Д.Булль, О.ДеМорган и Д.Венн. Основы современного представления о структуре логических формализмов (в частности, о синтаксисе и семантике исчислений высказываний и предикатов первого порядка) сформулировал во второй половине 19 века Ч.С.Пирс.

Работы Джорджа Булля, Огастаса ДеМоргана и Джона Венна обозначили новый шаг в развитии аристотелевой теории. У Аристотеля теория вывода состоит из трех частей:

- теории логических термов,
- теории предложений, и
- теории вывода.

В основе теории термов - существование так называемых категориальных термов лишь двух типов - субъектных или предикатных. (Назначение каждого категориального терма - разделить мир на две совокупности: на тех, кто соответствует терму, и тех, кто ему не соответствует). Суждение в теории Аристотеля есть предикатная комбинация двух термов (субъектного и предикатного) в рамках одного предложения[7]. Теория вывода - правила применения силлогизмов, т.е. формальных схем, определяемых следующими условиями:

- в силлогизме две категориальные посылки;
- в обеих посылках имеется всего три терма;
- каждый терм входит в силлогизм ровно два раза;
- нет терма, который входит два раза в одно и тоже суждение (предложение).

Нетрудно убедиться, что такого рода формализация не располагает средствами, например, для представления относительных понятий (в частности, **X** *РОДСТВЕННИК* **Y** и т.п.), формализуемых при помощи отношений. Кроме того, в теории Аристотеля нет средств для представления связей между понятиями (а также между суждениями), в то время как в науке используются достаточно сложные структуры предложений. Наконец, проблема правил вывода - схем порождения новых утверждений из уже имеющихся имеет в аристотелевой теории науки решение, которое, вообще говоря, трудно считать удовлетворительным.

Преодолению сформулированных ограничений и были посвящены основные усилия названных нами выше исследователей. Так, развивая формальные средства логики, Д.Булль, О.ДеМорган и Д.Венн в своих работах нацелились на две основные идеи - на переход к использованию термов как *совокупностей* (и представлению этих совокупностей в виде специальных графов - *диаграмм*), а также на переход к использованию *отношений* и их *композиций*. Тем не менее, проблема порождения новых знаний здесь либо повторяла аристотелеву схему, либо (в частности - у Д.Булля) относилась “в ведение” теории вероятностей.

---

[6] Здесь, по-видимому, достаточно привести лишь несколько примеров, в частности: определение понятия “*сила*” в общей физике, понятие «*гамильтониан*» в теоретической механике др. .

[7] Таким образом, фактически, характеризуются лишь количество и качество, при этом имеется всего четыре вида суждений (так называемый логический квадрат):

1) общие утверждения количества или качества (например, *ВСЕ* **X** *ИМЕЮТ* **Y**);

2) частные утверждения количества или качества (например, *НЕКОТОРЫЕ* **X** *ИМЕЮТ* **Y**);

3) общие отрицания утверждения количества или качества (например, *НЕВЕРНО, ЧТО ВСЕ* **X** *ИМЕЮТ* **Y**);

4) частные отрицания утверждения количества или качества (например, *НЕВЕРНО, ЧТО НЕКОТОРЫЕ* **X** *ИМЕЮТ* **Y**).

Чарльз Сандерс Пирс обратил внимание на различия *аналитических* и *синтетических* суждений у Канта (см,, например, [*Кант 66*]). Он стал рассматривать *дедуктивные* и *недедуктивные* выводы, придя к заключению, что проблема науки в получении *новых понятий* и *законов*. В 1860 г. Пирс устанавливает различия между *expliative* и *ampliative* (т.е. между *объяснительными* и *расширительными*) суждениями. Согласно Пирсу в expliative выводе заключение только объясняет смысл посылки, в то время как в ampliative выводе заключение расширяет смысл посылки. В 1865 г. Пирс говорит уже о двух видах ampliative выводов - абдукции и индукции, а также о expliative выводе как дедукции. В 1867 г. в работе "On the natural classification of arguments" [*Peirce 67, Peirce 67-79*] Пирс вместе с уже представленной классификацией выводов предлагает также ставшую классической схему номологического объяснения *от закона к результату* (так называемая *covering law model*[8]). Позднее эту схему можно найти у К.Поппера [*Popper 58*] и К.Гемпела [*Hempel* 48], ее использовали также Карнап [*Carnap 36*] и Рейхенбах [*Reichenbach 49*]. В самом общем виде предложенная Пирсом схема классификации выводов имеет следующую форму:

- заключение "*от общего к частному*" (или "*от правила к результату*")

все C-A
все B-C
∴все B-A

(этот вид вывода Пирс называет *дедукцией*);

- заключение "от частного к общему" (или "*от результата к правилу*")

все B-A
все B-C
∴все C-A

или

результат
условие
∴правило (закон)

(этот вид вывода Пирс называет *индукцией*);

- заключение "от общего и частного к конкретному случаю"

все C-A
все B-A
∴все B-C

или

правило
результат
∴условие

(этот вид вывода Пирс называет *абдукцией*).

Фактически Пирс модифицировал известную со времен Аристотеля теорию вывода и фокусировал внимание на следующих трех важнейших обстоятельствах:

- язык для представления утверждений;
- правила порождения новых утверждений из уже имеющихся (правила вывода); и

---

[8] *Covering law model* - схема научного объяснения следующего вида:

закон (*правило*)
условие (*обстоятельства*)
∴результат (*событие*)

в которой из большой посылки - природного (естественного) закона - и малой посылки - особенного обстоятельства (условия) - порождается заключение - результат, который характеризует специфическое событие.

- система оценивания (критерии принятия результатов применения правил вывода).

Предложенный им логический формализм для описания исчисления высказываний и исчисления предикатов первого порядка (см., например, реконструкцию соответствующих логических идей и построений Пирса, выполненную R.W.Burch'ем [*Burch 91*]) стал основой современных представлений о логике первого порядка. Идея разделения фигур вывода (дедуктивных, индуктивных и абдуктивных правил порождения новых предложений из уже имеющихся) в соответствии с теми или иными "правилами принятия" и "характером" (т.е. степенью "достоверности", "надежности" и т.п.) порождаемых результатов обусловила интерес к порождению адекватных процедур оценивания результатов вывода. Вновь востребованными оказались сформулированные еще Д.Юмом [*Hume 88*] принципы эпистемологии (в частности, принцип индукции Д.Юма, утверждающий невозможность верификации индуктивных обобщений исключительно логическими средствами, однако допускающий принятие гипотез на основе "ассоциации идей"). Наконец, сложилось представление о необходимости разделения процессов формирования гипотез (новых утверждений, получаемых, в частности, применением некоторых правил вывода) и процессов их верификации (т.е. подтверждения и принятия, либо опровержения и отбрасывания как ложных, не подтверждающихся).

Так развивая представления об абдукции как механизме порождения гипотез, Пирс обратился к анализу механизмов причинной связи (понимая причинность в соответствии с предложенным еще Юмом толкованием[9]). А в более детальном анализе абдукции и индукции он разделил их как две независимые стадии процесса научного исследования (а не только как две различные формы вывода). Согласно Пирсу абдукция - это вывод *к гипотезе*, объясняющей данные наблюдений (т.е. к гипотезе о законе, причине, правиле), имеющий вид:

(1) мы имеем наблюдения O_1, O_2, ...
(2) если гипотеза H истинна, то H объясняет O_1, O_2, ...
∴разумно считать, что гипотеза H может быть истинной

(Таким образом, абдукция призвана объяснять что-либо).

С другой стороны, индукция (по **Пирсу**) - это процедура "проверки" гипотезы, осуществляемой в соответствии со схемой:

(1) если гипотеза H истинна, то должны быть истинны
также и предложения P_1, P_2, ...
(2) наблюдения O_1, O_2, ... указывают на то,
что предложения P_1, P_2, ... действительно истинны

(2') наблюдения O_1, O_2, ... указывают на то,
что хотя бы одно из предложений P_1, P_2, ... действительно ложно
∴(3) гипотеза H имеет большую степень правдоподобия, чем ранее

***∴*(3') гипотеза *H* просто ложна**

[9] По определению Д.Юма причина - это явление, имеющее три свойства:

1. причина и следствие (т.е. причина и сопутствующее ей другое явление, называемое следствием) неизменно встречаются вместе (т.е. выполняется условие *неизменной совместимости - constant conjunction*). Таким образом, если случается причина, то неизменно встречается и следствие; кроме того, если встречается следствие, то непременно имеется и причина;
2. причина и следствие соседствуют друг с другом (т.е. выполняется условие *близости, соседства - contiguity*: они близки в физическом пространстве и физическом времени. Причина всегда предшествует во времени, а следствие - явление, которое всегда следует во времени за причиной);
3. причина и следствие связаны друг с другом так, что появление причины необходимо влечет за собой появление следствия (условие *необходимой связи - necessary connection*).

(Здесь при вхождении в посылки утверждения **(2)** в заключение входит утверждение **(3)**, а при вхождении в посылки утверждения **(2')**, соответственно, утверждение **(3')**. Одновременное вхождение в посылки утверждений **(2)** и **(2')** не допускается). Следует отметить, что правило ***(1), (2)*** $\rightarrow$ **(3)** в последней схеме представляет собой прообраз схемы подтверждения (*confirmation*) в смысле Рудольфа Карнапа [*Carnap 36*] (правда в отличие от его формализации здесь еще нет исчисления: Пирс говорит лишь о возрастании степени правдоподобия). Правило ***(1),*** **(2')** $\rightarrow$ **(3')**, в свою очередь, есть прообраз схемы фальсификации (*corroboration*) в смысле Карла Поппера [*Popper 58*]. Таким образом, индукция у Пирса - либо подтверждение, либо фальсификация (т.е. - *тестирование*) гипотезы. В свою очередь, абдукция у него - не всегда порождение гипотезы. Это - вывод *к гипотезе* (т.к. гипотеза появляется в заключении вывода. Более того, допускается, что она даже может существовать заранее - см. появление гипотезы во второй посылке абдукции, демонстрирующее, что гипотеза, вообще говоря, вовсе не должна порождаться собственно в процессе абдукции). В сложившейся ситуации понятно, что вопрос

Откуда возникают гипотезы?

достаточно удовлетворительного ответа не получил также и у Пирса: его указания на использование инстинкта и/или инсайта уместны в деятельности человека-исследователя, однако мало что дают в инструментальном смысле для создания интеллектуальных компьютерных систем, способных осуществлять процесс формирования гипотез.

Проблема *метода* как "механизма" порождения нового знания после Аристотеля, пожалуй, впервые нетривиальным образом была рассмотрена Ф.Бэконом (см., например, [*Бэкон 35*] и др.). Провозгласив целью науки способность увеличивать власть человека над природой, Бэкон обратил внимание, что этой цели наука способна достичь лишь постигая "истинные причины" явлений. Скептически относясь к предыдущему знанию (страдавшему по его мнению "догматизмом" или "эмпиризмом"), Бэкон тем не менее признавал возможность получения достоверного знания. Для того, чтобы обеспечить достижение истины, он предложил реформировать *метод* научного познания. Он считал, что после устранения ложных воззрений ("идолов" ума) возможно построение новой науки на основе истинного метода. Эта наука должна рациональным образом перерабатывать полученные из опыта факты. Посылками ее выводов должны также стать суждения, основанные на понятиях, полученных в рамках индукции или обобщений средствами метода. В свою очередь предпосылки индукции - аналитическое понимание результатов эксперимента.

Бэкон, сформулировав идею *таблиц присутствия* и *таблиц отсутствия*, предложил формализацию индуктивных выводов посредством таблиц причин. Ему также принадлежит идея организации эмпирической структурной индукции, цель которой - получение эмпирических зависимостей в виде индуктивных обобщений. Такие индуктивные обобщения должны получаться в ходе сравнения объектов, имеющих структуру и входящих в явления, которые должны быть представлены примерами и контрпримерами. Бэкон впервые указал в теории индукции на роль отрицательных примеров (так называемых "отрицательных инстанций"): подбор случаев противоречащих ранее сделанным обобщениям (в ситуации наличия таких случаев) говорит о необходимости пересмотра таких обобщений как недостаточно обоснованных (ср. с требованиями *достаточности оснований* для принятия суждений у Поппера, Есенина-Вольпина и др.).

Д.С.Милль (см., например [*Милль 00*] и др.) развил идеи Бэкона, предложив известные *методы сходства, различия, сопутствующих изменений* и *остатков*. (О структуре и свойствах миллевских построений мы подробно поговорим чуть ниже).

В последующее время предпринимался ряд попыток построить формальные имитации миллевских методов средствами математической логики (см., например, работы [*von Wright 60, Los 77, Czerwinski 60, Магинскас 63*] и др.). Однако все эти исследования были ориентированы на сугубо философские цели, использовали для формализации

аппарат двузначной логики и не дали сколь бы то заметных результатов, повлиявших на создание компьютерных систем поддержки рассуждений в рамках исследований по искусственному интеллекту.

Среди множества работ, посвященных совершенствованию средств вывода новых суждений из уже имеющихся, отметим также работы Д.Пойа ([*Пойа 70,75*] и др.), развивающие идеи вывода по *аналогии*. Здесь сформулированы два важных принципа:

- суждение становится *более правдоподобным* в ситуации, когда *истинным* оказывается *аналогичное* ему суждение; и
- суждение становится *несколько более правдоподобным* (по отношению к исходному состоянию) в ситуации, когда *правдоподобным* становится *аналогичное* ему суждение.

Наконец, В.К.Финном и его учениками (см., например, [*Финн 88, ИНТ 91, Финн 10*] и др.) был предложен машинно-ориентированный вариант формализации миллевских методов в рамках так называемых *квази-аксиоматических теорий*. Логические средства формализации здесь опираются на аппарат специальных многозначных логик, позволяющий в рамках разделения соответствующих внутренних и внешнего языков формализации реализовать также идеи эпистемологии в стиле Д.Юма (результат не являющегося достоверным вывода имеет истинностные оценки подобные классическим истинностным оценкам двузначной логики, но отличные от них). Осуществленная здесь логическая формализация миллевских схем вывода далее воплощена в алгоритмическую конструкцию, ставшую основой для создания серии компьютерных интеллектуальных систем, некоторые из которых эффективно используются в приложениях.

Существование различных типов выводов (т.е. различных фигур вывода и порождаемых с их помощью результатов) фокусируют внимание исследователей на проблеме принятия (оценивания) получаемых их применением заключений. Нетрудно представить себе, что при переходе к открытым предметным областям обсуждаемая проблема оценивания заведомо переходит в разряд нетривиальных: большинство традиционных методов оценивания, использующих те или иные формализации понятия *мера*, оказываются, вообще говоря, неадекватными (т.к. в условиях неполноты информации в открытой предметной области - ПО - данных об "*измеримости*" пополнений анализируемой ПО новой информацией может попросту не быть)[10].

Особая роль в характеристике формализованных средств вывода новых суждений (в том числе – из фактов) отводится их *рациональности*. Те или иные формальные уточнения этого термина (см,, например, [*Hajek 78, Финн 88*] и др.) позволяют получать конкретные версии соответствующего инструментария. Например, П.Гаек и Т.Гавранек [*Hajek 78*] предлагают считать рациональным такое правило вывода, которое при истинных посылках порождает не являющееся истинным заключение лишь в множестве случаев, имеющих (в так или иначе формализуемом смысле) "*малую меру*".

А.С Есенину-Вольпину (см., например, [*Есенин 93*] и др.) принадлежит идея так называемого *эвристического принципа доверия* - методического приема, определяющего правила принятия (или непринятия) утверждений в рамках научного исследования. Постулируется, что допустимыми могут быть лишь те суждения, которые имеют *достаточные основания* для их принятия. В свою очередь вера понимается как нарушение *закона достаточного основания* для принятия суждения. Под доказательством подразумевается любой честный прием, делающий суждение неоспоримым. Более того, *оспоримость* понимается как синоним *недоказанности*, а *не*оспоримость - *не* недоказанности. Таким образом, корректное суждение - это суждение, способное

[10] Так, например, отдельного обсуждения в каждом конкретном случае требует при работе с открытыми ПО техника статистического оценивания порождаемых гипотез: в общем случае открытый характер ПО означает, что мы располагаем лишь данными о частотах соответствующих событий, но, вообще говоря, не имеем достоверной информации об их вероятностях.

преодолеть любую критику (аналогично и понятие корректного доказательства). Есенин-Вольпин вводит понятие *фундаментализма* в науке, связанного в первую очередь с интересом к вопросу ПОЧЕМУ? и, конечно же, способностью получать ответы на такие вопросы. Эта точка зрения представляется естественным развитием обсуждавшихся нами чуть выше взглядов на роль в науке причинного анализа, осуществляемого в рамках рассуждения.

Для экспериментальных наук эвристический принцип доверия может быть сформулирован в следующем виде:

если *максимально доступное развитие критики доверия не приводит к акту недоверия*,
то *это можно считать (апостериорным) обоснованием акта доверия*.

Таким образом, в области экспериментальных наук *эвристический принцип доверия* требует учесть в процедуре оценке суждения все его возможные фальсификаторы. Лишь отсутствие таковых можно считать достаточным основанием для принятия рассматриваемого суждения. (В случае работы с такими ПО, в которых “вера может оказаться неистребимой”, предлагается строить соответствующие *логики доверия*, где в явном виде был бы зафиксирован *источник доверия* а также “*способ толкования его показаний*” [*Есенин 93*]). В такой ситуации науки (и прежде всего - эмпирические науки) естественно понимать как *системы ответов на вопросы* (см. выше понятие фундаментализма в науке), полученных с помощью *метода*[11].

Необходимость различать в таком процессе многие сорта истин естественным образом приводит к использованию в процессе формализации рассуждений математической техники многозначных логик. Д.А.Бочвар (см., например, [*Бочвар 38*] и др.) предложил разделить логические средства описания семантики анализируемой ПО и логические средства, собственно, формализации рассуждения. Таким образом появилась возможность разделить суждения о тех или иных *фактах* в исследуемой ПО, суждения об *оценках* этих фактов (построенных в рамках соответствующей эпистемологии), и наконец, средства формализации *рассуждения* о фактах и их оценках в анализируемой ПО. Оформленные (средствами специальных многозначных логик) в виде внутреннего (предназначенного для представления семантики ПО) и внешнего (предназначенного для формализации рассуждения о ПО и являющегося метаязыком по отношению к внутреннему) логических языков, математические средства логического анализа рассуждения оказались эффективным инструментарием исследования научных теорий[12].

Теперь, обсудив исторически сложившиеся основания для разделения процесса порождения нового знания на стадии:

- выбора языка формализации,
- выбора правил порождения новых утверждений из уже имеющихся (т.е. выбора правил вывода),
- выбора средств оценивания порождаемых результатов, и
- выбора критериев принятия порождаемых результатов (критериев анализа фальсификаторов и оценки достаточности оснований для принятия результатов вывода),

мы располагаем необходимым инструментарием для осуществления классификации возможных типов выводов. Следуя [*Финн 88*], мы разделим выводы на *достоверные* (т.е. всегда обеспечивающие переход от истинных посылок к истинным заключениям) и *недостоверные* (т.е., вообще говоря, допускающие и переход от истинных посылок к не

[11] Фактически, понимаемого как набор правил вывода и руководство по их применению.

[12] Примерами такого использования могут служить, в частности, анализ логических и семантических парадоксов, успешно осуществленный Д.А.Бочваром (см. работу [*Бочвар 38*]), или же предложенный Ф.Рамсеем [*Ramsey 50*] механизм логического анализа научных теорий.

являющемуся истинным заключению). В свою очередь недостоверные выводы мы классифицируем по способу "обзора" и характеру множеств их возможных фальсификаторов. (При этом приближенные выводы - например, выводы в рамках парадигмы *soft computing*[13] и т.п. - мы сознательно оставим несколько в стороне, как требующие отдельного рассмотрения). Далее предметом нашего основного интереса будут, преимущественно, те недостоверные выводы, для которых множество их потенциальных фальсификаторов не пусто, и существует эффективная процедура анализа этого *открытого* (нас ведь интересует возможность оперировать с открытыми предметными областями!) множества потенциально возможных фальсификаторов.

Итак, мы будем понимать под *правдоподобным выводом* такой вариант правила перехода от одних суждений к другим, в рамках которого

> *из истинных посылок выводятся заключения, которые, вообще говоря, могут оказаться и ложными, но степень нашей уверенности в которых достаточно высока.*

Нетрудно обнаружить параллели между только что данным определением и приводившимся нами ранее понятием *рационального* (по Гаеку и Гавранеку [*Hajek 78*]) правила вывода:

> *правило вывода считается рациональным, если множество порождаемых с его помощью из истинных посылок ложных заключений имеет малую меру* (понимаемую в том или ином формальном смысле).

Следует, однако, заметить, что оперируя в открытых предметных областях, можно оказаться в ситуации, когда с определением естественным образом интерпретируемой меры (корректно определенной на *любой* подобласти анализируемой ПО) могут возникнуть определенные трудности[14]. Именно поэтому полезным может также оказаться другое понимание рациональности, например, как *нефальсифицируемости доступными средствами* порождаемых результатов (ср. с обсуждавшимся выше *эвристическим принципом доверия* Есенина-Вольпина). Именно эти соображения и будут положены в основу подробно рассматриваемой нами ниже формализации правдоподобных рассуждений средствами квази-аксиоматических теорий.

При уточнении понятия *достоверный вывод* формальной конструкцией, используемой для определения понятия *дедуктивное рассуждение*, являются *аксиоматические теории* (***AT***). В ситуации, когда задан соответствующий логический язык (алфавит и правила построения формул), аксиоматическая теория определяется парой

$$\Gamma = \langle \Sigma, \Re \rangle$$

где

Σ - множество аксиом, а

$\Re$ - множество правил вывода, гарантирующих переход от истинных посылок всегда к истинному заключению[15].

В свою очередь *дедуктивное рассуждение* (которое мы будем понимать как синоним логического вывода) определяется как цепочка формул

$$\varphi_1, \varphi_2, \ldots, \varphi_n, \psi$$

---

[13] *Soft computing* ("мягкие" вычисления) - направление исследований в области computer science и ее приложений, основанное на использовании так называемых *нечетких множеств* (*fuzzy sets*) и построенных на их основе вычислительных конструкций.

[14] См., например, уже приводившиеся нами выше соображения по поводу проблемы корректного перехода в открытых ПО от частот к вероятностям событий.

[15] Т.е. множество правил достоверного (дедуктивного) вывода.

где ψ - цель рассуждения, а каждая из формул φ_i (для каждого $i \in \{1,2, \ldots, n\}$) - либо элемент множества Σ, либо (как и формула ψ) получена из предыдущих элементов этой цепочки применением правил вывода из множества $\mathfrak{R}$.

При уточнении понятия *правдоподобный вывод* формальной конструкцией, используемой (см., например, [*Финн 88*] и др.) для определения общего понятия *рассуждение* (а вместе с ним и понятия *правдоподобное рассуждение*) являются *квази-аксиоматические теории* (КАТ). В ситуации, когда задана соответствующая иерархия[16] логических языков (алфавиты и правила построения формул), квази-аксиоматическая теория определяется тройкой

$$\Gamma = <\Sigma,\Sigma',\mathfrak{R}>$$

где

Σ - множество аксиом, характеризующих текущее состояние анализируемой ПО. (Грубо говоря, это множество локальных зависимостей, имеющих место в текущем состоянии анализируемой ПО. Следует подчеркнуть, что это множество порождает лишь неполное аксиоматическое описание ПО, которое, вообще говоря, может быть пополнено новыми зависимостями, порождаемыми в ходе анализа вновь поступивших данных об анализируемой ПО);

Σ' - множество эмпирических утверждений (элементарных фактов), характеризующих эмпирические знания о текущем состоянии анализируемой ПО. (При этом следует подчеркнуть, что элементы множества Σ', вообще говоря, не являются логическим следствием утверждений из множества Σ);

$\mathfrak{R}$ - множество правил вывода, состоящее из двух частей

$$\mathfrak{R} = \mathfrak{R}_r \cup \mathfrak{R}_p ,$$

где

$\mathfrak{R}_r$ - множество правил достоверного (дедуктивного) вывода, а

$\mathfrak{R}_p$ - множество правил правдоподобного вывода (т.е. формализованных версий абдуктивных и/или индуктивных выводов, правил вывода по аналогии, рассуждений в духе “здравого смысла”[17] и др.).

Теперь, вновь следуя [*Финн 88*], мы имеем возможность определить *рассуждение* как цепочку формул

$$\varphi_1, \varphi_2, \ldots, \varphi_n, \psi$$

где ψ - цель рассуждения, а каждая из формул φ_i (для каждого $i \in \{1,2, \ldots, n\}$) - либо элемент множества Σ, - либо элемент множества Σ', либо (как и формула ψ) получена из предыдущих элементов этой цепочки применением правил вывода из множества $\mathfrak{R}$.

Предложенное уточнение понятия *рассуждение* позволяет

- включить в среду логического рассуждения конструктивные средства порождения гипотез (в частности, средства конструктивной абдукции и индуктивных обобщений в рамках рассуждений по аналогии[18]);
- соединить в унифицированной среде логического рассуждения средства выдвижения гипотез (правила правдоподобного вывода, в частности, абдуктивные и индуктивные

---

[16] Следует напомнить, что в соответствии с представленными выше принципами эпистемологии для построения адекватной формализации правдоподобных рассуждений необходимо, по крайней мере, разделить внутренний и внешний языки, выбираемые, соответственно, для представления фактов в анализируемой ПО и для представления рассуждений о фактах из анализируемой ПО.

[17] Так называемых *common sense reasoning* (см., например, [*AI 80*] и др.)

[18] Т.е. не только правила вывода “*к гипотезе*” (см. выше рассмотрение вариантов абдукции у Пирса), но и средства конструктивного порождения гипотез из имеющихся эмпирических данных - в частности, формальные уточнения правдоподобных рассуждений в стиле Бэкона-Милля, осуществленные в рамках ДСМ-метода автоматического порождения зависимостей из данных.

выводы, выводы по аналогии и др.) и средства верификации порождаемых гипотез проверкой их дедуктивной совместимости с аксиомами из множества **Σ;**

- включать в среду логического рассуждения использование тех или иных метасредств (от проверки на логическую непротиворечивость порождаемых правдоподобным выводом гипотез с привлекаемыми в процессе рассуждений посылками и использования тех или иных критериев отбора релевантных посылок для правдоподобного вывода[19], до использования тех или иных вычислительных процедур для порождения промежуточных результатов рассуждения[20]);
- предложить новый эффективный инструментарий представления знаний в интеллектуальных системах - квази-аксиоматические теории;
- учесть (на процедурном уровне[21]), вообще говоря, немонотонный характер развития научного знания (и прежде всего - в открытых предметных областях);
- осуществить (по аналогии с продвижением от формальных описаний аксиоматических теорий к компьютерным системам автоматического доказательства теорем и интеллектуальным системам - дедуктивным решателям задач) переход к созданию компьютерных систем поддержки рассуждений - *интеллектуальных систем*[22], способных объединить в унифицированной среде логического рассуждения средства числового (распознавание образов, статистика и другие вычислительные методы) и символьного (дедукция и правдоподобные рассуждения) анализа данных, порождения и верификации (характеризующих эти данные) эмпирических зависимостей.

Завершая обзор идей, послуживших основанием для сегодняшней точки зрения на принципы организации, архитектуру и функциональные возможности интеллектуальных систем автоматического порождения зависимостей, выделим[23] критически важные, на наш взгляд, составляющие сложившейся ситуации:

- античная теория науки, силлогистика (*Аристотель*);
- анализ причинности в рамках совершенствования научного метода (*Бэкон*);
- принципы эпистемологии (*Юм*);
- аналитические и синтетические суждения (*Кант*);
- логический формализм (*Булль*, *ДеМорган*, *Венн*, *Пирс*);
- каноны индуктивной логики (*Милль*);
- абдуктивные и индуктивные выводы, разграничение правдоподобных и достоверных выводов (*Пирс*);
- принципы верификации научных теорий (*Поппер*);
- многозначные логики и их приложения в логике научного познания (*Бочвар*, *Рамсей*);
- эвристический принцип доверия, логики естественных наук (*Есенин-Вольпин*);
- немонотонные рассуждения в интеллектуальных системах (*МакКарти*);
- логический формализм квази-аксиоматических теорий (*Финн*).

---

[19] Поучительным примером конкретной процедуры такого отбора может быть используемое в ***ДСМ-методе*** (см., например, [*Финн 88*] и др.) *правило направленного расширения обучающей выборки* в ситуации *невыполнения* критерия *достаточности оснований* для корректного вывода в рамках ДСМ-рассуждения.

[20] Это могут быть, например, методы распознавания образов или методы статистического анализа данных, используемые для предварительной - грубой - кластеризации исходных данных перед их детальным детерминистским причинным анализом в ходе формализованного рассуждения. Это могут быть также методы квантово-химических расчетов (см., например, [*Раевский 88*]) и др..

[21] Т.е. располагая возможностями моделировать эту немонотонность как на уровне синтаксических конструкций в рамках соответствующих квази-аксиоматических теорий, так и на уровне реализации тех или иных конкретных выводов в приложениях интеллектуальных компьютерных систем.

[22] См. **Раздел 1.4 Главы 1**.

[23] Упорядочив их в соответствии с естественной хронологией.

3.2.3. Общие свойства ДСМ-правил правдоподобного вывода.

Переходя к описанию свойственных ДСМ-методу механизмов вывода (как правдоподобного так и достоверного) дадим краткую характеристику логической базы реализуемых в нем рассуждений. Речь пойдет о квази-аксиоматических теориях рассматриваемого нами конкретного вида, мы начнем с детального представления правил правдоподобного вывода (или - в сокращенном варианте - ППВ) ДСМ-метода. Прежде всего нас будут интересовать их общие (для работы с различными структурами данных) свойства, а именно:

- условия применимости;
- рациональность;
- инвариантность так называемого "ядра" ДСМ-рассуждения и "*настраиваемость*[1]" его "окружения" при варьировании типов (структур) данных, обрабатываемых в рамках ДСМ-рассуждения; и, наконец,
- критерий достаточности оснований для принятия результатов ДСМ-рассуждения.

Условия применимости ДСМ-метода позволяют выделить те ситуации, в которых имеет смысл использовать описываемый нами формализм, и состоят в следующем:

1. Исследуются случаи *неполноты информации*, т.е. ситуации, когда имеются достоверные сведения лишь о частичном описании анализируемой ситуации, а оставшаяся часть требует доопределения (теми или иными средствами прогнозирования). Другими словами, речь идет о ситуациях, когда полное и корректное описание объекта исследования аксиоматическими средствами не является возможным[2]. (Часто такие предметные области называют *открытыми*[3], подразумевая при этом возможность пополнения их текущего описания новыми фактами или зависимостями).
2. Данные, которыми представлен объект исследования, хорошо *структурированы* (т.е. имеется возможность провести алгебраические построения, необходимые для корректного определения и анализа сходства в таких данных - см. выше **Раздел 3.1**), а знания *плохо формализованы* (т.е. представляющие их зависимости носят лишь частичный[4] характер).
3. В исследуемой предметной области существуют факторы, *причинным* влиянием которых (в детерминистском смысле) обусловлено как наличие, так и отсутствие анализируемых эффектов. (Другими словами, существуют *причинные зависимости*, детерминирующие проявление и\или отсутствие изучаемых явлений).
4. В текущем описании анализируемой предметной области присутствуют как *примеры* возникновения[5], так и примеры невозникновения[6] изучаемых явлений (т.е. имеется как информация для порождения интересующих нас зависимостей, так и сведения, которые могли бы быть использованы в процессе проверки *фальсифицируемости* порождаемых результатов).

Рациональность правил правдоподобного вывода ДСМ-метода будет вытекать из следующих особенностей схемы их построения:

- мы будем порождать результаты "параллельно" как на множестве примеров, так и на множестве контрпримеров, рассматривая каждое из них как исходный "материал" при

---

[1] На конкретный тип данных.

[2] Например, в ситуациях, когда исследуемая предметная область описывается открытым (т.е., возможно, пополняемым с поступлением новой информации о ее свойствах) множеством аксиом, и также открытым (т.е., также, возможно, пополняемым) множеством фактов, не являющихся логическим следствием этих аксиом.

[3] В противоположность *замкнутым* предметным областям, допускающим описание в том числе и средствами *аксиоматических теорий*.

[4] Т.е. дающий лишь неполное описание анализируемой предметной области.

[5] Мы будем называть их также (+)-*примерами*

[6] Мы будем называть их также (-)-*примерами* или *контрпримерами*.

порождении фальсификаторов для результатов, получаемых на основании данных из другого;

- при этом принятие той или иной причинной зависимости (порождаемой на множестве примеров) возможно тогда и только тогда, когда ту же самую зависимость невозможно породить на множестве контрпримеров, и наоборот.

Другими словами, правила правдоподобного вывода ДСМ-метода будут устроены так, что их заключения могут быть приняты тогда и только тогда, когда они не фальсифицируются порождаемыми из примеров "противоположного знака" результатами.

Идея *инвариантности* ядра ДСМ-рассуждения и *настраиваемости* его окружения при варьировании структур обрабатываемых данных в самой общей форме сводима к следующему:

- при работе с различными типами данных основные[7] элементы формальной конструкции ДСМ-рассуждения и задействованных в нем правил правдоподобного вывода - учет количества примеров и контрпримеров, служащих основанием для построения правдоподобных выводов; проверка условия исчерпываемости при анализе локального и глобального сходств или же в организации ДСМ-вывода по аналогии; учет дополнительных условий (например, так называемого запрета на контрпримеры и т.п.) при порождении гипотез о причинах возникающих явлений и др. - имеют один и тот же вид для всех обрабатываемых средствами ДСМ-рассуждения типов данных. Эту часть конструкции ДСМ-рассуждения мы также будем называть "*ядром*";
- другую часть формальной конструкции ДСМ-рассуждения и задействованных в нем правил правдоподобного вывода - собственно вычисление сходства примеров и контрпримеров; анализ вложимости в конкретные анализируемые объекты тех или иных конкретных фрагментов (подобъектов) и т.п. - мы будем для каждого конкретного типа данных определять специфическим образом (учитывая те требования, которые предъявляет здесь формальная алгебраическая природа сходства, уже рассмотренная нами подробно выше в **Разделе 3.1**). Мы будем также называть эту - специфицируемую (другими словами - настраиваемую на анализируемый тип данных) в каждом конкретном случае[8] - часть формальной конструкции ДСМ-рассуждения "*окружением*" ядра.

Критерий *достаточности оснований* для принятия результатов ДСМ-рассуждения - основной инструмент оценки степени "доверия" к получаемым в процессе рассуждения заключениям. В своем самом общем виде он учитывает три основополагающих обстоятельства:

- каждый из элементов исходной выборки примеров и контрпримеров должен быть "*объясним*[9]" порождаемыми из нее каузальными зависимостями. Это требование мы будем называть условием *каузальной полноты*;
- каждый из элементов исходной выборки примеров и контрпримеров должен быть *непротиворечивым* образом "объясним" порождаемыми из нее каузальными зависимостями. (При этом непротиворечивость объяснения понимается следующим образом: примеры должны быть объяснимы лишь позитивными[10] причинными зависимостями, а контрпримеры - лишь негативными). Это требование мы будем называть условием *каузальной непротиворечивости*;

---

[7] Т.е. отражающие ключевые моменты представленной нами ранее *эвристики Бэкона-Милля* - см. ранее **Разделы 2.2** и **2.3.**

[8] Т.е. при работе с каждым конкретным типом данных.

[9] Т.е. каждый из элементов исходной выборки примеров и контрпримеров должен быть порождаем (ДСМ-выводим) средствами ДСМ-рассуждения, стартующего именно с этой выборки.

[10] Здесь мы следуем терминологии, учитывающей, что примеры - это описания ситуаций, когда исследуемое явление имело место, а контрпримеры - описания ситуаций, когда это было не так. В свою очередь, позитивные причинные зависимости характеризуют причины наличия исследуемого явления, а негативные - причины его отсутствия.

- порождаемые каузальные зависимости должны быть логически совместимы с уже имеющейся базой знаний[11]: они не должны противоречить уже (возможно) имеющейся в базе знаний априорной информации.

В случае, когда последнее из названных условий не выполнено, необходим специальный (неформальный анализ) оснований возникновения такой ситуации. Это прежде всего - анализ содержимого базы знаний и\или анализ множества фактов (примеров и контпримеров), послуживших основанием для правдоподобного вывода. В частности, возможно следовать тактике, когда имеющееся множество фактов (примеров и контрпримеров) пересматривается либо в сторону уменьшения (например, когда некоторые из исходных фактов удаляются из обучающей выборки как “не заслуживающие доверия”), либо в сторону расширения (когда обучающая выборка пополняется новыми фактами, релевантными цели рассуждения), и процесс порождения зависимостей реализуется вновь[12].

В ситуациях, когда условия каузальной полноты и\или непротиворечивости оказываются невыполненными[13], мы либо

- будем считать, что ДСМ-вывод не может быть сделан на достаточных основаниях (т.е. степень доверия к результатам, полученным в ходе ДСМ-рассуждения, ниже, чем в случае *достаточности оснований*); либо
- будем пополнять исходную выборку обучающих примеров и контрпримеров новыми элементами[14], и вновь порождать причинные зависимости (уже на расширенном множестве прецедентов для обучения). Двигаясь по этому пути, мы либо придем к ситуации *достаточности оснований* для принятия результатов ДСМ-рассуждения, либо же остановимся в стадии отсутствия таковых (уже охарактеризованной нами выше).

Множество $\mathfrak{R}_{\mathbf{P}}$ правил правдоподобного вывода ДСМ-метода образовано элементами вида

$$\frac{\tau(\varphi),\ \ \mathbf{DP}(\varphi)}{\nu(\varphi)},$$

где

$\tau(\varphi)$ говорит о том, что о выполнимости φ (например, о том, обладает ли объект *O* свойством *P*) нет информации;

DP(φ) – это специальное условие (логическая комбинация решающих предикатов[15]), формализующее возможность найти подходящую эмпирическую зависимость и распространить ее на φ ("доопределить" φ); наконец,

$\nu(\varphi)$ отражает результат "доопределения" φ в ходе ДСМ-рассуждения.

---

[11] Напомним (см. общее описание структуры квази-аксиоматических теорий выше в **Разделе 3.2.1**), что текущее состояние анализируемой предметной области - ПО - характеризуется как множеством элементарных фактов **Σ‘** (представленных в базе фактов), так и множеством аксиом **Σ**, представленных в базе знаний (т.е. множеством локальных зависимостей, имеющих место в текущем состоянии анализируемой ПО.

[12] Решение о прекращении итераций представленного вида, разумеется, может быть принято лишь “внешними” средствами (например, в ситуации “насыщения” обучающей выборки при ее пополнении новыми фактами, когда никаких “новых” зависимостей уже не удается породить в ходе ДСМ-рассуждения).

[13] Например, в случае, когда часть примеров и\или контрпримеров исходной обучающей выборки остались “необъясненными” в смысле критерия каузальной полноты.

[14] Так в случае, представленном в предыдущей сноске, это должны быть в первую очередь элементы, сходные с необъясненными элементами исходной обучающей выборки.

[15] Используемая система ППВ представляет собой вариант формализации известных индуктивных схем Д.С.Милля (осуществляемый средствами специальных многозначных логик). Термин "решающий предикат" как раз и использован для того, чтобы подчеркнуть связь конкретных ППВ с соответствующими "методами" Д.С.Милля.

Множество $\mathfrak{R}$ правил правдоподобного вывода ДСМ-метода состоит из двух (разнотипных) подмножеств:

$$\mathfrak{R} = \mathfrak{R}_{\mathbf{I}} \cup \mathfrak{R}_{\mathbf{II}} ,$$

Таким образом, мы используем п.п.в. двух типов:

- правила порождения зависимостей (о причинах - т.е. о "носителях"- изучаемых свойств), собранные в множество $\mathfrak{R}_{\mathbf{I}}$, и •правила переноса найденных зависимостей (там, где это возможно осуществить корректными средствами) на исходно недоопределенные случаи (описания новых объектов), собранные в множество $\mathfrak{R}_{\mathbf{II}}$.

С другой стороны, каждое из множеств $\mathfrak{R}_{\mathbf{I}}$ и $\mathfrak{R}_{\mathbf{II}}$ содержит два подобным образом организованных подмножества:

$$\mathfrak{R}_{\mathbf{I}} = \mathfrak{R}^{(+)}_{\mathbf{I}} \cup \mathfrak{R}^{(-)}_{\mathbf{I}} \quad \text{и}$$

$$\mathfrak{R}_{\mathbf{II}} = \mathfrak{R}^{(+)}_{\mathbf{II}} \cup \mathfrak{R}^{(-)}_{\mathbf{II}},$$

где

(+)-подмножество использует для порождения зависимостей в первую очередь "позитивную" информацию (в нашей терминологии - примеры, т.е. описания прецедентов, когда исследуемое явление действительно имело место), а

(-)-подмножество использует для порождения зависимостей в первую очередь "негативную" информацию (т.е., в нашей терминологии - контрпримеры).

При этом оборот "*в первую очередь*" означает, что хотя стартовой информацией для применения конкретного правила правдоподобного вывода ДСМ-метода являются описания прецедентов одного знака, в процессе порождения результата (в рамках механизма проверки фальсифицируемости порождаемых выводов) задействованы также описания прецедентов противоположного знака.

Фактически, в представляемой схеме ДСМ-рассуждения организовано одновременное порождение эмпирических зависимостей обоих знаков ("позитивных[16]" и "негативных[17]"). Таким образом, вместе с порождением конкретной зависимости организовано порождение всех возможных (в рамках имеющихся базы фактов и базы знаний) ее фальсификаторов. Этот механизм фальсификации обеспечивает принятие результата лишь в ситуации, когда у конкретной гипотезы либо вообще нет фальсификаторов, либо же они порождены более слабыми (в смысле упорядочения по "силе[18]") средствами ДСМ-рассуждения.

С формальной точки зрения возможные варианты "взаимоотношений" гипотезы и ее (возможных фальсификаторов) отражают четыре конкретных варианта правил правдоподобного вывода рассматриваемых нами типов. Разбиение на варианты обусловлено типом истинностных значений порождаемых результатов:

- в ситуации, когда "аргументы" **за** принятие гипотезы (в случае порождения как позитивных, так и негативных зависимостей) "сильнее" "аргументов" **против** (т.е. когда фальсификаторы порождены более "слабыми" средствами ДСМ-рассуждения, либо вообще отсутствуют), порождение окончательного результата осуществляется конкретным правилом правдоподобного вывода, приписывающим заключению истинностное значение соответствующего знака (типа *фактическая истина* или *фактическая ложь*);
- в ситуации, когда "аргументы" **за** и **против** "равносильны", порождение конкретного результата осуществляется правилом, приписывающим заключению истинностное значение *фактическое* (или *эмпирическое*) *противоречие*; и, наконец,

[16] Т.е. зависимостей о "причинах" изучаемых явлений.
[17] Т.е. зависимостей об "антипричинах" изучаемых явлений.
[18] Более подробно возможности такого упорядочения мы обсудим чуть позже.

- в ситуации, когда нет ни “аргументов” **за**, ни “аргументов” **против**, конечный результат порождается правилом, приписывающим заключению истинностное значение типа “*фактическая недоопределенность*”.

Упорядочение по “силе” используемых в рамках ДСМ-рассуждения правил правдоподобного вывода осуществляется на основе сравнения логической силы использованных в них решающих предикатов. Так если для решающих предикатов $\mathbf{M_1}$ и $\mathbf{M_2}$ имеет место импликация $\mathbf{M_1} \rightarrow \mathbf{M_2}$, то считается, что решающий предикат $\mathbf{M_1}$ имеет большую силу чем решающий предикат $\mathbf{M_2}$. В противном случае, когда дополнительно также неверно, что $\mathbf{M_2} \rightarrow \mathbf{M_1}$, они считаются несравнимыми. Таким образом удается породить частичный порядок на подмножествах решающих предикатов каждого знака и каждого типа. Соотношение по “силе” правил правдоподобного вывода в ДСМ-рассуждении полностью определяется соотнесением по “силе” задействованных в них решающих предикатов. Дополнительным фактором, учитываемым при сравнении ”аргументов” **за** и **против**, является учет (в явном виде) номера шага ДСМ-рассуждения, на котором получено соответствующее выражение: полученным на соответствующем шаге ДСМ-рассуждения выражениям приписывается истинностная оценка вида

$$<\alpha,n> \text{, где } \alpha \in \{1,-1,0,\tau\},$$

и это, в свою очередь, означает, что каждое выражение с истинностной оценкой вида $<\alpha,n>$ имеет большую “силу” по отношению ко всем выражениям, имеющим истинностную оценку вида $<\alpha,m>$ при $n<m$.

3.3. Решающие предикаты, правила правдоподобного вывода и стратегии ДСМ-рассуждения.

Собственно «инструменты» анализа данных в ДСМ-методе автоматического порождения зависимостей формируются по следующей схеме:

- базовым компонентом каждого ДСМ-правила правдоподобного вывода являются соответствующие *решающие предикаты* (формальные уточнения эвристических процедур, базирующихся на идеях Бэкона -Милля[1]);
- каждое из *правил правдоподобного вывода* использует соответствующие комбинации решающих предикатов противоположных знаков (+ и -), что позволяет соблюсти условия достаточности оснований[2] для принятия результатов ДСМ-рассуждения;
- те или иные последовательности применения правил правдоподобного вывода характеризуют *стратегии* ДСМ-рассуждений.

3.3.1 Правила правдоподобного вывода ДСМ-метода: работа с данными, представленными в виде множеств.

Явный вид правил правдоподобного вывода - ППВ - мы, прежде всего, рассмотрим на конкретном примере их уточнения и реализации: при обработке данных, представленных множествами[3]. Описание ППВ мы начнем с анализа *правил первого рода* - правил порождения зависимостей о причинах изучаемых явлений. Определяющим (как здесь, так и при описании *правил второго рода*, используемых для переноса найденных зависимостей на новые, изначально недоопределенные, случаи) будет представление структуры задействованных в ППВ решающих предикатов. Строение же последних мы подробно разберем на примере предиката $\boldsymbol{M^{+}_{a}}$ - простого положительного сходства[4].

Итак, исходным материалом для применения ППВ (как первого, так и второго типа) являются:

- множество $\boldsymbol{O}$ объектов O_1, O_2, ..., и множество $\boldsymbol{P}$ свойств P_1, P_2,..., которыми обладают (или не обладают) объекты из множества $\boldsymbol{O}$;
- каждый из объектов в множестве $\boldsymbol{O}$ - непустое подмножество некоторого универсума $\mathbf{U^{(1)}}$ - множества “образующих” $\mathbf{U^{(1)}} = \{o_1,o_2,...,o_n\}$, а каждое из множеств свойств в множестве $\boldsymbol{P}$ – есть непустое подмножество некоторого универсума свойств $\mathbf{U^{(2)}} = \{p_1,p_2,...,p_m\}$;
- пара частично определенных бинарных отношений $==>_1$ и $==>_2$, первое из которых используется для того, чтобы характеризовать *наличие* (или *отсутствие*) *свойств* у объектов; второе используется для характеризации подобъектов, представляющих *причины наличия* (или, соответственно, *отсутствия*) *свойств* у исследуемых объектов;

---

[1] А также отражающих «логику здравого смысла», характеризующую формирование подходящих классов эквивалентности в процессе восстановления и проверки переносимости (скрытых) зависимостей с группы уже изученных объектов на новый (предложенный для анализа) объект – см. **Раздел 2.1**.

[2] В частности – *неоспариваемости* результатов (зависимостей), порожденных (на конкретных исходных данных) правилами одного знака, результатами (зависимостями), которые могут быть порождены (на тех же исходных данных) с помощью правил правдоподобного вывода противоположного знака.

[3] Работы по ДСМ-методу развивались так, что именно этот тип данных был первым, рассмотренным как в теоретических построениях, так и в конкретных приложениях - см., например, работы [*Финн 83, Забежайло 82, Забежайло 83*]и др..

[4] Здесь термин *простого* отражает тот факт, что рассматривается *бинарное* отношение причинности. Термин *положительного* означает, что речь пойдет о поиске “позитивных” зависимостей, порождаемых из примеров - описаний прецедентов, когда анализируемое явление *имело место*. Термин *сходства*, наконец, отражает то обстоятельство, что рассматриваемый нами решающий предикат строится для формализации *миллевского метода сходства* - см. **Раздел 2.2**.

- три заранее заданных множества объектов (описаний прецедентов) ***Exm*$^{+}$, *Exm*$^{-}$** и ***Exm*$^{\tau}$**, представляющих, соответственно,

  - примеры объектов, обладающих интересующими нас свойствами, т.е. описываемые выражениями вида[5]

  $$\mathbf{J}_{<1,0>}(O_i ==>_1 P_i)^{[6]};$$

  - примеры объектов, не обладающих интересующими нас свойствами, т.е. описываемые выражениями вида

  $$\mathbf{J}_{<-1,0>}(O_i ==>_1 P_i);$$

  и, наконец, объекты, наличие или отсутствие заданных свойств у которых и предполагается установить в процессе ДСМ-рассуждения, т.е. те, что представляются выражениями вида

  $$\mathbf{J}_{<\tau,0>}(O_i ==>_1 P_i).$$

Результатом применения ППВ типа **I** является заключение (утверждение) о том, что некоторый подобъект V (т.е. множество образующих из универсума $\mathbf{U}^{(1)}$) является причиной присутствия (в случае обучения на положительных примерах) или причиной отсутствия (в случае обучения на отрицательных примерах) некоторого множества свойств W (представляющего собой непустое подмножество универсума $\mathbf{U}^{(2)}$) у объектов из множества ***O*** Помимо этого возможны ситуации, когда сделать такое заключение непротиворечивым образом не представляется возможным. Также возможны ситуации, когда для порождения заключения о причинной связи V и W просто нет оснований (случай наследования недоопределенности на следующий шаг ДСМ-рассуждения).

Синтаксически эти результаты (т.е. четыре возможных исхода при порождении заключения средствами ППВ типа **I**) выражаются с помощью частично определенного отношения $==>_2$: в исходном состоянии, когда нет информации о причинной связи V и W, имеет место $\mathbf{J}_{<\tau,0>}(V ==>_2 W)$. После применения ППВ реализуется одна из четырех ситуаций[7]:

а) $\mathbf{J}_{<1,r>}(V ==>_2 W)$, в случае, когда установлена связь (положительная эмпирическая зависимость) между V и W (т.е. когда у нас появились основания[8] считать подобъект V причиной наличия у объектов из множества ***O*** множества свойств W);

б) $\mathbf{J}_{<-1,r>}(V ==>_2 W)$, когда установлено ее отсутствие (т.е. отрицательная ЭЗ, характеризующая множество V как причину отсутствия у объектов из множества ***O*** множества свойств W);

в) $\mathbf{J}_{<0,r>}(V ==>_2 W)$, когда есть доводы как в пользу положительной, так и в пользу отрицательной связи V и W, причем нет возможности предпочесть одну другой - случай "эмпирического противоречия";

г) $\mathbf{J}_{<\tau,r>}(V ==>_2 W)$, когда ни один из случаев а)-в) не реализуется, т.е. сохраняется наследуемое состояние недоопределенности на следующий шаг рассуждения.

Предикат простого положительного сходства $\boldsymbol{M}^{+}_{a}$ представляет в рассматриваемой формализации *миллевский метод сходства* (см. выше **Раздел 2.2** предыдущей **Главы 2**[9]), примененный к положительным примерам - описаниям прецедентов наличия у изучаемых

---

[5] Более точно: истинностные значения вида $<\alpha,0>$ имеют место лишь на первом шаге ДСМ-рассуждения. Для всех последующих шагов здесь следует использовать выражения вида $<\alpha,r\text{-}1>$,где r - номер текущего шага ДСМ-рассуждения.

[6] Определение функции $\mathbf{J}_{\nu}(\varphi)$ мы уже дали ранее в **Разделе 3.2.**

[7] Здесь **r** есть номер шага ДСМ-рассуждений, на котором получено соответствующее заключение.

[8] Формализованные в рамках ДСМ-рассуждения.

[9] А также обсуждение процедур выделения классов эквивалентности из классов сходства в **Разделе 2.1.**

объектов избранных для анализа свойств. Ключевые элементы миллевской конструкции, нашедшие отражение в ДСМ-формализации:

- рассматривается множество (попарно-различных!) примеров, представляющих наличие анализируемых свойств у изучаемых объектов;
- всякий раз, когда в структуре анализируемых объектов встречается некоторый (непустой) фрагмент, в свойствах этих же объектов содержится некоторое непустое общее подмножество;
- в множество примеров, отобранных для анализа сходства, включены все прецеденты, релевантные цели проводимого причинного анализа (т.е. в обучающей выборке учтены все объекты, свойства которых содержат обозначенное выше общее подмножество);
- основанием для заключения о причинах появления обсуждаемых свойств может быть анализ не менее двух (различных-!) примеров.

Формальным представлением этой схемы средствами системы логических языков $\boldsymbol{L^0_i} \subset \boldsymbol{L_i} \subset \boldsymbol{L_e} \subset \boldsymbol{L'_e} \subset \boldsymbol{L''_e}$ (см. выше **Раздел 3.2.1**) является выражение

$$\boldsymbol{M^+_a}(V,W) = \exists k \; \underline{\boldsymbol{M^+_a}}(V,W,k) =$$

$$= \exists k \; \exists O_{i1} ... \exists O_{ik} \; \exists P_{i1} ... \exists \; P_{ik} \; \mathbf{(} \; [\; \mathbf{J}_{<1,r>}(O_{i1} ==>_1 P_{i1}) \; \&$$

$$\& \; \forall P \; (\; \mathbf{J}_{<1,r>}(O_{i1} ==>_1 P) \rightarrow (P \subseteq P_{i1})) \;] \; \& \; ...$$

$$... \; \& \; [\; \mathbf{J}_{<1,r>}(O_{ik} ==>_1 P_{ik}) \; \&$$

$$\& \; \forall P \; (\; \mathbf{J}_{<1,r>}(O_{ik} ==>_1 P) \rightarrow (P \subseteq P_{ik})) \;] \; \&$$

$$\& \; [\; (\bigcap_{i=1}^{k} O_i = V) \; \& \; (V \neq \varnothing) \;] \; \&$$

$$\& \; [\; \forall i \forall j \; (\; [(i \neq j) \& (1 \leq i,j \leq k)] \rightarrow (O_i \neq O_j) \;) \;] \; \&$$

$$\& \; \forall X \; \forall Y \; \mathbf{[} \; [\mathbf{J}_{<1,r>}(X ==>_1 Y) \; \&$$

$$\& \; \forall A \; (\; \mathbf{J}_{<1,r>}(X ==>_1 P) \rightarrow (P \subseteq Y) \;) \; \& \; (V \subset X)] \rightarrow$$

$$\rightarrow [\; (W \neq \varnothing) \; \& \; (W \subseteq Y) \; \& \; ((X = O_1) \vee ... \vee (X = O_k)) \;] \; \mathbf{]} \; \&$$

$$\& \; (k \geq 2) \; \mathbf{)} \; ,$$

где соответствующие подформулы в представлении $\boldsymbol{M^+_a}$ есть в точности совместно выполняющиеся условия (***cn1.-cn7.***), такие что

cn1.) из множества положительных примеров отобраны ровно k штук

$$\boldsymbol{Exm^+}(V,W) = \; = \{ \; \mathbf{J}_{<1,r>}(O_{i1} ==>_1 P_{i1}) \; ,...$$

$$..., \; \mathbf{J}_{<1,r>}(O_{ik} ==>_1 P_{ik}) \; \};$$

cn2.) все примеры в $\boldsymbol{Exm^+}(V,W)$ попарно различны (условие невырожденности сходства);

cn3.) множество V выражает сходство всех примеров из множества $\boldsymbol{Exm^+}(V,W)$:

$$V = \bigcap_{i=1}^{k} \{O_i\};$$

cn4.) сходство примеров из $\boldsymbol{Exm^+}(V,W)$ не тривиально: $V \neq \varnothing$;

cn5.) множество V порождено максимальным (по вложению) множеством объектов в $\boldsymbol{Exm^+}(V,W)$, т.е. множество V есть действительно “носитель” глобального сходства

объектов из множества ***O*** (и все порождающие V объекты учтены в множестве $Exm^{+}(V,W)$ - оно нерасширяемо);

cn6.) W есть непустая общая часть (сходство) множеств свойств P_i (т.е.

$$W = \bigcap_{i=1}^{k} P_{ji} = P \neq \varnothing$$

в рассматриваемом нами простейшем случае);

cn7.) $k \geq 2$, т.е. порождаемая причинная зависимость выражает сходство различных (по крайней мере двух) примеров.

Подформула

$$\forall X\ \forall Y\ \mathbf{[}\ [\mathbf{J}_{<1,r>}(X ==>_1 Y)\ \&$$

$$\&\ \forall P\ (\ \mathbf{J}_{<1,r>}(X ==>_1 P) \rightarrow (P \subseteq Y)\ )\ \&\ (V \subset X)] \rightarrow$$

$$\rightarrow [\ (W \neq \varnothing)\ \&\ (W \subseteq Y)\ \&\ ((X = O_1) \vee \ldots \vee (X = O_k))\ ]\ \mathbf{]}$$

дает явный вид *эмпирической зависимости*, которая и выражает восстанавливаемое (другими словами - доопределяемое) предлагаемыми средствами ДСМ-рассуждения *причинно-следственное* отношение $V ==>_2 W$. (Напомним, что поиск неявным образом представленных в исходных данных эмпирических зависимостей, а также распространение средствами ДСМ-рассуждения их там, где это возможно на новые, исходно недоопределенные случаи, собственно, и является центральным моментом обсуждаемой нами формальной конструкции).

Аналогичным образом (хотя и с набольшими отличиями[10] - см. подробнее работы [*Финн 83, Забежайло 82, Забежайло 83, Финн 88, Финн 91, Финн 10*] и др.) строятся версии предиката отрицательного сходства, используемые при порождении зависимостей о контрпричинах (т.е. обстоятельствах, препятствующих наличию) исследуемого свойства.

Пополняя при порождении позитивных зависимостей представленное семейство {***cn1***.,...,***cn7***.} дополнительными условиями, такими, например, как:

cn8.1) неверно, что в множестве отрицательных примеров Exm^{-} или же в множестве недоопределенных примеров Exm^{τ} найдется хотя бы один пример

$$(\mathbf{J}_{<-1,r>}(O_0 ==>_1 P_0)),$$

сходный (в смысле вложимости V в O_0) с порождающими носитель V анализируемого глобального сходства примерами из $Exm^{+}(V,W)$ так, что это сходство выражает то же самое множество V. Формально это выражается условием:

$$\forall X \forall Y [\mathbf{(}(V \subset X)\ \&\ (W \subseteq Y)\mathbf{)} \rightarrow$$

$$\rightarrow \mathbf{(}\ \mathbf{J}_{<1,r>}(X ==>_1 Y) \vee \mathbf{J}_{<\tau,r>}(X ==>_1 Y)\ \mathbf{)}]$$

(которое мы также будем называть ***запретом на контрпримеры***),

cn8.2) неверно, что в множестве отрицательных примеров Exm^{-} найдется хотя бы один пример $(\mathbf{J}_{<-1,r>}(O_0 ==>_1 P_0))$, сходный (в смысле вложимости V в O_0) с

[10] Помимо воспроизведения условий ***cn1***.-***cn7***. для работы с отрицательными примерами можно предложить более тонкие формулировки предиката отрицательного сходства M^{-}_{a}, учитывающие, например, то обстоятельство, что может существовать несколько независимых способов "разрушения" механизма, обуславливающего "проявление" искомого свойства, или же, например, могут вообще отсутствовать причинные механизмы его проявления.

порождающими носитель V анализируемого глобального сходства примерами из $\boldsymbol{Exm}^{+}$(V,W) так, что это сходство выражает то же самое множество V. Формально это выражается условием:

$$\forall X \forall Y[\mathbf{(}(V \subset X) \,\&\, (W \subseteq Y)\mathbf{)} \rightarrow$$
$$\rightarrow \mathbf{(}\ \mathbf{J}_{<1,r>}(X ==>_1 Y) \vee$$
$$\vee\ \mathbf{J}_{<0,r>}(X ==>_1 Y) \vee$$
$$\vee\ \mathbf{J}_{<\tau,r>}(X ==>_1 Y)\ \mathbf{)}]$$

(которое мы также будем называть ***ослабленным запретом на контрпримеры***),

сп8.3) невхождение V - “причины” множества свойств W - в структуру объекта ведет к утрате им рассматриваемого множества свойств, и подтверждение этому можно найти хотя бы на одном объекте из множества отрицательных примеров $\boldsymbol{Exm}^{-}$. Формально это выражается условием:

$$\exists X \exists Y[\ (V \not\subset X) \,\&\, \mathbf{J}_{<-1,r>}(X ==>_1 Y) \,\&\, (W \subseteq Y)\ ]$$

(которое мы также будем называть условием ***существенности причины***),

сп8.4) невхождение V - “причины” множества свойств W - в структуру объекта ведет к утрате им рассматриваемого множества свойств, и подтверждение этому можно найти хотя бы на одном объекте в множестве отрицательных примеров $\boldsymbol{Exm}^{-}$, либо в множестве недоопределенных примеров $\boldsymbol{Exm}^{\tau}$. Формально это выражается условием:

$$\exists X \exists Y[(V \not\subset X) \,\&\, (\mathbf{J}_{<-1,r>}(X ==>_1 Y) \vee \mathbf{J}_{<\tau,r>}(X ==>_1 Y)) \,\&\, (W \subseteq Y)]$$

(которое мы также будем называть ***ослабленным*** условием ***существенности причины***),

сп8.5) удаление V - “причины” множества свойств W - из структуры объекта непременно ведет к утрате им рассматриваемого множества свойств, т.е. соответствующие объекты могут быть найдены лишь в множестве отрицательных примеров $\boldsymbol{Exm}^{-}$, либо в множестве недоопределенных примеров $\boldsymbol{Exm}^{\tau}$. Формально это выражается условием:

$$\forall X \forall Y[\mathbf{(}(V \not\subset X) \,\&\, (W \subseteq Y)\mathbf{)} \rightarrow$$
$$\rightarrow \mathbf{(}\ \mathbf{J}_{<-1,r>}(X ==>_1 Y) \vee \mathbf{J}_{<\tau,r>}(X ==>_1 Y)\ \mathbf{)}]$$

(которое мы также будем называть условием ***единственности причины***),

сп8.6) удаление V - “причины” множества свойств W из структуры объекта непременно ведет к утрате им рассматриваемого множества свойств. Формально это выражается условием:

$$\forall X\ \forall Y\ \exists O\ \exists V_0\ \exists W_0\ [\ \mathbf{(}\ \mathbf{J}_{<1,r>}(X ==>_1 Y) \,\&\, (V \subset X) \,\&\, (W \subseteq Y)\mathbf{)} \rightarrow$$
$$\rightarrow \mathbf{(}\ C = ((X \setminus V) \cup V_0) \,\&$$
$$\&\ (V \not\subset V_0) \,\&\, (V_0 \neq \varnothing) \,\&$$
$$\&\ (\mathbf{J}_{<-1,r>}(O ==>_1 W \cup W_0) \vee \mathbf{J}_{<\tau,r>}(O ==>_1 W \cup W_0))\ \mathbf{)}\ ]$$

(которое представляет миллевский ***метод различий***),

и другими[11] (проводя также, если это необходимо, для порождения негативных зависимостей аналогичные построения), можно сформировать частично упорядоченные (в смысле логического следования) множества $\mathbf{M}^+$ и $\mathbf{M}^-$ решающих предикатов обоих знаков.

Рассматривая логические зависимости на представленном множестве дополнительных условий {***cn8.1***,...,***cn8.6***}, нетрудно убедиться, что (при порождении как позитивных, так и негативных зависимостей) справедлива

3.3.1.1. Лемма

Для представляющих условия ***cn8.1***,...,***cn8.6*** формул имеют место следующие соотношения:

1) ***cn8.1*** $\rightarrow$ ***cn8.2***
2) ***cn8.3*** $\rightarrow$ ***cn8.4***
3) $\neg$[(***cn8.1*** & ***cn8.3***) $\rightarrow$ ***cn8.5***]
4) $\neg$[(***cn8.1*** & ***cn8.3***) $\rightarrow$ ***cn8.6***]

Доказательство.
1-2). Первые два соотношения очевидным образом следуют из явного вида условий ***cn8.1*** - ***cn8.4***.

3). Для опровержения импликации [***cn8.1*** & ***cn8.3***) $\rightarrow$ ***cn8.5***] обратимся к конкретному примеру следующего вида:

$\boldsymbol{Exm}^+ = \{ AB ==>_1 ab,\ AC ==>_1 ac,\ EF ==>_1 abc \}$

$\boldsymbol{Exm}^- = \{ BC ==>_1 abc \}$

Легко видеть, что

$A \not\subset EF$

$a \subseteq abc$

и, тем не менее, $EF ==>_1 abc$ лежит в $\boldsymbol{Exm}^+$, а не в $\boldsymbol{Exm}^-$ или в $\boldsymbol{Exm}^{\tau}$.

4). Для опровержения импликации [***cn8.1*** & ***cn8.3***) $\rightarrow$ ***cn8.6***] обратимся к конкретному примеру следующего вида:

$\boldsymbol{Exm}^+ = \{ AB ==>_1 ab,\ AC ==>_1 ac,\ BE ==>_1 ae \}$

$\boldsymbol{Exm}^- = \varnothing$

Легко видеть, что для $BE = C = (AB \backslash A) \cup E$, $W = a$, $W_0 = e$ и $V_0 = E$, имеем: $C ==>_1 ac$ лежит в $\boldsymbol{Exm}^+$.

o

Располагая представленными **Леммой 3.3.1.1** соотношениями можно естественным образом (т.е. на основании логического следования соответствующих множеств условий) определить отношение частичного порядка на множестве всех подмножеств множества ***cn8.1***,...,***cn8.6***. Таким образом, каждое из множеств $\mathbf{M}^+$ и $\mathbf{M}^-$ представляет собой решетку, где роль наименьшего элемента выполняют, соответственно, предикаты простого сходства $\boldsymbol{M}^+_{\boldsymbol{a}}$ и $\boldsymbol{M}^-_{\boldsymbol{a}}$.

Упорядочение по “силе” используемых в ДСМ-рассуждении решающих предикатов позволяет также организовать упорядочение по “силе” порождаемых ими зависимостей: результатом использования более “сильного” предиката оказывается более “сильная” эмпирическая зависимость. Таким образом, появляется возможность более тонкого учета

[11] Например, условием единственности сходства (см. [*Финн 10*] и др.).

промежуточных результатов в процедуре ДСМ-фальсификации получаемых эмпирических зависимостей. (Детальное описание ДСМ-метода, учитывающего отношение частичного порядка на порождаемых в процессе ДСМ-рассуждения эмпирических зависимостях можно найти, например, в работе [*Финн 91*]).

Теперь нам осталось лишь привести формулировки явного вида правил правдоподобного вывода I-го типа. В соответствии с рассмотренным выше механизмом приписывания одного из четырех истинностных значений это будут выражения следующих четырех типов (здесь они представлены на примере использования решающих предикатов простого сходства обоих знаков):

$$\boldsymbol{I}^{+}_{\mathbf{a}} = \frac{\mathbf{J}_{<\tau,r>}(V ==>_2 W),\ \boldsymbol{M}^{+}_{a}(V,W)\ \&\ \neg\boldsymbol{M}^{-}_{a}(V,W)}{\mathbf{J}_{<1,r+1>}(V ==>_2 W),}$$

$$\boldsymbol{I}^{-}_{\mathbf{a}} = \frac{\mathbf{J}_{<\tau,r>}(V ==>_2 W),\ \neg\boldsymbol{M}^{+}_{a}(V,W)\ \&\ \boldsymbol{M}^{-}_{a}(V,W)}{\mathbf{J}_{<-1,r+1>}(V ==>_2 W),}$$

$$\boldsymbol{I}^{0}_{\mathbf{a}} = \frac{\mathbf{J}_{<\tau,r>}(V ==>_2 W),\ \boldsymbol{M}^{+}_{a}(V,W)\ \&\ \boldsymbol{M}^{-}_{a}(V,W)}{\mathbf{J}_{<0,r+1>}(V ==>_2 W),}$$

$$\boldsymbol{I}^{\tau}_{\mathbf{a}} = \frac{\mathbf{J}_{<\tau,r>}(V ==>_2 W),\ \neg\boldsymbol{M}^{+}_{a}(V,W)\ \&\ \neg\boldsymbol{M}^{-}_{a}(V,W)}{\mathbf{J}_{<\tau,r+1>}(V ==>_2 W),}$$

Использование при построении ППВ ДСМ-метода остальных решающих предикатов из множеств $\mathbf{M}^{+}$ и $\mathbf{M}^{-}$ осуществляется по аналогичному принципу.

Правила II-го типа - правила переноса найденных зависимостей на новые объекты (там, где это можно сделать средствами ДСМ-рассуждения) также формулируются при помощи специальных решающих предикатов.

Результатом применения ППВ типа **II** является заключение (утверждение) о том, что первоначально неизученный объект (т.е. один из тех, что до применения соответствующего ППВ описывался выражением $\mathbf{J}_{<\tau,r>}(O_j ==>_1 P_j)$) обладает (в случае переноса положительных ЭЗ) или не обладает (в случае переноса отрицательных ЭЗ) исследуемыми свойствами. Структуру решающих предикатов для этого случая поясним на примере предиката Π^{+}, реализующего принцип индуктивного обобщения [*Финн 83, Забежайло 82-83, Финн 88, Финн 91* и др.] положительных ЭЗ на описания заранее не изученных объектов.

<u>Принцип индуктивного обобщения</u>.

Наличие в доопределяемом объекте подобъектов, являющихся "причинами" проявления некоторого набора свойств, при отсутствии в нем подобъектов, являющихся "причинами" отсутствия этого же набора свойств (либо присутствие лишь более "слабых" в смысле некоторого упорядочения "причин" по "силе"), позволяет выдвинуть гипотезу о том, что анализируемый объект также обладает этим набором свойств[12].

Формальное представление приведенного принципа средствами ДСМ-рассуждения требует одновременной выполнимости следующих условий:

[12] Или же, по крайней мере утверждать, что с точностью до имеющегося множества фальсификаторов, ничто не препятствует тому, чтобы изучаемый объект обладал рассматриваемым набором свойств.

сп9.) среди причинных зависимостей, порожденных посредством ППВ типа **I** на множестве положительных примеров (т.е. среди положительных ЭЗ), имеется по крайней мере одна такая, что анализируемый объект O_j содержит в себе "носитель" соответствующего глобального сходства V;

сп10.) среди вкладывающихся в O_j "носителей" глобальных сходств V_i, представляющих соответствующие ЭЗ, имеются лишь те, которые получены обучением на положительных примерах;

сп11.) объединение W_i по всем таким сходствам V_i есть в точности P_j (в рассматриваемом нами простейшем случае $W_i = \bigcup_i W_i = W = P_j$).

Явный вид (+,-)-решающих предикатов в данном случае представляет выражение

$$\Pi_{\alpha}(O,P) = \exists k\, \Pi_{\alpha}(O,P,k) =$$

$$= \exists W_1 ... \exists W_k \, \mathbf{(} \, \exists V_1 [\, \mathbf{J}^{<\alpha,r+1>}(V_1 ==>_2 W_1) \, \& \, (V_1 \subset O) \,] \, \& ...$$

$$... \, \& \, \exists V_k [\, \mathbf{J}^{<\alpha,r+1>}(V_k ==>_2 W_k) \, \& \, (V_k \subset O) \,] \, \&$$

$$\& \, (\bigcup_i W_i = W = A_j) \, \&$$

$$\& \, \forall W \, (\exists V \, [\, (\mathbf{J}^{<\alpha,r+1>}(V ==>_2 W) \, \& \, (V \subset O) \,] \rightarrow$$

$$\rightarrow [\, (V = V_1) \vee ... \vee (V = V_k) \,]) \, \&$$

$$\& \, \forall P_0 \, [(\, (P_0 \subseteq P) \& (\, P_0 \neq \varnothing) \,) \rightarrow$$

$$\rightarrow \neg \exists V_0 \, (\mathbf{J}^{<-\alpha,r+1>}(V_0 ==>_2 P_0) \, \& \, (V_0 \subset O) \, ,$$

где везде в нем для переноса позитивных зависимостей $\alpha = 1$, а для переноса негативных зависимостей $\alpha = -1$.

В случае "конкуренции" причин противоположного знака соответствующий решающий предикат

$$\Pi_0(C,P) =$$

$$= \exists V_1 \exists W_1 \exists V_2 \exists W_2 \, \mathbf{(} \, \mathbf{J}^{<1,r+1>}(V_1 ==>_2 W_1) \, \&$$

$$\& \, \mathbf{J}^{<-1,r+1>}(V_2 ==>_2 W_2) \, \&$$

$$\& \, [(W_1 \cap W_2) \neq \varnothing] \, \&$$

$$\& \, (V_1 \subset O) \, \& \, (V_2 \subset O) \, \& \, (W_1 \subseteq P) \, \& \, (W_2 \subseteq P) \, \mathbf{)}$$

Наконец, состояние недоопределенности наследуется на следующий шаг ДСМ-рассуждения, если ни один из решающих предикатов $\Pi_1(O,P)$, $\Pi_{-1}(O,P)$ и $\Pi_0(O,P)$ не выполним. Таким образом:

$$\Pi_{\tau}(O,P) = \neg [\, \Pi_1(O,P) \vee \Pi_{-1}(O,P) \vee \Pi_0(O,P) \,]$$

Явный вид ППВ переноса зависимостей о наличии или отсутствии изучаемых свойств - ППВ типа **II** - представляют следующие четыре соотношения:

$$\boldsymbol{II}^{+} = \frac{\mathbf{J}^{<\tau,r>}(\, O ==>_1 A), \, \Pi_1(O,P)}{\mathbf{J}^{<1,r+1>}(O ==>_1 P)} \, ,$$

$$II^{-} = \frac{\mathbf{J}_{<\tau,r>}(O ==>_1 P),\ \Pi_{-1}(O,P)}{\mathbf{J}_{<-1,r+1>}(O ==>_1 P)},$$

$$II^{0} = \frac{\mathbf{J}_{<\tau,r>}(O ==>_1 P),\ \Pi_{0}(O,P)}{\mathbf{J}_{<0,r+1>}(O ==>_1 P)},$$

$$II^{\tau} = \frac{\mathbf{J}_{<\tau,r>}(O ==>_1 P),\ \Pi_{\tau}(O,P)}{\mathbf{J}_{<\tau,r+1>}(O ==>_1 P)}.$$

Процесс обработки средствами ДСМ-рассуждения конкретной обучающей выборки - множества примеров, множества контрпримеров и множества объектов, свойства которых требуют прогноза (доопределения), - удобно проследить, наблюдая за преобразованиями матриц ***Mtx₁*** и ***Mtx₁*** отношений, соответственно, $==>_1$ и $==>_2$.

Матрица ***Mtx₁*** бинарного отношения $==>_1$ (в предположении, что n - общее количество различных объектов, а m - общее количество различных множеств свойств в множествах $\boldsymbol{Exm^{+}}$, $\boldsymbol{Exm^{-}}$ и $\boldsymbol{Exm^{\tau}}$)

$==>_1$	P_1	...	P_j	...	P_m
O_1	β_{11}		β_{1j}		β_{1m}
⋮	⋮		⋮		⋮
O_i	β_{i1}	...	β_{ij}	...	β_{im}
⋮	⋮		⋮		⋮
O_n	β_{n1}	...	β_{nj}	...	β_{nm}

в исходном состоянии (т.е. на нулевом шаге ДСМ-рассуждения) организована следующим образом:

- для тех пар $<O_i,P_j>$, которые лежат в множестве $\boldsymbol{Exm^{+}}$, значение $\beta_{ij} = <1,0>$;
- для тех пар $<O_i,P_j>$, которые лежат в множестве $\boldsymbol{Exm^{-}}$, значение $\beta_{ij} = <-1,0>$;
- для тех пар $<O_i,P_j>$, которые лежат в множестве $\boldsymbol{Exm^{\tau}}$, значение $\beta_{ij} = <\boldsymbol{\tau},0>$.

Аналогичным образом, матрица ***Mtx₂*** бинарного отношения $==>_2$ (в предположении, что p - мощность множества всех непустых подмножеств для объединения всех объектов, а q - мощность множества всех непустых подмножеств для объединения свойств всех объектов в множествах $\boldsymbol{Exm^{+}}$, $\boldsymbol{Exm^{-}}$ и $\boldsymbol{Exm^{\tau}}$)

$==>_2$	W_1	...	W_j	...	W_q
V_1	γ_{11}		γ_{1j}		γ_{1q}
⋮	⋮		⋮		⋮
V_i	γ_{i1}	...	γ_{ij}	...	γ_{iq}
⋮	⋮		⋮		⋮
V_p	γ_{p1}	...	γ_{pj}	...	γ_{pq}

в исходном состоянии (т.е. на нулевом шаге ДСМ-рассуждения) организована следующим образом:

- все[13] значения $\gamma_{ij} = <\tau,0>$.

Далее, на первом шаге ДСМ-рассуждения, - применением ППВ типа **I** к представленным в матрице ***Mtx₁*** примерам и контрпримерам - порождается некоторое множество причинных зависимостей вида $\mathbf{J}_{<\alpha,r>}(V==>_2 W)$, где $\alpha \in \{1,-1,0,\tau\}$, и, соответственно, в матрице ***Mtx₂*** происходит замена представляющих их истинностных значений на значения вида $<\alpha,1>$ (соответственно "знаку" порождаемых причинных зависимостей).

Далее в процесс ДСМ-рассуждения включаются ППВ типа **II**: в ходе (использующего найденные к текущему шагу и представленные в матрице ***Mtx₂*** зависимости) ДСМ-анализа недоопределенных на текущем шаге рассуждения элементов матрицы ***Mtx₁*** те из элементов последней, которые доопределяются на данном шаге ДСМ-рассуждения, получают новые истинностные значения. В матрице ***Mtx₁*** при этом производится замена соответствующих истинностных значений на новые,

- вида <1,r[14]> для тех пар $<O_i,P_j>$, которые доопределены правилами вида $\boldsymbol{II}^+$;
- вида <-1,r> для тех пар $<O_i,P_j>$, которые доопределены правилами вида $\boldsymbol{II}^-$;
- вида <0,r> для тех пар $<O_i,P_j>$, которые доопределены правилами вида $\boldsymbol{II}^0$; и наконец
- вида $<\tau$,r> для тех пар $<O_i,P_j>$, которые доопределены правилами вида $\boldsymbol{II}^\tau$ (т.е. остались недоопределенными на данном шаге ДСМ-рассуждения).

Далее[15] процесс анализа матриц ***Mtx₁*** и ***Mtx₂*** итерируется до тех пор, пока не окажется, что либо значений вида $<\tau$,r> в матрице ***Mtx₁*** больше не осталось (т.е. удалось доопределить средствами ДСМ-рассуждения все исходно требовавшие прогноза объекты), либо же новые шаги ДСМ-рассуждения не порождают больше изменений в матрице ***Mtx₂*** (и, соответственно, в матрице ***Mtx₁***) - т.е. новые[16] зависимости уже не порождаются из имеющегося "обучающего материала". Во втором случае мы будем говорить о ситуации *стабилизации* ДСМ-рассуждения.

Важнейшим элементом управления ДСМ-рассуждением является уже представленный нами *выше критерий достаточности оснований* для принятия результатов правдоподобного вывода. В рассматриваемой нами ситуации следовать ему - значит убедиться, что каждый раз, когда порождено некоторое множество причинных зависимостей вида $V==>_2 W$, требуется, чтобы **все** элементы исходной обучающий выборки (т.е. элементы множеств $\boldsymbol{Exm}^+$ и $\boldsymbol{Exm}^-$) были **непротиворечивым** образом доопределены (с использованием этого множества эмпирических зависимостей) средствами ППВ типа **II**. Лишь в этом случае можно говорить о *достаточности оснований* для принятия порождаемых результатов ДСМ-рассуждения. (В ситуации, когда помимо фактов - примеров и контрпримеров, используемых правилами правдоподобного вывода ДСМ-метода - описание изучаемой предметной области содержит также некоторые дополнительные зависимости - аксиомы, проверка достаточности оснований для принятия

---

[13] Мы здесь рассматриваем простейший случай, когда в исходном состоянии (на нулевом шаге ДСМ-рассуждения) множество эмпирических зависимостей (представляющее собой фрагмент базы знаний) *пусто*. Тем не менее, не составляет труда рассмотреть (по аналогичной схеме) и ситуацию, когда на нулевом шаге ДСМ-рассуждения уже имеются некоторые причинные зависимости. Здесь в матрице ***Mtx₂*** на соответствующих позициях будут находиться истинностные значения вида <1,0> для позитивных зависимостей, вида <-1,0> - для негативных, и вида <0,0> - для противоречивых.

[14] Здесь r, как и ранее, - номер текущего шага ДСМ-рассуждения.

[15] На этом месте возможна проверка критерия достаточности оснований для принятия текущих результатов ДСМ-рассуждения, более подробно о которой мы поговорим чуть ниже.

[16] Здесь следовало бы уточнить, о какой из двух версий ДСМ-рассуждения - *монотонной* или *немонотонной* - идет речь: в данном случае речь идет о так называемой *монотонной* версии.

результатов ДСМ-рассуждения предполагает также и проверку этих результатов на совместимость[17] с имеющимися аксиомами ПО).

В ситуации, когда условия критерия достаточности оснований не выполнены, одним из наиболее привлекательных способов преодоления возникших трудностей становится *расширение* исходной обучающий выборки. Важнейшей отличительной особенностью обсуждаемого здесь нами подхода - по сравнению с другими методами анализа данных и машинного обучения - является возможность *управлять* таким *расширением* обучающей выборки. Действительно, невыполнимость критерия достаточности оснований означает, что в исходной обучающей выборке имеются примеры (или контрпримеры), которые не удается “объяснить” с помощью порождаемых (на данной обучающей выборке) ДСМ-зависимостей. Для преодоления возникших трудностей представляется естественным в складывающейся ситуации пополнять обучающую выборку ($\boldsymbol{Exm}^{+} \cup \boldsymbol{Exm}^{-}$) в первую очередь объектами, *сходными с необъясненными*[18].

[17] Т.е. на дедуктивную непротиворечивость объединения порождаемых зависимостей и уже имеющегося множества аксиом предметной области.

[18] К особенностям процедуры такого пополнения мы еще вернемся позже.

3.3.2 Правила правдоподобного вывода ДСМ-метода: работа с данными, представленными в виде числовых векторов.

При работе с числовыми векторами структура представляемой нами конструкции также полностью соответствует общей схеме[1], уже представленной ранее в **Разделе 3.2**. Идея ***инвариантности*** ядра ДСМ-рассуждения и ***настраиваемости*** его окружения при обработке числовых векторов (как и при обработке данных, представленных множествами) выражается, прежде всего, в том, что "спецификации" (т.е. настройке на конкретный тип данных) подлежат те фрагменты конструкции ДСМ-рассуждения, которые специфичны именно для этого типа данных: - эти средства формального представления сходства и та часть правил правдоподобного вывода, которая "отвечает" за фальсификацию сходств одного знака[2] сходствами противоположного знака. Общая же структура решающих предикатов и правил правдоподобного вывода (в частности, учет количества примеров и контрпримеров, служащих основанием для построения правдоподобных выводов; проверка условия исчерпываемости как при анализе локального и глобального сходств, так и в организации ДСМ-вывода по аналогии; учет запрета на контрпримеры и т.д. при порождении гипотез о причинах возникающих явлений и доопределении новых объектов) остается неизменной.

Описание явного вида решающих предикатов и ППВ в случае работы с числовыми векторами мы начнем с обсуждения необходимых модификаций синтаксических средств ДСМ-рассуждения - системы логических языков

$$\boldsymbol{L^0_i \subset L_i \subset L_e \subset L'_e \subset L''_e}.$$

Фактически, нам необходимы выразительные средства для представления:

- объектов - монотонных матриц (см. **Раздел 3.1.2**),
- сходства между ними (формализованного средствами **Алгоритма *CONCAT*** и операции $\boldsymbol{\pitchfork}$ покомпонентного соединения м-матриц) и, наконец,
- факта вложимости подобъекта (м-матрицы) в объект (м-матрицу).

Легко видеть, что для представления м-матриц - декартовых произведений соответствующих множеств столбцов и строк матрицы-универсума **X** - вполне достаточно уже названных нами в **Разделе 3.2.1** выразительных средств (потребуется лишь новый бинарный функциональный символ X для представления декартова произведения множеств). После выбора для представления бинарной операции покомпонентного соединения м-матриц соответствующего бинарного функционального символа $\boldsymbol{\pitchfork}$, нам остается позаботиться лишь о представимости вложимости м-матриц. Не составляет труда добиться этого, рассматривая вложимости соответствующих множеств столбцов (в множества столбцов) и множеств строк (в множества строк) матрицы-универсума **X**.

Явный вид правил правдоподобного вывода- ППВ - мы, как и в случае рассмотренной в **Разделе 3.3.1** обработки данных, представленных множествами, мы начнем анализа правил первого рода - правил порождения зависимостей о причинах изучаемых явлений. Центральное место в их формулировке отводится (как и ранее) решающим предикатам. Строение же последних мы подробно разберем на примере предиката $\boldsymbol{M^+_a}$ - простого положительного сходства[3].

---

[1] В том числе – исходному постулату о поиске *причинных* зависимостей, неявным образом *содержащихся* в анализируемых числовых данных. (Задача «инструментов» ДСМ-ИАД – восстановить *явный вид* таких зависимостей).

[2] Т.е., например, сходств, порожденных на примерах, сходствами, порожденными на контрпримерах, и наоборот.

[3] Здесь термин *простого* отражает тот факт, что рассматривается *бинарное* отношение причинности. Термин *положительного* означает, что речь пойдет о поиске "позитивных" зависимостей, порождаемых из примеров - описаний прецедентов, когда анализируемое явление *имело место*. Термин *сходства*, наконец, отражает то

Итак, исходным материалом для применения ППВ (как первого, так и второго типа) являются:

- множество ***O*** объектов O_1, O_2, ..., представляемое множеством **C** всех C_1, C_2, ... - столбцов матрицы-универсума **X**, и множество[4] ***P*** свойств P_1, P_2,..., которыми обладают или не обладают объекты (столбцы) из множества **C**;
- каждый из объектов в множестве **C** - декартово произведение одноэлементного подмножества множества $\mathbf{U^{(1)}} = \{c_1, c_2,..., c_n\}$ всех столбцов и множества R(**X**) всех строк матрицы-универсума **X**; а каждое из множеств свойств в множестве **A** - непустое подмножество некоторого универсума свойств $\mathbf{U^{(2)}} = \{a_1, a_2,..., a_m\}$;
- пара частично определенных бинарных отношений $==>_1$ и $==>_2$, первое из которых используется для того, чтобы характеризовать *наличие* (или *отсутствие*) *свойств* у объектов; второе используется для характеризации подобъектов[5], представляющих *причины наличия* (или, соответственно, *отсутствия*) *свойств* у исследуемых объектов;
- три заранее заданных множества объектов (описаний прецедентов) ***Exm*$^+$, *Exm*$^-$** и ***Exm*$^\tau$**, представляющих, соответственно,

  - примеры объектов, обладающих интересующими нас свойствами, т.е. описываемые выражениями вида[6]

$$\mathbf{J}_{<1,0>}(C_i ==>_1 P_i)^{7};$$

  - примеры объектов, не обладающих интересующими нас свойствами, т.е. описываемые выражениями вида

$$\mathbf{J}_{<-1,0>}(C_i ==>_1 P_i);$$

  и, наконец, объекты, наличие или отсутствие заданных свойств у которых и предполагается установить в процессе ДСМ-рассуждения, т.е. те, что представляются выражениями вида

$$\mathbf{J}_{<\tau,0>}(C_i ==>_1 P_i).$$

Результатом применения ППВ типа **I** является заключение (утверждение) о том, что некоторый подобъект V (т.е. м-матрица в универсуме **X**) является причиной присутствия (в случае обучения на положительных примерах) или причиной отсутствия (в случае обучения на отрицательных примерах) некоторого множества свойств W (представляющего собой непустое подмножество универсума $\mathbf{U^{(2)}}$) у объектов из множества **C**. Возможны ситуации, когда сделать такое заключение непротиворечивым образом не представляется возможным. Также возможны ситуации, когда для порождения заключения о причинной связи V и W просто нет оснований (случай наследования недоопределенности на следующий шаг ДСМ-рассуждения).

Синтаксически эти результаты (т.е. четыре возможных исхода при порождении заключения средствами ППВ типа **I**) также, как, например, при обработке данных, представленных множествами - см. **Раздел 3.3.2**, - выражаются с помощью частично

обстоятельство, что рассматриваемый нами решающий предикат строится для формализации *миллевского метода сходства*, однако с учетом идей, заложенных также и в формулировку миллевского метода *сопутствующих изменений* - см. **Раздел 2.2**.

[4] Ради простоты изложения здесь мы рассматриваем случай, когда свойства объектов представлены множествами. Не составляет труда реализовать ту же конструкцию ДСМ-рассуждения и для случая представления свойств, например, числовыми векторами. (Чуть подробнее мы поговорим об этом позже - в **Разделе 3.5**).

[5] Здесь уместно напомнить, что каждый подобъект - м-матрица, порождаемая декартовым произведением некоторого подмножества множества $\mathbf{U^{(1)}} = \{c_1, c_2,..., c_n\}$ и некоторого подмножества множества R(**X**).

[6] Более точно: истинностные значения вида $<\alpha,0>$ имеют место лишь на первом шаге ДСМ-рассуждения. Для всех последующих шагов здесь следует использовать выражения вида $<\alpha,r\text{-}1>$,где r - номер текущего шага ДСМ-рассуждения.

[7] Определение функции $\mathbf{J_v(\varphi)}$ мы уже дали ранее в **Разделе 2.4**.

определенного отношения $==>_2$: в исходном состоянии, когда нет информации о причинной связи V и W, имеет место $\mathbf{J}_{<\tau,0>}(V==>_2 W)$. После применения ППВ реализуется одна из четырех ситуаций[8]:

а) $\mathbf{J}_{<1,r>}(V==>_2 W)$, в случае, когда установлена связь (положительная эмпирическая зависимость) между V и W (т.е. когда у нас появились основания[9] считать подобъект V причиной наличия у объектов из множества **C** множества свойств W);

б) $\mathbf{J}_{<-1,r>}(V==>_2 W)$, когда установлено ее отсутствие (т.е. отрицательная ЭЗ, характеризующая множество V как причину отсутствия у объектов из множества **C** свойств W);

в) $\mathbf{J}_{<0,r>}(V==>_2 W)$, когда есть доводы как в пользу положительной, так и в пользу отрицательной связи V и W, причем нет возможности предпочесть одну другой - случай "эмпирического противоречия";

г) $\mathbf{J}_{<\tau,r>}(V==>_2 W)$, когда ни один из случаев а)-в) не реализуется, т.е. сохраняется наследуемое состояние недоопределенности на следующий шаг рассуждения.

При обработке числовых векторов предикат простого положительного сходства $\boldsymbol{M^+_a}$ также представляет *миллевский метод сходства* (см. выше **Раздел 2.2**), примененный к положительным примерам - описаниям прецедентов наличия у изучаемых объектов избранных для анализа свойств. Ключевые элементы миллевской конструкции, нашедшие отражение в ДСМ-формализации:

- рассматривается множество (попарно-различных!) примеров, представляющих наличие анализируемых свойств у изучаемых объектов;
- всякий раз, когда в структуре анализируемых объектов встречается некоторый (непустой) фрагмент, в свойствах этих же объектов содержится некоторое непустое общее подмножество;
- в множество примеров, отобранных для анализа сходства, включены <u>все</u> прецеденты, релевантные цели проводимого причинного анализа (т.е. в обучающей выборке учтены <u>все</u> объекты, свойства которых содержат обозначенное выше общее подмножество);
- основанием для заключения о причинах появления обсуждаемых свойств может быть анализ не менее двух (различных-!) примеров.

Формальным представлением этой схемы средствами системы логических языков $\boldsymbol{L^0_i} \subset \boldsymbol{L_i} \subset \boldsymbol{L_e} \subset \boldsymbol{L'_e} \subset \boldsymbol{L''_e}$ (см. выше **Раздел 3.2.1**) является выражение

$$\boldsymbol{M^+_a}(V,W) = \exists k\ \underline{\boldsymbol{M^+_a}}(V,W,k) =$$

$$= \exists k\ \exists C_{i1}...\exists C_{ik}\ \exists P_{i1}...\exists P_{ik}\ \mathbf{(}\ [\ \mathbf{J}_{<1,r>}(C_{i1}==>_1 P_{i1})\ \&$$

$$\&\ \forall P\ (\ \mathbf{J}_{<1,r>}(C_{i1}==>_1 P) \rightarrow (P \subseteq P_1))\ ]\ \&\ ...$$

$$...\ \&\ [\ \mathbf{J}_{<1,r>}(C_{ik}==>_1 P_{ik})\ \&$$

$$\&\ \forall A\ (\ \mathbf{J}_{<1,r>}(C_{ik}==>_1 P) \rightarrow (P \subseteq P_k))\ ]\ \&$$

$$\&\ \mathbf{[}V \in \prod_{i=1}^{k}\{C_i\}\ \&;$$

$$\&\ (V \neq \varnothing)\ \mathbf{]}\ \&$$

$$\&\ [\ \forall i \forall j\ (\ (i \neq j)\&(1 \leq i,j \leq k) \rightarrow (C_i \neq C_j)\ )\ ]\ \&$$

[8] Здесь **r** есть номер шага ДСМ-рассуждений, на котором получено соответствующее заключение.

[9] Формализованные в рамках ДСМ-рассуждения.

$$\& \forall X \forall Y \, \mathbf{[} \, [\mathbf{J}_{<1,r>}(X ==>_1 Y) \,\&$$

$$\& \forall P \, (\, \mathbf{J}_{<1,r>}(X ==>_1 P) \rightarrow (P \subseteq Y) \,) \,\& \, (\{V\} \pitchfork \{X\} = \{V\})] \rightarrow$$

$$\rightarrow [\, (W \neq \varnothing) \,\& \, (W \subseteq Y) \,\& \, ((X = C_1) \vee \ldots \vee (X = C_k)) \,] \, \mathbf{]} \,\&$$

$$\& \, (k \geq 2) \, \mathbf{)} \ ,$$

где соответствующие подформулы в представлении $\boldsymbol{M}^{+}_{\boldsymbol{a}}$ есть совместно выполняющиеся условия (***cn1.-cn7.***), такие что

cn1.) из множества положительных примеров $\boldsymbol{Exm}^{+}$ отобраны ровно k штук (обладающих необходимым нам свойством)

$$\boldsymbol{Exm}^{+}(V,W) = \ = \{ \, \mathbf{J}_{<1,r>}(C_{i1} ==>_1 P_{i1}) \, , \ldots$$

$$\ldots, \, \mathbf{J}_{<1,r>}(C_{ik} ==>_1 P_{ik}) \, \};$$

cn2.) все примеры в множестве $\boldsymbol{Exm}^{+}(V,W)$ попарно различны (условие невырожденности сходства);

cn3.) м-башня V выражает сходство всех примеров из множества $\boldsymbol{Exm}^{+}(V,W)$:

$$V \in \pitchfork_{i=1}^{k} \{C_i\};$$

cn4.) сходство примеров из $\boldsymbol{Exm}^{+}(V,W)$ не тривиально: $V \neq \varnothing$;

cn5.) м-матрица V порождена максимальным (по вложению) множеством столбцов (объектов) в $\mathbf{X}^{+}$, т.е. м-матрица V есть действительно м-башня (и <u>все</u> порождающие ее столбцы учтены в соответствующем множестве $\boldsymbol{Exm}^{+}(V,W)$ - оно нерасширяемо);

cn6.) множество W есть непустая общая часть - сходство[10] - множеств свойств P_i, т.е.

$$W = \bigcap_{i=1}^{k} P_i = P \neq \varnothing$$

в рассматриваемом нами простейшем случае;

cn7.) $k \geq 2$, т.е. порождаемая причинная зависимость выражает сходство <u>различных</u> (по крайней мере двух) примеров.

Следует особо подчеркнуть, что подформула

$$\forall X \forall Y \, \mathbf{[} \, [\mathbf{J}_{<1,r>}(X ==>_1 Y) \,\&$$

$$\& \forall P \, (\, \mathbf{J}_{<1,r>}(X ==>_1 P) \rightarrow (P \subseteq Y) \,) \,\& \, (\{V\} \pitchfork \{X\} = \{V\})] \rightarrow$$

$$\rightarrow [\, (W \neq \varnothing) \,\& \, (W \subseteq Y) \,\& \, ((X = C_1) \vee \ldots \vee (X = C_k)) \,] \, \mathbf{]}$$

дает явный вид *эмпирической зависимости*, которая и выражает восстанавливаемое (другими словами - доопределяемое) средствами ДСМ-рассуждения *причинно-следственное* отношение $V ==>_2 W$. (Напомним, что поиск неявным образом представленных в исходных данных эмпирических зависимостей, а также распространение средствами ДСМ-рассуждения их там, где это возможно на новые, исходно недоопределенные случаи, собственно, и является центральным моментом обсуждаемой нами формальной конструкции).

---

[10] В данном случае (т.е. в представлении свойств объектов множествами).

Аналогичным образом (хотя и с набольшими отличиями[11]- см. подробнее [*Забежайло 93, Забежайло 93-а*]) - строятся версии предиката отрицательного сходства, используемые при порождении зависимостей о контрпричинах (т.е. фактах, препятствующих наличию) исследуемого свойства.

Пополняя (как и в **Разделе 3.3.1**) представленное семейство {***cn1***.,...,***cn7***.} дополнительными условиями, такими, например, как:

cn8.) неверно, что в множестве отрицательных примеров найдется хотя бы один пример ($\mathbf{J}_{<1,r>}(C_0 ==>_1 P_0)$), сходный с порождающими м-башню V примерами из множества ***Exm***$^{+}$(V,W) так, что это сходство выражает м-матрица, образованная теми же строками универсума X, что и уже построенная м-башня V;

и др. - см. выше **Раздел 3.3.1** - можно построить (более подробно – см. [*Забежайло 93*, *Забежайло 93-а*]) частично упорядоченные (как и ранее - в смысле логического следования) множества $\mathbf{M}^{+}$ и $\mathbf{M}^{-}$ решающих предикатов обоих знаков ($\alpha \in \{+,-\}$), для которых будет справедлив набор соотношений, подобный представленному **Леммой 3.3.1.1**.

В свою очередь, упорядочение по “силе” используемых в ДСМ-рассуждении решающих предикатов позволяет (по уже представленной в **Разделе 3.3.1** схеме) организовать упорядочение по “силе” порождаемых ими зависимостей: результатом использования более “сильного” предиката оказывается более “сильная” эмпирическая зависимость. Таким образом, появляется возможность более тонкого учета промежуточных результатов в процедуре ДСМ-фальсификации получаемых эмпирических зависимостей.

Процедура фальсификации сходств при обработке числовых векторов имеет некоторые особенности по сравнению, например, с аналогичной процедурой для обработки множеств (см. **Раздел 3.3.1**). Действительно, характер данных - столбцы матрицы универсума **X** и м-матрицы, образованные подмножествами ее строк и столбцов, - требует при обсуждаемой фальсификации сходств учесть и случаи, когда имеет место “вложимость[12]” объектов знака $\alpha \in \{+,-\}$ в м-башни знака $-\alpha$ (причем требующая учета “вложимость” может быть формализована как с учетом попадания рассматриваемого столбца “между” столбцами соответствующей м-башни, так и с учетом возможности “присоединить” данный столбец в качестве крайнего слева или крайнего справа к соответствующей м-башне[13]).

Таким образом ситуация фальсифицируемости эмпирической зависимости вида

$$V^{\alpha} = \mathbf{C}^{\alpha}(V) \times \mathbf{R}(V),$$

где $\alpha \in$ принимает значение (+) или (-), а

$$\mathbf{C}^{\alpha}(V) \subseteq \mathbf{C}^{\alpha} \text{ и } \mathbf{R}(V) \subseteq \mathbf{R}$$

и

$$\mathbf{X}^{\alpha} = \mathbf{C}^{\alpha} \times \mathbf{R} = \{C^{\alpha}_1, C^{\alpha}_2, ..., C^{\alpha}_n \alpha\} \times \{R_1, R_2, ..., R_m\}, \text{ для } \alpha \in \{+,-\},$$

определяется порождением одновременно м-башен (см. **Раздел 3.1**) V^{+} и V^{-}, сформированных на одном и том же множестве строк (числовых значений параметров)

$$\mathbf{R}(V^{+}) = \mathbf{R}(V) = \mathbf{R}(V^{-}),$$

---

[11] Помимо воспроизведения условий ***cn1***.-***cn7***. для работы с отрицательными примерами можно предложить более тонкие формулировки предиката отрицательного сходства $\boldsymbol{M^{-}_{a}}$, учитывающие, например, то обстоятельство, что может существовать несколько независимых способов "разрушения" механизма, обуславливающего "проявление" искомого свойства.

[12] В смысле упорядочения столбцов м-башни вдоль ее оси монотонности (см. **Раздел 3.1.2**).

[13] Подробный анализ ситуаций названного типа приведен в работе [*Забежайло 93*].

и таких множествах столбцов (объектов из исходных данных) $\mathbf{C}^+$ и $\mathbf{C}^-$, что

$$V' = \{\mathbf{C}^+(V) \cup \mathbf{C}^-(V)\} \times \mathbf{R}(V),$$

представляет собою м-матрицу в объединенном универсуме

$$\mathbf{X}'_{\mathbf{R}(V)} = \mathbf{X}^+_{\mathbf{R}(V)} \cup \mathbf{X}^-_{\mathbf{R}(V)} = \{\mathbf{C}^+ \cup \mathbf{C}^-\} \times \mathbf{R}(V),$$

т.е. иными словами: области *ко-монотонного* изменения числовых значений параметров из выделенного множества **R**(V) для позитивной V^+ и негативной V^- эмпирических зависимостей конфигурируются (в случае возникновения *эмпирического противоречия*) одним из двух возможных способов:

- *пересекаются*, либо же
- объекты одного знака V^{α} могут быть «*присоединены*» (с сохранением ко-монотонности изменения числовых значений параметров) «справа» или же слева к м-башне противоположного знака $V^{-\alpha}$.

Нетрудно убедиться, что формализация подобных соображений в случае проверки условия *запрета на контрпримеры* сводится к проверке «вложимости» столбца (примера) знака α в м-башню знака $-\alpha$ (в том числе – возможности «присоединить» его к ней «слева» или же «справа» с сохранением всех необходимых условий ко-монотонности).

Представление о возможностях использования такого условия дает

3.3.2.1. Пример

Пусть множество положительных примеров определяется универсумом

$\mathbf{X}^+$:

	C_1	C_2	C_3	C_4	C_5	C_6
R_1	1	6	7	4	2	3
R_2	11	16	10	14	12	13
R3	29	26	21	24	22	23

,

а в множестве столбцов матрицы $\mathbf{X}^-$ имеется столбец

	C_7
R_1	5
R_2	15
R_3	35

В множестве м-башен, построенных в $\mathbf{X}^+$ содержатся V_1,V_2 и V_3 такие, что:

V_1:

	C_1 C_5	C_6 C_4	C_2
R_1			
R_2	1 2 11 12	3 4 13 14	6 16
R_3			

V_2:

	C_3 C_5	C_6 C_4	C_2
R_1			
R_2	10 12	13 14	16
R_3	21 22	23 24	26

		C_5 C_6	C_4 C_2
R_1		2 3	4 6
R_2	V_3:	12 13	14 16
R_3		22 23	24 26

Однако, зависимости вида $\mathbf{J}_{<1,r+1>}(V_i ==>_2 W)$ порождаются применением ППВ вида **I** лишь для V_2 и V_3 (V_1 исключается рассматриваемым условием).

о

Теперь нам осталось лишь привести формулировки явного вида правил правдоподобного вывода I-го типа. В соответствии с рассмотренным выше механизмом приписывания одного из четырех истинностных значений это будут выражения следующих четырех типов (здесь они представлены на примере использования решающих предикатов простого сходства обоих знаков):

$$\boldsymbol{I}^{+}_{\mathbf{a}} = \frac{\mathbf{J}_{<\tau,r>}(\{\mathbf{C}^{+}(V) \times \mathbf{R}(V)\} ==>_2 W),\ \boldsymbol{M}^{+}_{a}(\{\mathbf{C}^{+}(V) \times \mathbf{R}(V)\}, W) \ \&\ \neg\, \boldsymbol{M}^{-}_{a}(\{\mathbf{C}^{-}(V) \times \mathbf{R}(V)\}, W)}{\mathbf{J}_{<1,r+1>}(\{\mathbf{C}^{+}(V) \times \mathbf{R}(V)\} ==>_2 W),}$$

$$\boldsymbol{I}^{-}_{\mathbf{a}} = \frac{\mathbf{J}_{<\tau,r>}(\{\mathbf{C}^{+}(V) \times \mathbf{R}(V)\} ==>_2 W),\ \neg\boldsymbol{M}^{+}_{a}(\{\mathbf{C}^{+}(V) \times \mathbf{R}(V)\}, W) \ \&\ \boldsymbol{M}^{-}_{a}(\{\mathbf{C}^{-}(V) \times \mathbf{R}(V)\}, W)}{\mathbf{J}_{<-1,r+1>}(\{\mathbf{C}^{+}(V) \times \mathbf{R}(V)\} ==>_2 W),}$$

$$\boldsymbol{I}^{0}_{\mathbf{a}} = \frac{\mathbf{J}_{<\tau,r>}(\{\mathbf{C}^{+}(V) \times \mathbf{R}(V)\} ==>_2 W),\ \boldsymbol{M}^{+}_{a}(\{\mathbf{C}^{+}(V) \times \mathbf{R}(V)\}, W) \ \&\ \boldsymbol{M}^{-}_{a}(\{\mathbf{C}^{-}(V) \times \mathbf{R}(V)\}, W)}{\mathbf{J}_{<0,r+1>}(\{\mathbf{C}^{+}(V) \times \mathbf{R}(V)\} ==>_2 W),}$$

$$\boldsymbol{I}^{\tau}_{\mathbf{a}} = \frac{\mathbf{J}_{<\tau,r>}(\{\mathbf{C}^{+}(V) \times \mathbf{R}(V)\} ==>_2 W),\ \neg\boldsymbol{M}^{+}_{a}(\{\mathbf{C}^{+}(V) \times \mathbf{R}(V)\}, W) \ \&\ \neg\, \boldsymbol{M}^{-}_{a}(\{\mathbf{C}^{-}(V) \times \mathbf{R}(V)\}, W)}{\mathbf{J}_{<\tau,r+1>}(\{\mathbf{C}^{+}(V) \times \mathbf{R}(V)\} ==>_2 W).}$$

Использование в случае работы с числовыми векторами при построении ППВ ДСМ-метода остальных решающих предикатов из множеств $\mathbf{M}^{+}$ и $\mathbf{M}^{-}$ осуществляется по аналогичному принципу.

Правила II-го типа - правила переноса найденных зависимостей на новые объекты (там, где это можно сделать средствами ДСМ-рассуждения) также как и ранее формулируются при помощи соответствующих решающих предикатов.

Аналогично уже рассмотренным случаям - см. **Раздел 3.3.1** - результатом применения ППВ типа **II** является заключение (утверждение) о том, что первоначально неизученный объект (т.е. один из тех, что до применения соответствующего ППВ описывался выражением $\mathbf{J}_{<\tau,r>}(C_j ==>_1 P_j)$) обладает (в случае переноса положительных ЭЗ) или не обладает (в случае переноса отрицательных ЭЗ) исследуемыми свойствами. Структуру решающих предикатов для этого случая представим на примере предиката $П^{+}$,

реализующего уже формулировавшийся нами в **Разделе 3.3.1** принцип индуктивного обобщения.

Формализация названного принципа средствами ДСМ-рассуждения требует одновременной выполнимости следующих условий:

cn9.) среди причинных зависимостей, порожденных с помощью ППВ типа **I** на множестве положительных примеров (т.е. среди положительных ЭЗ), имеется по крайней мере одна такая, что анализируемый объект C_j содержит в себе "носитель" соответствующего глобального сходства V (т.е. соответствующий столбец матрицы-универсума **X** вкладывается в соответствующую м-башню V, или же, другими словами, анализируемый объект сходен в смысле операции $\pitchfork$ с порождающей эту эмпирическую зависимость м-башней V);

cn10.) среди вкладывающихся в C_j "носителей" глобального сходства V_i (т.е. среди сходных с C_j м-башен, представляющих соответствующие ЭЗ,), имеются лишь те, которые получены ДСМ-обучением на положительных примерах;

cn11.) объединение W_i по всем таким башням V_i есть в точности P_j (в рассматриваемом нами простейшем случае $W_i = \bigcup_i W_i = W = P_j$).

Явный вид (+,-)-решающих предикатов в данном случае представляет выражение

$$\Pi_{\alpha}(C,P) = \exists k\, \Pi_{\alpha}(C,P,k) =$$

$$= \exists W_1 \ldots \exists W_k \,\mathbf{(}\, \exists V_1 [\, \mathbf{J}_{<\alpha,r+1>}(V_1 ==>_2 W_1) \,\&$$

$$\& \,(V_1 \in \{\, \{C\} \pitchfork \{V_1\}\,\}\,)\,\mathbf{]}\,\& \ldots$$

$$\ldots \,\&\, \exists V_k [\, \mathbf{J}_{<\alpha,r+1>}(V_k ==>_2 W_k)\,\&$$

$$\& \,(V_k \in \{\, \{C\} \pitchfork \{V_k\}\,\}\,)\,\mathbf{]}\,\&$$

$$\&\, (\bigcup_i W_i = W = P_j)\,\&$$

$$\&\, \forall W\, (\exists V\, [\, (\mathbf{J}_{<\alpha,r+1>}(V ==>_2 W) \& (V \in \{\, \{C\} \pitchfork \{V\}\,\})\,] \rightarrow$$

$$\rightarrow [\, (V = V_1) \vee \ldots \vee (V = V_k)\,])\,\&$$

$$\&\, \forall P_0\, [(\,(P_0 \subseteq P) \& (\,P_0 \neq \varnothing)\,) \rightarrow$$

$$\rightarrow \neg \exists V_0\, (\mathbf{J}_{<-\alpha,r+1>}(V_0 ==>_2 P_0) \,\&\, (V_0 \in \{\, \{C\} \pitchfork \{V_0\}\,\}))\,,$$

где везде в нем для переноса позитивных зависимостей $\alpha = 1$, а для переноса негативных зависимостей $\alpha = -1$.

В случае "конкуренции" причин противоположного знака соответствующий решающий предикат

$$\Pi_0(C,P) = \exists V_1 \exists W_1 \exists V_2 \exists W_2\, \mathbf{(}\, \mathbf{J}_{<1,r+1>}(V_1 ==>_2 W_1)\,\&$$

$$\&\, \mathbf{J}_{<-1,r+1>}(V_2 ==>_2 W_2)\,\&$$

$$\&\, [(W_1 \cap W_2) \neq \varnothing]\,\&$$

$$\&\, (V_1 \in \{\, \{C\} \pitchfork \{V_1\}\,\}\,)\,\&$$

$$\&\, (V_2 \in \{\, \{C\} \pitchfork \{V_2\}\,\}\,)\,\&$$

$$\&\, (W_1 \subseteq P)\,\&\,(W_2 \subseteq P)\,\mathbf{)}$$

Наконец, состояние недоопределенности наследуется на следующий шаг ДСМ-рассуждения, если ни один из решающих предикатов $\Pi_1(C,P)$, $\Pi_{-1}(C,P)$ и $\Pi_0(C,P)$ не выполним. Таким образом:

$$\Pi_{\tau}(C,P) = \neg[\ \Pi_1(C,P) \vee \Pi_{-1}(C,P) \vee \Pi_0(C,P)\ ]$$

Как и при обработке данных, представленных множествами, явный вид ППВ переноса зависимостей о наличии или отсутствии изучаемых свойств - ППВ типа **II** - представляют следующие четыре соотношения:

$$\boldsymbol{II}^{+} = \frac{\mathbf{J}_{<\tau,r>}(\ C{=}{=}{>}_1 P),\ \Pi_1(C,P)}{\mathbf{J}_{<1,r+1>}(C{=}{=}{>}_1 P)},$$

$$\boldsymbol{II}^{-} = \frac{\mathbf{J}_{<\tau,r>}(C{=}{=}{>}_1 P),\ \Pi_{-1}(C,P)}{\mathbf{J}_{<-1,r+1>}(C{=}{=}{>}_1 P)},$$

$$\boldsymbol{II}^{0} = \frac{\mathbf{J}_{<\tau,r>}(C{=}{=}{>}_1 P),\ \Pi_0(C,P)}{\mathbf{J}_{<0,r+1>}(\ C{=}{=}{>}_1 P)},$$

$$\boldsymbol{II}^{\tau} = \frac{\mathbf{J}_{<\tau,r>}(C{=}{=}{>}_1 P),\ \Pi_{\tau}(C,P)}{\mathbf{J}_{<\tau,r+1>}(\ C{=}{=}{>}_1 P)}.$$

В точном соответствии с уже рассматривавшимся в **Разделе 3.3.1** набором действий, при обработке числовых векторов процесс анализа средствами ДСМ-рассуждения конкретной обучающей выборки - множества примеров, множества контрпримеров и множества объектов, свойства которых требуют прогноза (доопределения), - также можно проследить, наблюдая за преобразованиями матриц ***Mtx₁*** и ***Mtx₁*** отношений, соответственно, $=\!=\!>_1$ и $=\!=\!>_2$.

Действие критерия достаточности оснований для принятия результатов ДСМ-рассуждения при обработке числовых векторов также полностью воспроизводит механизм, уже описанный выше в **Разделе 3.3.1**.

3.3.3. Правила правдоподобного вывода ДСМ-метода: некоторые дополнительные замечания..

Теперь, опираясь на представленный в **Разделах 3.3.1 - 3.3.3** фактический материал, мы имеем возможность возвратиться к обсуждению общих свойств правил правдоподобного вывода ДСМ-метода. Предметом дальнейшего обсуждения станут:

- виды (т.е. различные варианты) эмпирических зависимостей, порождаемых в рамках ДСМ-рассуждения;
- обзор возможностей обработки средствами ДСМ-рассуждения различных (в том числе - комбинированных) типов данных;
- варианты стратегий ДСМ-рассуждения; и, наконец,
- возможности анализа средствами ДСМ-рассуждения небинарного отношения причинности.

Как уже было продемонстрировано выше, ключевым элементом ДСМ-рассуждения является поиск так называемых *эмпирических зависимостей* - ЭЗ. Фундаментальная роль ЭЗ обусловлена прежде всего тем, что (в силу используемого в ДСМ-рассуждении механизма *фальсификации* результатов правдоподобного вывода) выполнимость соответствующего набора решающих предикатов (и сформированного на их основе правила правдоподобного вывода) означает, что для выражаемого такой ЭЗ утверждения (гипотезы) нет фальсификаторов в имеющемся для ДСМ-обучения исходном материале. Действительно, не составляет труда убедиться, что представляющая каждую ЭЗ формула представляет собой логическое следствие соответствующего решающего предиката. Так, например, легко видеть, что выражение

$$\forall X\ \forall Y\ \mathbf{[}\ [\mathbf{J}_{<1,r>}(X ==>_1 Y)\ \&$$
$$\&\ \forall P\ (\ \mathbf{J}_{<1,r>}(X ==>_1 P) \rightarrow (P \subseteq Y)\ )\ \&\ (V \subset X)] \rightarrow$$
$$\rightarrow [\ (W \neq \varnothing)\ \&\ (W \subseteq Y)]\ \mathbf{]}$$

представляет собою очевидное логическое следствие выражения для $\boldsymbol{M}^{+}_{\boldsymbol{a}}(V,W)$:

$$\boldsymbol{M}^{+}_{\boldsymbol{a}}(V,W) =$$
$$= \exists k\ \exists O_{i1}...\exists\ O_{ik}\ \exists P_{i1}...\exists P_{ik}\ \mathbf{(}\ [\ \mathbf{J}_{<1,r>}(O_{i1} ==>_1 P_{i1})\ \&$$
$$\&\ \forall P\ (\ \mathbf{J}_{<1,r>}(O_{i1} ==>_1 P) \rightarrow (P \subseteq P_{i1}))\ ]\ \&\ ...$$
$$\&\ [\ \mathbf{J}_{<1,r>}(O_{ik} ==>_1 P_{ik})\ \&$$
$$\&\ \forall P\ (\ \mathbf{J}_{<1,r>}(O_{ik} ==>_1 P) \rightarrow (P \subseteq P_{ik}))\ ]\ \&$$
$$\&\ [\ (\bigcap_{i=1}^{k} O_i = V)\ \&\ (V \neq \varnothing)\ ]\ \&$$
$$\&\ [\ \forall i \forall j\ (\ [(i \neq j) \& (1 \leq i,j \leq k)] \rightarrow (O_i \neq O_j)\ )\ ]\ \&$$
$$\&\ \forall X\ \forall Y\ \mathbf{[}\ [\mathbf{J}_{<1,r>}(X ==>_1 Y)\ \&$$
$$\&\ \forall P\ (\ \mathbf{J}_{<1,r>}(X ==>_1 P) \rightarrow (P \subseteq Y)\ )\ \&\ (V \subset X)] \rightarrow$$
$$\rightarrow [\ (W \neq \varnothing)\ \&\ (W \subseteq Y)\ \&\ ((X = O_1) \vee ... \vee (X = O_k))\ ]\ \mathbf{]}\ \&$$
$$\&\ (k \geq 2)\ \mathbf{)}\ ,$$

В представленном здесь примере реализуется поиск *позитивных* эмпирических зависимостей - ЭЗ, порождаемых на базе анализа сходства примеров (т.е. позитивных прецедентов изучаемого явления), причем в форме “от причины к следствию”, т.е. вида:

ПРИЧИНА ==> СЛЕДСТВИЕ.

В случае поиска *негативных* ЭЗ, порождаемых на базе анализа сходства контр-примеров (т.е. негативных прецедентов - случаев, когда исследуемое явление не наступало), можно рассмотреть по крайней мере 2 сорта эффектов, требующих формализации.

В первом формализуется ситуация, когда осуществлению анализируемых свойств *препятствует* некоторая совокупность факторов (т.е. имеется негативная причинность отсутствия наблюдаемого явления). На поиск а также выделение явного вида такой "*антипричинности*" и ориентирована конструкция решающих предикатов и ППВ ДСМ-метода в данном случае.

Однако, также возможна и ситуация, когда для реализации анализируемых свойств у тех или иных объектов попросту нет достаточных оснований (т.е. когда в структуре объектов не оказывается ни "причин", ни "антипричин"[1] для проявления этих свойств. Нетрудно понять, что в данном случае явный вид искомых ЭЗ и соответствующих им решающих предикатов будет несколько отличен от представленной выше ситуации поиска "антипричин".

Детальные формулировки и обзор результатов использования различных версий решающих предикатов (и соответствующих ППВ) для порождения негативных эмпирических зависимостей можно найти, например, в работах [*Финн 88, Финн 10* и др.]. Расширяющая ранее сформулированные представления оригинальная идея построения негативных ЭЗ предложена в работе [*Финн 99-а*].

Еще одним важным типом эмпирических зависимостей, порождаемых средствами ДСМ-рассуждения, являются зависимости "от следствия к причине", имеющие вид

ПРИЧИНА <==[2] СЛЕДСТВИЕ.

ДСМ-инструментарий такого типа называется *обратными* методами (в противоположность ранее представленным *прямым* методам). Использование в анализе причинно-следственных связей формальных средств такого типа оказывается адекватным в тех предметных областях, где природа изучаемых явлений отражена в эмпирическом материале "феноменологического" характера (в частности, в задачах медицинской и\или технической диагностики, где, обычно, по описаниям проявления того или иного эффекта требуется сформулировать выводы о возможных причинах его появления). Для более детального знакомства с формальными средствами ДСМ-рассуждения, используемыми в этом случае, можно обратиться, например, к работам [*Ивашко 84, Ивашко 86, Аншаков 87* и др.].

Переходя к обсуждению ДСМ-средств обработки других типов данных (а нас здесь, в дополнение к уже представленным конструкциями, в первую очередь интересует - см. **Раздел 3.1.2** - возможность оперировать с мультимножествами, кортежами множеств и векторами числовых интервалов), следует подчеркнуть, что в целом структура формальных средств ДСМ-рассуждения для этих типов данных в точности воспроизводит те механизмы, которые мы уже представили выше в **Разделах 3.3.1 - 3.3.2**. Общая схема подхода выглядит так: рассмотренные в **Раздел 3.1.2** механизмы формализации сходства используются сперва для модификации синтаксической платформы ДСМ-рассуждения, а затем для уточнения (спецификации) "окружения ядра" в решающих предикатах и правилах правдоподобного вывода.

По этой же схеме может быть организована и ДСМ-обработка смешанных типов данных, когда объекты (и\или их свойства) представлены *кортежами* так, что каждый из элементов такого кортежа - элемент данных определенного типа (например, когда объекты - химические структуры[3], сопровождаемые также и набором физико-химических

---

[1] Т.е. ни факторов, способствующих (в рамках предполагаемой неявной модели причинности) -, ни факторов, препятствующих проявлению целевых свойств у исследуемых объектов.

[2] Обозначение <== выбрано для наглядности: "обращение" обсуждавшегося ранее отношения причинности требует соответствующей модификации формальных средств ДСМ-рассуждения.

[3] Структурные формулы которых представлены, например, средствами дескрипторного языка (т.е. конкретный тип данных - *множества*). Пример такого представления можно найти в **Приложении П.2** .

характеристик - дозами, концентрациями, температурными параметрами фазовых переходов и т.п[4].; свойства объектов в данном случае могут быть представлены как множествами[5], так мультимножествами[6], или же векторами чисел[7], а возможно - и числовых интервалов[8]). Определяя покомпоненто механизмы ДСМ-обработки таких кортежей, мы имеем возможность оперировать в ДСМ-рассуждении *комплексными* типами данных. (Актуальность подобного рода инструментария заметно растет в последние годы в связи с необходимостью создавать системы управления для больших объектов\структур комплексного характера, где, часто - в режиме *реального времени*, необходимо обрабатывать в унифицированной среде совершенно разнородные данные - показания приборов, описания структурных характеристик фрагментов объекта управления[9], текстовую информацию из больших коллекций полнотекстовых документов и др., - эффективно порождая при этом разумные альтернативы для принятия решений - см., например, работы [*AS'95, SA'96, ISAS'97* и др.]).

При обработке числовых данных механизмы представляемого класса позволяют, помимо прочего, соединить качественные[10] и традиционные количественные методы анализа данных. Так, например, возможности при ДСМ-обработке числовых данных использовать для уточнения явного вида числовых зависимостей внутри м-башен традиционные средства регрессионного анализа, либо системы *Автоматизированного Открытия* (*Automated Discovery & Deriving Laws Through Analysis of Processes and Equations*) - см., в частности, [*Zytkow 86, Langley 89, Zytkow 96*] (см. подробнее **Раздел 1.4**), позволяют естественным образом погрузить названные процедуры численного анализа данных в унифицированную логическую среду ДСМ-рассуждения. Это, в свою очередь, позволяет говорить о теоретических основаниях построения интеллектуальных систем, объединяющих рассуждения - дедукцию, индукцию, абдукцию и вывод по аналогии - с традиционными вычислительными методами. Таким образом, говоря о логических средствах квази-аксиоматических теорий как о платформе для синтеза познавательных процедур (см., например, [*Финн 99* и др.]), можно включить в такое объединение и вычислительные методы (в частности, методы распознавания образов и методы статистического анализа данных).

Еще одно заслуживающее обсуждения интересное обстоятельство связано с возможностью соединения метрических и "дометрических" (т.е. качественных) характеристик числовых данных в рамках ДСМ-рассуждения. Предложенная в **Разделе 3.3.2** система правил правдоподобного вывода (для числовых векторов) может быть модифицирована за счет включения в нее качественных оценок (например, типа "*близко-далеко*", "*ближе границы - дальше границы*" и т.п.) для соответствующих количественных величин, характеризующих расстояния между исследуемыми объектами.

Можно предложить (см., например, работу [*Забежайло 93*]) по крайней мере два варианта введения названных метрических соображений в ППВ типа **I**:

1. При вычислении сходства объектов будем оценивать взаимную "близость" столбцов в порождаемой м-башне (т.е. будем считать допустимым, чтобы м-башню образовывали лишь столбцы, расположенные на расстоянии не более заданной величины ρ друг от друга).

---

[4] Понятно, что тип данных для этих характеристик - *числовые* вектора (или даже вектора числовых интервалов, учитывающие, помимо прочего, и ошибки измерений соответствующих величин).

[5] В ситуации, когда учитывается лишь наличие или отсутствие соответствующей физиологической активности.

[6] В ситуации, когда учитываются степени градации соответствующих физиологических активностей.

[7] В ситуации, когда степень градации соответствующих физиологических активностей представлена конкретным числом.

[8] В ситуации, когда степень градации соответствующих физиологических активностей представлена конкретной числовой величиной (с учетом ошибок ее измерения).

[9] В частности - структурных формул физиологически активных химических соединений.

[10] А средства ДСМ-рассуждения следует отнести именно к таким инструментам анализа данных.

2. При проверке выполнимости условия ***запрета на контрпримеры*** будем учитывать лишь те контрпримеры (столбцы “противоположного” знака в матрице-универсуме **X**), что расположены от рассматриваемой м-башни на расстоянии, не превышающем ρ.

В определении ППВ типа **II** также возможен учет метрических соображений, осуществляемый в рамках подобной схемы:

1. При поиске контрпримеров рассматриваются продолжения уже построенных м-башен лишь на те “новые” объекты-столбцы, которые расположены не далее, чем на расстоянии ρ от одного из столбцов, образующих эту м-башню.
2. При подборе м-башен, “продолжаемых” на доопределяемый столбец-объект рассматриваются лишь те, которые расположены от него на расстоянии, не превышающем ρ.

Таким образом, введение в формулировки решающих предикатов нового метрического параметра ρ позволяет построить соответствующее семейство *параметрических* решающих предикатов и *параметрических* ППВ как для работы с числовыми векторами, так и с векторами числовых интервалов. (Легко видеть, что при $\rho = \infty$ обе версии - параметрическая и непараметрическая - решающих предикатов и ППВ будут удовлетворяться одними и теми же объектами).

Перейдем к обсуждению эффектов немонотонности в ДСМ-рассуждениях. Традиционные формальные системы, изучающие проблемы выводимости формул из фиксированных множеств аксиом посредством использования корректных правил – правил *достоверного вывода* в принятой нами ранее терминологии, - обладают свойством монотонности (по длине вывода) в том смысле, что все теоремы какой-либо теории такого типа содержатся в множестве теорем всякого расширения этой теории. При таком подходе, фактически, аксиомы и правила вывода – “компактная” форма представления полной информации о теоремах соответствующей формальной системы.

В свою очередь, исследования в области интеллектуальных систем со времени возникновения этого направления связаны с проблемами анализа рассуждений по “здравому смыслу” (ЗС-рассуждений – см., например, [*Winograd 80*] и др.). Значительная часть этих работ посвящена попыткам формализовать ЗС-рассуждения в рамках традиционных логических систем. Однако при построении прикладных интеллектуальных систем для открытых предметных областей подобного рода формализации оказались неадекватными: решающим обстоятельством оказалась необходимость принимать решения и действовать в условиях ограниченности знаний о “реальном мире” - т.е. в условиях *неполноты* информации. Практические приемы, используемые в такого рода решениях в прикладных компьютерных системах, как правило, основаны на аналогии с действиями человека в подобных ситуациях: допустимыми часто провозглашаются “следствия”, для которых нет точных доказательств, но которые, тем не менее, получены “естественным путем” и выглядят достаточно “правдоподобными”. Вполне естественно, что эти “следствия”, полученные на основе *неполной* информации, могут впоследствии оказаться “невыводимыми” на базе более полных данных. Необходимость учитывать сформулированный принцип *немонотонности* вывода в рамках ЗС-рассуждений привела, в частности, к развитию соответствующих нестандартных логических средств (см., например, обзор работ по немонотонным рассуждениям, выполненным группой Д.Маккарти в уже цитированном номере журнала Artificial Intelligence – [*AI 80*]). Аналогичные эффекты имеют место и в системах реализации правдоподобных рассуждений: при пополнении множества **Σ‘** фактов квази-аксиоматической теории **Γ** = <**Σ**,**Σ‘**,**ℜ**> новыми элементами, вообще говоря, возможны и ситуации, когда ряд зависимостей в множестве **Σ**, выполнявшихся до такого пополнения, оказывается невыполнимыми после него. Ниже (в **Разделе 6.1** отдельной **Главы 6**) мы приведем пример названного эффекта *немонотонности* ДСМ-рассуждения.

Предъявленные соображения позволяют вести речь по крайней мере о двух принципиально различных версиях ДСМ-рассуждения:

- *монотонной* (когда каждое из ранее порожденных, либо имевшихся в исходном состоянии процесса рассуждений истинностных значений) не пересматривается (не ревизуется) на последующих шагах ДСМ-рассуждения. При этом если одна и та же пара <*O*,*P*> может выводиться на разных шагах рассуждения, то это не приводит к противоречию, т.к. считается, что разное число шагов ДСМ-вывода порождает разные *степени правдоподобия*; и
- *немонотонной*, когда из всех истинностных значений, приписываемых каждой паре вида <*O*,*P*> на различных шагах ДСМ-рассуждения, актуальным (и, возможно, пересматриваемым на последующих шагах[11]) считается лишь значение, порождаемое на текущем шаге.

Касаясь вкратце[12] возможных *стратегий* ДСМ-рассуждения - комбинаций ППВ, обеспечивающих ту или иную конкретную форму цепочки рассуждения, и, соответственно, те или иные конкретные результаты рассуждения, отметим, что в зависимости от цели рассуждения допустим ряд "стандартных" стратегий, наиболее часто используемым среди которых являются:

- стратегия "все гипотезы" (при использовании которой сперва - в порядке возрастания их силы, т.е. от $\mathbf{I}^{+}_{\mathbf{a}}$ и далее - применяются все возможные правила типа **I**, затем в аналогичном порядке[13] применяются правила типа **II**, после чего вновь полученные результаты присоединяются к обучающей выборке, и процесс итерируется по описанной выше в **Главе 2** схеме); и
- стратегия "полезные гипотезы" (при использовании которой делается попытка порождать лишь те гипотезы, которые могут быть использованы для доопределения требующих прогноза элементов исходной выборки примеров и контрпримеров).

Завершая представление используемых в ДСМ-рассуждении правил правдоподобного вывода, обратим внимание на еще одно соображение фундаментального характера. Все приведенные выше построения были осуществлены для работы с *бинарным* отношением причинности: $V ==>_2 W$.

Тем не менее, следуя предложенной выше схеме, можно рассмотреть и более сложный вариант - *тернарное* отношение причинности вида

$$(V,S) ==>_2 W$$

где V рассматривается как причина для W лишь в отсутствии "препятствий" из множества S. В целом схема уже представленной конструкции ДСМ-рассуждения может быть воспроизведена и в этом случае: определив сходство объектов, можно перейти к формулировке решающих предикатов и ППВ, завершив весь процесс характеристикой стратегий ДСМ-рассуждения. Разумеется, что такое развитие выразительных возможностей на техническом уровне потребует достаточно серьезных усилий (здесь, в частности, будут необходимы и корректировка критерия достаточности оснований для принятия результатов правдоподобного вывода, и построение механизмов порождения "препятствий" для выводимых стандартными ДСМ-средствами причин и др.). Возникающий таким образом (и весьма обширный - !) круг вопросов выходит далеко за рамки предпринятого в настоящей работе анализа и потому заслуживают отдельного обстоятельного обсуждения, значительная часть которого может быть найдена в *работах* [*Финн 83, Михеенкова 86, Михеенкова 87, Финн 93-а* и др.]

---

[11] Разумеется, до достижения состояния стабилизации ДСМ-рассуждения.

[12] Подробное обсуждение стратегий ДСМ-рассуждения можно найти, например, в работе [*Финн 10*].

[13] Т.е. стартуя с использования ЭЗ, порожденных наиболее слабыми правилами, и продвигаясь в сторону их "усиления".

Глава 4

Алгоритмы и оценки сложности вычислений в рамках ДСМ-ИАД.

4.1 Основные элементы ДСМ-алгоритмики восстановления эмпирических зависимостей из данных. Проблема быстрого роста объемов исчерпывающего ДСМ-перебора вариантов.

Принимая во внимание конкретную структуру правил правдоподобного вывода ДСМ-метода (т.е. учитывая необходимость перебирать возможные пары вида <СТРУКТУРА ОБЪЕКТА, СВОЙСТВА ОБЪЕКТА>), легко понять, что ключевой элемент алгоритмики ДСМ-рассуждения - это механизмы порождения “гипотез” - “кандидатов”, для которых далее должно быть осуществлено соответствующее тестирование (т.е. проверка, удовлетворяет ли конкретный “кандидат” требованиям того или иного решающего предиката, того или иного правила правдоподобного вывода - см. выше **Главу 3**).

Здесь возможны, по крайней мере, две “базовые” модели порождения гипотез (назовем их условно моделями **M1** и **M2**) - см. **Схемы 4.1.1** и **2**:

Схема 4.1.1

Стратегия **M1:**

все, что можно — получаемые результаты

Реализует следующую последовательность действий:

о

Схема 4.1.2

Стратегия **M2**:

перспективные варианты — их, формируя рекомендации по порождению новых вариантов — направленным образом, т.е. с учетом предыдущих шагов, - новые перспективные варианты

Реализует следующую последовательность действий:

o

Каждая из представленных моделей, в свою очередь, в той или иной форме реализует обход "лабиринта" (более точно – *диаграммы* [1]всех локальных ДСМ-сходств, порождаемых на заданном множестве примеров соответствующего знака) - см. **Схему 4.1.3**.

Рассмотрим более детально обсуждаемую диаграмму всех возможных сходств в случае ДСМ-обработки данных, представленных множествами (см. также ранее **Разделы 3.1** и **3.2**) В данном случае корневым вершинам деревьев, входящих в эту диаграмму, сопоставлены множества образующих, характеризующие структуры исходно заданных объектов из множества $\boldsymbol{O} = \{O_1, O_2, \ldots, O_n\}$. (При этом все образующие берутся из соответствующего универсума $\mathbf{U^{(1)}} = \{c_1, c_2, \ldots, c_m\}$). Далее каждая из вершин первого уровня, имеющая индекс **i.j** , порождается при вычислении локального сходства (т.е. в данном случае - пересечения) объектов (в данном случае - множеств) O_i и O_j , представленного соответствующим подобъектом (в данном случае - множеством[2]) $O_{i.j}$. Затем по той же схеме порождаются вершины второго уровня, имеющие индекс вида **i.j.l** , и т.д., пока не будет порождена самая последняя вершина с индексом **1.2. ... .n** , соответствующая подобъекту $O_{1.2. \ldots .n}$.

Перебор всех возможных вершин построенной диаграммы всех локальных ДСМ-сходств осложнен двумя очевидными обстоятельствами:

- число вершин этой диаграммы, вообще говоря, *экспоненциальным образом* зависит от **n** - размера исходного множества объектов[3] (т.е. растет *экспоненциально быстро* с ростом размера исходного множества примеров);
- в зависимости от конкретной структуры исходных объектов в множестве $\boldsymbol{O} = \{O_1, O_2, \ldots, O_n\}$ тем или иным промежуточным вершинам этой диаграммы будет соответствовать пустое множество (а значит все "дочерние[4]" вершины для вершин такого типа не следует рассматривать при порождении ДСМ-гипотез, т.к. им также будет сопоставлено пустое множество).

Целью исчерпывающего перебора нетривиальных (т.е. таких, которым сопоставлено непустое локальное сходство) вершин рассматриваемой диаграммы будет поиск тех из них, у которых сопоставленное им локальное сходство оказывается также *глобальным сходством* (в смысле определений, данных в **Разделе 3.1**)[5]. Именно они и порождают те объекты, которые удовлетворяют предикату простого[6] сходства и могут служить исходным материалом для проверки остальных условий, используемых при формулировке решающих предикатов ДСМ-метода. Таким образом, из диаграммы всех локальных ДСМ-сходств требуется выделить под-диаграмму, характеризующую взаимную вложимость замыканий

---

[1] Еще более точно – диаграммы взаимной вложимости всех локальных ДСМ-сходств соответствующего знака, порождаемой на заданном множестве примеров.

[2] В ряде случаев - пустым!

[3] Нетрудно убедиться, что число вершин этой диаграммы определяется величиной $2^n\text{-}1 \leq O(2^n)$.

[4] Т.е. порождаемые вычислением *локального сходства* с ними.

[5] Т.е. *замкнутым* множеством в смысле определенного в **Разделе 3.1** понятия замыкание Галуа.

[6] В данном рассмотрении - положительного сходства.

Галуа (глобальных сходств), порождаемых на соответствующем множестве примеров. Именно эта диаграмма замыканий далее и дает возможность проверять выполнимость на ее элементах дополнительных логических условий, с помощью которых и формулируются решающие предикаты (а с ними – и правила правдоподобного вывода) ДСМ-метода.

Те или иные конкретные механизмы сокращения перебора при поиске ДСМ-гипотез (глобальных сходств) как раз и ставят свое целью стремление “просматривать”, по возможности, лишь *значимые* (т.е. порождающие непустые локальные сходства) вершины этого дерева, в свою очередь обходя стороной те, которым сопоставлено пустое множество образующих.

За несколько десятилетий развития ДСМ-технологий восстановления зависимостей из данных был предложен достаточно разнообразный набор конкретных алгоритмов, реализующих тот или иной способ порождения диаграммы всех глобальных сходств (см. например, соответствующие публикации из *Перечня*, приведенного в *Приложении 1* к [*Финн 11*], а также [*ДСМ АПГ*] и др.) доопределения с их помощью новых (предложенных для прогноза их свойств) примеров. Если свести вместе основные идеи[7], использованные при построении этих алгоритмов, то можно было бы получить алгоритм примерно следующего вида:

Алгоритм Г-Ц-СЛ-ЗНС[8] (неформальное описание).

Реализуется идея последовательного направленного выбора образующих, совокупность которых и порождала бы искомую ДСМ-гипотезу (а также осуществляемое с ее помощью доопределение текущего нового объекта). Анализ ДСМ-доопределимости обоих знаков - как позитивной, так и негативной - организован симметричным образом.

Для каждого из доопределяемых новых объектов, сперва рассматривается то подмножество множества минимальных глобальных сходств соответствующего знака, элементы которого вкладываются в этот объект. В ситуации, когда таковых нет (т.е. когда ни одно из минимальных сходств не вкладывается в текущий доопределяемый объект) осуществляется переход к новому требующему доопределения объекту.

Если же среди них найдется хотя бы один элемент (набор образующих), который не вкладывается ни в один из контрпримеров, то искомая цель достигнута - найдена необходимая ДСМ-гипотеза, порождающая соответствующее доопределение текущего нового объекта.

В противном случае осуществляется перебор образующих, которые могли бы быть добавлены к уже найденным, порождая при этом невложимость в контрпримеры вместе с вложимостью в текущий доопределяемый объект. Для этого, прежде всего порождаются те элементы верхней границы Диаграммы Всех Глобальных Сходств (т.е. максимальные глобальные сходства), которые содержат в себе (как подмножеста образующих) уже отобранные нами элементы нижней границы этой Диаграммы (т.е. соответствующие минимальные глобальные сходства). Те из отобранных элементов верхней границы, для которых не выполняется условие Запрета на Контрпримеры (т.е. для которых имеет место вложимость по крайней мере в один из контрпримеров)

---

[7] А это, в первую очередь:
- выделение ***Границ*** (верхней и нижней по вложению) в Диаграмме Всех Глобальных Сходств;
- анализ ***Цепей*** частичного порядка (по вложимости) в Диаграмме Всех Глобальных Сходств;
- учет структуры так называемых - **С**топ **Л**истов (для контрпримеров);
- перебор замыканий (Галуа) с учетом лексикографического порядка формирующих их исходно заданных объектов;
- выполнение исчерпывающего перебора вариантов при поиске зависимостей, начиная с ***З**амыкания **Н**аименьших* (обеспечивающих вложимость гипотезы в доопределяемый объект) ***С**ходств* (при поиске образующих, составляющих каждую искомую ДСМ-гипотезу, пригодную для доопределения текущего нового объекта.

[8] **Г-Ц-СЛ-ЗНС** – сокращение для «формулы» ***Г**раницы-**Ц**епи-**С**топ **Л**исты-**З**амыкания **Н**аименьших **С**ходств.*

отбрасываются. (При этом пустота оставшегося множества элементов верхней границы - сигнал для перехода к анализу следующего доопределяемого объекта).

Далее, принимая во внимание, что для каждого минимального сходства любое из содержащих его максимальных сходств определяет и соответствующее множество цепей[9] Диаграммы Всех Глобальных Сходств, поиск подходящей (для доопределения) ДСМ-гипотезы - некоторого множества образующих - можно вести уже перебором лишь вдоль таких цепей (т.е. лишь внутри) подмножества образующих, определяемых разностью максимального и минимального элементов всех этих цепей.

Более того, найденное таким образом подмножество образующих должно быть очищено (по понятным соображениям) от тех элементов, которые не вкладываются в текущий доопределяемый объект. (Переход при такой “очистке” к пустому множеству образующих - сигнал к смене анализируемого множества цепей, а при завершении их перебора - к смене текущего анализируемого объекта).

Наконец, перебор элементов “очищенного” представленным выше способом множества образующих, подразумевающий вычисление ДСМ-замыкания (и, разумеется, проверку вложимости найденного замыкания в доопределяемый новый объект) для объединения образующих из соответствующего минимального глобального сходства и вновь отбираемых «расширений» (дополнительных образующих) к нему, удобно вести с учетом встречаемости таких добавляемых в каждом конкретном случае образующих в СТОП-листах[10] контрпримеров.

o

Детальным анализом алгоритмики восстановления диаграммы всех глобальных сходств, порождаемых из заданного множества примеров мы займемся позднее (см. **Главу 5**), а здесь сфокусируем наше внимание на *комбинаторных характеристиках* – оценках *сложности вычислений* при порождении - как этого объекта в целом, так и некоторых его специальных структурных компонентов – максимальных и минимальных (по вложению соответствующих множеств образующих) глобальных сходств, а с ними - нижней и верхней границ такой диаграммы, на оценках числа всех вершин такой диаграммы, а также на сложности поиска в диаграмме вершин с заданными характеристиками (например, сформированных точно заданным числом образующих или же точно заданным числом объектов из исходного множества ДСМ-примеров).

К сожалению (и это будет показано ниже в данной **Главе**) ряд переборных задач в только что перечисленном нами списке принадлежит к известным классам трудных комбинаторных проблем. Демонстрация подобных фактов – весомый аргумент в пользу актуальности[11] проблематики управления перебором при реализации алгоритмики ДСМ-ИАД (о чем мы отдельно и поговорим далее – в **Главе 5**).

Переходя к формальному описанию перечисленных переборных задач, не будем забывать, что каждая из них естественным образом возникает[12] в рамках следующей неформальной схемы (контекста):

Даны:

1) множество (примеров) *объектов*, обладающих целевыми *свойствами*, и множество[13] (контрпримеров) объектов, не обладающих целевыми свойствами,

а также

---

[9] В смысле задаваемого вложимостью глобальных сходств отношения доминирования.

[10] СТОП-листом для конкретной образующей мы условились называть то множество контрпримеров, в которые она не входит.

[11] И в первую очередь – для использования обсуждаемой техники ДСМ-ИАД в реальных приложениях.

[12] Может быть сформулирована как естественным образом возникающий вопрос о тех или иных комбинаторных характеристиках процедур реализации требуемых такой схемой действий.

[13] В исходном состоянии, - возможно, пустое.

2) *новый объект* (или же некоторое явным образом представленное множество таких объектов).

Требуется:

оценить *наличие* (или отсутствие) *целевых свойств* у нового *объекта* (новых объектов из заданного множества), т.е.

- дать соответствующий *прогноз* (о наличии целевых свойств) и
- предъявить *основания* (неоспоримые *аргументы*), позволяющие *принять* этот прогноз.

Ранее в **Разделе 2.1** уже был предложен вариант уточнения сформулированной выше схемы, в основе которого –техника восстановления соответствующих классов сходства исходно заданных объектов а также покрывающих их классов эквивалентности. Общий вид используемой здесь понятийно-процедурной конструкции может быть описан следующим образом:

1. Каждый представленный в исходных данных объект может быть охарактеризован набором признаков (например, определенным набором тех или иных последовательностей 0 и 1 в его «теле»).
2. Все множество таких признаков может быть собрано в (пополняемый и расширяемый) каталог – «алфавит» («атомарных») признаков. Каждый уже идентифицированный ранее объект может быть «закодирован» (представлен) определенным набором (множеством) таких признаков.
3. На множестве идентификационных признаков могут быть заданы отношения (т.е. описания исходно заданных объектов могут стать не множествами признаков, а, например, графами, цепочками и т.п.). Описания изучаемых объектов могут быть расширены также числовыми значениями тех или иных параметров (*множества признаков + отношения + числовые характеристики*).
4. В случае пустоты исходно заданного множества контрпримеров, к имеющимся коллекциям объектов (примерам изучаемых объектов) путем дополнительного анализа[14] изучаемой предметной области могут быть добавлены также наборы контрпримеров – комплексов признаков из исходного «алфавита», которые тем не менее на являются объектами, обладающими целевыми признаками.
5. На множествах примеров и контрпримеров может быть определена (алгебраическая) операция сходства *, удовлетворяющая трем условиям:

   - a*a = a
   - a*в = в*a
   - (a*в)*c = a*(в*c) = a*в*c

   Например: * определяемая как пересечение множеств, или же как выделение множества всех максимальных общих подграфов двух графов, множества всех максимальных общих подцепочек двух цепочек и т.п.
6. С помощью такой операции * на примерах и контрпримерах строятся классы сходства: для каждого заданного объекта исходной коллекции примеров (и, соответственно, контрпримеров) класс сходства - это множество всех известных объектов, которые сходны[15] с ним.
7. Диагностика (идентификация) новых объектов производится с помощью порождаемых таким образом классов сходства: возможность отнесения нового – диагностируемого - объекта к какому-либо из уже реконструированных (по примерам[16]) классов сходства позволяет утверждать, что это – еще один объект,

[14] Проводимого с привлечением экспертов по изучаемой предметной области.
[15] Т.е. операция сходства в каждом таком случае дает непустой (ненулевой) результат.
[16] Или же контрпримерам.

аналогичный (см. класс сходства) ранее уже наблюдавшимся (в исходной обучающей выборке примеров и контрпримеров).

8. Более детально идея техники отнесения нового объекта к классу описывается следующим образом:

- класс сходства есть объединение некоторых классов эквивалентности[17],
- по построенному классу сходства можно восстановить (с учетом имеющегося контекста – т.е. множества примеров и контрпримеров, задействованных в обучении) порождающие его классы эквивалентности;
- каждый класс сходства при восстановлении классов эквивалентности, объединением которых он является, будет характеризоваться непустыми (ненулевыми) результатами вычисления уже заданной операции сходства. Таким образом каждый восстанавливаемый здесь класс эквивалентности будет определяться теми примерами из исходной выборки, результат вычисления операции сходства для которых будет одним и тем же[18] (например, в случае простого сходства на множествах – всех примеров, содержащих общее подмножество образующих); следовательно
- каждый класс эквивалентности в таком случае – это так называемая орбита (множество примеров-родителей) для соответствующего сходства - ДСМ-гипотезы. (При этом, разумеется, один и тот же пример из исходной ДСМ-выборки может попадать в несколько пересекающихся, но не совпадающих классов эквивалентности[19]);
- в дополнение к этому порождаемые на исходном множестве примеров и контрпримеров классы сходства а также соответствующие классы эквивалентности могут характеризоваться выполнимостью дополнительных логических условий – в частности: запрета на контрпримеры, условия единственности причины и т.п..

Теперь предметом нашего основного интереса будет собственно построение *оценок вычислительной сложности* ряда задач порождения (по заданному множеству примеров) обсуждаемых ДСМ-классов эквивалентности с заданными характеристиками, а также задачи о вычислении мощности множества всех таких классов. Механизм построения соответствующих оценок вычислительной сложности мы сперва подробно рассмотрим на примере работы с исходными объектами, представленными как множества «признаков» - образующих из некоторого наперед заданного «алфавита». Операция сходства в данном случае будет формализована как операция пересечения множеств ∩. Далее будут представлены оценки сложности соответствующих переборных задач, возникающих в случае операций и с другими (уже обсуждавшимися ранее в **Разделе 3.1**) типами данных.

---

[17] Порождаемых всеми возможными в данном случае версиями соответствующего отношения эквивалентности.

[18] Именно этим фактом дополнительно к рефлексивности (идемпотентности соответствующей операции сходства) и симметричности обеспечивается также и транзитивность соответствующего (уже используемой операции сходства) отношения сходства (суженного представленным дополнительным условием до эквивалентности).

[19] Для, разумеется, различных – см. различные ДСМ-гипотезы – конкретных вариантов отношений эквивалентности, участвующих в формировании соответствующих классов сходства.

4.2. Оценки сложности вычислений при работе с булевскими данными.

Начнем с некоторых базовых понятий, ранее уже обсуждавшихся нами в **Разделе 3.1**. Пусть $\mathbf{U} = \{a_1, a_2, \dots, a_n\}$ универсум образующих[1], а $\mathbf{\Omega} = \{O_1, O_2, \dots, O_m\} \subseteq 2^{\mathbf{U}}\backslash\varnothing$ - множество объектов (т.е. непустых множеств образующих), построенных над универсумом $\mathbf{U}$[2]**.**

4.2.1. Определение

Множество $O_{i1,i2,..,is} = O_{i1} \cap O_{i2} \cap \dots \cap O_{is}$ мы будем называть ***исчерпывающим сходством*** объектов из множества $\mathbf{\Omega}$ тогда и только тогда, когда:

(1) $O_{i1,i2,..,is} \neq \varnothing$, и

(2) $(\forall \mathbf{j})\ (([(O_j \in \mathbf{\Omega}) \,\&\, \neg(O_j \in \{O_{i1}, O_{i2}, \dots, O_{is}\})] \supset [(O_j \cap O_{i1,i2,..,is}) \subset O_{i1,i2,..,is}])$[3]

o

Пусть $O_{i1,i2,..,is} = O_{i1} \cap O_{i2} \cap \dots \cap O_{is} = \{a_{i1}, a_{i2}, \dots, a_{il}\}$ – исчерпывающее сходство (пересечение). Мы будем называть объекты $O_{i1}, O_{i2}, \dots, O_{is}$ - ***родителями*** исчерпывающего сходства $O_{i1,i2,..,is}$, множество $\{O_{i1}, O_{i2}, \dots, O_{is}\}$ - ***орбитой*** исчерпывающего сходства $O_{i1,i2,..,is}$, а образующие $a_{i1}, a_{i2}, \dots, a_{il}$ - ***образующими*** исчерпывающего сходства $O_{i1,i2,..,is}$. Множество всех исчерпывающих сходств, порождаемых над универсумом $\mathbf{U}$ из множества объектов $\mathbf{\Omega}$, мы будем обозначать посредством ***ИС***$(\mathbf{\Omega})$.

4.2.2. Определение

Длиной исчерпывающего сходства будем называть мощность множества его образующих. ***Размером орбиты*** исчерпывающего сходства будем называть мощность множества его родителей.

o

Пусть $\mathbf{U} = \{a_1, a_2, \dots, a_n\}$ - универсум образующих, а

$$\mathbf{\Omega} = \{O_1, O_2, \dots, O_m\} \subseteq 2^{\mathbf{U}}\backslash\varnothing$$

есть множество объектов (т.е. непустых множеств образующих), построенных над универсумом $\mathbf{U}$[4]**.** Рассмотрим универсум $\mathbf{U^*} = \{a^*_1, a^*_2, \dots, a^*_m\}$ и множество объектов

$$\mathbf{\Omega^*} = \{O^*_1, O^*_2, \dots, O^*_n\} \subseteq 2^{\mathbf{U^*}}\backslash\varnothing\ ,$$

построенных над универсумом $\mathbf{U^*}$ по следующим правилам:

- пусть $O_i = \{a_{i1}, a_{i2}, \dots, a_{is}\} \subseteq \mathbf{U}$ - i-тый элемент множества $\mathbf{\Omega}$**:** $O_i \in \mathbf{\Omega}$**.**
- Каждому объекту O_i из множества $\mathbf{\Omega}$ сопоставляется образующая a^*_i из универсума $\mathbf{U^*}$, а каждому объекту O^*_i из множества $\mathbf{\Omega^*}$ - образующая a_i из универсума $\mathbf{U}$.
- Каждый объект O^*_i в множестве $\mathbf{\Omega^*}$ сформирован теми и только теми образующими a^*_i из множества $\mathbf{U^*}$, для которых сопоставленные им объекты O_i содержат сопоставленную объекту O^*_i образующую a_i.

4.2.3. Определение

Пусть заданы универсум $\mathbf{U}$ и $\mathbf{\Omega}$ - некоторое множество объектов над ним. Построенные по описанным выше правилам множества $\mathbf{U^*}$ и $\mathbf{\Omega^*}$ будем называть (соответственно) сопряженным универсумом и сопряженным множеством объектов над ним.

o

Рассмотрим некоторые свойства сопряженных множеств объектов над сопряженными универсумами. Очевидным следствием **Определения 4.2.3** является

---

[1] При этом, не теряя общности, мы будем считать, что в $\mathbf{U}$ нет одинаковых элементов.

[2] Т.е. множество объектов, построенных из образующих $a_1, a_2, \dots, a_n$.

[3] Другими словами, каждое исчерпывающее сходство $O_{i1,i2,..,is}$ - не пусто и образовано в точности всеми теми объектами из $\mathbf{\Omega}$, которые при пересечении (вычислении сходства) дают именно множество $O_{i1,i2,..,is}$.

[4] Т.е. множество объектов, построенных из образующих $a_1, a_2, \dots, a_n$.

4.2.4. Лемма

Пусть заданы **U** и **Ω**. Тогда:

$$[\mathbf{U^*}]^* = \mathbf{U} \text{ и } [\mathbf{\Omega^*}]^* = \mathbf{\Omega}.$$

o

Более сложные взаимоотношения сопряженных универсумов представляет

4.2.5. Теорема (*о двойственности***).**

Пусть заданы: универсум **U**, множество объектов над ним - **Ω** и натуральное число k. В множестве ***ИС***(**Ω**) всех исчерпывающих сходств объектов из **Ω** существует исчерпывающее сходство длины k тогда и только тогда, когда в множестве ***ИС***(**Ω***) всех исчерпывающих сходств объектов из **Ω*** найдется исчерпывающее сходство с орбитой размера k.

Доказательство.

==>. Предположим противное. Тогда в ***ИС***(**Ω**) существует некоторый элемент O_0 вида $O_0 = \{a_{i1}, a_{i2}, \dots, a_{ik}\}$, но при этом в ***ИС***(**Ω***) не существует элемента O'_0 вида

$$O'_0 = \{ O^*_{l1} \cap O^*_{l2} \cap \dots \cap O^*_{lk}\}.$$

Рассмотрим множество O^*_0, представляющее собой пересечение объектов $O^*_{i1}, O^*_{i2}, \dots, O^*_{ik}$ из **Ω*.** Т.к.

$$O_0 \neq \varnothing, \text{ то } O^*_{i1} \cap O^*_{i2} \cap \dots \cap O^*_{ik} = O^*_0 \neq \varnothing,$$

кроме того, т.к. в **U** нет повторяющихся элементов, то O^*_0 - исчерпывающее сходство, а размер его орбиты - ровно k элементов. Это противоречит сделанному нами предположению (об отсутствии A'_0 заданного вида: достаточно положить $O'_0 = O^*_0$).

<==. Предположим противное. Тогда в ***ИС***(**Ω***) найдется элемент вида

$$O^*_0 = O^*_{i1} \cap O^*_{i2} \cap \dots \cap O^*_{ik},$$

но при этом в ***ИС***(**Ω**) не должно оказаться элемента длины ровно k. Рассмотрим множество

$$O_0 = \{a_{i1}, a_{i2}, \dots, a_{ik}\}.$$

Для каждого **j** из множества {**i1**, **i2**, ... , **ik**} рассмотрим соответствующее подмножество $\mathbf{\Omega_j}$ объектов из множества **Ω**, причем таких, что каждый из них содержит соответствующую образующую $\mathbf{a_j}$. Т.к. $A^*_0 \in$ ***ИС***(**Ω***), то каждое из таких подмножеств $\mathbf{\Omega_j}$ не пусто, а их пересечение есть в точности O_0. Это противоречит сделанному нами предположению.

oo

Из **Теоремы 4.2.5** и **Леммы 4.2.4** очевидным образом вытекает

4.2.6. Следствие

Пусть заданы: универсум **U**, множество объектов над ним - **Ω** и натуральное число k. В множестве исчерпывающих сходств объектов из **Ω** существует исчерпывающее сходство с орбитой размера k тогда и только тогда, когда в множестве исчерпывающих сходств объектов из **Ω*** найдется исчерпывающее сходство длины k.

o

Полезным для последующих рассмотрений утверждением является также

4.2.7. Лемма

Пусть заданы **U** и **Ω**. Пусть также $O_{i1,i2,..,is} = O_{i1} \cap O_{i2} \cap \dots \cap O_{is}$ есть произвольное исчерпывающее сходство объектов $O_{i1}, O_{i2}, \dots, O_{is}$ из **Ω**, формирующих множество родителей этого исчерпывающего сходства $O_{i1,i2,..,is}$. Тогда множество $\{O_{i1}, O_{i2}, \dots, O_{is}\}$ есть *класс эквивалентности* на **Ω**, сформированный по отношению ***E***, определяемому как «*иметь (содержать) общую часть* $O_{i1,i2,..,is}$ ».

Доказательство.

Рефлексивность и симметричность отношения ***E*** есть следствие наличия этих свойств у используемой нами операции сходства – пересечения множеств ∩. Транзитивность ***E*** есть

следствие *исчерпываемости* сходства (см. выше - **Определение 4.2.1**): если $O_{\mathbf{i,j,k}}$ есть общая часть объектов $O_{\mathbf{i}}$ и $O_{\mathbf{j}}$, и одновременно $O_{\mathbf{i,j,k}}$ есть общая часть объектов $O_{\mathbf{j}}$ и $O_{\mathbf{k}}$, то $O_{\mathbf{i,j,k}}$ есть общая часть также и объектов $O_{\mathbf{i}}$ и $O_{\mathbf{k}}$.

o

Таким образом, оперируя с исчерпывающими сходствами, мы фактически имеем дело с двумя характеристиками каждого соответствующего класса эквивалентности $\boldsymbol{E}(O_{\mathbf{i1,i2,..,is}})$ (участвующего в покрытии порождаемых на множестве примеров $\boldsymbol{\Omega}$ операцией $\cap$ классов сходства):

- *числом* его *элементов* (размером множества родителей соответствующего исчерпывающего сходства $O_{\mathbf{i1,i2,..,is}}$) и
- *числом образующих*, которые формируют рассматриваемое (и, фактически, порождающее этот класс эквивалентности) исчерпывающее сходство $O_{\mathbf{i1,i2,..,is}}$ – его длиной (см. выше - **Определение 4.2.2**).

Определим пару отображений $\boldsymbol{s}$ и $\boldsymbol{g}$, связывающих множества $\mathbf{U}$ и $\boldsymbol{\Omega}^5$ следующим образом:

$$\boldsymbol{s} : 2^{\boldsymbol{\Omega}} \to 2^{\mathbf{U}}$$

и для

$$\forall \mathbf{X} = \{O_{\mathbf{i1}}, O_{\mathbf{i2}}, \dots , O_{\mathbf{is}}\} \in 2^{\boldsymbol{\Omega}}$$

имеем

$$\boldsymbol{s}(\mathbf{X}) = O_{\mathbf{i1}} \cap O_{\mathbf{i2}} \cap \dots \cap O_{\mathbf{is}};$$

кроме того

$$\boldsymbol{g} : 2^{\mathbf{U}} \to 2^{\boldsymbol{\Omega}}$$

и для

$$\forall \mathbf{Y} = \{a_{\mathbf{i1}}, a_{\mathbf{i2}}, \dots , a_{\mathbf{ik}}\} = \{O^*_{\mathbf{i1}}, O^*_{\mathbf{i2}}, \dots , O^*_{\mathbf{ik}}\} \in 2^{\mathbf{U}}$$

имеем

$$\boldsymbol{g}(\mathbf{Y}) = O^*_{\mathbf{i1}} \cap O^*_{\mathbf{i2}} \cap \dots \cap O^*_{\mathbf{ik}}$$

Легко видеть, что отображения $\boldsymbol{s}$ и $\boldsymbol{g}$ удовлетворяют соотношениям:

$$[\mathbf{X_1} \subseteq \mathbf{X_2}] \supset [\boldsymbol{s}(\mathbf{X_2}) \subseteq \boldsymbol{s}(\mathbf{X_1})]$$
$$[\mathbf{Y_1} \subseteq \mathbf{Y_2}] \supset [\boldsymbol{g}(\mathbf{Y_2}) \subseteq \boldsymbol{g}(\mathbf{Y_1})],$$

т.е. представляют собой соответствие Галуа [*Гусакова 87*, *Кон 68*]. Несложно также убедиться, что отображения $\boldsymbol{s}^*\boldsymbol{g}(\mathbf{X}) = \boldsymbol{g}(\boldsymbol{s}(\mathbf{X}))$ и $\boldsymbol{g}^*\boldsymbol{s}(\mathbf{Y}) = \boldsymbol{s}(\boldsymbol{g}(\mathbf{Y}))$ являются операторами замыкания, соответственно, на множествах $\boldsymbol{\Omega}$ и $\mathbf{U}$. Теперь, наконец, не составляет труда проверить, что **Определение 4.2.1** выделяет замкнутые (в смысле операторов $\boldsymbol{s}^*\boldsymbol{g}$ и $\boldsymbol{g}^*\boldsymbol{s}$) подмножества, соответственно в универсумах $\mathbf{U}^*$ (а вместе с ним - и в множестве $\boldsymbol{\Omega}$) и $\mathbf{U}$. Итак, множество неподвижных точек замыкания Галуа $\boldsymbol{s}^*\boldsymbol{g}$ (и, соответственно, $\boldsymbol{g}^*\boldsymbol{s}$) над множеством $\boldsymbol{\Omega}$ (и, соответственно, над $\boldsymbol{\Omega}^*$) определяет множество всех исчерпывающих сходств объектов из $\boldsymbol{\Omega}$ (и, соответственно, из $\boldsymbol{\Omega}^*$).

"Наглядное" представление о природе исчерпывающих сходств пересечений дает следующая *комбинаторная интерпретация*:

- пусть задан универсум образующих $\mathbf{U} = \{a_1, a_2, \dots , a_n\}$ и множество объектов $\boldsymbol{\Omega} = \{O_1, O_2, \dots , O_m\} \subseteq 2^{\mathbf{U}} \backslash \varnothing$ над универсумом $\mathbf{U}$;
- рассмотрим $\mathbf{n}$-мерный куб $\boldsymbol{BC}_{\mathbf{n}}$, пометив на нем $\mathbf{m}$ вершин, сопоставленных элементам множества $\boldsymbol{\Omega}$ по следующему закону:

  - каждому $O_{\mathbf{i}} = \{a_{\mathbf{i1}}, a_{\mathbf{i2}}, \dots , a_{\mathbf{is}}\} \in \boldsymbol{\Omega}$ сопоставляется вершина из $\boldsymbol{BC}_{\mathbf{n}}$, характеризуемая таким набором единиц и нулей, что единицы в ней стоят на позициях $\mathbf{i1}, \mathbf{i2}, \dots , \mathbf{is}$ и только на них;

  - если среди элементов множества $\boldsymbol{\Omega}$ есть одинаковые, то им сопоставляется одна и та же вершина куба $\boldsymbol{BC}_{\mathbf{n}}$;

---

[5] Более точно: связывающих множества $\mathbf{U}$, $\mathbf{U}^*$, $\boldsymbol{\Omega}$ и $\boldsymbol{\Omega}^*$.

- пересечению элементов любого подмножества мощности $\mathbf{m} \geq \mathbf{k} \geq 2$ в множестве $\mathbf{\Omega}$ сопоставляются (и помечаются):
  - вершина <0,0, ... ,0> в ситуации когда рассматриваемое пересечение пусто;
  - вершина, порождаемая единицами на тех позициях, номера которых - номера образующих из **U**, попавших в вычисленное пересечение;
- на множестве помеченных вершин $\mathbf{DV}(\boldsymbol{BC}_\mathbf{n},\mathbf{\Omega})$ естественным образом (по вложению сопоставленных помеченным вершинам множеств образующих из **U**) определяется отношение частичного порядка $\propto$;
- теперь помеченный фрагмент **n**-мерного куба $\boldsymbol{BC}_\mathbf{n}$ представляет диаграмму частично упорядоченного множества $<\mathbf{DV}(\boldsymbol{BC}_\mathbf{n},\mathbf{\Omega}),\propto>$, в которой (непустому) исчерпывающему сходству объектов $O_{i1}, O_{i2}, \dots, O_{is}$ из множества $\mathbf{\Omega}$ соответствует $O_{i1,i2,..,is} = O_{i1} \cap O_{i2} \cap \dots \cap O_{is}$ - точная нижняя граница (в смысле частичного порядка $\propto$) множества вершин **n**-мерного куба $\boldsymbol{BC}_\mathbf{n}$, сопоставленных объектам $O_{i1}, O_{i2}, \dots, O_{is}$, выделенная вместе со всеми O_0 - максимальными элементами частично упорядоченного множества $<\mathbf{DV}(\boldsymbol{BC}_\mathbf{n},\mathbf{\Omega}),\propto>$, для которых выполняется вложимость:

$$\varnothing \subset O_{i1,i2,..,is} \subseteq O_0.$$

Пусть заданы универсум **U** и множество объектов над ним - $\mathbf{\Omega}$. Обратимся к более подробному анализу свойств множества $\boldsymbol{GC}(\mathbf{U},\mathbf{\Omega})$ исчерпывающих сходств (неподвижных точек определенных нами замыканий Галуа) для заданных множеств **U** и $\mathbf{\Omega}$. Предварительные представления о возможном "объеме" множества $\boldsymbol{GC}(\mathbf{U},\mathbf{\Omega})$ дает

4.2.8. Лемма

Существуют $\mathbf{U_0}$ и $\mathbf{\Omega_0}$ такие, что мощность множества $\boldsymbol{GC}(\mathbf{U_0},\mathbf{\Omega}_0)$ растет экспоненциально быстро с ростом размеров множеств $\mathbf{U_0}$ и $\mathbf{\Omega_0}$.

Доказательство.

Пусть $\mathbf{U_0} = \{a_1, a_2, \dots, a_n\}$, а $\mathbf{\Omega_0} = \{O_1, O_2, \dots, O_n\}$, где

$$O_i = \mathbf{U_0} \setminus \{a_i\} = \{a_1, a_2, \dots, a_n\} \setminus \{a_i\}.$$

Легко видеть, что сходство (пересечение) любого непустого подмножества объектов из $\mathbf{\Omega}$ представляет собой соответствующее (т.е. порожденное именно этими объектами и только ими) исчерпывающее сходство. Число таких исчерпывающих сходств

$$S = \left| \boldsymbol{GC}(\mathbf{U_0},\mathbf{\Omega_0}) \right| = \sum_{i=1}^{n-2} C^{\mathbf{n}}_{\mathbf{i}} = 2^{\mathbf{n}} - n - 2 \leq O(2^{\mathbf{n}})$$

(здесь перечислены исчерпывающие сходства, порожденные каждыми **i** образующими из универсума $\mathbf{U_0}$).

o

4.2.9. Определение

Задачей ***ЧИСЛО КЛАССОВ ЭКВИВАЛЕНТНОСТИ*** **(**или ***ЧИСЛО КЭ)*** будем называть следующую задачу:

Дано: множества **U** и $\mathbf{\Omega}$.

Найти: число элементов в множестве $\boldsymbol{GC}(\mathbf{U},\mathbf{\Omega})$

$$S = \left| \boldsymbol{GC}(\mathbf{U},\mathbf{\Omega}) \right|$$

o

Точное представление о комбинаторной природе множества $\boldsymbol{GC}(\mathbf{U},\mathbf{\Omega})$ дает

4.2.10. Теорема

Задача ***ЧИСЛО КЭ*** перечислительно полна:

$$\textbf{\textit{ЧИСЛО КЭ}} \in \textbf{\textit{\#PC}}^{6}$$

---

[6] ***#PC*** представляет собой класс перечислительно полных (см., например, [*Гэри 82, Simon 77, Valiant 79, Valiant 79-a*] и др.) комбинаторных задач.

Доказательство.

Идея доказательства - продемонстрировать полиномиальную сводимость известной (см, например, [*Valiant 79*]) переборной задачи ***МОНОТОННАЯ ВЫПОЛНИМОСТЬ***[7] к рассматриваемой нами задаче ***ЧИСЛО КЭ***. Мы реализуем следующую схему сведения:

- по заданной функции **Ф** полиномиально быстро строится пара множеств $\mathbf{U_Ф}$,$\mathbf{\Omega_Ф}$, такая, что исчерпывающие пересечения объектов из $\mathbf{\Omega_Ф}$ над универсумом $\mathbf{U_Ф}$ "перечисляют" все нули функции **Ф**;
- размер множества единиц функции **Ф** (т.е. множества наборов, выполняющих эту функцию) есть размер дополнения множества нулей функции **Ф** до множества всех возможных наборов значений переменных $x_1,x_2, \dots ,x_n$.

Теперь более подробно. Пусть **Ф** имеет вид

$$\mathbf{Ф}(x_1,x_2, \dots ,x_n) = \mathbf{Ф_1} \,\&\, \mathbf{Ф_2} \,\&\, \dots \,\&\, \mathbf{Ф_m} = (x_{11} \vee x_{12}) \,\&\, (x_{21} \vee x_{22}) \,\&\, \dots \,\&\, (x_{m1} \vee x_{m2}).$$

Не теряя общности, будем считать, что в записи функции **Ф** все конъюнкты $\mathbf{Ф_i}$ различны[8].

Примитивным нулем номер i функции

$$\mathbf{Ф}(x_1,x_2, \dots ,x_n) = \mathbf{Ф_1} \,\&\, \mathbf{Ф_2} \,\&\, \dots \,\&\, \mathbf{Ф_i} \,\&\, \dots \,\&\, \mathbf{Ф_m}$$

будем называть такой булевский вектор $\mathbf{\sigma_i}$, что

	x_1	x_2	…	x_{i1}	x_{i1+1}	…	x_{i2-1}	x_{i2}	x_{i2+1}	…	x_{n-1}	x_n
σ_i	**1**	**1**	…	**0**	**1**	…	**1**	**0**	**1**	…	**1**	**1**

Первым сопряженным с примитивным нулем $\mathbf{\sigma_i}$ будем называть всякий такой булевский вектор $\mathbf{\sigma^*_i}$, что ровно одна из n-2 переменных множества

$$\mathbf{X_i} = \{x_1,x_2, \dots ,x_n\} \setminus \{x_{i1}, x_{i2}\}$$

принимает значение 0. (Несложно убедиться, что среди первых сопряженных с каждым примитивным нулем могут оказаться и некоторые другие примитивные нули функции **Ф**, а любой из нулей этой функции, не являющийся примитивным или первым сопряженным с примитивным, может быть представлен как суперпозиция некоторых примитивных и первых сопряженных с ними нулей).

Положим

$$\mathbf{U_Ф} = \{a^1, a^2, \dots , a^n\},$$

а множество $\mathbf{\Omega_{Ф,i}}$ для каждого $\mathbf{\sigma_i}$ определим как

$$\mathbf{\Omega_{Ф,i}} = \{\mathbf{U_Ф}, A^i_0, A^i_1, A^i_2 , \dots A^i_n\},$$

где

$$A^i_0 = \mathbf{U_Ф} \setminus \{a^{i1}, a^{i2}\},$$
$$A^i_1 = \mathbf{U_Ф} \setminus \{a^{i1}, a^{i2}\} \setminus \{a^1\},$$
$$A^i_2 = \mathbf{U_Ф} \setminus \{a^{i1}, a^{i2}\} \setminus \{a^2\},$$
$$\dots$$
$$A^i_n = \mathbf{U_Ф} \setminus \{a^{i1}, a^{i2}\} \setminus \{a^n\},$$

при этом, как нетрудно убедиться, окажется, что для каждого i имеет место равенство

$$A^i_0 = A^i_{i1} = A^i_{i2} = \mathbf{U_Ф} \setminus \{a^{i1}, a^{i2}\}.$$

В рассматриваемой ситуации всякий нуль функции **Ф** , который сопряжен с заданным примитивным нулем $\mathbf{\sigma_i}$, может быть представлен как исчерпывающее сходство[9] (а вместе с

---

[7] Речь идет о задаче о выполнимости монотонной булевской функции, представленной в виде 2-КНФ:
Дано: монотонная булевская функция **Ф** в виде 2-КНФ
$$\mathbf{Ф}(x_1,x_2, \dots ,x_n) = \mathbf{Ф_1} \,\&\, \mathbf{Ф_2} \,\&\, \dots \,\&\, \mathbf{Ф_m} = (x_{11} \vee x_{12}) \,\&\, (x_{21} \vee x_{22}) \,\&\, \dots \,\&\, (x_{m1} \vee x_{m2})$$
Найти: число наборов значений переменных $x_1,x_2, \dots ,x_n$, выполняющих функцию **Ф**.

[8] В противном случае функция **Ф** может быть приведена к нужному нам виду сортировкой (полиномиально сложной процедурой - !) образующих ее конъюнктов.

[9] См. выше – **Раздел 3.1**.

этим – и как класс эквивалентности по отношению[10] эквивалентности $\mathbf{E}^{\cap}$, в свою очередь порождаемому по заданному операцией $\cap$ пересечения множеств из $\mathbf{\Omega_{\Phi,i}}$ отношению сходства $\mathbf{R}^{\cap}$) некоторого подмножества элементов соответствующего заданному $\mathbf{\sigma_i}$ множества $\mathbf{\Omega_{\Phi,i}}$. Действительно, (с точностью до совпадения A^i_0, A^i_{i1} и A^i_{i2}) любое двух- и более-элементное подмножество множества $\mathbf{\Omega_{\Phi,i}}$ порождает уникальное множество общих для всех входящих в $\mathbf{\Omega_{\Phi,i}}$ элементов из $\mathbf{U_\Phi}$ (и обусловлено это «диагональной» - по вхождению в него элементов из $\mathbf{U_\Phi}$ - структурой множества $\mathbf{\Omega_{\Phi,i}}$).

При этом имеет место и обратное соответствие: каждое исчерпывающее сходство элементов множества $\mathbf{\Omega_{\Phi,i}}$ может быть сопоставлено одному и только одному из нулей исходной булевской функции $\mathbf{\Phi}$. Однозначность задействованного сопоставления есть следствие используемого здесь способа «кодировки» единиц каждого восстанавливаемого (по соответствующему исчерпывающему сходству) булевского вектора $\mathbf{\sigma}$ в ситуации, когда который этот вектор $\mathbf{\sigma}$ формируется соответствующими элементами из $\mathbf{U_\Phi}$ так, что каждый из используемых элементов a^j множества $\mathbf{U_\Phi}$ в свою очередь входит в порождаемое исчерпывающее сходство некоторых элементов множества $\mathbf{\Omega_{\Phi,i}}$, которое и сопоставляется этому восстанавливаемому вектору $\mathbf{\sigma}$.

Пусть далее

$$\mathbf{\Omega_\Phi} = \bigcup_i \mathbf{\Omega_{\Phi,i}} ,$$

считая, что i пробегает все значения из множества {1,2, … , m}, т.е. что по индексу i перечисляется все конъюнкты $\mathbf{\Phi_i}$ из $\mathbf{\Phi}$ и соответствующие пары переменных $\langle x_{i1}, x_{i2}\rangle$ из $\mathbf{\Phi_i}$.

Таким образом, нуль каждого конъюнкта $\mathbf{\Phi_i}$ из 2-КНФ представления функции $\mathbf{\Phi}$ «кодируется» n+1 множеством $\mathbf{A^i} = \{A^i_0, A^i_1, A^i_2, \dots, A^i_n\}$ по следующей схеме:

- A^1_0 представляет ситуацию, когда $x_{j1} = x_{j2} = 0$, а все остальные переменных функции $\mathbf{\Phi}$ принимают значение 1;
- каждое из множеств A^i_j, где $j \in \{1,2, \dots ,n\}$ характеризует ситуацию, когда одна отличная от x_{i1} и x_{i2} переменная x_j (и только она) обращается в 0 (одновременно с $x_{i1} = x_{i2} = 0$),
- по значениям индекса i подобных коллекций множеств вида A^i_j (где $j \in \{1,2, \dots ,n\}$) будет ровно m (по числу конъюнктов $\mathbf{\Phi_i}$ в 2-КНФ исходно заданной функции $\mathbf{\Phi}$),
- множество всех нулей функции $\mathbf{\Phi}$, являющихся одновременно и нулем конъюнкта $\mathbf{\Phi_i}$, будет перечисляться множеством всех исчерпывающих пересечений (сходств) объектов из множества $\mathbf{\Omega_{\Phi,i}}$. (Можно показать, например, индукцией по числу пересекаемых элементов из $\mathbf{\Omega_{\Phi,i}}$, что с точностью до равенства

$$A^i_0 = A^i_{i1} = A^i_{i2}$$

каждое пересечение объектов из множества $\mathbf{\Omega_{\Phi,i}}$ будет исчерпывающим).

В ситуации, когда функция $\mathbf{\Phi}$ обращается в нуль, по крайней мере один из ее конъюнктов $\mathbf{\Phi_i}$ непременно также обращается в нуль. Нетрудно убедиться, что каждый из нулей функции $\mathbf{\Phi}$ может быть рассмотрен как суперпозиция[11] примитивных и сопряженных с ними нулей.

Разумеется, в объединенном множестве $\mathbf{\Omega_\Phi}$ будут встречаться одинаковые элементы (такие повторы порождаются для сопряженных с примитивными нулей, каждый из которых формируется суперпозициями примитивных нулей). Однако, очевидно, что наличие таких «повторов» не может расширить множество исчерпывающих сходств (пересечений) элементов сводного множества $\mathbf{\Omega_\Phi}$, сформированного объединением исчерпывающих сходств по всем $\mathbf{\Omega_{\Phi,i}}$ (для $i \in \{1,2, \dots ,m\}$).

Итак, каждое исчерпывающее пересечение (сходство) объектов из множества $\mathbf{\Omega_\Phi}$ будет представлять какой-либо из нулей функции $\mathbf{\Phi}$. И, наоборот, каждый из нулей

---

[10] См. выше – **Раздел 2.1.**

[11] Т.е. ситуация, когда одновременно реализуются соответствующие *примитивные* нули функции $\mathbf{\Phi}$.

функции **Ф** - как соответствующая суперпозиция примитивных и сопряженных с ними нулей функции **Ф** - может быть представлен (исчерпывающим) сходством объектов из множества $\mathbf{\Omega_Ф}$. Таким образом, множество нулей функции **Ф** и множество исчерпывающих сходств объектов из множества $\mathbf{\Omega_Ф}$ могут быть взаимно однозначным образом сопоставлены друг другу. Т.е. зная число исчерпывающих сходств

$$S_Ф = \left| \boldsymbol{GC}(\mathbf{U_Ф},\mathbf{\Omega_Ф}) \right|,$$

мы имеем возможность полиномиально быстро определить число наборов, выполняющих функцию **Ф**.

Для завершения доказательства не составляет труда убедиться, что множества $\mathbf{U_Ф}$ и $\mathbf{\Omega_Ф}$ полиномиально быстро порождаются по заданной функции **Ф**. Действительно:

- в $\mathbf{U_Ф}$ мы имеем $n = |\{x_1, x_2, \ldots, x_n\}|$ элементов, а
- в $\mathbf{\Omega_Ф}$ содержится не более чем $[(n+1) \times m]$ элементов, каждый из которых содержит не более чем $n-2$ элементов из $\mathbf{U_Ф}$.

оо

Экспоненциально быстрый (с ростом **U** и **Ω**) рост размеров множества ***GC***(**U**,**Ω**) заставляет обратиться к анализу достаточно просто обозримых частных характеристик множества исчерпывающих пересечений, порождаемых над заданным универсумом **U** из множества примеров **Ω**. Одной из таких характеристик естественно считать границы множества ***GC***(**U**,**Ω**) - т.е. множества максимальных и минимальных по вложению элементов в ***GC***(**U**,**Ω**).

4.2.11. Определение

1. Задача ***ВСЕ ИСЧЕРПЫВАЮЩИЕ СХОДСТВА МАКСИМАЛЬНОГО РАЗМЕРА*** определяется следующими условиями:
   Дано: множества **U** и **Ω**.
   Найти: множество всех максимальных по вложению элементов в ***GC***(**U**,**Ω**).
2. Задача ***ВСЕ ИСЧЕРПЫВАЮЩИЕ СХОДСТВА МИНИМАЛЬНОГО РАЗМЕРА*** определяется следующими условиями:
   Дано: множества **U** и **Ω**.
   Найти: множество всех минимальных по вложению элементов в ***GC***(**U**,**Ω**).
3. Задача ***ВСЕ КЛАССЫ ЭКВИВАЛЕНТНОСТИ МАКСИМАЛЬНОГО РАЗМЕРА*** определяется следующими условиями:
   Дано: множества **U** и **Ω**.
   Найти: множество всех элементов в ***GC***(**U**,**Ω**), характеризуемых максимальным по вложению размером порождаемых ими классов эквивалентности.
4. Задача ***ВСЕ КЛАССЫ ЭКВИВАЛЕНТНОСТИ МИНИМАЛЬНОГО РАЗМЕРА*** определяется следующими условиями:
   Дано: множества **U** и **Ω**.
   Найти: множество всех элементов в ***GC***(**U**,**Ω**), характеризуемых минимальным по вложению размером порождаемых ими классов эквивалентности.
5. Задача ***ИСЧЕРПЫВАЮЩЕЕ СХОДСТВО РАЗМЕРА НЕ МЕНЕЕ k*** определяется следующими условиями:
   Дано: натуральное число ***k***, множества **U** и **Ω**.
   Найти: элемент множества ***GC***(**U**,**Ω**), содержащий не менее ***k*** образующих из множества **U**.
6. Задача ***ИСЧЕРПЫВАЮЩЕЕ СХОДСВО РАЗМЕРА НЕ БОЛЕЕ k*** определяется следующими условиями:
   Дано: натуральное число ***k***, множества **U** и **Ω**.
   Найти: элемент множества ***GC***(**U**,**Ω**), содержащий не более ***k*** образующих из множества **U**.

7. Задача ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА НЕ МЕНЕЕ k*** определяется следующими условиями:
   Дано: натуральное число ***k***, множества **U** и **Ω**.
   Найти: элемент множества ***GC*(U,Ω)**, которому соответствует класс эквивалентности размера не менее ***k*** (элементов из множества **Ω**).
8. Задача ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА НЕ БОЛЕЕ k*** определяется следующими условиями:
   Дано: натуральное число ***k***, множества **U** и **Ω**.
   Найти: элемент множества ***GC*(U,Ω)**, которому соответствует класс эквивалентности размера не более ***k*** (элементов из множества **Ω**).
9. Задача ***ИСЧЕРПЫВАЮЩЕЕ СХОДСТВО РАЗМЕРА РОВНО k*** определяется следующими условиями:
   Дано: натуральное число ***k***, множества **U** и **Ω**.
   Найти: элемент множества ***GC*(U,Ω)**, содержащий ровно ***k*** образующих из множества **U**.
10. Задача ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА РОВНО k*** определяется следующими условиями:
    Дано: натуральное число ***k***, множества **U** и **Ω**.
    Найти: элемент множества ***GC*(U,Ω)**, которому соответствует класс эквивалентности размера ровно ***k*** (элементов из множества **Ω**).

o

Вычислительную сложность порождения верхней границы каждого из множеств исчерпывающих сходств (а с ними - и соответствующих классов эквивалентности) характеризует

4.2.12. Теорема

Задача ***ВСЕ ИСЧЕРПЫВАЮЩИЕ СХОДСТВА МАКСИМАЛЬНОГО РАЗМЕРА*** полиномиально разрешима[12].

Доказательство.

Легко видеть, что каждое исчерпывающее сходство максимального размера порождается не менее чем парой объектов из множества **Ω**. Таким образом, для порождения множества всех исчерпывающих сходств максимального размера достаточно построить соответствующие замыкания Галуа для всех пар объектов из множества **Ω**, удалив затем из построенного множества те замыкания Галуа, которые вкладываются по крайней мере в один из других (отличных от него самого) элементов построенного множества.

oo

Вычислительную сложность порождения нижней границы каждого множества исчерпывающих сходств (а с ними - и соответствующих классов эквивалентности) характеризует

4.2.13. Теорема

Задача ***ВСЕ ИСЧЕРПЫВАЮЩИЕ СХОДСТВА МИНИМАЛЬНОГО РАЗМЕРА*** полиномиально разрешима.

Доказательство.

Легко видеть, что каждое исчерпывающее сходство минимального размера содержит по крайней мере одну образующую из универсума **U**. Таким образом, для порождения множества всех исчерпывающих сходств минимального размера достаточно для каждой образующей из универсума **U** найти сходство (пересечение) всех объектов из **Ω**, содержащих данную образующую, удалив затем из полученного множества те элементы, в которые вкладывается по крайней мере один из других (оставшихся) элементов этого же (полученного) множества - см. соображения о так называемых наклеенных образующих в работах [*Забежайло 14*].

oo

[12] Т.е. существует алгоритм, порождающий ее решение в ходе вычислений, объем которых ограничен сверху полиномом от характерных размеров входных данных.

Принимая во внимание **Теорему 4.2.5 (***о двойственности***)**, легко убедиться, что **Теорема 4.2.12** и **Теорема 4.2.13** имеют

4.2.14. Следствие

Задача ***ВСЕ КЛАССЫ ЭКВИВАЛЕНТНОСТИ МАКСИМАЛЬНОГО РАЗМЕРА*** полиномиально разрешима.

o

4.2.15. Следствие

Задача ***ВСЕ КЛАССЫ ЭКВИВАЛЕНТНОСТИ МИНИМАЛЬНОГО РАЗМЕРА*** полиномиально разрешима.

o

Кроме этого из **Утверждений 4.2.12-15** вытекают также **Следствия 4.2.16-19**:

4.2.16. Следствие

Задача ***ИСЧЕРПЫВАЮЩЕЕ СХОДСТВО РАЗМЕРА НЕ МЕНЕЕ k*** полиномиально разрешима.

Доказательство.

Достаточно рассмотреть элементы верхней границы множества $\boldsymbol{GC}(\mathbf{U},\boldsymbol{\Omega})$, остановив рассмотрение на первом, удовлетворяющем необходимому условию. Нетрудно понять, что в случае отсутствия такового среди элементов верхней границы, его нет и во всем множестве $\boldsymbol{GC}(\mathbf{U},\boldsymbol{\Omega})$.

o

4.2.17. Следствие

Задача ***ИСЧЕРПЫВАЮЩЕЕ СХОДСТВО РАЗМЕРА НЕ БОЛЕЕ k*** полиномиально разрешима.

Доказательство.

Достаточно рассмотреть элементы нижней границы множества $\boldsymbol{GC}(\mathbf{U},\boldsymbol{\Omega})$, остановив рассмотрение на первом, удовлетворяющем необходимому условию. Нетрудно понять, что в случае отсутствия такового среди элементов нижней границы, его нет и во всем множестве $\boldsymbol{GC}(\mathbf{U},\boldsymbol{\Omega})$.

o

4.2.18. Следствие

Задача ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА НЕ МЕНЕЕ k*** полиномиально разрешима.

Доказательство.

Искомый результат - следствие **Теоремы 4.2.5** и **Следствия 4.2.16**.

o

4.2.19. Следствие

Задача ***КЛАСС ЭКВИВАЛЕТНОСТИ РАЗМЕРА НЕ БОЛЕЕ k*** полиномиально разрешима.

Доказательство.

Искомый результат - следствие **Теоремы 4.2.5** и **Следствия 4.2.17**.

o

Комбинаторную природу задачи ***ИСЧЕРПЫВАЮЩЕЕ СХОДСТВО РАЗМЕРА РОВНО k*** характеризует

4.2.20. Теорема

Задача ***ИСЧЕРПЫВАЮЩЕЕ СХОДСТВО РАЗМЕРА РОВНО k*** принадлежит классу **NP**-полных задач.

Доказательство.

Пусть задана произвольная булевская функция

$$\boldsymbol{\Phi}(x_1,x_2,\ldots,x_n) = \boldsymbol{\Phi}_1 \,\&\, \boldsymbol{\Phi}_2 \,\&\, \ldots \,\&\, \boldsymbol{\Phi}_m =$$

$$= (x^{\alpha 11}{}_{11} \vee x^{\alpha 12}{}_{12} \vee \ldots \vee x^{\alpha 1\, l1}{}_{1\, l1}) \,\&\, (x^{\alpha 21}{}_{21} \vee x^{\alpha 22}{}_{22} \vee \ldots \vee x^{\alpha 2\, l2}{}_{2\, l2}) \,\&\, \ldots$$

$$\ldots \,\&\, (x^{\alpha m1}{}_{m1} \vee x^{\alpha m2}{}_{m2} \vee \ldots \vee x^{\alpha m\, lm}{}_{m\, lm}),$$

где $x^{\alpha i}{}_{i} = x_i$ при $\alpha_i = 1$, и $x^{\alpha i}{}_{i} = \neg x_i$ при $\alpha_i = 0$.

Наша цель - построить универсум $\mathbf{U}_{\boldsymbol{\Phi}}$ и множество объектов над ним $\boldsymbol{\Omega}_{\boldsymbol{\Phi}}$ так, чтобы

- отображение **f: Φ** $\rightarrow$ $<\mathbf{U_Φ},\mathbf{Ω_Φ}>$ позволяло эффективно (т.е. с полиномиальными затратами ресурсов - размеров памяти и объема вычислений) порождать множества $\mathbf{U_Φ}$ и $\mathbf{Ω_Φ}$ по заданной функции **Φ**;
- для функции **Φ** выполняющий ее набор значений булевских переменных существовал тогда и только тогда, когда в множестве $\boldsymbol{GC}(\mathbf{U_Φ},\mathbf{Ω_Φ})$ существовал бы элемент, порожденный ровно $\boldsymbol{k}$ образующими из множества $\mathbf{U_Φ}$.

Для каждой входящей в конъюнктивную нормальную форму функции **Φ** переменной x_i (и ее отрицания $\neg x_i$) перечислим все те конъюнкты $\mathbf{Φ_i}$ из $\mathbf{Φ}(x_1,x_2, \dots ,x_n)$, которые выполняются при обращении x_i (и соответственно $\neg x_i$) в единицу. Обозначим множество конъюнктов, выполняемых при выполнении литеры x_i, через $\mathbf{Q}(x_i)$, а множество конъюнктов, выполняемых при выполнении литеры $\neg x_i$, через $\mathbf{q}(\neg x_i)$. Рассмотрим множества

$$\mathbf{Q^*}(x_i) = \{\mathbf{Φ_1},\mathbf{Φ_2}, \dots ,\mathbf{Φ_m}\} \setminus \{\mathbf{Q}(x_i)\}$$

и

$$\mathbf{q^*}(\neg x_i) = \{\mathbf{Φ_1},\mathbf{Φ_2}, \dots ,\mathbf{Φ_m}\} \setminus \{\mathbf{q}(\neg x_i)\}.$$

Для построения искомого отображения **f**, прежде всего воспользуемся следующим очевидным обстоятельством:

если набор $<\alpha_1, \alpha_2 , \dots , \alpha_n>$ *значений переменных* $<x_1, x_2, \dots , x_n>$ *выполняет функцию* $\boldsymbol{Φ}$, *то*

$$\bigcup_{i=1}^{n} \boldsymbol{Z_i}(x^{\alpha i}{}_i) = \{\boldsymbol{Φ_1},\boldsymbol{Φ_2}, \dots ,\boldsymbol{Φ_m}\}$$

и, соответственно,

$$\bigcap_{i=1}^{n} \boldsymbol{Z_i}^*(x^{\alpha i}{}_i) = \varnothing$$

для

$$Z_i = Q \text{ при } \alpha_i = 1, \text{ и } Z_i = q \text{ при } \alpha_i = 0.$$

Пусть заданы множества

$$\boldsymbol{A} = \{\mathbf{A_1}, \mathbf{A_2}, \dots , \mathbf{A_r}\} \quad \text{и} \quad \boldsymbol{a} = \{\mathbf{a_1}, \mathbf{a_2}, \dots , \mathbf{a_r}\}.$$

4.2.21. Определение

Множество $\boldsymbol{B} \subset (\boldsymbol{A} \cup \boldsymbol{a})$ будем называть *ИС-допустимым*, тогда и только тогда, когда

1) $|\boldsymbol{B}| = r$

2) $\neg\exists j\ ((1 \le j \le r) \,\&\, (A_j \in \boldsymbol{B}) \,\&\, (a_j \in \boldsymbol{B}))$

Другими словами, ИП-допустимое подмножество объединения двух равномощных конечных множеств (элементы которых пронумерованы) мощности **r** содержит ровно **r** элементов, среди которых нет ни одной пары элементов, имеющих один и тот же номер.

о

Несложно убедиться, что тот факт, что набор

$$<\alpha_1, \alpha_2, \dots , \alpha_n>, \text{ где } \alpha \in \{0,1\} \text{ а } i \in \{1, \dots , n\}$$

выполняет функцию **Φ**, означает, что объединение элементов соответствующего ИС-допустимого множества, построенного из соответствующих $\mathbf{Q_i}$ и $\mathbf{q_i}$ совпадает с множеством $\{\mathbf{Φ_1},\mathbf{Φ_2}, \dots ,\mathbf{Φ_m}\}$. Более точно: выполнимость **Φ** означает, что объединение элементов соответствующего ИС-допустимого множества $\boldsymbol{B} = \{\mathbf{b_1},\mathbf{b_2}, \dots , \mathbf{b_n}\}$, полученного из множеств

$$\boldsymbol{A} = \{\mathbf{Q_1}, \mathbf{Q_2}, \dots , \mathbf{Q_n}\} \quad \text{и} \quad \boldsymbol{a} = \{\mathbf{q_1}, \mathbf{q_2}, \dots , \mathbf{q_n}\}$$

по правилу

$$\mathbf{b_i} = \begin{cases} \mathbf{Q_i}, \text{ при } \alpha_i = 1 \\ \mathbf{q_i}, \text{ при } \alpha_i = 0. \end{cases} \qquad (\boldsymbol{Пр})$$

дает множество

$$\{\mathbf{\Phi_1},\mathbf{\Phi_2}, \dots ,\mathbf{\Phi_m}\}.$$

Таким образом, справедлива

4.2.22. Лемма

Пусть $\boldsymbol{B}(\underline{\alpha})= \{\mathbf{b_1},\mathbf{b_2}, \dots , \mathbf{b_n}\}$ - ИС-допустимое множество, полученное по правилу (***Пр***) из

$$\boldsymbol{A} = \{\mathbf{Q_1}, \mathbf{Q_2}, \dots , \mathbf{Q_n}\} \text{ и } \boldsymbol{a} = \{\mathbf{q_1}, \mathbf{q_2}, \dots , \mathbf{q_n}\}$$

по набору $\underline{\alpha} = \langle\alpha_1, \alpha_2, \dots , \alpha_n\rangle$ (где $\alpha\in\{0,1\}$ а $i\in\{1, \dots , n\}$), выполняющему функцию **Φ**. Тогда

$$\bigcup_{i=1}^{n} \mathbf{b_i} = \{\mathbf{\Phi_1},\mathbf{\Phi_2}, \dots ,\mathbf{\Phi_m}\}.$$

o

Нетрудно убедиться, что из **Леммы 4.2.22** с учетом соотношения

$$\backslash(A \cup B) = \backslash A \cap \backslash B$$

где \X есть дополнение множества X, вытекает

4.2.23. Лемма

Пусть $\boldsymbol{B}^*(\underline{\alpha})= \{\mathbf{b^*_1},\mathbf{b^*_2}, \dots , \mathbf{b^*_n}\}$ - ИС-допустимое множество, полученное по правилу (***Пр****)

$$\mathbf{b^*_i} = \begin{cases} \mathbf{Q^*_i}, \text{ при } \alpha_i = 1 \\ \mathbf{q^*_i}, \text{ при } \alpha_i = 0. \end{cases} \qquad (\boldsymbol{Пр}^*)$$

из

$$\boldsymbol{A}^* = \{\mathbf{Q^*_1}, \mathbf{Q^*_2}, \dots , \mathbf{Q^*_n}\} \quad \text{и} \quad \boldsymbol{a}^* = \{\mathbf{q^*_1}, \mathbf{q^*_2}, \dots , \mathbf{q^*_n}\}$$

по набору $\underline{\alpha} = \langle\alpha_1,\alpha_2, \dots , \alpha_n\rangle$ (где $\alpha\in\{0,1\}$ а $i\in\{1, \dots ,n\}$), выполняющему функцию **Φ**. Тогда

$$\bigcap_{i=1}^{n} \mathbf{b^*_i} = \varnothing.$$

o

Рассмотрим множество $\boldsymbol{\Gamma}$, образованное **n** парами множеств $\mathbf{T_i},\mathbf{t_i}$ ($i\in\{1, \dots , n\}$)

$$\boldsymbol{\Gamma} = \{\mathbf{T_1},\mathbf{t_1}, \mathbf{T_2},\mathbf{t_2}, \dots , \mathbf{T_n},\mathbf{t_n}\}$$

где

$$\begin{array}{lllll}
\mathbf{T_1} = & \{ \Delta^1_1, \delta^1_1, & \Delta^1_2, \Delta^2_2, \dots , \Delta^{n+1}_2, & \dots & \Delta^1_n, \Delta^2_n, \dots , \Delta^{n+1}_n\} \\
\mathbf{t_1} = & \{ \delta^1_2, \Delta^2_1, & \Delta^1_2, \Delta^2_2, \dots , \Delta^{n+1}_2, & \dots & \Delta^1_n, \Delta^2_n, \dots , \Delta^{n+1}_n\} \\
\mathbf{T_2} = & \{\Delta^1_1, \Delta^2_1, \dots , \Delta^{n+1}_1, & \Delta^1_2, \delta^2_1, & \dots & \Delta^1_n, \Delta^2_n, \dots , \Delta^{n+1}_n\} \\
\mathbf{t_2} = & \{\Delta^1_1, \Delta^2_1, \dots , \Delta^{n+1}_1, & \delta^2_2, \Delta^2_2, & \dots & \Delta^1_n, \Delta^2_n, \dots , \Delta^{n+1}_n\} \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
\mathbf{T_n} = & \{\Delta^1_1, \Delta^2_1, \dots , \Delta^{n+1}_1, & \Delta^1_2, \Delta^2_2, \dots , \Delta^{n+1}_2, & \dots & \Delta^1_n, \delta^n_1 \quad \} \\
\mathbf{t_n} = & \{\Delta^1_1, \Delta^2_1, \dots , \Delta^{n+1}_1, & \Delta^1_2, \Delta^2_2, \dots , \Delta^{n+1}_2, & \dots & \delta^n_2, \Delta^2_n \quad \}
\end{array}$$

Здесь в каждое множество $\mathbf{T_i}$ элемент δ^i_1 входит лишь в позиции номер (i-1)(n+1)+2, а элемент δ^i_2 в множества $\mathbf{t_i}$ - лишь в позиции номер (i-1)(n+1)+1. Все элементы, имеющие различные индексы, различны.

4.2.24. Лемма

Пусть $\boldsymbol{\iota} = \{\iota_{i1}, \iota_{i2}, \dots , \iota_{ir}\}$ - некоторое множество элементов из $\boldsymbol{\Gamma}$. Пусть также

$$\varsigma = \bigcap_{j=1}^{r} \iota_{ij}$$

Множество ς содержит ровно n элементов тогда и только тогда, когда $\boldsymbol{\iota}$ есть некоторое ИС-допустимое множество, построенное из множеств

$$\boldsymbol{T} = \{\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_n\} \text{ и } \boldsymbol{t} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n\}$$

Доказательство.

==>. Пусть число элементов в множестве ς есть n. Покажем, что r = n и в множестве $\boldsymbol{\iota}$ нет $\mathbf{T}_j$ и $\mathbf{t}_j$ с одинаковым индексом j (1≤j≤n).

а) Предположим, что r < n.

Пусть r_1 - число индексов j, таких, что $\mathbf{T}_j$ и $\mathbf{t}_j$ входят одновременно в $\boldsymbol{\iota}$. Тогда число элементов в множестве ς есть

$$S \geq (n+1)(n-r) + (r-2r_1) + r_1(n+1) = (n+1)(n-r) + r + r_1(n-1) \geq n+1.$$

(Здесь первое слагаемое отвечает за элементы множества $\boldsymbol{\Gamma}$, которые не вошли в множество $\boldsymbol{\iota}$; второе - за вошедшие в $\boldsymbol{\iota}$ элементы, которым соответствуют различные индексы j; третье - за элементы с одинаковым индексом j). Таким образом, мы имеем противоречие с исходным предположением, что S=n.

б) Предположим, что r = n, но число "повторов" $r_1 \neq 0$.

В этом случае

$$S \geq (n-2r_1) + r_1(n+1) = n + r_1(n-1) \geq n+1.$$

(Здесь первое слагаемое отвечает за элементы из множества $\boldsymbol{\iota}$, второе - за остальные, которых будет ровно r_1 штук). Таким образом, мы вновь имеем противоречие с исходным предположением, что S=n.

в) Предположим, что r > n.

в1) задействованы все n индексов j из множества {1, ... , n}).

В этом случае имеется r-n "повторов" в индексах, т.е.

$$S = n - (r-n) = 2n - r < n-1.$$

Таким образом, мы вновь имеем противоречие с исходным предположением, что S = n.

в2) задействованы r_1 из всех n индексов.

В этом случае (аналогично случаю **а**) имеем:

$$S = (n+1)(n-r_1) + r_1 - (r-r_1) \geq n+1 \text{ (ввиду того, что } r_1 - (r-r_1) \geq 0 \text{).}$$

(Здесь первый член отвечает за те элементы множества $\boldsymbol{\Gamma}$, которые не вошли в множество $\boldsymbol{\iota}$; второй - за те, что вошли; третий - за повторы индексов среди элементов множества $\boldsymbol{\iota}$). Вновь имеем противоречие с исходным предположением, что S=n.

г). Предположим, что r=n и среди индексов j у элементов множества $\boldsymbol{\iota}$ нет повторов.

В этом случае каждое из n штук $\delta^{\mathbf{p}}_{\mathbf{g}}$ ($p \in \{1, 2, \ldots, n)\}$, $g \in \{1,2\}$) в пересечении по соответствующей позиции в множествах из $\boldsymbol{\iota}$ даст ровно ∅, т.е. из 2n "кандидатов на вхождение" в множество ς останется только n, т.е.

$$S = |\varsigma| = n.$$

<==. Если $\boldsymbol{\iota}$ - ИП-допустимое множество, то аналогично случаю **г** из предыдущей части (т.е. ==>-доказательства) множество ς будет содержать ровно n символов.

o

Вернемся к рассмотрению множеств $\mathbf{Q^*_i}$ и $\mathbf{q^*_i}$ ($i \in \{1, \ldots, n\}$). Каждому $\mathbf{Q^*_i}$ и $\mathbf{q^*_i}$ сопоставим, соответственно, множества

$$\{\mathbf{T^1_i}, \mathbf{T^2_i}, \ldots, \mathbf{T^{m+1}_i}\} \text{ и } \{\mathbf{t^1_i}, \mathbf{t^2_i}, \ldots, \mathbf{t^{m+1}_i}\},$$

полагая, что смена верхнего индекса у **T** и **t** означает замену соответствующих Δ и δ на новые, но по способу, описанному выше.

Рассмотрим множество $\boldsymbol{\omega} = \{Y_1, Y_2, \ldots, Y_{2n}\}$, где

$$Y_1 = \mathbf{Q^*_1} \cup \mathbf{T^1_1} \cup \mathbf{T^2_1} \cup \ldots \cup \mathbf{T^{m+1}_1}$$
$$Y_2 = \mathbf{q^*_1} \cup \mathbf{t^1_1} \cup \mathbf{t^2_1} \cup \ldots \cup \mathbf{t^{m+1}_1}$$
$$Y_3 = \mathbf{Q^*_2} \cup \mathbf{T^1_2} \cup \mathbf{T^2_2} \cup \ldots \cup \mathbf{T^{m+1}_2}$$

$$Y_4 = \mathbf{q^*}_2 \cup \mathbf{t^1}_2 \cup \mathbf{t^2}_2 \cup \ldots \cup \mathbf{t^{m+1}}_2$$
$$\vdots$$
$$Y_{2n-1} = \mathbf{Q^*}_n \cup \mathbf{T^1}_n \cup \mathbf{T^2}_n \cup \ldots \cup \mathbf{T^{m+1}}_n$$
$$Y_{2n} = \mathbf{q^*}_n \cup \mathbf{t^1}_n \cup \mathbf{t^2}_n \cup \ldots \cup \mathbf{t^{m+1}}_n$$

Пусть

$$w = \bigcup_{i=1}^{2n} Y_i.$$

Следующая цель наших построений - продемонстрировать взаимосвязь исчерпывающих сходств (пересечений) элементов множества $\boldsymbol{\omega}$ и наборов $\underline{\boldsymbol{\alpha}} = <\alpha_1, \alpha_2, \ldots, \alpha_n>$ (где $\alpha \in \{0,1\}$ а $i \in \{1, \ldots, n\}$), выполняющих функцию **Ф**.

Рассмотрим задачу ***ИСЧЕРПЫВАЮЩЕЕ СХОДСТВО РАЗМЕРА РОВНО k*** - Ξ **ис**(w,$\boldsymbol{\omega}$,k). Пусть отображение

$$\varphi : \Xi \text{вып}(m,n) \rightarrow \Xi \text{ис}(w,\boldsymbol{\omega},k),$$

по каждой задаче ***ВЫПОЛНИМОСТЬ БУЛЕВСКОЙ ФУНКЦИИ*** - Ξ**вып**(m,n) - о выполнимости булевской функции

$$\mathbf{\Phi}(x_1,x_2, \ldots ,x_n) = \mathbf{\Phi_1} \,\&\, \mathbf{\Phi_2} \,\&\, \ldots \,\&\, \mathbf{\Phi_m} =$$
$$= (x^{\alpha 11}{}_{11} \vee x^{\alpha 12}{}_{12} \vee \ldots \vee x^{\alpha 1\, l1}{}_{1\, l1}) \,\&\, (x^{\alpha 21}{}_{21} \vee x^{\alpha 22}{}_{22} \vee \ldots \vee x^{\alpha 2\, l2}{}_{2\, l2}) \,\&\, \ldots$$
$$\ldots \,\&\, (x^{\alpha m1}{}_{m1} \vee x^{\alpha m2}{}_{m2} \vee \ldots \vee x^{\alpha m\, lm}{}_{m\, lm}),$$

(где $x^{\alpha i}{}_i = x_i$ при $\alpha_i = 1$, и $x^{\alpha i}{}_i = \neg x_i$ при $\alpha_i = 0$) с параметрами **m** и **n** описанным выше (см. представленную выше последовательность порождения множеств $Y_1, Y_2, \ldots, Y_{2n}$) способом строит задачу Ξ **ис**(w,$\boldsymbol{\omega}$,k) о существовании среди исчерпывающих сходств элементов множества $\boldsymbol{\omega}$ такого исчерпывающего сходства, в которое входили бы ровно k = n(m+1) элементов из множества w.

4.2.25. Лемма

Порождение

$$\Xi \text{ис}(w,\boldsymbol{\omega},k) = \Xi \text{ис}[w(m,n),\boldsymbol{\omega}(m,n),k(m,n)] = \Xi \text{ис}(\mathbf{U_\Phi},\mathbf{\Omega_\Phi},k)$$

по **Ф** посредством φ требует объема вычислений, ограниченного сверху полиномом от размеров задачи Ξ**вып**(m,n).

Доказательство.

Утверждение **Леммы** следует из того, что

а) число элементов, необходимых для задания множеств $\mathbf{T^1}_i$, $\mathbf{t^1}_i$, $\mathbf{T^2}_i$, $\mathbf{t^2}_i$, ... , $\mathbf{T^n}_i$, $\mathbf{t^n}_i$ не превосходит 2n+2n=4n;

б) число элементов в каждом из множеств $\mathbf{Q^*}_i$ и $\mathbf{q^*}_i$ (где $i \in \{1, \ldots, n\}$) не превосходит m;

в) наконец, $|w| = m + 4n(m+1)$.

o

4.2.26. Лемма

Каждому решению **Iвып** задачи Ξ**вып**(m,n) соответствует единственное решение **Iис** задачи Ξ**ис**(w,$\boldsymbol{\omega}$,k) = Ξ**ис**[w(m,n),$\boldsymbol{\omega}$(m,n),k(m,n)] = Ξ**ис**(**U**$_\Phi$,Ω_Φ,k) и наоборот.

Доказательство.

==>. Набор $\underline{\boldsymbol{\alpha}} = <\alpha_1, \alpha_2, \ldots, \alpha_n>$ выполняющий функцию **Ф**, однозначно определяет ИС-допустимое множество, построенное из $\mathbf{Q^*}_i$ и $\mathbf{q^*}_i$ (где $i \in \{1, \ldots, n\}$), которое в свою очередь, однозначно определяет исчерпывающее сходство размера ровно k(m,n) = n(m+1), получаемое из элементов множества $\boldsymbol{\omega}$(**Ф**(m,n)). (Действительно, в случае непустоты пересечения Y_j ($1 \le j \le s$, $s \le n$) по элементам "$\mathbf{Q^*}_i$-$\mathbf{q^*}_i$ -части", число элементов в получаемом

аналогично множеству ς (из **Леммы 4.2.24**) пересечении ς* не может быть кратным m+1, а значит и равным n(m+1)).

<u>⇐=</u>. Исчерпывающее сходство - множество ς*, полученное из элементов множества **ω**, однозначным образом определяет ИС-допустимое множество, состоящее из $\mathbf{Q^*_i}$ и $\mathbf{q^*_i}$ (где $i \in \{1, \dots, n\}$), которое, в свою очередь, однозначно определяет набор

$$\underline{\alpha} = <\alpha_1(\mathbf{Q^*_1},\mathbf{q^*_1}), \alpha_2(\mathbf{Q^*_2},\mathbf{q^*_2}), \dots, \alpha_n(\mathbf{Q^*_n},\mathbf{q^*_n}), >,$$

выполняющий функцию **Φ**:

$$\alpha_i = \begin{cases} 1, \text{ если в ИП-допустимое множество взято } \mathbf{Q^*_i} \\ 0 \text{ если в ИП-допустимое множество взято } \mathbf{q^*_i}. \end{cases}$$

o

Теперь очевидным образом справедливо

4.2.27. Следствие

1. Множества решений задач

$$\Xi_{\text{вып}}(m,n) \text{ и}$$

$$\Xi_{\text{ИС}}(w,\omega,k=n(m+1)) = \Xi_{\text{ИС}}[w(m,n),\omega(m,n),k(m,n)=n(m+1)] = \Xi_{\text{ИС}}(\mathbf{U_Φ},\mathbf{Ω_Φ},k)$$

равномощны.

2. Отображение φ консервативно в смысле [*Гэри 82*] и [*Simon 77*].

o

Этим доказательство **Теоремы 4.2.20** завершено.

oo

Продемонстрированное выше (в доказательстве **Теоремы 4.2.20**) взаимно однозначное соответствие (мы только что убедились, что используемая сводимость

$$\mathbf{f: Φ \rightarrow <U_Φ,Ω_Φ>}$$

носит консервативный характер в смысле [*Гэри 82*] и [*Simon 77*]) элементов множества выполняющих функцию **Φ** наборов значений булевских переменных и множества исчерпывающих пересечений размера ровно ***k*** в $\mathbf{\Gamma C}(\mathbf{U_Φ},\mathbf{Ω_Φ})$ доказывает, что имеет место

4.2.28. Следствие

Задача ***ИСЧЕРПЫВАЮЩЕЕ СХОДСТВО РАЗМЕРА РОВНО k*** принадлежит классу **#PC** - перечислительно полных задач.

o

Комбинаторную природу задачи ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА РОВНО k*** характеризует

4.2.29. Следствие

1. Задача ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА РОВНО k*** принадлежит классу **NPC** (**NP**-полных задач).
2. Задача ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА РОВНО k*** принадлежит классу **#PC** – перечислительно полных задач.

<u>Доказательство.</u>

Искомый результат - следствие **Теоремы 4.2.5** и **Следствий 4.2.27-28**.

o

Принимая во внимание установленную **Следствиями 4.2.28-29** перечислительную полноту задач ***ИСЧЕРПЫВАЮЩЕЕ СХОДСТВО РАЗМЕРА РОВНО k*** и ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА РОВНО k***, мы имеем возможность убедиться также в перечислительной полноте и соответствующего множества переборных задач из **Определение 4.2.11,** получаемых заменой равенства ...=***k*** на неравенства ...≥***k*** и ...≤***k***. (Действительно, при каждой из таких замен множество решений получаемой новой переборной задачи содержит в себе множество решений задачи для равенства ...=***k***, принадлежащей, как только что было показано, **NPC** - классу **NP**-полных задач). Таким образом, имеет место

4.2.30. Следствие

Задачи

- ***ИСЧЕРПЫВАЮЩЕЕ СХОДСТВО РАЗМЕРА НЕ МЕНЕЕ k***
- ***ИСЧЕРПЫВАЮЩЕЕ СХОДСТВО РАЗМЕРА НЕ БОЛЕЕ k***
- ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА НЕ МЕНЕЕ k***
- ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА НЕ БОЛЕЕ k***

принадлежат классу ***#PC*** перечислительно полных задач.

o

Итак, здесь мы можем говорить об аналогии со сложностными характеристиками известных задач о выполнимости произвольной булевской функции и о выполнимости булевской функции в виде монотонной 2-КНФ. В случае названных переборных задач переход от **NP** -полноты первой задачи к ее полиномиально разрешимому частному случаю - второй задаче - не оказывает влияния на принадлежность обеих рассматриваемых задач к классу перечислительно полных (см., например, [*Valiant 79*] и др.). В рассмотренном нами только что выше случае подобным же образом ослабление равенства ...=***k*** до любого из неравенств ...≥***k*** или ...≤***k*** переводит соответствующую переборную задачу (см. **Теорему 4.2.20** и **Следствия 4.2.28-29**) из класса **NP** -полных в класс полиномиально разрешимых задач. Тем не менее, такое упрощение поиска какого-либо одного решения соответствующей рассматриваемой задачи не сказывается на экспоненциально большом (в общем случае) размере множества всех таких решений. Модифицированные (и, вообще говоря, - упрощенные) варианты рассматриваемых переборных задач остаются по-прежнему перечислительно полными.

Особого внимания заслуживает консервативный характер предложенной в доказательстве **Теоремы 4.2.20** сводимости задачи ***ВЫПОЛНИМОСТЬ БУЛЕВСКОЙ ФУНКЦИИ*** к задаче ***ИСЧЕРПЫВАЮЩЕЕ СХОДСТВО РАЗМЕРА РОВНО k***. Здесь каждому набору значений булевских переменных, выполняющих функцию **Ф**, взаимно однозначным образом сопоставляется соответствующее исчерпывающее пересечение из ***GC***(**U**$_{\Phi}$,**Ω**$_{\Phi}$). При этом на вид функции **Ф**, вообще говоря не накладывается никаких дополнительных ограничений. Иными словами, процедура сводимости корректна и для частных случаев представлений функции **Ф**. Таким образом, рассмотрев конкретный частный случай, когда **Ф** представлена в виде 2-**КНФ**, мы “автоматически” убеждаемся, что имеет место (полностью совпадающее с полученным совершенно иным путем утверждением **Теоремы 4.2.20**)

4.2.31. Следствие

Задача ***ЧИСЛО ИС*** перечислительно полна:

$$\textit{ЧИСЛО ИС} \in \#PC$$

oo

Опираясь на уже полученные результаты, обсудим свойства еще одного частного вида элементов во множестве ***GC***(**U**,**Ω**) над заданными **U** и **Ω** - исчерпывающих пересечений, для которых зафиксирована сумма двух характеризующих их параметров - длины исчерпывающего пересечения и размера его орбиты. При этом нам пригодятся в том числе и уже рассматривавшиеся выше соображения о двойственности (см. выше **Определение 4.2.3, Лемму 4.2.4** и **Теорему 4.2.5**).

Пусть **Ξ**$_{\text{ип}}$(**U**,**Ω**,k) - задача ***ИСЧЕРПЫВАЮЩЕЕ СХОДСТВО РАЗМЕРА РОВНО k***. Рассмотрим двойственную ей задачу **Ξ***$_{\text{ип}}$(**U**,**Ω**,k) = **Ξ**$_{\text{ип}}$(**U***,**Ω***,l). Задачу **Ξ**$^{0}_{\text{ип}}$(**U**^{0},**Ω**^{0},m) определим следующими условиями:

1) **U**^{0} = **U** ∪ **U***
2) **Ω**^{0} = **Ω** ∪ **Ω***
3) m = k + l.

4.2.32. Лемма

Задача $\Xi^0_{ИС}(\mathbf{U}^0,\Omega^0,m)$ строится по задачам $\Xi_{ИС}(\mathbf{U},\Omega,k)$ и $\Xi^*_{ИС}(\mathbf{U},\Omega,k)$ за полиномиальное от размеров задачи $\Xi_{ИС}(\mathbf{U},\Omega,k)$ число операций.

Доказательство.

Утверждение Леммы следует (с учетом **Леммы 4.2.4, Теоремы 4.2.5** и **Следствия 4.2.6**) из того, что

1) $|\mathbf{U}^0| = |\mathbf{U}| + |\Omega|$

2) Ω^0 строится по Ω и Ω^* не более чем за $|\Omega| + |\Omega^*|$ шагов.

o

Сопоставим решения задач $\Xi_{ИС}(\mathbf{U},\Omega,k)$ и $\Xi^0_{ИС}(\mathbf{U}^0,\Omega^0,m)$:
если **I** - решение задачи $\Xi_{ИП}(\mathbf{U},\Omega,k)$, такое, что его размер есть k, то сопоставим ему решение $\mathbf{I}^0$ задачи $\Xi^0_{ИС}(\mathbf{U}^0,\Omega^0,m)$, такое, что его размер есть k+l.

4.2.33. Лемма

Каждому решению задачи $\Xi_{ИС}(\mathbf{U},\Omega,k)$ соответствует одно и только одно решение задачи $\Xi^0_{ИС}(\mathbf{U}^0,\Omega^0,m)$ и наоборот.

Доказательство.

Утверждение рассматриваемой **Леммы** следует из необходимости иметь в каждом решении задачи $\Xi^0_{ИС}(\mathbf{U}^0,\Omega^0,m)$ решение одновременно задач $\Xi_{ИС}(\mathbf{U},\Omega,k)$ и $\Xi^*_{ИС}(\mathbf{U},\Omega,k)$, декартово произведение которых и задает решение задачи $\Xi^0_{ИС}(\mathbf{U}^0,\Omega^0,m)$.

o

Пусть ϕ - отображение, порождающее $\Xi^0_{ИП}(\mathbf{U}^0,\Omega^0,m)$ по $\Xi_{ИП}(\mathbf{U},\Omega,k)$.

4.2.34. Лемма

Отображение ϕ - консервативно.

Доказательство.

Утверждение **Леммы** следует из утверждений **Лемм 4.2.32-33**.

o

Теперь мы имеем все основания утверждать, что имеет место

4.2.35. Теорема

Задача $\Xi^0_{ИС}(\mathbf{U}^0,\Omega^0,m)$ лежит в классах **NPC** и **#PC**.

Доказательство.

Утверждение **Теоремы** следует из **Теоремы 4.2.20** с учетом **Леммы 4.2.32**.

oo

В заключение приведем еще несколько соображений, характеризующих множество всех исчерпывающих сходств, формируемых на заданных алфавите **U** и исходном множестве примеров Ω. Ранее (в **Разделе 2.1**) уже было показано, как на множестве Ω и отношении сходства $\mathbf{R}^{\otimes}$(а в рассматриваемой нами здесь ситуации, когда используемая операция $\otimes$ есть $\cap$ - операция пересечения множеств, это будет отношение $\mathbf{R}^{\cap}$) могут быть сформированы

- пространство толерантности $<\Omega,\mathbf{R}^{\cap}>$,
- множество $Eq_{<O',O>}(\Omega,\mathbf{R}^{\cap})$ всех пар вида $<O',O>$, характеризующих порождаемые на этом пространстве толерантности $<\Omega, \mathbf{R}^{\cap}>$ классов эквивалентности,

а также

- диаграммы $< Eq_{<O',O>}(\Omega,\mathbf{R}^{\cap})|_O , \subseteq >$ и $< Eq_{<O',O>}(\Omega,\mathbf{R}^{\cap})|_{O'} , \preccurlyeq >$ частичных порядков на множестве пар вида $<O',O>$, определяемые в первом случае отношением

$\subseteq$ взаимной вложимости (формирующих соответствующие классы эквивалентности) подмножеств примеров из множества **Ω**, а во втором случае – отношением $\preccurlyeq$ частичного порядка на соответствующих «подструктурах» структурных описаний объектов (примеров) из **Ω**.

Несложно убедиться, что в рассматриваемом нами здесь случае (когда объекты – примеры из **Ω** есть попросту подмножества образующих – элементов исходного алфавита **U**) отношение частичного порядка $\preccurlyeq$ может быть представлено отношением $\subseteq$ взаимной вложимости подмножеств исходного алфавита **U**, задающих в парах вида $<\boldsymbol{O'},\boldsymbol{O}>$ характеризующие «структуры» сходства подмножества вида $\boldsymbol{O'}$, которые взаимно однозначно сопоставляемы соответствующим подмножествам примеров из **Ω** .

Не составляет особого труда убедиться, что в обоих обсуждаемых здесь случаях мы имеем место с уже рассматривавшимися нами выше в текущем **Разделе** диаграммами взаимной вложимости неподвижных точек замыканий Галуа - $\boldsymbol{g}^*\boldsymbol{s}$ и $\boldsymbol{s}^*\boldsymbol{g}$:

- в первом случае – с диаграммой $\mathbf{D_}\boldsymbol{GC}_{\boldsymbol{g}^*\boldsymbol{s}}(\mathbf{U},\boldsymbol{\Omega})$ частично упорядоченного множества $<\boldsymbol{GC}_{\boldsymbol{g}^*\boldsymbol{s}}(\mathbf{U},\boldsymbol{\Omega}),\subseteq >$, сформированного неподвижными точками замыкания Галуа $\boldsymbol{g}^*\boldsymbol{s}$ и отношением их взаимной вложимости как подмножеств исходного множества примеров **Ω**;
- во втором случае – с диаграммой $\mathbf{D_}\boldsymbol{GC}_{\boldsymbol{s}^*\boldsymbol{g}}(\mathbf{U},\boldsymbol{\Omega})$ частично упорядоченного множества $<\boldsymbol{GC}_{\boldsymbol{s}^*\boldsymbol{g}}(\mathbf{U},\boldsymbol{\Omega}),\subseteq >$, сформированного неподвижными точками замыкания Галуа $\boldsymbol{s}^*\boldsymbol{g}$ и отношением их взаимной вложимости как подмножеств исходного алфавита **U.**

(Напомним, что отображение $\boldsymbol{s}$ в данном случае каждому множеству примеров из **Ω** сопоставляет множество всех их общих образующих из алфавита **U**:

$$\boldsymbol{s} : 2^{\boldsymbol{\Omega}} \to 2^{\mathbf{U}}$$

и для

$$\forall \mathbf{X} = \{O_{\mathbf{i1}}, O_{\mathbf{i2}}, \dots , O_{\mathbf{is}}\} \in 2^{\boldsymbol{\Omega}}$$

имеем

$$\boldsymbol{s}(\mathbf{X}) = O_{\mathbf{i1}} \cap O_{\mathbf{i2}} \cap \dots \cap O_{\mathbf{is}};$$

а отображение $\boldsymbol{g}$ каждому множеству образующих из алфавита **U** сопоставляет множество всех тех примеров из **Ω** , в которые эти образующие входят одновременно:

$$\boldsymbol{g} : 2^{\mathbf{U}} \to 2^{\boldsymbol{\Omega}}$$

и для

$$\forall \mathbf{Y} = \{a_{\mathbf{i1}}, a_{\mathbf{i2}}, \dots , a_{\mathbf{ik}}\} = \{O^*_{\mathbf{i1}}, O^*_{\mathbf{i2}}, \dots , O^*_{\mathbf{ik}}\} \in 2^{\mathbf{U}}$$

где каждый $O^*_{\mathbf{ij}}$ есть множество всех объектов из , в которые входит $a_{\mathbf{ij}}$, имеем

$$\boldsymbol{g}(\mathbf{Y}) = O^*_{\mathbf{i1}} \cap O^*_{\mathbf{i2}} \cap \dots \cap O^*_{\mathbf{ik}}.$$

Таким образом, для подмножеств объектов из **Ω** замыкание Галуа $\boldsymbol{g}^*\boldsymbol{s}$ позволяет сформировать неподвижные точки вида

$$\mathbf{X} = \boldsymbol{g}^*\boldsymbol{s}\ (\mathbf{X}) = \boldsymbol{g}(\boldsymbol{s}(\mathbf{X})),$$

а для подмножеств исходного алфавита **U** замыкание Галуа $\boldsymbol{s}^*\boldsymbol{g}$ позволяет сформировать неподвижные точки вида

$$\mathbf{Y} = \boldsymbol{s}^*\boldsymbol{g}\ (\mathbf{Y}) = \boldsymbol{s}(\boldsymbol{g}(\mathbf{Y})).$$

При этом, как было показано **Утверждением 2.1.7**, в каждой определяющей соответствующий класс эквивалентности паре $<\boldsymbol{O'},\boldsymbol{O}>$, множество образующих $\boldsymbol{O'} = \mathbf{Y}$ взаимно однозначным образом сопоставляется множеству объектов $\boldsymbol{O} = \mathbf{X}$).

Дополним диаграмму $\mathbf{D_}\boldsymbol{GC}_{\boldsymbol{s}^*\boldsymbol{g}}(\mathbf{U},\boldsymbol{\Omega})$ двумя элементами – множеством $\bigwedge(\boldsymbol{\Omega})$ (возможно - пустым) всех общих образующих всех примеров **Ω** и алфавитом **U** . Несложно убедиться, что для любого $\boldsymbol{O'}$ из пар $<\boldsymbol{O'},\boldsymbol{O}>$, образующих множество $\boldsymbol{GC}_{\boldsymbol{s}^*\boldsymbol{g}}(\mathbf{U},\boldsymbol{\Omega})$, выполняются соотношения:

$$\wedge(\Omega) \subseteq O' \text{ и } O' \subseteq U .$$

В таком случае, для расширенной представленным образом диаграммы **D*_*GC*$_{s*g}$(U,Ω)** любой паре ее элементов можно указать точные верхнюю и нижнюю границы. Другими словами, справедливо

4.2.36. Утверждение

Диаграмма **D*_*GC*$_{s*g}$(U,Ω)** представляет собою решетку.

o

Опираясь на уже использованные нами представления о двойственности можно, используя множество примеров **Ω** в качестве наибольшего элемента, а (возможно пустое) множество примеров из **Ω**, содержащих все образующие из алфавита **U**, сформировать абсолютно аналогичным образом и диаграмму **D*_*GC*$_{g*s}$(U,Ω)**, для которой также будет справедливо

4.2.37. Утверждение

Диаграмма **D*_*GC*$_{g*s}$(U,Ω)** представляет собою решетку.

o

Естественный вопрос о дистрибутивности[13] построенных решеток в общем случае имеет отрицательный ответ. Ниже приведены примеры таких структур множеств **U** и **Ω**, что порождаемые на них диаграммы **D*_*GC*$_{s*g}$(U,Ω)** изоморфны известным (см., например, [*Гретцер 82*] и др.) подрешеткам $\mathfrak{N}_5$ и $\mathfrak{M}_3$, определяющим недистрибутивность[14] содержащих их решеток.

4.2.38. Пример

Пусть **U** = {**a**$_1$, **a**$_2$, … , **a**$_{10}$ } а **Ω** = { ***h***$_1$, ***h***$_2$, … , ***h***$_5$ } ,

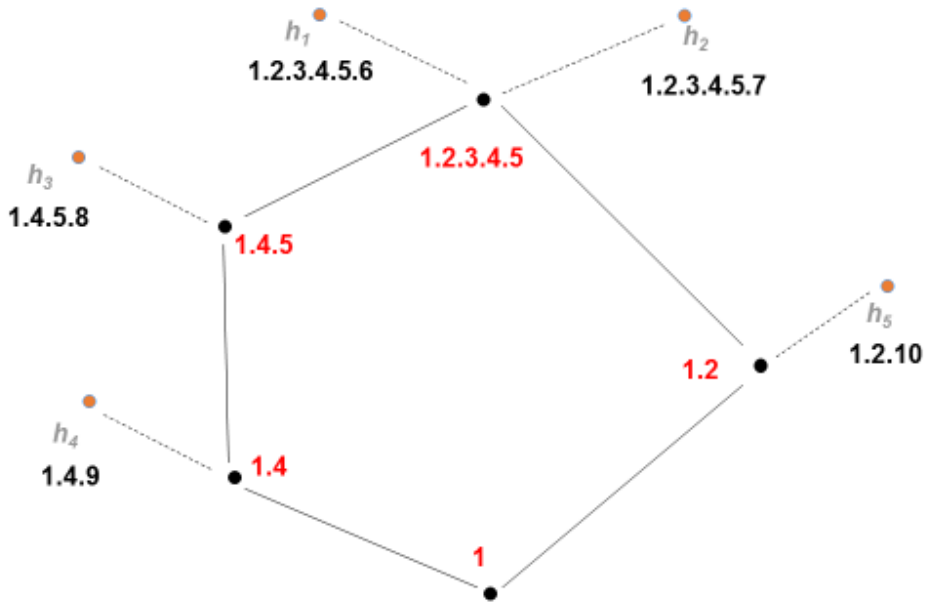


Рис. 4.2.1. Диаграмма **D_*GC*$_{s*g}$(U,Ω)** в **Примере 4.2.38**.

где ***h***$_1$ = {**a**$_1$, **a**$_2$, … , **a**$_6$ }

[13] Как и о модулярности.

[14] И , в соответствующих случаях, немодулярность.

$\boldsymbol{h_2}$ = {$\mathbf{a_1}$, $\mathbf{a_2}$, $\mathbf{a_3}$, $\mathbf{a_4}$, $\mathbf{a_5}$, $\mathbf{a_7}$ }
$\boldsymbol{h_3}$ = {$\mathbf{a_1}$, $\mathbf{a_4}$, $\mathbf{a_5}$, $\mathbf{a_8}$ }
$\boldsymbol{h_4}$ = {$\mathbf{a_1}$, $\mathbf{a_4}$, $\mathbf{a_9}$ }
$\boldsymbol{h_5}$ = {$\mathbf{a_1}$, $\mathbf{a_2}$, $\mathbf{a_{10}}$ }

Диаграмма $\mathbf{D^*_GC_{s^*g}(U,\Omega)}$ в данном случае представляет собою недистрибутивную решетку, содержащую подрешетку «*пентагон*[15]» $\mathfrak{R}_5$ (см. Рис. 4.2.1).

o

4.2.39. Пример

Пусть **U** = {$\mathbf{a_1}$, $\mathbf{a_2}$, … , $\mathbf{a_9}$ } а **Ω** = { $\boldsymbol{h_1}$, $\boldsymbol{h_2}$, … , $\boldsymbol{h_5}$ } ,

где $\boldsymbol{h_1}$ = {$\mathbf{a_1}$, $\mathbf{a_2}$, $\mathbf{a_3}$, $\mathbf{a_4}$, $\mathbf{a_9}$ }
$\boldsymbol{h_2}$ = {$\mathbf{a_1}$, $\mathbf{a_2}$, $\mathbf{a_3}$, $\mathbf{a_4}$, $\mathbf{a_8}$ }
$\boldsymbol{h_3}$ = {$\mathbf{a_1}$, $\mathbf{a_4}$, $\mathbf{a_7}$ }
$\boldsymbol{h_4}$ = {$\mathbf{a_1}$, $\mathbf{a_3}$, $\mathbf{a_6}$ }
$\boldsymbol{h_5}$ = {$\mathbf{a_1}$, $\mathbf{a_2}$, $\mathbf{a_5}$ }

Диаграмма $\mathbf{D^*_GC_{s^*g}(U,\Omega)}$ в данном случае представляет собою недистрибутивную решетку, содержащую подрешетку «*диамант*[16]» $\mathfrak{R}_3$ (см. Рис. 4.2.2):

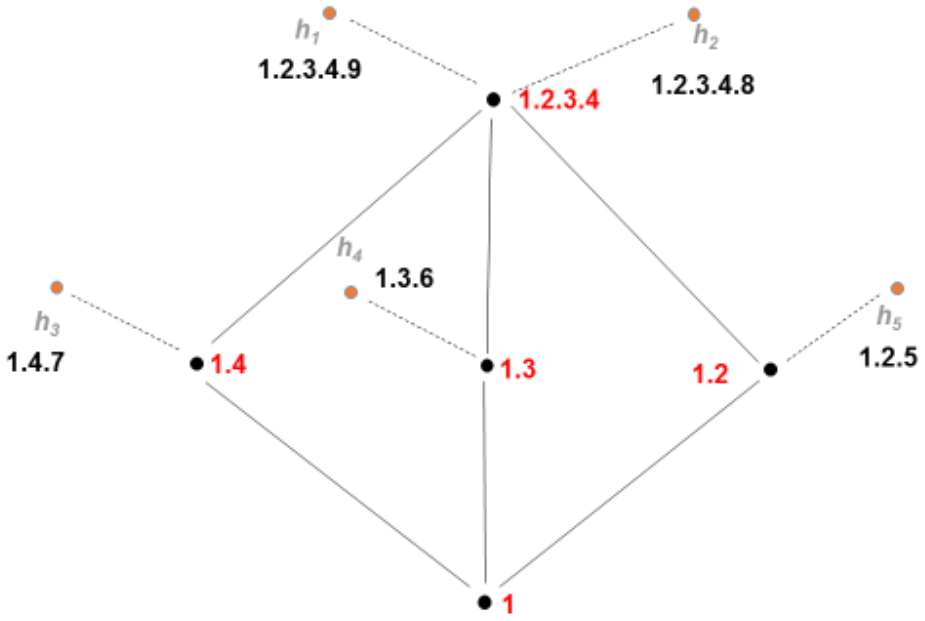


Рис. 4.2.2. Диаграмма $\mathbf{D_GC_{s^*g}(U,\Omega)}$ в **Примере 4.2.39**.

o

[15] См., например, [*Гретцер 82*] и др.
[16] См., например, [*Гретцер 82*] и др.

4.3. Некоторые оценки сложности вычислений для случаев обработки других типов данных.

Если взглянуть на обсуждаемую нами схему анализа данных в процессе обучения на примерах и контрпримерах как на одно из оснований для формирования платформы – *понятийной, методологической, дескриптивной* и *аргументационной, алгоритмической,* – позволяющей организовать интеллектуальный[1] анализ данных, то представляется вполне естественным проследить как именно уже рассмотренные выше на «простейшем[2]» примере (где используемый тип представления данных – множества признаков, а операция сходства – пересечение множеств) комбинаторные свойства проявляются при исполнении аналогичных[3] операций и с другими типами данных.

Здесь мы как и прежде следуем уже обсуждавшейся в **Главе 1** точке зрения на характерные особенности ИАД (в смысле действительно *интеллектуального* анализа данных), которые позволяют выделять соответствующие компьютерные системы-«*усилители*» интеллектуальных возможностей эксперта[4] из всего множества, в том числе - и «обычных»,- вычислительных инструментов компьютерного анализа данных. В обозначенном контексте представляется целесообразным обратить внимание, в частности, на следующие функциональные свойства используемых инструментов ИАД:

- возможности выбора адекватных средств представления данных об объекте анализа (т.е. возможности варьирования типов данных в рамках одного и того же класса алгоритмов интеллектуального анализа данных), например: множеств признаков, кортежей, графов, числовых векторов, …; а также
- возможности настройки локальных средств анализа эмпирических данных в изучаемой предметной области на работу с соответствующими типами данных (например, выбор соответствующей алгебраической операции сходства при формализации сходства анализируемых объектов – множеств признаков, кортежей, графов, числовых векторов, …).

Наличие (доступность исследователю, реализующему ИАД) таких средств позволяет

- формулировать классы однотипных формализованных моделей анализа данных, дающих возможности обеспечить варьирование инструментов принятия решений для их «настройки» на особенности данных в изучаемой предметной области (в том числе – варьирование как конкретных математических средств поддержки приятия решений, так и «подстройки» соответствующей алгоритмики);
- формировать (в частности – в процессе сравнения «наследуемости» порождаемых результатов, при переходе от более «простых» к более «сложным» способам представления данных) содержательные интерпретации результатов, получаемых применением используемых формализмов (в том числе – содержательной интерпретации порождаемых зависимостей – каузальных, ассоциативных и т.п.) а также оценки достаточности оснований для принятия полученных содержательных результатов (их обоснованности, корректной интерпретируемости и т.п.); и в случае неудовлетворительной ситуации в данном аспекте
- задействовать возможности анализа причин фиксируемой неадекватности, чтобы вернуться к исходным шагам проводимого интеллектуального анализа (выбору формальных средств представления данных и т.д. – см. выше) и провести, опираясь

---

[1] В классическом смысле этого термина – см., например, [*Turing 50*] и др.

[2] Далее мы обратимся к операциям с более сложными типами данных.

[3] Аналогичными в том смысле, что при их анализе используется та же понятийная схема (*примеры* и *контрпримеры => формализация сходства => классы сходства => классы эквивалентности* => …)

[4] См., например, [*Эшби 59*] и др. .

на выполненную «работу над ошибками», новую – более адекватную - «настройку» используемых решающих средств ИАД на критически важные особенности изучаемой предметной области.

Именно такой «цикл»:

- *обработка* данных,
- *оценка приемлемости* получаемых результатов и последующая
- *коррекция* (дополнительные «*настройки*») задействованного инструментария (математических моделей вместе с реализующими их алгоритмикой и программными системами)

дает *аргументированные* основания считать реализуемый подход действительно *интеллектуальным* анализом данных.

Теперь мы обратимся к обсуждению особенностей работы в рамках используемой схемы ИАД

исходные ***примеры*** и ***контрпримеры*** =>
=> формализация их ***представления*** (выбор средств описания данных) =>
=> формализация операции ***сходства*** =>
=> ***классы сходства*** =>
=> ***классы эквивалентности*** =>
=> ***прогноз свойств*** новых объектов

с другими типами данных:

- нумерованными множествами;
- кортежами множеств признаков;
- графами (множествами признаков с отношениями на таких множествах);
- цепочками (упорядоченными последовательностями символов);
- векторами числовых значений параметров.

Пусть $\mathbf{U}=\{a_1,a_2, \dots ,a_n\}$ есть как и ранее алфавит образующих[5], а $\mathbf{\Omega}=\{O_1, O_2,\dots, O_m\} \subseteq 2^{\mathbf{U}}\backslash\varnothing$ - множество объектов (т.е. непустых множеств образующих), построенных над $\mathbf{U}$[6]**.** Мы будем предполагать, что среди элементов множества **U** могут встречаться графически совпадающие (по аналогии с тем, что в множестве **Ω** также могут быть графически совпадающие элементы, имеющие тем не менее различные имена). Рассмотрим множество

$\mathbf{U^*} = \{a^*_1, a^*_2, \dots , a^*_{n^*}\} = \{<a_1,k_1>,<a_2,k_2>, \dots ,<a_n,k_n>\}$,

построенное из множества **U** по следующим правилам:

- для каждого $a^*_i = a_i$ выделяются все его вхождения в универсум **U***,
- число таких вхождений k_i вместе с самой образующей a_i будет порождать каждую соответствующую пару - элементов множества **U***.

4.3.1. Определение

1. Множества вида $\mathbf{U} = \{<a_1,k_1>,<a_2,k_2>, \dots ,<a_n,k_n>\}$, где каждая из входящих в него образующих характеризуется еще и числом (или частотой) таких вхождений, мы будем называть ***нумерованными множествами***.

2. Пусть заданы нумерованные множества $\mathbf{U_1} = \{<a_{11},k_{11}>, <a_{12},k_{12}> , \dots , <a_{n1},k_{n1}>\}$ и $\mathbf{U_2} = \{<a_{21},k_{21}>,<a_{22},k_{22}>, \dots ,<a_{n2},k_{n2}>\}$.

Их ***объединением*** назовем множество $\mathbf{U_{1\cup 2}}$, в которое входят все образующие множеств $\mathbf{U_1}$ и $\mathbf{U_2}$, при этом частота вхождения каждой соответствующей образующей в

[5] При этом, не теряя общности, мы будем считать, что в **U** нет одинаковых элементов.
[6] Т.е. множество объектов, построенных из образующих $a_1, a_2, \dots , a_n$.

объединение есть максимум из частот вхождения[7] этой образующей в компоненты объединения:

$$\mathbf{U_{1\cup 2}} = \{<a_i,k_i> \mid (<a_{1i},k_{1i}>\in\mathbf{U_1}) \,\&\, (<a_{2i},k_{2i}>\in\mathbf{U_2}) \,\&\, (k_i = \boldsymbol{max}(k_{1i},k_{2i}) \,\&\, (1 \le \mathbf{i} \le \mathbf{n_1}+\mathbf{n_2})\}$$

Их ***пересечением*** назовем множество $\mathbf{U_{1\cap 2}}$, в которое входят все общие образующие множеств $\mathbf{U_1}$ и $\mathbf{U_2}$, при этом частота вхождения каждой соответствующей образующей в пересечение есть минимум из частот вхождения этой образующей в компоненты объединения:

$$\mathbf{U_{1\cap 2}} = \{<a_i,k_i> \mid (<a_{1i},k_{1i}>\in\mathbf{U_1}) \,\&\, (<a_{2i},k_{2i}>\in\mathbf{U_2}) \,\&\, (k_i = \boldsymbol{min}(k_{1i},k_{2i}) \,\&\, (1 \le \mathbf{i} \le \mathbf{n_1}+\mathbf{n_2})\}$$

o

Вложимость нумерованного множества $\mathbf{U_1}$ в нумерованное множество $\mathbf{U_2}$ понимается естественным образом, т.е. $\mathbf{U_1} \subseteq \mathbf{U_2}$ означает, что $\mathbf{U_1} = \mathbf{U_{1\cap 2}} = \mathbf{U_1} \cap \mathbf{U_2}$.

Учитывая идемпотентность, коммутативность и ассоциативность операций взятия максимума и минимума из пары натуральных чисел, а также известные алгебраические свойства обычных булевских операций, нетрудно убедиться, что справедлива

4.3.2. Лемма

Заданные **Определением 4.3.1** операции объединения и пересечения нумерованных множеств определяют, соответственно, верхнюю полурешетку с единицей (роль которой выполняет множество образующих алфавита **U** с максимальными значениями частот встречаемости соответствующих образующих) и нижнюю полурешетку с нулем (роль которого выполняет пустое множество) на соответствующем универсуме нумерованных множеств.

o

4.3.3. Определение

Нумерованное множество $O_{i1,i2,..,is} = O_{i1} \cap O_{i2} \cap ... \cap O_{is}$, где $\cap$ - операция пересечения нумерованных множеств, мы будем называть ***исчерпывающим сходством*** (пересечением) объектов из нумерованного множества $\boldsymbol{\Omega}$ тогда и только тогда, когда:

(1) $O_{i1,i2,..,is} \neq \varnothing$, и

(2) $(\forall \mathbf{j})$ **(**$[(O_j \in \boldsymbol{\Omega}) \,\&\, \neg(O_j \in \{O_{i1}, O_{i2}, \ldots, O_{is}\})] \supset [(O_j \cap O_{i1,i2,..,is}) \subset O_{i1,i2,..,is}]$**)**[8]

o

Пусть заданы алфавит **U** и множество объектов (нумерованных множеств) над ним - $\boldsymbol{\Omega}$. По аналогии с **Определением 4.2.11** для данного случая также можно сформировать 10 соответствующих переборных задач.

Прежде чем обратиться к их детальному анализу рассмотрим возможность сводимости случая обработки нумерованных множеств к случаю обработки обычных множеств. Процедуру ***REFORM*****(**$\mathbf{Z_{num}}$**,**$\mathbf{Z(Z_{num})}$**)** такого сведения задачи $\mathbf{Z_{num}}$ обработки нумерованных множеств к порождаемой по ней (и аналогичной по формулировке) задаче $\mathbf{Z(Z_{num})}$ обработки обыкновенных множеств задают следующие условия:

---

[7] При этом считается, что не встречающаяся в нумерованном множестве образующая имеет нулевую частоту вхождения в него.

[8] Другими словами, каждое исчерпывающее пересечение $O_{i1,i2,..,is}$ - не пусто и образовано в точности всеми теми объектами из $\boldsymbol{\Omega}$, которые при пересечении дают именно множество $O_{i1,i2,..,is}$.

Процедура *REFORM*(Z_{num}, $Z(Z_{num})$)

Исходные данные:

- нумерованные множества $\mathbf{U}_{num}$[9] и $\mathbf{\Omega}_{num}$, условие задачи $\mathbf{Z}_{num}$.

Алгоритм сведения:

- в исходном состоянии положим множества **U** и **Ω** пустыми;
- для каждой образующей a_i из $\mathbf{U}_{num}$ найдем в множестве $\mathbf{\Omega}_{num}$ объект с максимальным значением $k_i = k_{i0}$ - частоты вхождений образующей a_i в элементы $\mathbf{\Omega}_{num}$; затем
- внесем в множество **U** элементы $a^1_i, a^2_i, \ldots, a^{ki0}_i$ (ровно k_{i0} штук - по числу возможных ненулевых частот вхождений образующей a_i в пересечения элементов из $\mathbf{\Omega}_{num}$); далее
- для каждого элемента X_{num} из множества $\mathbf{\Omega}_{num}$ построим (заменой в X_{num} каждого элемента вида $<a_i,k_i>$ на множество элементов вида $a^1_i, a^2_i, \ldots, a^{ki}_i$) новый объект X, затем включим каждый построенный таким образом X в множество **Ω**. Построение описания задачи **Z** - тройки <**U**,**Ω**, *условие задачи* $\mathbf{Z}_{num}$> - завершено.

Взаимоотношения задач **Z** и $\mathbf{Z}_{num}$ описывает

4.3.4. Теорема

Любому исчерпывающему пересечению объектов из множества $\mathbf{\Omega}_{num}$ над универсумом $\mathbf{U}_{num}$ может быть сопоставлено одно и только одно исчерпывающее пересечение объектов из множества **Ω** над универсумом **U**.

Доказательство.

Пусть ω_{num} - исчерпывающее пересечение объектов из множества $\mathbf{\Omega}_{num}$ над универсумом $\mathbf{U}_{num}$. Построим множество ω, порождаемое из множество ω_{num} применением процедуры сведения ***REFORM*** (т.е. “перекодируем” каждое из кратных вхождений каждой из образующих из универсума $\mathbf{U}_{num}$ соответствующими однократными вхождениями образующих из универсума **U**). Легко видеть, что при такой “перекодировке” каждое исчерпывающее пересечение объектов из множества $\mathbf{\Omega}_{num}$ может быть сопоставлено его “образу” - исчерпывающему пересечению объектов из множества **Ω**. При этом построенное соответствие взаимно однозначно.

оо

Таким образом, **Теорема 4.3.4** позволяет нам свести вопросы о комбинаторных свойствах обсуждаемых переборных задач к воспроизведению уже разобранных нами свойств соответствующих комбинаторных проблем, возникших при обработке данных типа ***множества***. Другими словами, справедливо:

4.3.5. Следствие

Для нумерованных множеств остаются в силе все справедливые для работы с обычными множествами результаты, перечисленные выше **Теоремами 4.2.5, 4.2.10, 4.2.12, 4.2.13, 4.2.20** и **4.2.35** (вместе с соответствующими **Леммами** и **Следствиями**).

о

Итак, свойства трудно-разрешимости соответствующих комбинаторных задач наследуется при переходе к обработке данных, представленных нумерованными множествами.

Абсолютно аналогичными рассуждениями формируется утверждение о наследуемости (полученных для обработки данных, которые представлены множествами) результатов и на случай работы с кортежами (в том числе и нумерованных) множеств признаков. (Работать с данными такого типа приходится, в частности, в случае анализа свойств физиологически активных веществ в ситуациях, когда химическая структура изучаемых

[9] Не теряя общности, мы будем считать, что универсум **U** есть обычное множество, т.е. образован парами вида $<a_i,1>$.

соединений представлена, например, кодом ФКСП[10], а вместе с нею обрабатываются данные о тех или иных релевантных эффекту физико-химических характеристиках (например, концентрациях, растворимости, температурах и т.п.).

Здесь в каждом из «полей» соответствующего кортежа возникает ситуация уже рассмотренного нами выше типа (т.е. реализуется работа с обычными либо нумерованными множествами). Соответственно, на такие ситуации также распространяются (наследуются) свойства комбинаторной трудно-разрешимости соответствующих переборных задач.

В случае обработки данных, представленных цепочками (упорядоченными последовательностями, в том числе – с многократными вхождениями) символов из заданного исходного алфавита аналогичное (уже сформулированным выше в тексте) заключение о трудно-разрешимости соответствующих комбинаторных проблем несложно получить, опираясь на следующее очевидное обстоятельство:

(**!**) «обычное» множество (символов из некоторого алфавита **U**) можно взаимно однозначным образом представить в виде упорядоченной последовательности символов (т.е. цепочки), перенумеровав образующие исходного алфавита **U** и выстроив в описании исходного множества его образующие в соответствии с упорядочением их номеров. (Таким образом, можно рассматривать «обычные» множества как подмножество всех таких непустых цепочек символов из **U**, которые не содержат повторных вхождений образующих).

При этом сходство цепочек определяется как операция выделения всех максимальных по вложению общих подцепочек для заданного множества цепочек.

Таким образом, справедливо

4.3.6. Следствие

Для цепочек символов конечного алфавита остаются в силе все справедливые для работы с обычными множествами результаты, определяющие вычислительную сложность задач

- ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА РОВНО*** $\boldsymbol{k}$
- ***ЧИСЛО КЛАСС ЭКВИВАЛЕНТНОСТИ.***

o

В случае, когда ИАД по обсуждаемой нами схеме необходимо вести с объектами (примерами и контрпримерами), представленными в виде графов, уже ожидаемые результаты о сложности задач

- порождения класса эквивалентности фиксированного размера $\boldsymbol{k} = \boldsymbol{k_0}$ и
- вычисления числа классов эквивалентности (порождаемых на заданном множестве исходных примеров и контпримеров)

также могут быть получены на базе следующего простого соображения (о взаимной вложимости соответствующих классов переборных задач):

(**!!**) цепочки символов есть (специальный) частный вид (плоских) графов.

Таким образом, с учетом **Следствия 4.3.6** имеет место

4.3.7. Следствие

Для представления данных в виде графов остаются в силе все справедливые для работы с обычными множествами результаты, определяющие вычислительную сложность задач

- ***КЛАСС ЭКВИВАЛЕНТНОСТИ РАЗМЕРА РОВНО*** $\boldsymbol{k}$
- ***ЧИСЛО КЛАСС ЭКВИВАЛЕНТНОСТИ.***

o

[10] ФКСП (Фрагментарный Код Суперпозиции Подструктур) обеспечивает покрытие графа структуры изучаемого соединения соответствующими элементами множества стандартизованных подграфов, содержащего порядка 20 000 элементов (см., подробнее, [*ДСМ АПГ*] и др.).

Теперь обратимся к ситуации, когда ИАД ведется с объектами, представленными в виде векторов числовых значений соответствующих эмпирических параметров. Для уточнения операции сходства воспользуемся эвристикой Д.С.Милля, определяющей так называемый *метод сопутствующих изменений* (см. [*Милль 00*]). В основу используемой формализации положена идея выделения участков (т.е. максимальных по вложению подмножеств столбцов) *ко-монотонного изменения* числовых значений множеств параметров (т.е. максимальных по вложению подмножеств строк) в исходно заданной матрице (см. подробнее [*Забежайло 92*, *Забежайло 93*])

<**параметры** (строки)> x <**объекты** (столбцы)>,

где каждый анализируемый в рамках выполняемого ИАД объект представлен вектором числовых значений измеряемых характеристик (параметров). Здесь:

- в качестве (наделенных исследуемыми свойствами) *объектов* рассматриваются соответствующие *векторы численных значений параметров*;
- локально сходными объявляются каждые два объекта, для которых в рассматриваемом множестве *параметров* найдется непустое подмножество таких, что их значения от одного из выбранных объектов к другому (по всему множеству включенных в рассмотрение объектов) *изменяются монотонно*;
- классы сходства (и покрывающие их классы эквивалентности) формируются всеми такими объектами, которые попадают в соответствующий "*участок монотонности* изменения значений параметров" в матрице[11] всех исходно заданных для анализа объектов.

Пример ИАД-обработки данных такого типа при анализе нефте-газоносных свойств территорий представлен в работе [*Забежайло 97*] (см. также **Приложение П.4**).

Итак, пусть **X** - матрица действительных чисел, образованная n столбцами и m строками:

$$\mathbf{X} = \mathbf{C} \times \mathbf{R} = \{C_1, C_2, ..., C_n\} \times \{R_1, R_2, ..., R_m\}.$$

Будем называть матрицу **X** универсумом, ее столбцы C_i (где $i \in \{1,2, \ldots ,n\}$) - объектами, а множество всех ее столбцов $\mathbf{\Omega} = \mathbf{C} = \{C_1, C_2, ..., C_n\}$ множеством примеров (или объектов) над универсумом **X**.

Воспользуемся уже введенными ранее в **Разделе 3.1.2** понятиями
- монотонная под-матрица (*м-матрица*) в матрице-универсуме **X**,
- *м-башня* (максимальная – нерасширяемая по числу столбцов м-матрица) в универсуме **X**,
- бинарная операция $\pitchfork$, "поэлементного соединения" множеств м-матриц в универсуме **X**, которая позволяет определить понятие *сходства* на м-матрицах (а вместе с этим также и на объектах, представляемых *столбцами числовых значений параметров в матрице-универсуме **X**).*

Пусть $V = \mathbf{C_v} \times \mathbf{R_v} = \{C_{i1}, C_{i2}, \ldots, C_{is}\} \times \{R_{i1}, R_{i2}, \ldots, R_{il}\}$ - м-матрица. Мы будем называть объекты $C_{i1}, C_{i2}, \ldots, C_{is}$ - ***родителями*** м-матрицы V, множество $\mathbf{C_v} = \{C_{i1}, C_{i2}, ..., C_{is}\}$ - ***орбитой*** м-матрицы V, а элементы множества $\mathbf{R_v}$ (т.е. имена строк $R_{i1}, R_{i2}, ..., R_{il}$) - ***образующими*** м-матрицы V. Множество всех м-матриц V, порождаемых над универсумом **X** из множества объектов **Ω**, мы будем обозначать посредством ***Мм*(Ω)**. Его подмножество, содержащее лишь все м-башни, мы (как и ранее – в **Разделе 3.1.2**) будем обозначать посредством ***Tow*(Ω)**.

---

[11] Таким образом, в данном случае следует говорить о сходстве объектов в определенном *контексте* (множестве исходно заданных примеров). Именно такими *контекстными* зависимостями и будут обусловлены продемонстрированные далее отличия сложностных характеристик рассматриваемых задач ИАД для этого типа данных.

4.3.8. Определение

Шириной м-матрицы V будем называть число (мощность множества) ее строк. ***Размером орбиты*** м-матрицы V будем называть число (мощность множества) ее родителей (т.е. число образующих ее столбцов матрицы **X**).

o

Теперь мы имеем возможность обратиться собственно к анализу комбинаторных свойств элементов множеств ***Мм*(Ω)** и ***Tow*(Ω)**.

4.3.9. Определение

Задачей ***ПРОИЗВОЛЬНАЯ М-БАШНЯ*** (***М-БАШНЯ***) будем называть следующую задачу:

Дано: числовая матрица **X**.

Найти: какую-либо м-башню[12] в множестве ***Мм*(Ω)**

o

Точное представление о комбинаторной природе множества ***Tow*(Ω)** дают **Теоремы 4.3.10** и **4.3.11.**

4.3.10. Теорема

Задача ***ПРОИЗВОЛЬНАЯ М-БАШНЯ*** при обработке числовых матриц лежит в классе полиномиально разрешимых задач.

Доказательство.

Легко видеть, что любой элемент множества ***Tow*(Ω)** может быть получен посредством следующей процедуры:

- выберем в универсуме **X** минор, порожденный пересечением ровно двух строк и двух столбцов (т.е. образованный четверкой чисел);
- будем перебирать последовательно все оставшиеся столбцы (добавляя их к выбранному минору), стремясь пополнить имеющееся множество столбцов, сохранив при этом монотонность изменения числовых значений вдоль строк расширяемого таким образом минора. Получив первый из столбцов, сохраняющий ко-монотонность изменения числовых значений вдоль строк минора, продолжим расширение до состояния нерасширяемости (по столбцам), затем
- будем перебирать последовательно все оставшиеся строки матрицы **X**, повторив и здесь (по аналогии со столбцами) процедуру пополнения выбранного нами минора.

Результат применения процедуры для каждого исходного минора - некоторая м-башня (содержащая исходный минор), либо доказательство того факта, что исходный минор не может быть расширен в универсуме **X** до какой-либо м-башни.

Учитывая, что число “стартовых” миноров для представленной процедуры в универсуме **X** не превосходит величины

$$C^{n}_{2} \text{ x } C^{m}_{2} \text{ x (n-2) x (m-2)} \simeq O(\frac{n(n-1)(n-2)m(m-1)(m-2)}{4}) \simeq O(n^{3} \text{ x } m^{3}),$$

а пополнение каждого из “стартовых” миноров (до завершения представленной процедуры) может быть осуществлено однократным просмотром множеств строк и столбцов матрицы **X**, мы приходим к искомому утверждению.

oo

4.3.11. Теорема

Задача ***ПРОИЗВОЛЬНАЯ М-БАШНЯ*** при обработке числовых матриц перечислительно полна:

$$\textit{М-БАШНЯ} \in \#\textbf{РС}$$

Доказательство.

Идея доказательства - продемонстрировать полиномиальную сводимость известной (см, например, [*Valiant 79*]) переборной задачи ***МОНОТОННАЯ ВЫПОЛНИМОСТЬ*** к

[12] При этом мы будем рассматривать лишь нетривиальные случаи, т.е. м-башни, имеющие не менее трех родителей.

рассматриваемой нами задаче ***ПРОИЗВОЛЬНАЯ М-БАШНЯ***. Мы воспользуемся той же схемой сведения, которую мы использовали при доказательстве аналогичных утверждений при работе с данными в виде множеств: мы будем перечислять нули соответствующей монотонной булевской функции **Ф**, представленной в виде 2-КНФ, и кодировать их с помощью м-башен в некотором специально построенном универсуме (числовой матрице).

Теперь более подробно. Пусть **Ф** имеет вид

$$\mathbf{\Phi}(x_1,x_2, \dots ,x_n) = \mathbf{\Phi_1} \,\&\, \mathbf{\Phi_2} \,\&\, \dots \,\&\, \mathbf{\Phi_s} = (x_{11} \vee x_{12}) \,\&\, (x_{21} \vee x_{22}) \,\&\, \dots \,\&\, (x_{s1} \vee x_{s2}).$$

Не теряя общности, будем считать, что в представлении функции **Ф** все конъюнкты $\mathbf{\Phi_i}$ различны[13]. Положим, что универсум для задачи о числе исчерпывающих пересечений множеств $\mathbf{U_\Phi} = \{a_1,a_2, \dots ,a_n\}$, а множество объектов над выбранным универсумом $\mathbf{\Omega_\Phi} = \{O_1, O_2, \dots , O_m, O_{m+1}\}$ определяется условиями:

$$O_j = \{a_1,a_2, \dots ,a_n\} \setminus \{a_{i1},a_{i2}\} \quad \text{для } \mathbf{i}\in\{\mathbf{1,2, \dots ,n}\} \text{ и } \mathbf{j}\in\{\mathbf{1,2, \dots ,s}\},$$

$$O_{s+1} = \{a_1,a_2, \dots ,a_n\}.$$

Фактически, каждый из объектов O_i (для $1\le i\le s$) сопоставляется нулю соответствующего конъюнкта $\mathbf{\Phi_i}$ из 2-КНФ представления функции **Ф**. При этом для каждого такого объекта O_j значения переменных x_{i1} и x_{i2} в сопоставляемых ему обращающих $\mathbf{\Phi_i}$ в 0 наборах есть 0, а значения каждой из оставшихся n-2 переменных x_j , где

$$j\in\{1,2, \dots, i_1\text{-}1, i_1\text{+}1, \dots . i_2\text{-}1, i_2\text{+}1, \dots n\},$$

может быть любым из двух элементов пары <0,1>. Именно эти два обстоятельства (фиксированность значений «характеристических» для каждого $\mathbf{\Phi_i}$ переменных и варьируемость значений всех оставшихся переменных функции **Ф** для выбранного $\mathbf{\Phi_i}$) мы и используем для порождения соответствующих матрицы-универсума **X** и сопоставляемых каждому нулю функции **Ф** м-матриц.

Итак, матрицу-универсум

$$\mathbf{X} = \mathbf{C} \times \mathbf{R} = \{C_{1,1}, C_{1,2}, C_{2,1}, C_{2,2}, \dots , C_{n,1}, C_{n,2}\} \times \{R_1,R_2,\dots,R_{s+1}\}$$

сформируем из 2n столбцов (по 2 на каждую из переменных функции **Ф**) и s+1 строк (по одной для каждого конъюнкта $\mathbf{\Phi_i}$ и одной дополнительной (*выделенной*) строки специального вида[14]). При этом

- для (сопоставленных переменным x_{i1} и x_{i2} из конъюнкта $\mathbf{\Phi_i}$) столбцов на пересечении со строкой номер i_1 размещены числа i_1+1 и i_1+2 (в столбцах, соответственно, $C_{i1,1}$ и $C_{i1,2}$), а со строкой i_2 - размещены числа i_2+1 и i_2+2 (в столбцах, соответственно, $C_{i2,1}$ и $C_{i2,2}$);
- во всех остальных парах столбцов $C_{j,1}$ и $C_{j,2}$, сопоставляемых остальным переменным функции **Ф**, располагаются по 2 одинаковых числа – j+1 и j+1 .

Таким образом, зафиксированному значению соответствующей переменной сопоставлен ровно один вариант упорядочения соответствующей пары столбцов вдоль оси монотонности, а для ситуации, допускающей выбор из множества {0,1} – два, подразумевающий возможность перестановки соответствующей пары столбцов без нарушения порядка вдоль оси монотонности). При этом наличие в сопоставленной конкретной переменной x_j функции **Ф** паре столбцов $<C_{j,1}, C_{j,2}>$ хотя бы одной безальтернативной (т.е. имеющей вид <j+1,j+2>) строки означает, что во всех остальных строках, включаемых в ту или иную подматрицу исходного универсума **X** при

---

[13] В противном случае функция **Ф** может быть приведена к нужному нам виду сортировкой (полиномиально сложной процедурой - !) образующих ее конъюнктов.

[14] Которая позволит формировать 2-строчные м-матрицы, где соответствующее **R*** = **{**сопоставленная заданному $\mathbf{\Phi_i}$ одна строка (*как ось монотонности*)**;** *выделенная* строка**}**

формировании соответствующих областей ко-монотонности, вариативность (возможность перестановки столбцов $C_{j,1}$, $C_{j,2}$) будет исключена.

Строка R_{s+1} формируется только из 0, что позволяет формировать отдельные ко-монотонности даже для примитивных[15] и сопряженных с ними нулей функции **Ф**.

Таким образом, матрица-универсум будет иметь вид

	…	C_{2i-1}	C_{2i}	…
…	…	…	…	…
$R_{<0>}$	…	**2i-1**	**2i**	…
$R_{<0,1>}$	…	2i-1	2i-1	…
…	…	…	…	…
R_{s+1}	**0**	**0**	**0**	**0**

Выбирая для столбцов, сопоставляемых «варьируемым» (выбираемым из {0,1}) значениям соответствующей переменной, тот или иной вариант упорядочения (т.е. приравнивая ее 0 или же оставляя «варьируемой») и комбинируя такие упорядочения по всем имеющимся (сопоставленным переменным функции **Ф**) парам столбцов, мы получаем возможность перебрать все варианты означивания всех переменных функции **Ф**. При этом уже заданные избранной формой строк матрицы **X** условия (см. сопоставление строки нулю соответствующего конъюнкта функции **Ф**) сужают такой перебор до перечисления лишь нулей исходной функции **Ф**. Корректность используемой процедуры обеспечивается выбором всех тех и только тех строк матрицы-универсума **X**, которые сохраняют согласованность[16] упорядочений в выбранных строках по всем столбцам матрицы-универсума **X**.

Нетрудно убедиться, что каждому из 0 функции **Ф** будет соответствовать ровно одна[17] м-матрица в **X**, содержащая ровно все 2n столбцов и от 2 до, возможно, s+1 строк. При этом, в соответствии с процедурой построения матицы **X** по функции **Ф** (исчерпывающим перечислением нулей ее конъюнктов $\mathbf{Ф_i}$) в структуре этой матрицы-универсума **X** учтены все возможные нули функции **Ф**. Аналогичным образом, верно и обратное: каждая м-матрица описанного вида в универсуме **X**, содержащая все 2n столбцов, «кодирует» (выбором упорядочений на соответствующих парах столбцов) ровно один набор значений переменных функции **Ф**, обращающий последнюю в 0. Таким образом, множество м-матриц, порождаемых на матрице-универсуме **X**(**Ф**), содержит перечислительно полное подмножество (м-матриц, образованных всеми столбцами матрицы **X**)

Для завершения доказательства заметим, что для порождения матрицы-универсума **X** по функции **Ф** достаточно полиномиально сложной процедуры (распределения не более чем 2n+1 целых чисел по 2n столбцами и s+1 строками).

оо

Экспоненциально быстрый в (общем случае) рост числа элементов множества ***Tow***(**X**) м-башен в универсуме **X** обусловлен возможными повторениями чисел в строках матрицы **X**. Исключив подобные повторения, мы получим быстро обозримый частный случай, представление о котором дает

---

[15] См. **Теорему 4.2.10**.

[16] Т.е. если на некоторой паре столбцов в выбранных строках уже имеется «неварьируемая» строка (и вместе с нею - пара чисел в ней, «кодирующая» 0 соответствующей переменной), то каждая формируемая м-матрица, которая будет содержать эту строку, не допускает изменения порядка на этой паре столбцов.

[17] Ведь она кодирует (соответствующими перестановками столбцов матрицы **X**) распределение значений 0 и 1 на переменных функции **Ф**, в котором каждая из переменных может принимать лишь одно значение – либо 0, либо 1.

4.3.12. Теорема

Задача ***ПРОИЗВОЛЬНАЯ М-БАШНЯ*** при обработке числовых матриц в случае отсутствия в строках матрицы-универсума **X** одинаковых элементов имеет полиномиально большое множество решений.

Доказательство.

Легко видеть, что в рассматриваемой ситуации каждая из осей монотонности имеет ровно один вариант упорядочения как по возрастанию, так и по убыванию. Представленная в доказательстве **Теоремы 4.3.10** процедура "выращивания" м-башен из 4-х-элементных миноров позволяет породить любой элемент множества ***Tow***(**X**). Однозначный вариант упорядочения элементов каждой из строк матрицы **X** делает справедливой оценку сверху

$$C^{n}_{2} \text{ x } C^{m}_{2} \text{ x (n-2) x (m-2)} \simeq O(\frac{n(n-1)(n-2)m(m-1)(m-2)}{4}) \simeq O(n^3 \text{ x } m^3) \geq Nx$$

для числа Nx искомых м-башен.

oo

Теперь, по аналогии с описанием комбинаторных свойств соответствующих переборных задач в случае обработки ранее рассмотренных типов данных, мы имеем возможность сформулировать

4.3.13. Определение

1. Задача ***ВСЕ М-БАШНИ МАКСИМАЛЬНОЙ ШИРИНЫ*** при обработке числовых матриц определяется условиями:
   Дано: универсум **X** (а вместе с ним и множество объектов **Ω**).
   Найти: множество всех максимальных по вложению образующих их множеств строк элементов в ***Tow***(**X**).
2. Задача ***ВСЕ М-БАШНИ МИНИМАЛЬНОЙ ШИРИНЫ*** при обработке числовых матриц определяется условиями:
   Дано: универсум **X** (а вместе с ним и множество объектов **Ω**).
   Найти: множество всех минимальных по вложению образующих их множеств строк элементов в ***Tow***(**X**).
3. Задача ***ВСЕ М-БАШНИ С ОРБИТОЙ МАКСИМАЛЬНОГО РАЗМЕРА*** при обработке числовых матриц определяется условиями:
   Дано: универсум **X** (а вместе с ним и множество объектов **Ω**).
   Найти: множество всех элементов в ***Tow***(**X**), характеризуемых максимальным по вложению размером орбит.
4. Задача ***ВСЕ М-БАШНИ С ОРБИТОЙ МИНИМАЛЬНОГО РАЗМЕРА*** при обработке числовых матриц определяется условиями:
   Дано: универсум **X** (а вместе с ним и множество объектов **Ω**).
   Найти: множество всех элементов в ***Tow***(**X**), характеризуемых минимальным по вложению размером орбит.
5. Задача ***М-БАШНЯ ШИРИНЫ НЕ МЕНЕЕ k*** при обработке числовых матриц определяется условиями:
   Дано: натуральное число ***k***, универсум **X** (а вместе с ним и множество объектов **Ω**).
   Найти: элемент множества ***Tow***(**X**), образованный не менее ***k*** строками из множества $\mathbf{R}_X$.
6. Задача ***М-БАШНЯ ШИРИНЫ НЕ БОЛЕЕ k*** при обработке числовых матриц определяется условиями:
   Дано: натуральное число ***k***, универсум **X** (а вместе с ним и множество объектов **Ω**).
   Найти: элемент множества ***Tow***(**X**), образованный не более ***k*** строками из множества $\mathbf{R}_X$.
7. Задача ***М-БАШНЯ С ОРБИТОЙ РАЗМЕРА НЕ МЕНЕЕ k*** при обработке числовых матриц определяется условиями:

Дано: натуральное число ***k***, универсум **X** (а вместе с ним и множество объектов **Ω**).
Найти: элемент множества ***Tow***(**X**), орбита которого содержит не менее ***k*** элементов из множества **Ω**.

8. Задача ***М-БАШНЯ С ОРБИТОЙ РАЗМЕРА НЕ БОЛЕЕ k*** при обработке числовых матриц определяется условиями:
   Дано: натуральное число ***k***, универсум **X** (а вместе с ним и множество объектов **Ω**).
   Найти: элемент множества ***Tow***(**X**), орбита которого содержит не более ***k*** элементов из множества **Ω**.
9. Задача ***М-БАШНЯ ШИРИНЫ РОВНО k*** при обработке числовых матриц определяется условиями:
   Дано: натуральное число ***k***, универсум **X** (а вместе с ним и множество объектов **Ω**).
   Найти: элемент множества ***Tow***(**X**), образованный ровно ***k*** строками из множества **R**x.
10. Задача ***М-БАШНЯ С ОРБИТОЙ РАЗМЕРА РОВНО k*** при обработке числовых матриц определяется условиями:
    Дано: натуральное число ***k***, универсум **X** (а вместе с ним и множество объектов **Ω**).
    Найти: элемент множества ***Tow***(**X**), орбита которого содержит ровно ***k*** элементов из множества **Ω**.

o

Вычислительную сложность порождения границ каждого множества м-башен характеризуют **Теоремы 4.3.14** и **4.3.16** и **Следствия 4.3.15** и **4.3.17**:

4.3.14. Теорема

Задача ***ВСЕ М-БАШНИ МАКСИМАЛЬНОЙ ШИРИНЫ*** при обработке числовых матриц полиномиально разрешима[18].

Доказательство.

Легко видеть, что каждая м-башня максимальной ширины всегда содержит тройку находящихся внутри нее рядом (в смысле упорядочения вдоль некоторой оси монотонности) столбцов. Поиск такой тройки столбцов и позволит нам выделить (достраивая добавлением новых столбцов найденные тройки до максимального возможного количества объектов) все искомые м-башни.

Рассмотрим множество троек столбцов из **Ω**. Оставим в нем лишь те тройки, которые порождают (на подмножествах своих строк) некоторые м-матрицы. Нетрудно убедиться, что таких троек не более

$$C^{m}_{3} = m(m-1)(m-2)\backslash 6$$

штук. Легко видеть, что для каждой такой тройки столбцов можно предложить не более (3!) x m = 6m упорядочений (вариантов выбора оси монотонности) элементов в строках. К каждому из не более чем 6m вариантов упорядочения каждой из отобранных троек столбцов добавим (перебором по оставшимся m-3 столбцам) по одному столбцу. Построим на каждой из построенных четверок столбцов все возможные упорядочения (их не более 4!xm = 24m), и оставим среди полученных четверок лишь те, которые представляют собой расширение (приписыванием четвертого столбца справа или слева) одной из ранее построенных троек столбцов, сохраняющее характер упорядочений (там, где оно имело место) на строках выбранной тройки столбцов. (Другими словами, выделим на тройках столбцов те, которые порождают некоторые м-матрицы, убедившись в том, что это именно м-матрицы, проанализировав все возможные расширения троек столбцов до четверок столбцов). Нам осталось среди отобранных троек столбцов убрать те, для которых множество строк содержащейся в тройке столбцов м-матрицы вкладывается в множество

[18] Т.е. существует алгоритм, порождающий ее решение в ходе вычислений, объем которых ограничен сверху полиномом от характерных размеров входных данных.

строк содержащейся в некоторой другой тройке столбцов м-матрицы (т.е. удалить “немаксимальные” элементы). Процедура проверки вложимости имеет характер сортировки и, как нетрудно убедиться, также полиномиально реализуема.

oo

4.3.15. Следствие

Задача ***ВСЕ М-БАШНИ С ОРБИТОЙ МИНИМАЛЬНОГО РАЗМЕРА*** при обработке числовых матриц полиномиально разрешима.

Доказательство.

Легко видеть, что каждая м-башня с орбитой минимального размера имеет максимальное по вложению множество строк.

o

4.3.16. Теорема

Задача ***ВСЕ М-БАШНИ МИНИМАЛЬНОЙ ШИРИНЫ*** при обработке числовых матриц полиномиально разрешима.

Доказательство.

Для доказательства достаточно рассмотреть в множестве троек столбцов, использованных в доказательстве **Теорема 4.3.14**, минимальные по вложимости множества строк м-матрицы.

oo

4.3.17. Следствие

Задача ***ВСЕ М-БАШНИ С ОРБИТОЙ МАКСИМАЛЬНОГО РАЗМЕРА*** при обработке числовых матриц полиномиально разрешима.

Доказательство.

Легко видеть, что каждая м-башня с орбитой максимального размера имеет минимальное по вложению множество строк.

o

Вычислительную сложность порождения фрагментов каждого множества м-башен, характеризуемых числовыми ограничениями на их размеры, представляют

4.3.18. Теорема

Следующие переборные задачи при обработке числовых матриц полиномиально разрешимы:

1. ***М-БАШНЯ ШИРИНЫ НЕ МЕНЕЕ k***
2. ***М-БАШНЯ ШИРИНЫ НЕ БОЛЕЕ k***
3. ***М-БАШНЯ С ОРБИТОЙ РАЗМЕРА НЕ МЕНЕЕ k***
4. ***М-БАШНЯ С ОРБИТОЙ РАЗМЕРА НЕ БОЛЕЕ k***
5. ***М-БАШНЯ ШИРИНЫ РОВНО k***

Доказательство.

Декларируемые свойства задач 1,2,5 определяются, как и при доказательстве **Теоремы 4.3.14**, наличием в м-башне, обладающей искомыми свойствами, трех расположенных рядом столбцов, которые содержат м-матрицу (трехстолбцовую!) с искомыми свойствами. Таким образом, полиномиальный алгоритм перебора трех- и четырехстолбцовых фрагментов матрицы-универсума, использованный при доказательстве **Теоремы 4.3.14**, способен привести нас к цели и в рассматриваемом случае.

Декларируемые свойства задач 3 и 4 определяются анализом множества м-башен с орбитами максимального и минимального размера. Отсутствие среди м-башен с орбитами максимального размера такой, которая была бы образована не менее ***k*** объектами (а это по **Следствию 4.3.17** может быть выяснено полиномиально быстро) гарантирует отсутствие таковых вообще. Аналогично, отсутствие среди м-башен с орбитами минимального размера

такой, которая была бы образована не более ***k*** объектами (а это по **Следствию 4.3.15** может быть выяснено полиномиально быстро) также гарантирует отсутствие таковых вообще.

оо

4.3.19. Теорема

Задача ***М-БАШНЯ С ОРБИТОЙ РАЗМЕРА РОВНО k*** при обработке числовых матриц принадлежит классу **NP**-полных задач.

Доказательство.

Прежде всего убедимся, что по заданным произвольным образом матрице-универсуме X и некоторой числовой матрице V можно полиномиально быстро проверить, является ли V м-башней с орбитой размера ровно ***k*** в универсуме X. Действительно, для этого достаточно проверить

- вложимость столбцов матрицы V (как подстолбцов) в столбцы матрицы X;
- возможность пополнить (последовательным перебором оставшихся строк матрицы X) матрицу V до м-матрицы, содержащей большее чем в V число строк;
- возможность пополнить (последовательным перебором оставшихся столбцов матрицы X) матрицу V до м-матрицы, содержащей большее число столбцов.

Далее в доказательстве мы воспользуемся тем же механизмом сводимости, который был нами задействован при доказательстве **Теоремы 4.2.20**. Так же, как и ранее мы воспользуемся сводимостью задачи ***ВЫПОЛНИМОСТЬ БУЛЕВСКОЙ ФУНКЦИИ*** к задаче ***ИСЧЕРПЫВАЮЩЕЕ СХОДСТВО РАЗМЕРА РОВНО k***. Далее, как и при доказательстве **Теоремы 4.3.11**, матрицу-универсум $\mathbf{X_Ф}$ мы построим из строк, сопоставленных всем отдельным переменным функции **Ф** и их отрицаниям, и столбцов, сопоставленных образующим универсума $\mathbf{U_Ф}$ (см. образующие множества $\omega = \{Y_1, Y_2, \ldots, Y_{2n}\}$ из доказательства **Теоремы 4.2.20**), порождаемого по виду исходной булевской функции **Ф**, заданной нам в форме 3-КНФ. Все перечисленные строки и столбцы в матрице $\mathbf{X_Ф}$ заполнены 0 (в случае, когда соответствующая образующая не входит в соответствующее множество образующих, представленное соответствующей строкой матрицы $\mathbf{X_Ф}$) и 1 (в случае, когда соответствующая образующая не входит в соответствующее множество образующих, представленное соответствующей строкой матрицы $\mathbf{X_Ф}$). Дополнительная, 2n+1-я строка универсума $\mathbf{X_Ф}$ образована по следующему закону:

- во всех столбцах, соответствующих образующим, которые не войдут в возможные исчерпывающие пересечения размера ровно n, находятся 0;
- в оставшихся 2n столбцах по строке номер 2n+1 расставлены числа 1,2, ... ,2n-2,2n-1, 2n-1 (в первом и втором - 1 и 2; в третьем и четвертом - 3 и 4, и т.д. до соответствующих значений в предпоследней паре столбцов и чисел 2n-1 в двух последних столбцах).

Легко видеть, что перенумеровав (что может быть сопоставлено выбору подходящей оси монотонности и последующему построению необходимой м-башни) соответствующим образом строки и столбцы матрицы $\mathbf{X_Ф}$, можно каждое исчерпывающее пересечение размера ровно n представить минором размера n(столбцов) на n+1(строк), где в строке номер n+1 монотонное возрастание значений от 1 до 2n-1, прерывается 0 на первом же столбце с номером более n+1. (В столбце номер n+1 в строках с номерами меньшими чем 2n+1 мы найдем нули и единицы, что не изменит направления монотонности “невозрастание” в каждой из этих строк. В строке же номер 2n+1 в этом столбце будет стоять число 2n-1, превосходящее любое из чисел из множества {0,1,2, ... , 2n-2}, расположенных в этой строке на пересечении со столбцами с номером более n+1). Таким образом, каждому исчерпывающему пересечению размера ровно n в универсуме $\mathbf{U_Ф}$ взаимно однозначным образом сопоставляется м-башня, построенная на n строках и n+1 столбце матрицы-универсума $\mathbf{X_Ф}$.

Для завершения доказательства **Теоремы 4.3.19** нам достаточно убедиться, что матрица-универсум $\mathbf{X_Ф}$ строится полиномиально быстро по заданной функции **Ф** (что

действительно справедливо, т.к. матрица-универсум **$X_{Ф}$** всего одной строкой отличается от множества объектов над универсумом **$U_{Ф}$**, для которых исчерпывающие пересечения необходимого размера перечисляют выполняющие функцию **Ф** наборы значений булевских переменных).

oo

* * *

Завершая обзор оценок сложности вычислений, характеризующих обсуждаемую нами схему ИАД, обратим внимание на некоторые отличия ситуации с обработкой данных в виде векторов числовых значений переменных (в частности – результат **Теоремы 4.3.18** относительно *полиномиальной* разрешимости задачи ***М-БАШНЯ ШИРИНЫ РОВНО k***). Здесь существенную роль играет уже отмечавшийся нами выше эффект учета *контекста*, в рамках которого и производится вычисление соответствующих сходств: сходство объектов (столбцов исходной матрицы-универсума **X**) определяется наличием «*пути*» (последовательности промежуточных объектов, на соответствующих значениях параметров которых сохраняется условие *ко-монотонности* [19]изменений их числовых значений). Именно это дополнительное обстоятельство – учет *контекста* – дает возможность избежать здесь экспоненциально быстрого комбинаторного роста требующих анализа вариантов. Однако, и здесь (см. обозначенный **Теоремой 4.3.19** результат о **NP**–полноте задачи ***М-БАШНЯ С ОРБИТОЙ РАЗМЕРА РОВНО k***) уже знакомая нам по работе с другими, ранее рассмотренными, типами данных ситуация трудно-разрешимости задачи о порождении класса эквивалентности заданного размера вновь оказывается актуальной характеристикой требующих решения комбинаторных проблем.

Таким образом, в ситуации возникновения подробно рассмотренных нами трудностей с вычислительной сложностью соответствующих алгоритмов ИАД оказывается весьма полезно для применения в реальных индустриальных приложениях иметь соответствующие «процедурные» возможности

- быстро выделять эффективно разрешимые подклассы решаемых переборных задач восстановления зависимостей, а также
- реализовать (в том или иной виде формализованные) *приближенные* вычисления.

Именно такой путь предложен в работах [*Забежайло 14, Забежайло 14-б*], где обсуждаются возможности специальной проблемно-ориентированной техники оптимизации перебора в рамках так называемого *приближенного* ДСМ-метода, позволяющего целенаправленно управлять объемами вычислений, которые необходимы для реализации ДСМ-рассуждений. В основе предлагаемого механизма оптимизации ИАД-расчетов лежит специальным образом организованная технология гибкой настройки и управления «навигацией» в множестве соответствующих ДСМ-сходств (целенаправленное управление перебором элементов множества потенциально порождаемых в каждом конкретном случае ДСМ-гипотез). Именно таким образом и сформирована технология восстановления эмпирических зависимостей из данных, позволяющая вести средствами ДСМ-метода ИАД в том числе и на больших массивах. Этому и посвящена следующая **Глава 5.**

[19] Напомним, что в рамках выбранной нами семантики именно этим условием характеризуется Миллевское (см. подробнее [*Милль 00*]) условие *сопутствующих изменений* наблюдаемых «каузальных» параметров.

Глава 5

Приближенный ДСМ-метод

Теперь наша цель - перейти к обсуждению некоторых специальных алгоритмических аспектов ДСМ-ИАД, предоставляющих возможности оптимизировать объемы вычислений, которые необходимы для реализации ДСМ-рассуждений. Подобная оптимизация будет выполнена за счет организации целенаправленного управления перебором элементов множества ДСМ-сходств, которые потенциально могут быть восстановлены[1] из имеющихся исходных данных в каждом конкретном случае.

5.1. Общие представления о методах и приемах сокращения перебора при реализации ДСМ-ИАД.

Экспоненциальные оценки сложности вычислений, характеризующие (как мы в этом только что убедились в предыдущей **Главе 4**) ряд ключевых переборных задач ДСМ-ИАД – очевидный аргумент в пользу развития специальной процедурной техники управления и оптимизации ДСМ-перебора, ориентированной в первую очередь на работу с большими выборками исходных данных при порождении зависимостей в реальных приложениях.

Обычно в ситуациях, когда поиск точного решения анализируемой проблемы по каким-либо причинам затруднен, предлагаются те или иные способы поиска *приближенных* решений. При этом термин *приближенный*, как правило, понимается в следующем смысле:

- точное решение задачи (S) и
- ее приближенное решение (S_{app})

разделяет некоторая величина

- Δ, которая и

характеризует *точность* используемого *приближения* S_{app}.

При организации приближенных ДСМ-вычислений нас будет интересовать лишь такая ситуация, когда можно *целенаправленно управлять* «объемом» и «структурой» параметра Δ. Ниже - в **Разделах 5.2-3** - будет предложена процедурная конструкция именно этого типа, которая позволяет:

- начиная с *полиномиально быстрого* порождения такого подмножества множества всех глобальных ДСМ-сходств, которое позволяет выполнить все возможные (при заданном исходном множестве примеров и контрпримеров) доопределения новых объектов из исходно заданного множества $\mathbf{\Omega_\tau}$,

далее (в процессе все более и более *вычислительно сложных* расчетов)[2]

- последовательно строить – причем в процессе *управляемой* и *целенаправленной*[3] навигации в диаграмме всех глобальных ДСМ-сходств – все более и более точные приближенные описания этой диаграммы.

Учитывая предоставляемые ДСМ-методом (как платформой ведения ИАД) возможности варьировать при поиске эмпирических зависимостей и закономерностей используемые средства описания данных, представляется естественным найти соответствующие приемы управления перебором как для простейших – дескрипторных

---

[1] Потенциально порождаемых на заданном множестве примеров и контрпримеров.

2 Фактически – за *точность* описания диаграммы всех сходств приходится платить *сложностью* вычислений, необходимых для формирования соответствующих ее фрагментов.

3 Позволяющей ориентироваться на различные цели – например: скорейшее доопределение новых объектов их множества $\Omega\tau$, проверку условий каузальной полноты, поиск эмпирических зависимостей определенного вида (в частности – содержащих заданные образующие и т.п.) и др. .

(теоретико-множественных), - так и более сложных языков описания анализируемых средствами ДСМ-ИАД примеров и конртпримеров.

Здесь при работе с более выразительными (однако более «тяжелыми» в смысле сложности вычислений) языками, например, в случае представления исходных данных для ДСМ-анализа графами или же цепочками символов весьма эффективным приемом, позволяющим существенным образом сократить перебор при поиске соответствующего решения, оказывается «*сужение*» множества анализируемых такими средствами прецедентов до «*релевантного*» поиску конкретной эмпирической зависимости *подмножества* исходно заданных примеров и контрпримеров. А выделять такие подмножества целенаправленным образом позволяет использование приближенных описаний анализируемых объектов более простыми средствами (сперва – пусть и «грубым», однако, достаточно быстро «обсчитываемым» теоретико-множественным языком, а затем - его уточнением средствами языка нумерованных множеств). Так, например, при работе с графовым представлением данных, представив каждый граф в виде пары множеств (*кортежа* – см. **Раздел 3.1**) вида

<*множество* вершин, *множество* ребер>,

можно провести «предварительный» ДСМ-анализ, в ходе которого буду выделены как воспроизводимые в языке собственно графов результаты, так и, возможно, некоторый дополнительный «шум» (характеризуемый множествами ребер, не совместимыми с соответствующими множествами вершин, выделенными в ходе такого анализа). «Орбиты»[4] порожденных здесь соответствующих теоретико-множественных ДСМ-зависимостей - это, вообще говоря, те самые «усечения» исходного множества примеров и контрпримеров, о которых мы только что говорили выше. Абсолютно аналогично, при обработке данных представленных цепочками символов «огрубление» описания анализируемых объектов до использования языка нумерованных множеств также позволяет получить некоторые «стартовые» позиции для реализации исчерпывающего перебора уже на «усеченном» множестве релевантных прецедентов.

Таким образом, комбинация двух базовых процедур:

- «*сужения*» предложенного для ИАД исходного множества ДСМ-прецедентов (с последующей обработкой отдельно[5] каждого подобного подмножества) использованием более *простых*[6] *языков* описания данных и
- технологии *целенаправленного управляемого перебора* вариантов при обработке данных, представленных дескрипторными языками,

позволяет организовать приближенные ДСМ-вычисления при варьировании в рамках ДСМ-ИАД (**см. Главу 1**) как задействованных средств представления данных, так и соответствующих средств ДСМ-экспертизы.

Достаточно распространенным приемом, позволяющим существенным образом сократить объемы перебора в приложениях комбинаторных алгоритмов является «представление» исходной переборной задачи в виде набора подзадач меньшей размерности, допускающих поиск решения в каждой их них отдельно (независимо) друг от друга. Идея подобного разбиения множества всех глобальных ДСМ-сходств, которые могут быть порождены из соответствующего множества примеров, на такие «самостоятельные» части (подмножества), что

- объединение всех подобных «самостоятельных» частей совпадало бы с исходной диаграммой;
- каждая из таких самостоятельных частей порождалась бы «достаточно просто»;

---

4 См**. Главы 3** и **4**.

5 Причем обработку таких частных подзадач в промышленных приложениях можно при необходимости организовать и в параллельном режиме.

6 В смысле характерных для них оценок сложности вычислений.

и, наконец.

- число таких «самостоятельных частей» с ростом обучающей выборки и\или исходного алфавита росло бы «не очень быстро»,

может быть реализована за счет выделения в исходной диаграмме ДСМ-сходств «независимых» поддиаграмм, каждая из которых в корне имеет один из примеров исходной обучающей выборки. Ветки такой диаграммы ведут к каждой из вершин, образованной минимальными по вложению ДСМ-сходствами – минимальными по вложению неподвижными точками рассматриваемого замыкания Галуа. Мы будем именовать эти диаграммы *псевдо-деревьями*, подчеркивая, что поддиаграммы предлагаемого вида, вообще говоря, могут и не быть деревьями (например, за счет существования в них пар вершин во фрагментах типа β - см. ниже, - находящихся по крайней мере на двух различных ветках частичного порядка взаимной вложимости соответствующих замыканий Галуа), однако, могут быть сведены к виду дерева удалением тех или иных подмножеств входящих в них ребер. В свою очередь, каждое такое псевдо-дерево может быть представлено как объединение соответствующих деревьев, содержащих то же самое множество вершин и вкладывающихся в рассматриваемое псевдо-дерево.

Будет показано, что каждое из подобных псевдо-деревьев может быть реконструировано (восстановлено по имеющемуся множеству обучающих примеров) достаточно удобно устроенной процедурой. Далее мы покажем, как организовать анализ всего множества ДСМ-гипотез последовательным перебором вдоль веток частичного порядка в каждом из порожденных представленным образом псевдо-деревьев.

Предлагаемая техника организации перебора ДСМ-сходств предоставляет возможности для *выборочного* порождения ДСМ-гипотез (например, лишь тех, которые имеют отношение к ДСМ-доопределению заданных новых объектов, однако, удовлетворяют заданным структурным ограничениям).

Именно таким путем и реализован далее механизм *управляемых приближенных вычислений* в рамках (традиционного) ДСМ-метода: технология *направленного* анализа ДСМ-гипотез в процессе *управляемой навигации* вдоль веток частичного порядка в диаграмме взаимной вложимости ДСМ-сходств. В предельном случае (при анализе всех веток всех псевдо-деревьев частичного порядка, который может быть организован как в последовательном, так и в *параллельном* режиме) эта технология порождает те же результаты, что и обычно используемые в программной реализации ДСМ-метода алгоритмы (см., например, [*ДСМ АПГ*] и др.).

В **Разделе 5.2** наше внимание будет сфокусировано на способах формирования а также на структурных и комбинаторных свойствах псевдо-деревьев и их каркасов. В **Разделе 5.3** мы обратимся к детальному рассмотрению алгоритмики предлагаемого направленного перебора ДСМ-сходств, включая некоторые оценки сложности вычислений. Далее будут также обсуждены некоторые полезные особенности развиваемых алгоритмических конструкций, скоординированное использование которых и на описаниях ДСМ-объектов и на описаниях свойств этих объектов позволяют получить дополнительный выигрыш в части сокращения перебора. В **Разделе 5.4** будут представлены формальные обоснования корректности процедур целенаправленного управляемого перебора вариантов, реализуемых в рамках приближенного ДСМ-метода. В заключение мы обсудим некоторые возможности дополнительного ускорения ДСМ-обработки данных за счет использования параллельных алгоритмов, специального типа «облачных» вычислений, а также некоторых специальных проблемно-ориентированных аппаратно-схемных решений.

5.2. Псевдо-деревья и их каркасы.

Центральным элементом алгоритмической конструкции ДСМ-метода является построение множества всех глобальных сходств, порождаемых из заданного набора примеров (и контрпримеров). Фактически (см. подробнее **Главы 3** и **4**) речь здесь идет о построении диаграммы взаимных вложений неподвижных точек замыкания Галуа[1], порождаемого парой следующих соответствий Галуа:

<множество букв исходного алфавита => множество примеров из обучающей выборки, в которые одновременно входят все эти буквы>

и

<множество примеров из обучающей выборки => множество букв исходного алфавита, входящих одновременно во все эти примеры>.

Известно (см. например, соответствующие публикации из Перечня, приведенного в Приложении 1 к [*Финн 11*], а также [*Забежайло 15-а*] и др.), что даже для достаточно просто устроенных пар *<алфавит, множество построенных на его основе примеров>* соответствующая диаграмма взаимных вложений неподвижных точек представленного выше замыкания Галуа может иметь достаточно сложную структуру. В частности: число вершин (и уж тем более – число ребер) в такой диаграмме может расти экспоненциально быстро с ростом размеров исходного алфавита и\или числа примеров в обучающей выборке. Именно по этой причине обычно используемые в программной реализации ДСМ-метода алгоритмы нередко сталкиваются с проблемой сложности вычислений (быстрого роста объемов комбинаторного перебора вариантов при порождении ДСМ-гипотез), что в ряде случаев затрудняет (или же, вообще, делает невозможным) исчерпывающий анализ вариантов при поиске ДСМ-сходств на заданном для обучения наборе примеров.

В обсуждаемом контексте достаточно заманчиво выглядит идея разбиения рассматриваемой диаграммы взаимного вложения неподвижных точек замыкания Галуа на «самостоятельные» части (под-диаграммы) таким образом, что

- объединение всех подобных «самостоятельных» частей совпадало бы с исходной диаграммой;
- каждая из таких самостоятельных частей порождалась бы «достаточно просто»;

и, наконец.

- число таких «самостоятельных частей» с ростом обучающей выборки и\или исходного алфавита росло бы «не очень быстро».

Ниже будет показано, как подобная схема может быть реализована путем, например, выделения в исходной диаграмме отдельных поддиаграмм[2] (каждая из которых в корне имеет один из примеров, входящих в исходную обучающую выборку, а ее ветки ведут к каждой из вершин, образованной минимальными по вложению ДСМ-сходствами – всем однобуквенным[3] неподвижным точкам рассматриваемого замыкания Галуа для букв исходного алфавита, входящих в выбранный в качестве корневой вершины этой поддиаграммы объект – пример из обучающей выборки). Оказалось, что каждое из

---

[1] Точная формулировка будет дана ниже.

[2] Подграфов специального вида.

[3] Вообще говоря, здесь в качестве листьев рассматриваемой поддиаграммы могут встречаться и неоднобуквенные подмножества исходно заданного алфавита – так называемые «склеенные образующие» (исключительно вместе встречающиеся и во всех остальных вершинах этой поддиаграммы). Важно лишь заметить, что каждый такой лист есть минимальное (по вложению соответствующих подмножеств образующих) множество из сопоставленных вершинам рассматриваемой поддиаграммы букв исходно заданного алфавита. Можно показать, что существует быстрый алгоритм диагностики «склеенности» образующих в соответствующей поддиаграмме.

подобных псевдо-деревьев[4] может быть реконструировано (восстановлено по имеющемуся множеству обучающих примеров) достаточно удобно устроенной процедурой, где сперва полиномиально[5] быстро может быть построен так называемый каркас – т.е. некоторое специальное приближенное описание этого псевдо-дерева (его подграф, содержащий замыкания всех однобуквенных подмножеств находящегося в корне этого псевдо-дерева объекта). Далее реконструкция рассматриваемого псевдо-дерева может быть продолжена рекурсивным применением процедуры порождения каркаса на специальным образом выделяемых последовательно вложимых друг в друга подграфах анализируемого псевдо-дерева. При этом порождение множества вершин каждого такого псевдо-дерева может быть выполнено процедурой, вычислительная сложность которой ограничена сверху полиномиальными (от параметров исходных данных – размеров алфавита и обучающей выборки) оценками, а число подобных псевдо-деревьев, очевидным образом, совпадает с числом примеров в исходной обучающей выборке. Наконец, структура связей между вершинами в каждом таком псевдо-дереве полностью определяется специальными легко проверяемыми условиями, и имеет всего три варианта реализации:

- в одном (архитектура α-типа) число ребер (а с ними и число возможных веток частичного порядка взаимной вложимости множеств букв, сопоставленных вершинам этого псевдо-дерева) ограничено полиномом от числа (количества элементов реконструируемого полиномиально сложным алгоритмом множества) вершин соответствующего псевдо-дерева;
- в другом (архитектура β-типа) множество ребер (между элементами полиномиально быстро «узнаваемого» множества вершин, образующих[6] n-мерный куб без минимальной вершины) есть подмножество множества всех ребер соответствующего n-мерного куба, из которого удалены все такие ребра (ровно n-штук), которые связывают вершину <0,0,… ,0> с вершинами <1,0,0,… ,0>, <0,1,0,0, … , 0>, … , <0,0, …,0,1>;
- в третьем (архитектура γ-типа) соответствующее множество ребер формируется «наложением» фрагментов типа α или β на некоторые подмножества веток во фрагменте типа β и также как и в предыдущих случаях имеет либо полиномиально быстро растущий (с ростом числа вершин исходного псевдо-дерева) – как в случае архитектур типа α, либо полиномиально быстро узнаваемый (как в случае архитектур типа β) размер (и структуру) «накладываемых» фрагментов.

Таким образом, имеется возможность целенаправленно (порождая быстрыми алгоритмами каркас и затем последовательно достраивая его в случае необходимости до все более и более точных приближений соответствующего псевдо-дерева) выделять все три (α-, β- и γ-типа) части архитектуры соответствующего псевдо-дерева – поддиаграммы для диаграммы взаимной вложимости ДСМ-сходств. Используя это обстоятельство, можно организовать анализ всего множества ДСМ-гипотез последовательным (управляемым теми или иными содержательными соображениями) перебором вдоль веток частичного порядка в каждом из порожденных представленным образом псевдо-деревьев – под-диаграмм множества всех ДСМ-гипотез, формируемых на заданном множестве (обучающих) примеров. При этом, как будет показано ниже, порождаемые представленным образом псевдо-деревья частичного порядка ДСМ-гипотез удовлетворяют сформулированным выше «пожеланиям» в части

---

[4] Представляется уместным использовать этот термин, подчеркивая, что поддиаграммы предлагаемого вида, вообще говоря, могут и не быть деревьями (например, за счет существования в них пар вершин во фрагментах типа β - см. ниже, находящихся по крайней мере на двух различных ветках частичного порядка взаимной вложимости соответствующих замыканий Галуа), однако, могут быть сведены к виду дерева удалением тех или иных подмножеств входящих в них ребер. (В свою очередь, каждое такое псевдо-дерево может быть представлено как объединение соответствующих деревьев, содержащих то же самое множество вершин и вкладывающихся в рассматриваемое псевдо-дерево).

[5] В соотнесении с характерными размерами исходных алфавита и обучающей выборки примеров.

[6] Для соответствующего (также быстро диагностируемого) значения n.

декомпозиции исходной диаграммы ДСМ-сходств на «самостоятельные» фрагменты с «удобными» характеристиками.

Предлагаемая техника организации перебора ДСМ-сходств предоставляет возможности для *выборочного* порождения ДСМ-гипотез (в частности, лишь тех из них, которые имеют отношение к ДСМ-доопределению заданных новых объектов и при этом удовлетворяют тем или иным структурным ограничениям, например:

- на состав задействованных в их структуре образующих – букв исходного алфавита

или же

- на участие в порождении соответствующей гипотезы тех или иных выделенных «родителей» - конкретных примеров из исходной обучающей выборки,

и т.п.). Таким путем может быть реализована технология *управляемых приближенных вычислений* в рамках (традиционного) ДСМ-метода: технология *направленного* анализа ДСМ-гипотез, базирующаяся на *управляемой навигации* вдоль веток частичного порядка в диаграмме взаимной вложимости ДСМ-сходств, в предельном случае (при анализе[7] всех веток всех псевдо-деревьев частичного порядка) порождающая те же результаты, что и обычно используемые в программной реализации ДСМ-метода алгоритмы.

Сперва наше внимание будет сфокусировано на способах формирования а также на структурных и комбинаторных свойствах специально порождаемых для управления перебором при порождении ДСМ-гипотез объектов – так называемых псевдо-деревьев и их каркасов. Мы обсудим возможности *целенаправленного* порождения *приближенных* описаний множеств всех ДСМ-гипотез, формируемых из заданного множества примеров **Ω**. Далее (в **Разделе 5.3**) мы перейдем к детальному рассмотрению особенностей алгоритмики предлагаемого направленного перебора ДСМ-сходств, некоторым оценкам сложности вычислений а также обсуждению особенностей скоординированного использования развиваемых алгоритмических конструкций на обеих частях зависимостей вида < *объект* O $=>_1$ *свойства* $\boldsymbol{P}$ > (*объект* O *обладает множеством свойств* $\boldsymbol{P}$).

* * *

Обратимся к детальному обсуждению предлагаемого подхода. Пусть в нашем распоряжении имеются:

- исходный алфавит $\boldsymbol{U}$ (множество образующих для анализируемых примеров)

$$\boldsymbol{U} = \{a_1, a_2, \dots, a_n\},$$

- множество (примеров) **Ω** - объектов (т.е. непустых множеств образующих), построенных над универсумом $\mathbf{U}$[8]**:**

$$\boldsymbol{\Omega} = \{O_1, O_2, \dots, O_m\} \subseteq 2^{\mathbf{U}} \setminus \varnothing .$$

Располагая множествами $\boldsymbol{U}$ и **Ω,** определим два отображения f и φ:

$\forall \xi \in 2^{\mathbf{U}}$ (т.е. для каждого ξ - подмножества образующих из множества $\boldsymbol{U}$)

f (ξ) = {множество всех таких **Y** из **Ω** , что для $\forall \mathbf{x} \in \xi$ имеет место $\mathbf{x} \in \mathbf{Y}$ для каждого из этих выбранных **Y** (т.е. это множество всех примеров из **Ω,** в которые все образующие **x** из заданного ξ входят одновременно)}

$\forall \zeta \in 2^{\boldsymbol{\Omega}}$ (т.е. для каждого ζ - подмножества примеров из множества **Ω**)

φ (ζ) = {множество всех таких **x** из $\boldsymbol{U}$, что для $\forall \mathbf{Y} \in \zeta$ имеет место $\mathbf{x} \in \mathbf{Y}$ для каждого из этих выбранных **x** (т.е. это множество всех таких образующих **x** из $\boldsymbol{U}$**,** которые во все примеры из заданного ζ входят одновременно)}

---

[7] Последовательном или же параллельном - !

[8] Т.е. множество объектов, построенных из образующих $a_1, a_2, \dots, a_n$.

(при этом $2^{\mathbf{Z}}$ - множество всех подмножеств множества **Z** (полагая здесь $\mathbf{Z}\in\{\boldsymbol{U},\boldsymbol{\Omega}\}$)).

Несложно убедиться, что пара отображений <f, φ> представляет собою соответствие Галуа[9], а их произведения f(φ(ζ)) и φ(f(ξ)) - соответствующие замыкания Галуа[10]: $[_]_{\boldsymbol{U},\boldsymbol{\Omega}}$ и $[_]_{\boldsymbol{\Omega},\boldsymbol{U}}$. Посредством $\mathbf{GC}_{f,\varphi}(\boldsymbol{\Omega})$ и $\mathbf{GC}_{\varphi,f}(\boldsymbol{U})$ будем обозначать множества неподвижных точек соответствующих замыканий Галуа:

$$\mathbf{GC}_{f,\varphi}(\boldsymbol{\Omega}) = \{\ \zeta\in 2^{\boldsymbol{\Omega}} \text{ таких, что } f(\varphi(\zeta)) = \zeta\ \}$$

$$\mathbf{GC}_{\varphi,f}(\boldsymbol{U}) = \{\ \xi\in 2^{\boldsymbol{U}} \text{ таких, что } \varphi(f(\xi)) = \xi\ \}$$

Каждое из этих множеств можно рассматривать как частично упорядоченное в соответствии как со взаимным вложением соответствующих множеств объектов (подмножеств для **Ω**), так и подмножеств образующих (подмножеств для ***U***).

5.2.1. Определение.

Неподвижными точками замыкания Галуа $[_]_{\boldsymbol{U},\boldsymbol{\Omega}}$ будем называть все такие $[X]_{\boldsymbol{U},\boldsymbol{\Omega}}$, что:

$$[X]_{\boldsymbol{U},\boldsymbol{\Omega}} = X\ .$$

□

Для каждой пары множеств ***U*** (алфавит образующих) и **Ω** (некоторое множество примеров, построенных из образующих алфавита ***U***) будем рассматривать множество $\mathbf{GC}_{\varphi,f}(\boldsymbol{U})$ неподвижных точек замыкания Галуа $[_]_{\boldsymbol{U},\boldsymbol{\Omega}}$ как частично упорядоченное – по взаимному вложению соответствующих подмножеств образующих из ***U***. Соответствующую диаграмму вложимости обозначим как $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$.

Пусть $\boldsymbol{O} = \{O_{i1}, O_{i2}, \ldots, O_{is}\}\subseteq\boldsymbol{\Omega}$ - некоторое подмножество исходно заданного множества примеров **Ω**. Сужением

$$\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\ \big|_{\boldsymbol{O}}$$

диаграммы $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$ на множество ***O*** будем называть такой фрагмент этой диаграммы, который образован заданными объектами (примерами) $O_{i1}, O_{i2}, \ldots, O_{is}$ а также всеми теми ее элементами, которые, в свою очередь, сформированы образующими из множества

$$O_{i1}\cup O_{i2}\cup\ldots\cup O_{is}$$

и только ими. Так, например, сужение

$$\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\ \big|_{\{O_0\}}$$

диаграммы **D_GC** на отдельный объект O_0 из множества примеров **Ω** есть псевдо-дерево – под-диаграмма исходной диаграммы, сформированное всеми неподвижными точками соответствующего замыкания Галуа $[_]_{\boldsymbol{U},\boldsymbol{\Omega}}$, построенными из образующих объекта O_0 и только из них, имеющее при этом в своем корне объект O_0 (множество всех его образующих).

5.2.2. Замечание.

Несложно убедиться, что диаграммы $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\boldsymbol{O}}$ и $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{O})$, могут и не совпадать: пример такого несовпадения дает, в частности, конфигурация $\boldsymbol{U} = \{a_1,a_2,\ldots,a_5\}$, $\boldsymbol{\Omega} = \{\{a_1,a_2,a_5\},\{a_1,a_2,a_4\},\{a_1,a_3\}\}$ и $\boldsymbol{O} = \{\{a_1,a_2,a_4\},\{a_1,a_3\}\}$[11].

□

---

[9] См., например, [*Гусакова 87, Кон 68*] и др. .

[10] См. предыдущую сноску.

[11] Здесь для $\boldsymbol{O} = \{\{a_1,a_2,a_4\},\{a_1,a_3\}\}$ имеем: вершина $\{a_1,a_2\}$ имеется в диаграмме $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\boldsymbol{O}}$, однако, отсутствует в диаграмме $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{O})$.

Технологии восстановления подобных псевдо-деревьев, порождаемых в точности одним объектом (примером) из исходной обучающей выборки, мы и будем детально анализировать далее.

Прежде всего, рассмотрим имеющиеся возможности диагностировать принадлежность заданного множества $\boldsymbol{u}$ образующих из алфавита $\boldsymbol{U}$ к множеству вершин соответствующей диаграммы $\mathbf{D_GC}_{\varphi,f}$ ($\boldsymbol{U}$, $\boldsymbol{\Omega}$). Простейший вариант ответа на подобный вопрос – проверка (прямым вычислением соответствующего замыкания) выполнимости условия

$$[\boldsymbol{u}]_{\boldsymbol{U},\boldsymbol{\Omega}} = \boldsymbol{u}\ ^{12}.$$

Дополнительные возможности в «диагностике» замкнутости множеств образующих предлагает нижеследующая процедурная конструкция.

Пусть заданы

- множество примеров $\boldsymbol{\Omega} = \{O_1, O_2, \ldots, O_m\}$, построенных на алфавите $\boldsymbol{U} = \{a_1, a_2, \ldots, a_n\}$ и
- некоторое непустое подмножество $\mathbf{u} = \{u_1, u_2, u_3, \ldots, u_k\}$ исходного алфавита $\boldsymbol{U}$.

Рассмотрим следующую рекурсивную процедуру:

(**i**) строим по $\boldsymbol{\Omega}$ и $\boldsymbol{U}$ множество замыканий всех одноэлементных подмножеств алфавита $\boldsymbol{U}$:

$$\boldsymbol{B_GC}_{\boldsymbol{s*g}}(\mathbf{U},\boldsymbol{\Omega}) = \{[\{a_1\}],[\{a_2\}], \ldots,[\{a_n\}]\}$$

(**ii**) выбираем в $\mathbf{u}$ максимальный элемент $\boldsymbol{B_GC}_{\boldsymbol{s*g}}(\mathbf{U},\boldsymbol{\Omega})$. Не теряя общности можем считать, что это будет u_1 ;

(**iii**) выделяем в $\boldsymbol{\Omega}$ подмножество $\boldsymbol{\Omega}(u_1)$ всех содержащих $[u_1]_{\boldsymbol{U},\boldsymbol{\Omega}}$ примеров из $\boldsymbol{\Omega}$;

(**iv**) удаляем из каждого входящего в $\boldsymbol{\Omega}(u_1)$ примера образующие из $[u_1]_{\boldsymbol{U},\boldsymbol{\Omega}}$ и формируем (этими удалениями) новое – модифицированное - множество примеров $\boldsymbol{\Omega}^*(u_1)$. Таким образом в нашем распоряжении оказываются усеченное множество примеров $\boldsymbol{\Omega}^*(u_1)$ и использованный при его формировании усеченный алфавит $\boldsymbol{U}(u_1)$;

(**v**) удаляем из текущего $\mathbf{u}$ все образующие, вошедшие в $[u_1]_{\boldsymbol{U},\boldsymbol{\Omega}}$, после чего

- возвращаемся к построению $\boldsymbol{B_GC}_{\boldsymbol{s*g}}$, но теперь уже на усеченных множествах $\boldsymbol{\Omega}^*(u')$ и $\boldsymbol{U}(u')$, т.е. переходим к построению множества замыканий всех одноэлементных подмножеств усеченного алфавита $\boldsymbol{U}(u')$ относительно усеченного множества примеров $\boldsymbol{\Omega}^*(u')$, где текущее значение u' получено удалением из текущего $\mathbf{u}$ всех образующих из текущего $[u_1]_{\boldsymbol{U},\boldsymbol{\Omega}}$, после чего
- повторяем процедуру (**i**)-(**v**) на усеченных множествах $\boldsymbol{\Omega}^*(u')$ и $\boldsymbol{U}(u')$ рекурсивно до момента, когда в множестве $\mathbf{u}$ (в его текущем состоянии, также усекаемом на каждом цикле процедуры (**i**)-(**v**)) больше не останется ещё не удалённых образующих. При этом будем обозначать последовательность $u_1, u_2, \ldots, u_l$ (где $l \le k$) из $\mathbf{u}$ образующих исходного алфавита $\boldsymbol{U}$, удаляемых на каждом из шагов предпринимаемой рекурсии, как $T(\mathbf{u},\boldsymbol{\Omega},\boldsymbol{U})$ и называть эту последовательность *траекторией усечения* исходного множества $\mathbf{u}$.

Полезные свойства описанной рекурсивной процедуры представляет

5.2.3. Утверждение (*Теорема о представлении*[13]).

Множество образующих $\mathbf{u}$ замкнуто относительно множества примеров $\boldsymbol{\Omega}$ в алфавите $\boldsymbol{U}$ тогда и только тогда, когда объединение всех замыканий $[u']_{\boldsymbol{U},\boldsymbol{\Omega}}$ вдоль траектории $T(\mathbf{u},\boldsymbol{\Omega},\boldsymbol{U})$ совпадает с $\mathbf{u}$:

$$[\mathbf{u}] = \mathbf{u} \qquad \textit{тогда и только тогда, когда} \qquad \Big(\bigcup_{u'_i \in T(u,\Omega,U)} [u'_i]\Big) = \mathbf{u}.$$

---

[12] Что может быть исполнено (полиномиально – от соответствующих размеров $\boldsymbol{U}$ и $\boldsymbol{\Omega}$) быстро.

[13] О представлении «составных» (многоэлементных) неподвижных точек рассматриваемого замыкания Галуа с помощью замыканий одноэлементных подмножеств алфавита $\boldsymbol{U}$.

Доказательство.
=>. Несложно убедиться, что достаточным условием незамкнутости **u** = { a_1, a_2, a_3, … , a_k} будет строгое включение

$$\mathbf{u} \subset \bigcup_{i=1}^{k} [\{a_i\}].$$

Однако, если **u** замкнуто, то

$$\mathbf{u} = \bigcup_{i=1}^{k} [\{a_i\}],$$

а вместе с этим выполняется и равенство

$$\mathbf{u} = \Big(\bigcup_{u'_i \in T(u,\Omega,U)} [u'_i] \Big) = \mathbf{u},$$

т.к. в нем некоторые из [{a_i}] из множества {[{ u_1}],[{ u_2}], …,[{ u_k}]} будут иметь кратные вхождения в правую часть этого соотношения (отвечающего за объединение замыканий вдоль траектории T(**u**,**Ω**,***U***)), что, разумеется, не сможет вывести их объединение за пределы **u** (ведь **u** - замкнуто).
<=. Финальное множество примеров **Ω***, порожденное при последовательном «усечении» всего множества **u**, содержит (согласно процедуре своего построения) каждое множество [u'$_i$] из траектории T(**u**,**Ω**,***U***) в качестве максимального (на соответствующем шаге «усечения» - !) по вложению элемента. Таким образом:

$$\mathbf{u} \subseteq \bigcap_{Oj \in \Omega^*} O_j \, .$$

Если же при этом ни в одном из [u'$_i$] нет выходящих за пределы **u** образующих, то в правой части этого соотношения также нет находящихся вне **u** элементов исходного алфавита ***U***, что означает выполнение равенства [**u**] = **u** , т.к. в финальном **Ω*** оказываются собранными все те и только те O_j из исходного **Ω** , которые содержат в себе рассматриваемое **u**.

□

5.2.4. Пример

Положим ***U*** = {a_1, a_2, a_3,a_4, a_5, a_6, a_7}, а **Ω** = {h_1, h_2, h_3, h_4}, где

$h_1 = \{a_1, a_2, a_4\}$, $h_2 = \{a_1, a_2, a_5\}$,
$h_3 = \{a_1, a_3, a_6\}$, $h_4 = \{a_1, a_3, a_7\}$.

Несложно убедиться, что:

$$[\{a_1\}]_{U,\Omega} = \{a_1\},$$
$$[\{a_2\}]_{U,\Omega} = \{a_1, a_2\} \supset^{14} \{a_2\},$$
$$[\{a_3\}]_{U,\Omega} = \{a_1, a_3\} \supset \{a_3\},$$

тем не менее:

$$h_1 = [\{a_1, a_2, a_4\}]_{U,\Omega} = [\{a_4\}]_{U,\Omega} = \{a_1, a_2, a_4\} \supset \{a_4\},$$
$$h_2 = [\{a_1, a_2, a_5\}]_{U,\Omega} = [\{a_5\}]_{U,\Omega} = \{a_1, a_2, a_5\} \supset \{a_5\},$$
$$h_3 = [\{a_1, a_3, a_6\}]_{U,\Omega} = [\{a_6\}]_{U,\Omega} = \{a_1, a_3, a_5\} \supset \{a_6\},$$
$$h_4 = [\{a_1, a_3, a_7\}]_{U,\Omega} = [\{a_7\}]_{U,\Omega} = \{a_1, a_3, a_7\} \supset \{a_7\}.$$

□

Вернемся к идее разбиения диаграммы **D_GC**$_{\varphi,f}$ (***U*, Ω**) на «самостоятельно» анализируемые фрагменты. Обратимся к сужениям рассматриваемой диаграммы на одноэлементные множества примеров. Несложно убедиться, что для каждого объекта (примера)

---

[14] Т.е. множество {a_2} есть подмножество множества {a_1, a_2} , не совпадающее с последним.

$$O_0 \in \mathbf{\Omega} = \{O_1, O_2, \dots, O_m\}$$

каждое такое сужение – это псевдо-дерево:

а) с соответствующим O_0 в корне;

б) листьями в виде замкнутых одноэлементных[15] подмножеств O_0 (образующих из алфавита $\boldsymbol{U}$);

в) ветками частичного порядка, сформированными последовательно вкладывающимися (сверху – вниз по этому псевдо-дереву) замкнутыми непустыми подмножествами множества O_0.

Обратимся к уточнению характеристик обсуждаемого псевдо-дерева. Прежде всего, оценим длину и возможное число его веток, а также возможную длину каждой из них. Имеет место следующее

5.2.5. Утверждение.

Пусть заданы алфавит $\boldsymbol{U}$ и множество объектов (примеров) $\mathbf{\Omega}$, а также выбран объект $O_0 = \{ a_{i1}, a_{i2}, \dots, a_{i\,n0} \}$ из $\mathbf{\Omega}$, в который входит соответственно ровно n_0 образующих из $\boldsymbol{U}$.

1. Каждая ветка рассматриваемого нами частичного порядка (взаимной вложимости сопоставленных вершинам множеств образующих) в каждом псевдо-дереве $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\mathbf{\Omega})\big|_{\{O_0\}}$ содержит не более n_0 вершин.

2. На некоторых «этажах»[16] этого псевдо-дерева может располагаться множество вершин, размеры которого растут экспоненциально быстро с ростом значений $|O|$, $|\boldsymbol{U}|$ и $|\mathbf{\Omega}|$.

Доказательство.

При движении от корня этого псевдо-дерева к каждому из не более чем n_0 возможных листьев от находящегося в корне псевдо-дерева множества O_0 на каждом шаге должна «отщепляться» по крайней мере одна образующая, следовательно, таких «отщеплений» может быть не более n_0. Таким образом, первая часть **Утверждения 5.2.5** доказана.

Представление о ситуациях, обозначенных во второй части этого **Утверждения** дает

5.2.6. Пример

Пусть $\boldsymbol{U} = \{a_1, a_2, \dots, a_r\} = A_0$, при этом $\mathbf{\Omega} = \{O_1, O_2, \dots, O_r, O_0\}$ и

$O_1 = \{ a_2, a_3, \dots, a_r \}$

$O_2 = \{ a_1, a_3, \dots, a_r \}$

.....

$O_r = \{ a_1, a_2, \dots, a_{r-1} \}$.

Несложно убедиться, что в данном случае псевдо-дерево $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U}, \mathbf{\Omega})\big|_{\{O_0\}}$ будет представлять собою фрагмент r-мерного куба, образованный всеми его вершинами, кроме наибольшего и наименьшего элементов – т.е. вершин <0,0, … , 0> и <1,1, … , 1>, а также смежных с ними 2r ребер.

□

Однако, надежды на построение быстрой процедуры восстановления псевдо-деревьев рассматриваемого нами вида по заданным множествам $\boldsymbol{U}$ и $\mathbf{\Omega}$ позволяет сохранить возможность быстрой диагностики наличия в конкретной диаграмме $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U}, \mathbf{\Omega})\big|_{\{O\}}$ фрагментов только что представленного в **Примере 5.2.6** типа. Действительно, вычислим

---

[15] Или же приводимых к такому виду заменой в данном псевдо-дереве соответствующих «склеенных» образующих одной новой образующей.

[16] «Этаж» рассматриваемого псевдо-дерева – множество его вершин, достижимых (в части минимального пути при движении из корня) через одно и то же число ребер из корня этого псевдо-дерева (первый этаж – корень, второй – смежные с ним вершины, третий – вершины, смежные с вершинами второго этажа, и т.д.).

замыкания всех одноэлементных подмножеств множества $O = \{ a_1, a_2, \dots, a_n \}$ и выберем из них неподвижные точки[17]:

$$a_{i1}, a_{i2}, \dots, a_{ir}.$$

Рассмотрим далее все двухэлементные[18] (по-парные) подмножества этого множества. Те из них -

$$a_{j1}, a_{j2}, \dots, a_{js} \text{ (здесь } s \le r),$$

что также представляют собою неподвижные точки замыкания Галуа $[_]_{U,\Omega}$, далее, как нетрудно убедиться (легко видеть, что это происходит за счет формирования по-троечных, по-четверочных и т.д. подмножеств множества A в ситуации, когда все эти последовательно порождаемые подмножества образующих вкладываются в исходно выбранное множество O), порождают в диаграмме $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U}, \boldsymbol{\Omega})\big|_{\{O\}}$ соответствующий s–мерный куб (без наименьшей вершины <0,0, … ,0> и s исходящих из нее ребер). Таким образом, справедливо

5.2.7. Утверждение

Пусть заданы множества O, $\boldsymbol{U}$ и $\boldsymbol{\Omega}$. Объем вычислений при реализации процедуры, позволяющей диагностировать наличие в диаграмме $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U}, \boldsymbol{\Omega})\big|_{\{O\}}$ соответствующего фрагмента s–мерного куба (для некоторого $0 \le s \le |O|$) представленного выше вида, растет полиномиально быстро с ростом значений $|O|$, $|\boldsymbol{U}|$ и $|\boldsymbol{\Omega}|$.

□

Будем называть такие (комбинаторно сложные) фрагменты диаграммы $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U}, \boldsymbol{\Omega})\big|_{\{O\}}$ *архитектурами типа* β.

В свою очередь *архитектуры типа* α формируются всеми парами соседних вершин $<h_1, h_2>$ следующего вида:

(α_1) $\{a_1, a_2, \dots, a_k\} = h_1 \subset h_2 = \{a_1, a_2, \dots, a_k, a_{k+1}, \dots\}$,

(α_2) среди образующих – элементов множества h_2 – имеется такая a_0, что $h_2 = [a_0]$,

(α_3) для каждой вершины (множества образующих) h_1 множество h_2, формирующее вместе с нею рассматриваемую пару $<h_1, h_2>$, представляет собою минимальный элемент в $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U}, \boldsymbol{\Omega})\big|_{\{O\}}$, удовлетворяющий условиям (α_1) и (α_2) .

Наглядные представления об архитектурах типов α и β помогает сформировать

5.2.8. Пример

Пусть $\boldsymbol{U} = \{a_1, a_2, \dots, a_{20}\}$, $\boldsymbol{\Omega} = \{ h_1, h_2, \dots, h_{13}, h_{14}\}$ и

$h_1 = \{ a_1, a_2 \}$, $h_2 = \{ a_1, a_6, a_9 \}$, $h_3 = \{ a_2, a_3, a_4 \}$, $h_4 = \{ a_2, a_3, a_5 \}$, $h_5 = \{ a_2, a_4, a_5 \}$,
$h_6 = \{ a_3, a_4, a_5 \}$, $h_7 = \{ a_4, a_5 \}$, $h_8 = \{ a_1, a_2, \dots, a_6, a_{11}, a_{12}, \dots, a_{16} \}$,
$h_9 = \{ a_2, a_3, \dots, a_5, a_{10} \}$, $h_{10} = \{ a_1, a_2, \dots, a_6, a_{14}, a_{15} \}$,
$h_{11} = \{ a_1, a_2, \dots, a_6, a_{11}, a_{12}, \dots, a_{15}, a_{20} \}$, $h_{12} = \{ a_1, a_2, \dots, a_6, a_{12}, a_{17} \}$,
$h_{13} = \{ a_1, a_2, \dots, a_6, a_{13}, a_{18} \}$, $h_{14} = \{ a_1, a_2, \dots, a_6, a_{11}, a_{12}, \dots, a_{15}, a_{19} \}$.

Диаграмма $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U}, \boldsymbol{\Omega})$ в данном случае имеет представленный на Рис.5.2.1 вид (см. ниже).

Цепочки частичного порядка, образованные, например, его вершинами

$$< \{ a_1 \}, \{ a_1, a_6 \}, \{ a_1, a_2, \dots, a_6 \}, \{ a_1, a_2, \dots, a_6, a_{11} \}, \{ a_1, a_2, \dots, a_6, a_{11}, a_{16} \} >$$

---

[17] С учетом возможностей «склейки» образующих.

[18] Здесь и далее, если это не будет оговорено особо, следует учитывать возможный эффект «склейки» рассматриваемых образующих.

или

$< \{ a_1\}, \{ a_1, a_6\}, \{ a_1, a_2, \dots, a_6 \}, \{ a_1, a_2, \dots, a_6, a_{12} \},$
$\{ a_1, a_2, \dots, a_6, a_{11}, a_{12}, \dots, a_{15}\}, \{ a_1, a_2, \dots, a_6, a_{11}, a_{12}, \dots, a_{15}, a_{20} \} >,$

являются примерами архитектур типа α.

В свою очередь, выделенная темным фоном в овале на Рис.5.2.1 область (исключая вершины $h_3, h_4, \dots, h_7$) – пример архитектуры типа β.

Наконец, *архитектуры типа* γ формируются «наложением» фрагментов типа α и\или β на некоторые подмножества веток в уже построенном ранее[19] фрагменте типа β. Фактически здесь мы имеем дело либо с «одиночными» цепями частичного порядка[20] типа α, или же с множествами таких цепей, представленными фрагментами дополнительных[21] булевских кубов типа β, причем и те и другие «наложены» на некоторые ветки «ранее» (т.е. - в смысле избранного порядка восстановления псевдо-дерева: от его листьев к корню) уже сформированных фрагментов типа β в рассматриваемой нами исходной диаграмме вложимости замыканий Галуа $[_]_{U,\Omega}$.

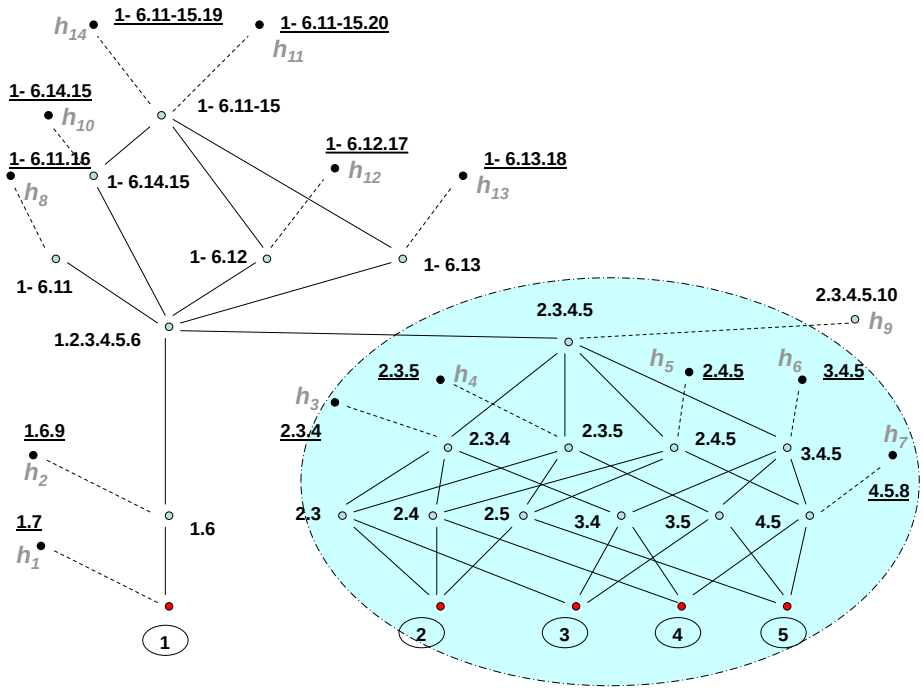


Рис. 5.2.1. Пример архитектур типов α и β .

Наглядные представления об архитектурах типа γ помогает сформировать, в частности,

5.2.9. Пример

Пусть $\boldsymbol{U} = \{a_1, a_2, \dots, a_9\}$, $\boldsymbol{\Omega} = \{ h_1, h_2, \dots, h_7\}$ и

$h_1 = \{ a_1, a_2, a_3, a_5 \}$, $h_2 = \{ a_1, a_2, a_4 \}$, $h_3 = \{ a_1, a_3, a_4 \}$, $h_4 = \{ a_1, a_2, \dots, a_5, a_8 \}$,
$h_5 = \{ a_1, a_2, \dots, a_5, a_9 \}$, $h_6 = \{ a_2, a_3, \dots, a_6 \}$, $h_7 = \{ a_1, a_2, \dots, a_5, a_7 \}$.

---

[19] При формировании соответствующего псевдо-дерева $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U}, \boldsymbol{\Omega})\big|_{\{O\}}$ *от* его *листьев* к *корню*.
[20] В смысле взаимной вложимости подмножеств образующих из алфавита $\boldsymbol{U}$.
[21] Полученные за счет выявления «новых» «склеенных» образующих.

Диаграмма $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U, \Omega})$ в данном случае имеет представленный на Рис.5.2.2 вид. Здесь цепочки частичного порядка, образованные, например, следующими вершинами (см. выделения жирным на Рис.5.2.2)

$$< \{ a_2, a_3, a_5 \}, \{ a_2, a_3, a_4, a_5 \}, \{ a_1, a_2, \dots, a_5 \}, \{ a_1, a_2, \dots, a_5, a_9 \} >$$

или

$$< \{ a_2, a_3, a_5 \}, \{ a_2, a_3, \dots, a_5 \}, \{ a_1, a_2, \dots, a_5, a_7 \} >,$$

являются примерами архитектур типа γ.

В свою очередь, выделенная темным фоном в овале на Рис.5.2.2 область (исключая вершины $h_1, h_2, \dots, h_7$) – пример архитектуры типа β.

Полезную информацию о внутренней структурной организации каждого из рассматриваемых псевдо-деревьев может предоставить его специальный подграф – каркас.

5.2.10. Определение.

Будем называть *каркасом* псевдо-дерева $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U,\Omega})\,|_{\{O\}}$ граф

$$GK_{\mathbf{D},O} = < \mathbf{K}_{\mathbf{D},O}, \mathbf{R}_{\mathbf{D},O} >,$$

множество $\mathbf{K}_{\mathbf{D},O}$ вершин которого образовано замыканиями всех образующих из выбранного множества (примера) $O = \{a_1, a_2, \dots, a_n\}$:

$$\mathbf{K}_{\mathbf{D},O} = \{[a_1]_{U,\Omega}, [a_2]_{U,\Omega}, \dots, [a_n]_{U,\Omega}\},$$

а множество ребер $\mathbf{R}_{\mathbf{D},O}$ отражает «ближайшие[22]» по-парные взаимные вложимости элементов из $\mathbf{K}_{\mathbf{D},O}$ друг в друга.

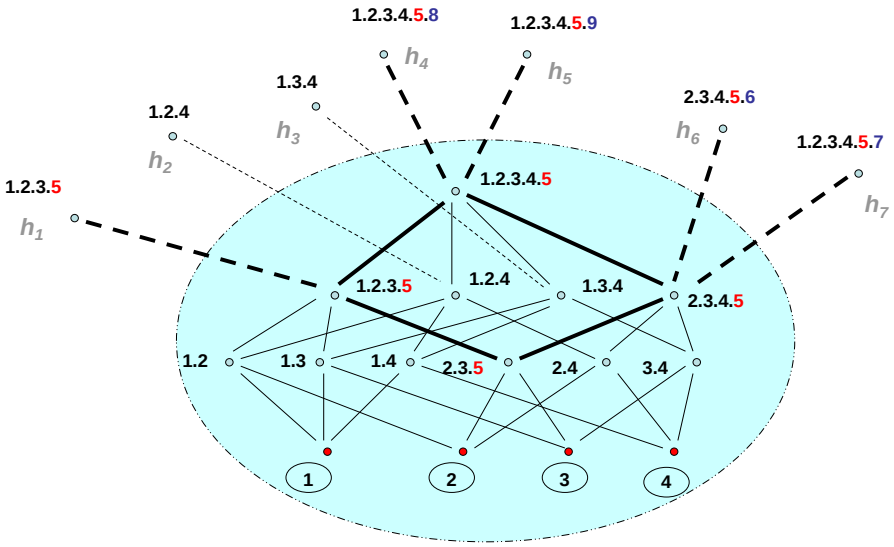


Рис. 5.2.2. Пример архитектуры типа γ .

Имеет место

[22] Т.е. если $\{\mathbf{a},\mathbf{b}\} \subseteq \mathbf{K}_{\mathbf{D},O}$, то ребро $< \mathbf{a},\mathbf{b} >$ принадлежит $\mathbf{R}_{\mathbf{D},O}$ тогда и только тогда, когда $\mathbf{a} \subseteq \mathbf{b}$, и неверно, что в $\mathbf{K}_{\mathbf{D},O}$ найдется (отличный от **b**) элемент **c** такой, что $\mathbf{a} \subseteq \mathbf{c} \subseteq \mathbf{b}$.

5.2.11. Утверждение.

При заданных O **,** $\boldsymbol{U}$ и $\boldsymbol{\Omega}$ каркас $GK_{\mathbf{D},O}$ может быть порожден процедурой полиномиальной (от $|\boldsymbol{U}|$ и $|\boldsymbol{\Omega}|$) сложности.

Доказательство.

В множестве $\mathbf{K}_{\mathbf{D},O}$ имеется не более $|\boldsymbol{U}|$ элементов, каждый из которых (как соответствующее замыкание Галуа $[_]_{\boldsymbol{U},\boldsymbol{\Omega}}$) порождается процедурой полиномиальной вычислительной сложности. В свою очередь проверка взаимной вложимости вершин множества $\mathbf{K}_{\mathbf{D},O}$ также требует не более $O(|\boldsymbol{U}|^2)$ операций.

□

Также справедливо

5.2.12. Утверждение.

Пусть заданы множества O**,** $\boldsymbol{U}$ и $\boldsymbol{\Omega}$**,** при этом в множестве $\boldsymbol{\Omega}$ нет дубликатов**.** В представленном случае каркас $\mathbf{K}_{\mathbf{D},O}$ определяет единственное псевдо-дерево $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$ в соответствующей диаграмме $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$.

Доказательство.

Предположим противное. Тогда в множестве $\boldsymbol{\Omega}$ должна найтись пара примеров O_1 и O_2 , таких что $\mathbf{K}_{\mathbf{D},O1} = \mathbf{K}_{\mathbf{D},O2}$. Следовательно

$$\mathbf{K}_{\mathbf{D},O1} = \{[a_{11}]_{\boldsymbol{U},\boldsymbol{\Omega}}, [a_{12}]_{\boldsymbol{U},\boldsymbol{\Omega}}, \dots, [a_{1n}]_{\boldsymbol{U},\boldsymbol{\Omega}}\} = \mathbf{K}_{\mathbf{D},O2} = \{[a_{21}]_{\boldsymbol{U},\boldsymbol{\Omega}}, [a_{22}]_{\boldsymbol{U},\boldsymbol{\Omega}}, \dots, [a_{2n}]_{\boldsymbol{U},\boldsymbol{\Omega}}\},$$

т.е. имеем $O_1 = O_2$ (как одно и то же множество образующих $\{a_{11},a_{12},\dots,a_{1n}\}=\{a_{21},a_{22},\dots,a_{2n}\}$), что противоречит исходным допущениям об отсутствии дубликатов в множестве $\boldsymbol{\Omega}$.

Дополнительно отметим, что

1) при наличии дубликатов O_1 и O_2 в множестве примеров $\boldsymbol{\Omega}$, ситуация, когда

$$\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O1\}} = \mathbf{D_1} \neq \mathbf{D_2} = \mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O2\}}$$

означает, что по крайней мере в одной из этих под-диаграмм найдется вершина, которой нет в другой диаграмме. Пусть это будет вершина $\boldsymbol{u}$ в $\mathbf{D_1}$, которая получена сходством некоторого подмножества примеров $\boldsymbol{\Omega}_{\boldsymbol{u}} \subseteq \boldsymbol{\Omega}$. Т.к. $\mathbf{D_1}$ и $\mathbf{D_2}$ формируются на одном и том же множестве примеров $\boldsymbol{\Omega}$, при построении $\mathbf{D_2}$ подмножество $\boldsymbol{\Omega}_{\boldsymbol{u}}$ множества всех примеров $\boldsymbol{\Omega}$ даст нам тот же результат $\boldsymbol{u}$ (т.е. предположение $\mathbf{D_1} \neq \mathbf{D_2}$ при $\mathbf{K}_{\mathbf{D},O1} = \mathbf{K}_{\mathbf{D},O2}$ невыполнимо);

2) при рассмотрении двух различных множеств примеров $\boldsymbol{\Omega}_1$ и $\boldsymbol{\Omega}_2$ ситуация, когда порождаемый объектом O_1 каркас $\mathbf{K}_{\mathbf{D}1,O1}$ псевдо-дерева $\mathbf{D_1}$ в диаграмме $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega}_1)$ совпадает с каркасом $\mathbf{K}_{\mathbf{D}2,O2}$ псевдо-дерева $\mathbf{D_2}$ в диаграмме $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega}_2)$, однако при этом $\mathbf{D_1} \neq \mathbf{D_2}$, возможна, что иллюстрирует приведенный ниже пример множеств $\boldsymbol{\Omega}_1$ и $\boldsymbol{\Omega}_2$:, построенных на одном и том же алфавите $\boldsymbol{U}$:

$$\boldsymbol{U} = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8\}$$

$\boldsymbol{\Omega}_1 = \{h_1, h_2, h_3, h_4, h_5, h_6, h_7, h_8\}$,
где $h_1 = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7\}$,
$h_2 = \{a_1, a_2, a_3, a_4, a_5, a_6, a_8\}$,
$h_3 = \{a_5, a_6\}$,
$h_4 = \{a_6\}$,
$h_5 = \{a_1, a_2, a_3, a_4\}$,
$h_6 = \{a_1, a_2, a_3\}$,
$h_7 = \{a_1, a_2, a_4\}$,
$h_8 = \{a_1, a_3, a_4\}$

$\boldsymbol{\Omega}_2 = \{h_1, h_2, h_3, h_4, h_5, h_9, h_{10}, h_{11}\}$,
где $h_1 = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7\}$,
$h_2 = \{a_1, a_2, a_3, a_4, a_5, a_6, a_8\}$,
$h_3 = \{a_5, a_6\}$,
$h_4 = \{a_6\}$,
$h_5 = \{a_1, a_2, a_3, a_4\}$,
$h_9 = \{a_1, a_2\}$,
$h_{10} = \{a_1, a_3\}$,
$h_{11} = \{a_1, a_4\}$.

Каркасы $\boldsymbol{K}$ для $\boldsymbol{\Omega}_1$ и $\boldsymbol{\Omega}_2$ строятся в обоих случаях из одного и того же множества элементов:

$[\{a_1\}] = \{a_1\}$

$[\{a_2\}] = \{a_1, a_2\}$
$[\{a_3\}] = \{a_1, a_3\}$
$[\{a_4\}] = \{a_1, a_4\}$
$[\{a_5\}] = \{a_5, a_6\}$
$[\{a_6\}] = \{a_6\}$
$[\{a_7\}] = \{a_1, a_2, a_3, a_4, a_5, a_6, a_7\}$
$[\{a_8\}] = \{a_1, a_2, a_3, a_4, a_5, a_6, a_8\}$,

однако, диаграммы $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega}_1)$ и $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega}_2)$ не совпадают.

□

Итак, в нашем распоряжении появился новый инструмент диагностики принадлежности заданного множества образующих множеству вершин диаграммы всех ДСМ-сходств, формируемых на заданном множестве примеров $\boldsymbol{\Omega}$. Этот инструмент – система замыканий (рассчитанных в «контексте» используемого для ДСМ-обучения множества примеров $\boldsymbol{\Omega}$) одноэлементных подмножеств используемого алфавита $\boldsymbol{U}$. Будем называть это множество *базисом замыканий*. Его главной особенностью при использовании в дальнейших построениях будет возможность, однократно сформировав его на заданных множествах $\boldsymbol{U}$ и $\boldsymbol{\Omega}$, далее многократно задействовать при быстрой диагностике принадлежности каждого текущего набора образующих из $\boldsymbol{U}$ множеству вершин диаграммы всех ДСМ-сходств $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$.

Обратимся к процедурам диагностики в заданном псевдо-дереве $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$ фрагментов архитектуры каждого из представленных выше трех типов. Прежде всего, не сложно убедиться, что существование (в соответствующем псевдо-дереве) любого фрагмента типа β может быть идентифицировано по его (этого фрагмента) нижней границе, - как уже было показано выше (при обосновании **Утверждения 5.2.7**), -вычислением замыканий для одноэлементных[23] множеств образующих[24], входящих в корень псевдо-дерева, и последующим анализом замыканий по-парных комбинаций тех из них, которые являются неподвижными точками замыкания Галуа $[_]_{\boldsymbol{U},\boldsymbol{\Omega}}$. В ситуации, например, представленной на Рис.5.2.1, это будет выглядеть следующим образом:

- убедившись, что одноэлементные множества образующих $\{a_2\}$, $\{a_3\}$, $\{a_4\}$ и $\{a_5\}$, являются неподвижными точками рассматриваемого нами замыкания Галуа $[_]_{\boldsymbol{U},\boldsymbol{\Omega}}$, и что вместе с ними
- все двухэлементные множества этих же образующих $\{a_2, a_3\}$, $\{a_2, a_4\}$, $\{a_2, a_5\}$, $\{a_3,a_4\}$, $\{a_3,a_5\}$ и $\{a_4,a_5\}$ также являются неподвижными точками, причем – вкладывающимися, например, в исходные примеры h_{9}, h_{11}, h_{14} (и др.),

можно заключить, что

- весь четырехмерный булевский куб, построенный на вершинах $\{a_2\}$,$\{a_3\}$,$\{a_4\}$ и $\{a_5\}$[25], вкладывается как в диаграмму $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$, так и, например, в псевдо-деревья $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{\mathbf{h9}\}}$, $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{\mathbf{h11}\}}$ и $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{\mathbf{h14}\}}$.

---

[23] См предыдущую сноску.

[24] При этом возможно формирование двух подтипов архитектур типа β:

(i) нижняя граница (множество вершин, в векторе координат каждой из которых имеется ровно одна единица) соответствующего многомерного куба образована одноэлементными неподвижными точками рассматриваемого замыкания Галуа;

(ii) рассматриваемая нижняя граница сформирована замыканиями вида $\{a_j \cup \{b,c, \ldots, d\}\}$, где a_j - пробегает множество образующих, «задействованных» собственно при порождении соответствующего s-мерного куба, а дополнительное множество образующих {b,c, …, d} характеризует «основание» (в случае (i) - пустое), на котором в дереве $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$ «располагается» (в смысле рассматриваемого частичного порядка – взаимной вложимости вершин дерева) этот куб.

[25] «Нижней границе» этого куба.

Однако, наряду с представленной на Рис.5.2.1 ситуацией полного совпадения соответствующего фрагмента типа β со «стандартным» булевским гиперкубом подходящей размерности (разумеется, с учетом удаления из него минимальной вершины и множества всех инцидентных ей ребер) несложно также привести примеры ситуаций, когда актуальный фрагмент типа β сопоставляется лишь части вершин и ребер такого гиперкуба. Восстановление точной структуры фрагментов типа β в таких случаях требует дополнительных вычислений. Наиболее существенные особенности таких процедур и станут предметом нашего дальнейшего рассмотрения.

С формальной точки зрения фрагменты типа β можно определить следующим образом: они формируются всеми такими вершинами рассматриваемого псевдо-дерева, которые

(β_1) составляют нижнюю границу соответствующего гиперкуба (т.е. при порождении замыканий одноэлементных подмножеств находящегося в корне рассматриваемого псевдо-дерева примера – множества образующих – и формировании из них каркаса этого псевдо-дерева эти вершины оказываются наименьшими вершинами одного уровня в смысле упорядочения по вложению множеств образующих, сопоставленных вершинам порожденного каркаса), или же

(β_2) находятся на более «высоких» (по отношению к только что обсуждавшейся нижней границе гиперкуба) уровнях, при этом

(β_3) в множестве образующих, сопоставленных каждой такой вершине, нельзя найти элемент, замыкание которого совпадает с этим (сопоставленным данной вершине) множеством образующих.

Другими словами: фрагменты типа β формируются за счет «равнозначных» (одноуровневых однотипных) вершин каркаса и нижней границы каждого соответствующего гиперкуба а также вершин, которым сопоставлены все допустимые комбинации множеств образующих, соответствующих только что выбранным на нижней границе вершинам каждого такого гиперкуба.

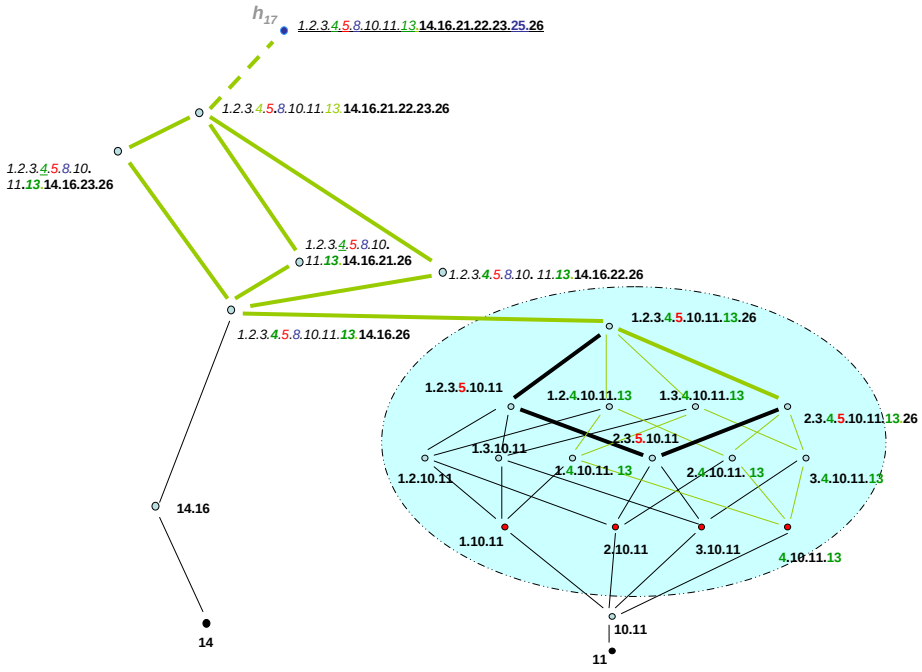


Рис. 5.2.3. Примеры встречаемости «склеенных» и «наклеенных» образующих:
a) «склеенные» образующие $\mathbf{a_4}$ и $\mathbf{a_{13}}$ – вершина <*4*.10.11.*13*>,
b) «наклеенная» (на образующие $\mathbf{a_2}$, $\mathbf{a_3}$,$\mathbf{a_{10}}$ и $\mathbf{a_{11}}$) образующая $\mathbf{a_5}$ -
- вершина <2.3.**5**.10.11>.

При этом в рассматриваемой классификации вершин анализируемого псевдо-дерева следует учесть также и случай, когда в тех или иных вершинах типа β, расположенных на нижней границе соответствующего гиперкуба, появляются одновременно по 2 (или более) «равноправных» образующих. Фактически, это случай – будем называть его эффектом «склеенных» образующих (см., например, вхождения образующих **a4** и **a13** в псевдо-дерево на Рис. 5.2.3), – когда каждую такую пару (или, соответственно более) образующих можно заменить одной вновь введенной в рассмотрение образующей, и при этом получить псевдо-дерево той же конфигурации[26].

Наконец, фрагменты типа γ формируются за счет наложения фрагментов типа α или β на фрагменты типа β . При этом, для β–β наложений условие β_3 в данном случае может быть упразднено, за счет чего могут возникать дополнительные конфигурации, в которых на те или иные множества образующих, сопоставленные вершинам типа β, оказываются «наклеенными» дополнительные образующие, среди которых имеются и те, для которых не выполняется (в том числе – при появлении «склеенных» образующих) условие β_3 .

Полезную информацию о возможностях быстрой диагностики типов[27] вершин рассматриваемого псевдо-дерева дает

5.2.13. Утверждение.

Пусть заданы алфавит $\boldsymbol{U}$, множество примеров $\boldsymbol{\Omega}$ и текущий (помещаемый в корень анализируемого псевдо-дерева) пример **A**. В соответствующем псевдо-дереве $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$

1. Все вершины типа α диагностируются полиномиально-сложной[28] вычислительной процедурой.
2. Основания всех содержащихся в псевдо-дереве $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$ гиперкубов формируются полиномиально-сложной вычислительной процедурой.
3. Перечень всех встречающихся в вершинах псевдо-дерева $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$ (двух- и более-элементных) множеств «склеенных» образующих формируется полиномиально-сложной вычислительной процедурой.
4. Перечень всех встречающихся в вершинах псевдо-дерева $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$ (одно- и более-элементных) множеств «наклеенных» образующих формируется полиномиально-сложной вычислительной процедурой.

Доказательство.

1-2. Первая и вторая части доказываемого **Утверждения** (как свойства процедур порождения каркаса $GK_{\mathbf{D},O}$) есть очевидные следствия **Утверждения 5.2.11**. Так, в частности, для первой части имеем: все фрагменты типа α идентифицируются анализом замыканий одноэлементных подмножеств образующих из исходного множества *O* (корня рассматриваемого псевдо-дерева). Таким образом, необходимо лишь проверить выполнимость условий α_1, α_2 и α_3 (см. выше). При этом условие α_3 требует для каждой образующей из множества *O* проверки не более чем |*O*|, замыканий образующих из этого же множества. Т.е., все требуемые здесь проверки могут быть также реализованы алгоритмом полиномиальной (не более чем квадратичной от величины $|\boldsymbol{U}|$) сложности.

3. Нетрудно убедиться, что третья часть доказываемого **Утверждения** есть следствие следующих соображений: для диагностики «склеенных» образующих достаточно выделить при порождении каркаса $GK_{\mathbf{D},O}$ *совпадающие* замыкания для пар *несовпадающих*

---

[26] Т.е. псевдо-дерево с тем же множеством ребер и с тем же (пусть и сформированным заменой «склеенных» образующих вновь введенной образующей) множеством вершин.

[27] В рамках предложенного перечня: вершины фрагментов типа α, β и γ, основания гиперкубов, склеенные образующие, наклеенные образующие.

[28] Здесь и далее в обсуждении текущего **Утверждения** - в соотнесении с характерными размерами исходных алфавита и обучающей выборки примеров.

образующих. Именно такие пары (или же их тройки, четверки и т.п. – как это будет следовать из анализа имеющихся исходных данных - множеств $\boldsymbol{U}$, $\boldsymbol{\Omega}$ и O) и характеризуют соответствующие подмножества «склеенных» образующих.

4. Последняя часть **Утверждения** есть следствие следующих соображений:
   a) «наклеенных» образующих не может быть больше, чем элементов в множестве $\boldsymbol{U}$;
   b) замыкание $[\mathbf{a}]_{\boldsymbol{U},\boldsymbol{\Omega}}$ каждой «наклеенной» переменной **a** содержит помимо ее самой также некоторое непустое множество $\boldsymbol{U}\big|_{\{\mathbf{a}\}} = \{\mathbf{b}^{\mathbf{a}}_{\mathbf{1}}, \ldots, \mathbf{b}^{\mathbf{a}}_{\mathbf{2}}\}$ образующих из $\boldsymbol{U}$, для каждой $\mathbf{b}^{\mathbf{a}}_{\mathbf{i}}$ из которых
   c) замыкание $[\mathbf{b}^{\mathbf{a}}_{\mathbf{i}}]_{\boldsymbol{U},\boldsymbol{\Omega}}$ либо совпадает с $[\mathbf{a}]_{\boldsymbol{U},\boldsymbol{\Omega}}$ (это - случай «склеенных» «наклеенных» образующих **a** и $\mathbf{b}^{\mathbf{a}}_{\mathbf{i}}$), либо не содержит **a** (причем наличие последних и определяет статус «наклеенности» рассматриваемой образующей **a**).

□

Используя только что представленную технику идентификации вершин рассматриваемого псевдо-дерева $\mathbf{D_GC}_{\varphi,\mathrm{f}}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$ (а также его каркаса $GK_{\mathbf{D},O}$), можно (быстрыми алгоритмами):

- восстановить всю α-часть архитектуры $\mathbf{D_GC}_{\varphi,\mathrm{f}}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$;
- восстановить нижние границы всех входящих в $\mathbf{D_GC}_{\varphi,\mathrm{f}}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$ гиперкубов, формирующих архитектуры типа β. Далее с их помощью реконструировать каждый из таких гиперкубов, чтобы затем
- размещать на подходящих вершинах таких гиперкубов диагностированные (в смысле приведенной выше классификации) вершины $GK_{\mathbf{D},O}$ и $\mathbf{D_GC}_{\varphi,\mathrm{f}}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$: сперва – анализируемые в ходе построения формирующих каркас $GK_{\mathbf{D},O}$ замыканий одноэлементных подмножеств образующих из соответствующего примера O , а далее – вершины, порождаемые парами, тройками и т.п. множеств образующих с нижней границы каждого соответствующего гиперкуба. Сюда же затем будут добавлены результаты диагностики как «наклеенных», так и «склеенных» образующих.

Таким образом формируется «технология» восстановления (в дополнение к сразу же порожденным при формировании каркаса $GK_{\mathbf{D},O}$ α-частям) β– и γ–частей реконструируемого псевдо-дерева $\mathbf{D_GC}_{\varphi,\mathrm{f}}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$ в направлении *снизу-вверх*[29].

Одновременно представляется целесообразным рассмотреть возможности восстановления точного вида β– и γ–частей реконструируемого псевдо-дерева $\mathbf{D_GC}_{\varphi,\mathrm{f}}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$ также и в направлении *сверху-вниз*.

Здесь прежде всего обратим внимание на особую роль, которую в обсуждаемом процессе реконструкции $\mathbf{D_GC}_{\varphi,\mathrm{f}}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$ по заданным $\boldsymbol{U}$, $\boldsymbol{\Omega}$ и O могут играть некоторые из исходных примеров. «Полнота» покрытия вершинами $GK_{\mathbf{D},O}$ и $\mathbf{D_GC}_{\varphi,\mathrm{f}}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$ вершин каждого сопоставленного архитектурам типа β гиперкуба $\boldsymbol{HC}_0(\boldsymbol{U},\boldsymbol{\Omega},O)$ естественным образом зависит от наличия в $\boldsymbol{\Omega}$ таких элементов $O_{\mathrm{j}1}, \ldots, O_{\mathrm{j}k}$, что для каждого

$$O_0 \in \{O_{\mathrm{j}1}, \ldots, O_{\mathrm{j}k}\}$$

найдется некоторый из входящих (как подмножество образующих) в O_0 элемент множества $\mathbf{D_GC}_{\varphi,\mathrm{f}}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$, который соответствует[30] одной из вершин этого гиперкуба $\boldsymbol{HC}_0(\boldsymbol{U},\boldsymbol{\Omega},O)$. Будем говорить в таких случаях, что *пример* O_0 (в *контексте* анализируемого псевдо-дерева $\mathbf{D_GC}_{\varphi,\mathrm{f}}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$) *раскрывает* соответствующую *вершину* гиперкуба $\boldsymbol{HC}_0(\boldsymbol{U},\boldsymbol{\Omega},O)$.

---

[29] В смысле последовательной реконструкции (*снизу-вверх* - от минимальных к максимальным по вложению множествам образующих, сопоставленных вершинам анализируемого псевдо-дерева) восстанавливаемых фрагментов псевдо-дерева $\mathbf{D_GC}_{\varphi,\mathrm{f}}(\boldsymbol{U},\boldsymbol{\Omega})\big|_{\{O\}}$, «лежащих» на соответствующих вершинах каждого соответствующего гиперкуба.

[30] Совпадает как подмножество образующих из множества O.

(Иллюстрацию ситуаций такого типа можно найти, в том числе, и на Рис.5.2.1 – см. раскрытие вершин <2.3.4>, <2.3.5>, <2.4.5> и <3.4.5> примерами $\mathbf{h_3}$ – $\mathbf{h_6}$ и вершины <2.3.4.5> примером h_9).

Пусть зафиксированы множества $\boldsymbol{U}$, $\boldsymbol{\Omega}$ и O, а также некоторый гиперкуб $\boldsymbol{HC}_j$ $(\boldsymbol{U,\Omega,}O)$, порождаемый при реконструкции по ним псевдо-дерева $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U,\Omega})\big|_{\{O\}}$.

5.2.14. Утверждение.

Для каждого j (номера гиперкуба $\boldsymbol{HC}_j$ $(\boldsymbol{U,\Omega,}O)$, идентифицированного в ходе реконструкции псевдо-дерева $\mathbf{D_GC}_{\varphi,f}$ $(\boldsymbol{U,\Omega})\big|_{\{O\}}$) подмножество $\mathbf{H}_j$ $(\boldsymbol{U,\Omega,}O)$ всех примеров $\mathbf{h_i}$ из $\boldsymbol{\Omega}$, раскрывающих (некоторые) вершины соответствующего гиперкуба, может быть перечислено процедурой полиномиальной вычислительной сложности.

Доказательство.

Идентифицируем в гиперкубе $\boldsymbol{HC}_j$ $(\boldsymbol{U,\Omega,}O)$ вершину, соответствующую наибольшему по вложению множеству образующих из O. Сделать это можно, например, следующей быстрой процедурой:

- строим (как уже было описано выше) нижнюю границу этого гиперкуба;
- объединяем множества образующих из O, идентифицированные при порождении этой границы. Результат такого объединения – множество $\mathbf{h_0}$ образующих, которое может быть сопоставлено идентифицируемой вершине гиперкуба.

Далее для каждого $\mathbf{h_i}$ из $\boldsymbol{\Omega}$ вычисляем

$$\mathbf{h_{i0}} = \mathbf{h_i} \cap \mathbf{h_0}.$$

В искомое множество $\mathbf{H}_j$ $(\boldsymbol{U,\Omega,A})$ будем помещать каждый такой пример $\mathbf{h_i}$, для которого на нижней границе рассматриваемого гиперкуба найдется по крайней мере один такой элемент $[\mathbf{a}]_{\boldsymbol{U,\Omega}}$, что

$$[\mathbf{a}]_{\boldsymbol{U,\Omega}} \subseteq \mathbf{h_{i0}}.$$

Учтем дополнительно, что $\mathbf{h_i}$ пробегает не более чем все множество примеров $\boldsymbol{\Omega}$, а при порождении соответствующей вершины (замыкания) из псевдо-дерева $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U,\Omega})\big|_{\{O\}}$ каждая пара $<\mathbf{h_i}, \mathbf{h_0}>$ и соответствующая ей пара $<\mathbf{h_i}, \mathbf{h^*_0}>$, где $\mathbf{h^*_0}$ – минимальный элемент каркаса $GK_{\mathbf{D},O}$, охватывающий множество $\mathbf{h_0}$ (т.е. такой, что $\mathbf{h_0} \subseteq \mathbf{h^*_0}$)[31], порождают одно и то же замкнутое множество образующих из $\mathbf{A}$, сопоставляемое соответствующей (*раскрываемой*) вершине гиперкуба $\boldsymbol{HC}_j$ $(\boldsymbol{U,\Omega,}O)$.

□

При «раскрытии» вершин каждого гиперкуба $\boldsymbol{HC}_j$ $(\boldsymbol{U,\Omega,}O)$ *сверху-вниз* (т.е. по схеме: сперва – вершины, «раскрываемые» подходящими исходными примерами, затем вершины, порождаемые с помощью только что раскрытых вершин, и т.д.) представляется естественным вновь вернуться к уже начатому ранее обсуждению условий «полноты» покрытия вершин рассматриваемого гиперкуба вершинами каркаса $GK_{\mathbf{D},O}$ и, конечно же, вершинами β–типа собственно псевдо-дерева $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U,\Omega})\big|_{\{O\}}$. Простейшие представления о ситуациях подобного типа а также полезную (для оптимизации процедур восстановления структуры анализируемого псевдо-дерева информацию) дает следующее несложное (однако, весьма полезное)

5.2.15. Утверждение.

Для заданных $\boldsymbol{U}$, $\boldsymbol{\Omega}$ и O порождаемый при реконструкции псевдо-дерева $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U,\Omega})\big|_{\{O\}}$ гиперкуб $\boldsymbol{HC}_j$ $(\boldsymbol{U,\Omega,}$ $O)$ полностью покрывается вершинами типа β соответствующей части восстанавливаемого псевдо-дерева тогда и только тогда, когда в исходном множестве

[31] А с ним и все порождающее его множество исходно заданных примеров из $\boldsymbol{\Omega}$.

примеров $\boldsymbol{\Omega}$ существует подмножество, элементы которого раскрывают все вершины верхней границы[32] рассматриваемого гиперкуба.

Доказательство.

<=. Если в множестве примеров $\boldsymbol{\Omega}$ имеется подмножество $\mathbf{H}_j$ ($\boldsymbol{U,\Omega,}O$), раскрывающее верхнюю границу гиперкуба $\boldsymbol{HC}_j$ ($\boldsymbol{U,\Omega,}O$), то далее сходства этих множеств образующих, сопоставленных вершинам верхней границы рассматриваемого гиперкуба (заметим – сходства, являющиеся соответственно элементами псевдо-дерева $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U,\Omega})\big|_{\{O\}}$), накрывают все[33] оставшиеся его вершины.

=>. В ситуации, когда псевдо-дерево $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U,\Omega})\big|_{\{O\}}$ накрывает все вершины гиперкуба $\mathbf{H}_j$ ($\boldsymbol{U,\Omega,}O$), оно естественным образом накрывает также и все вершины верхней границы рассматриваемого гиперкуба. Это означает, что эти вершины должны быть «раскрыты» некоторыми примерами из множества $\boldsymbol{\Omega}$ (в противном случае их не было бы среди сходств – соответствующих замыканий Галуа $[_]_{\boldsymbol{U,\Omega}}$ – в псевдо-дереве $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U,\Omega})\big|_{\{O\}}$).

□

Полезность этого **Утверждения**, прежде всего, в следующем: восстанавливая структуру псевдо-дерева $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U,\Omega})\big|_{\{O\}}$ на соответствующем гиперкубе $\boldsymbol{HC}_j$ ($\boldsymbol{U,\Omega,}O$), приходится прежде всего выделять подмножество $\mathbf{H}_j$ ($\boldsymbol{U,\Omega,}O$) раскрывающих его вершины примеров из $\boldsymbol{\Omega}$. При этом, если в нем найдутся примеры, раскрывающие всю верхнюю границу рассматриваемого гиперкуба, то дальнейших действий по его восстановлению (т.е. собственно вычислений сходств примеров из $\mathbf{H}_j$ ($\boldsymbol{U,\Omega,}O$)) уже не потребуется, т.к. в силу последнего **Утверждения** все оставшиеся вершины этого гиперкуба буду накрыты соответствующими элементами из $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U,\Omega})\big|_{\{O\}}$.Другими словами, (реализуемый быстрыми процедурами) поиск примеров, входящих в $\mathbf{H}_j$ ($\boldsymbol{U,\Omega,}O$), попутно позволяет быстро проверить выполнимость условия, обеспечивающее существенное сокращение перебора при восстановлении явного вида целевого псевдо-дерева $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U,\Omega})\big|_{\{O\}}$.

В ситуации, когда условия **Утверждения 5.2.15** не выполняются (т.е. когда есть необходимость последовательно - например, по-элементно: вершина за вершиной, - выделять ту часть структуры гиперкуба $\boldsymbol{HC}_j(\boldsymbol{U,\Omega,}O)$, которая полностью содержится в восстанавливаемом псевдо-дереве $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U,\Omega})\big|_{\{O\}}$), как нетрудно убедиться, уже найденное множество примеров $\mathbf{H}_j(\boldsymbol{U,\Omega,}O)$, вместе со специальным дополнительным примером – множеством $\mathbf{h_{j0}}$_ образующих из исходного алфавита $\boldsymbol{U}$, сопоставленным наибольшей по вложению образующих вершине гиперкуба $\boldsymbol{HC}_j(\boldsymbol{U,\Omega,}O)$, полностью определяют ту искомую часть восстанавливаемого псевдо-дерева, которая «накладывается» на этот гиперкуб. Таким образом, пара $<\boldsymbol{U,\Omega_{j0}}>$, в которой новое множество примеров

$$\boldsymbol{\Omega_{j0}} = \mathbf{H}_j(\boldsymbol{U,\Omega,A}) \cup \mathbf{h_{j0}},$$

полностью определяет восстанавливаемую часть псевдо-дерева $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U,\Omega})\big|_{\{O\}}$ как диаграмму $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U,\Omega_{j0}})$ взаимной вложимости замыканий Галуа $[_]_{\boldsymbol{U,\Omega}}$ для примеров из множества $\boldsymbol{\Omega_{j0}}$. Теперь покажем, что и множество образующих $\mathbf{h_{j0}}$ может быть порождено как сходство некоторых примеров из исходного $\boldsymbol{\Omega}$. Действительно, для вершин каркаса $GK_{\mathbf{D},O}$ возможны всего лишь две ситуации:

1) $\mathbf{h_{j0}}$ соответствует одной из вершин этого каркаса;
2) $\mathbf{h_{j0}}$ дополняется до одной (случай α–архитектуры) или нескольких (случай архитектуры типа β) вершин каркаса $GK_{\mathbf{D},O}$ несколькими дополнительными образующими

---

[32] Т.е. все его вершины, смежные наибольшему элементу в смысле взаимной вложимости множеств образующих, сопоставленных его вершинам.

[33] За исключением «наименьшей» вершины, которой сопоставляется пустое множество образующих из O.

из ***U***, причем это – ближайшие (в смысле взаимного вложения множеств образующих) содержащие $\mathbf{h_{j0}}$ множества образующих в вершинах каркаса $GK_{\mathbf{D},O}$.

В первом из только что представленных случаев $\mathbf{h_{j0}}$ (как элемент каркаса) есть замыкание некоторого подмножества образующих из находящегося в корне текущего псевдо-дерева множества образующих (и, одновременно, примера из $\mathbf{\Omega}$) O. Следовательно, $\mathbf{h_{j0}}$ может быть представлено как сходство соответствующих примеров из $\mathbf{\Omega}$. В обоих подслучаях второго случая вершины каркаса, до которых «достраивается» $\mathbf{h_{j0}}$, также (на основании аналогичных первому случаю рассуждений) могут быть представлены как сходства соответствующих примеров из $\mathbf{\Omega}$. Таким образом, в обоих рассматриваемых случаях восстанавливаемая (на рассматриваемом гиперкубе $\boldsymbol{HC}_j(\boldsymbol{U},\mathbf{\Omega},O)$) часть целевого псевдо-дерева $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\mathbf{\Omega})\big|_{\{O\}}$ будет изоморфна диаграмме сходств $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\mathbf{\Omega_j})$, порождаемой на исходном алфавите ***U*** и некотором подмножестве[34] $\mathbf{\Omega_j}$ исходного множества примеров $\mathbf{\Omega}$. Это обстоятельство позволяет организовать в $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\mathbf{\Omega})\big|_{\{O\}}$ для $\mathbf{\Omega_j} \subset \mathbf{\Omega}$ рекурсивную процедуру восстановления диаграммы $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\mathbf{\Omega})$, последовательно двигаясь от $\mathbf{\Omega}$ к соответствующим $\mathbf{\Omega_j}$, и, рекурсивно продолжая, далее.

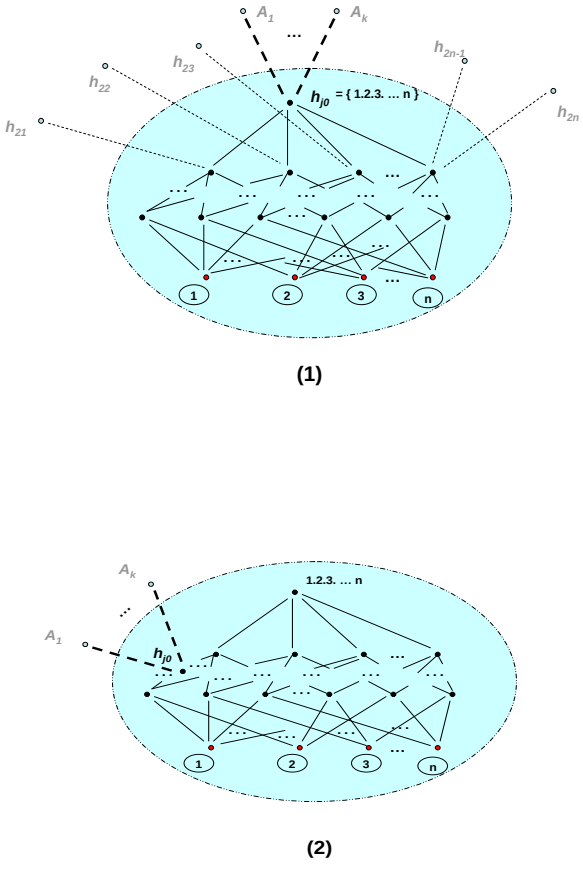


Рис. 5.2.4. Варианты раскрытия вершин гиперкуба примерами из исходного множества $\mathbf{\Omega}$.

Теперь проверим, в каких случаях $\mathbf{\Omega_j} = \mathbf{\Omega}$? Прежде всего, заметим, что если в восстанавливаемом псевдо-дереве имеются (а это выясняется при формировании его каркаса):

- более одного архитектурного фрагмента типа β или же

---

[34] Причем – быстро восстанавливаемом (см. **Утверждение 5.2.14** о сложности порождения подмножества примеров $\mathbf{H}_j(\boldsymbol{U},\mathbf{\Omega},O)$ из $\mathbf{\Omega}$, раскрывающих вершины гиперкуба $\boldsymbol{HC}_j(\boldsymbol{U},\mathbf{\Omega},O)$).

- содержащий по крайней мере 3 вершины[35] фрагмент типа α ,

то $\mathbf{\Omega_j} \subset \mathbf{\Omega}$. Действительно, второй (по отношению к анализируемому гиперкубу) фрагмент типа β как и (по крайней мере) одна из трех вершин во фрагменте типа α «задействуют» при порождении соответствующих сходств хотя бы один пример из **Ω**, который не войдет в текущее множество $\mathbf{\Omega_j}$. Таким образом, в нашем распоряжении остаются лишь два варианта конфигурации (см. Рис. 5.2.4), когда в восстанавливаемом псевдо-дереве располагается лишь один гиперкуб, а фрагменты типа α содержат каждый не более одной вершины за пределами этого гиперкуба. При этом $\mathbf{h_{j0}}$ – либо один из элементов исходного множества примеров **Ω**, либо расширяется до некоторых его элементов O_1, … , O_k добавлением образующих, которые в каждом из этих k случаев входят только в один (получаемый расширением $\mathbf{h_{j0}}$)[36] пример из **Ω** (в противном случае $\mathbf{h_{j0}}$ можно было бы пополнить каждой из таких «не уникальных» образующих).

Итак, единственный (с точностью до совмещения $\mathbf{h_{j0}}$ и O_1 при k=1) вид допустимых конфигураций рассматриваемого типа представлен на фрагменте **(1)** Рис.5.2.4 (случай, когда $\mathbf{h_{j0}}$ находится не в наивысшей вершине рассматриваемого гиперкуба – см. фрагмент **(2)** на Рис. 5.2.4 – не соответствует представлениям об $\mathbf{h_{j0}}$ как об объединении множеств образующих, лежащих на нижней границе этого гиперкуба). Для этой конфигурации при построении диаграммы $\mathbf{D_GC}_{\varphi,f}$ (***U*,Ω**) восстановлением содержащихся в ней псевдо-деревьев $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\mathbf{\Omega})\big|_{\{O\}}$ рассматриваемого здесь типа достаточно:

- сперва построить такие поддиаграммы для примеров из множества $\Omega \setminus \{O_1, \dots, O_k\}$ исходных примеров, не принадлежащих подмножеству $\{O_1, \dots, O_k\}$, затем
- объединить эти псевдо-деревья в общую диаграмму и добавить к ним подграф (см. Рис.5.2.4), образованный вершинами $O_1, \dots, O_k$ и h_{j0}.

При этом, как видно из только что приведенных рассмотрений, построение диаграммы $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\mathbf{\Omega})$ для текущего множества примеров из множества Ω может быть сведено к построению диаграмм того же вида для специальным образом порождаемых *подмножеств* исходного множества примеров Ω.

Таким образом, продемонстрирована возможность формирования *рекурсивной*[37] процедуры восстановления диаграммы сходств $\mathbf{D_GC}_{\varphi,f}$ (***U*,Ω**), порождаемой из заданного множества примеров **Ω** над алфавитом ***U***. Предлагаемая процедура позволяет переходить от исходного множества примеров **Ω** к его специальным – характеризующим соответствующие гиперкубы – подмножествам. Подробности расширения возможностей такого целенаправленного рекурсивного перехода к специальным подмножествам исходного множества примеров **Ω** мы рассмотрим далее - в **Разделе 5.3**.

---

[35] Одна из которых может быть, в том числе, и исходным примером из **Ω**, а другая – принадлежать множеству вершин анализируемого гиперкуба.

[36] Т.е. образующие из каждого $O_i \setminus \mathbf{h_{j0}}$ (дополнения $\mathbf{h_{j0}}$ до O_i) содержатся только в O_i.

[37] Позволяющей применять ее сперва к исходным данным, затем (по ходу их обработки) к определенным подмножествам исходных данных, и далее (продолжая такие рекурсивные «вложения») - до анализа соответствующих простейших конфигураций (объединение которых и дает исходные анализируемые данные). Таким образом формируется естественный вариант декомпозиции исходной решаемой задачи в набор подзадач меньшей сложности.

5.3. Процедуры направленного перебора в диаграммах взаимной вложимости классов эквивалентности

В предыдущем **Разделе 5.2** нами уже была продемонстрирована возможность формирования *рекурсивной* процедуры восстановления диаграммы сходств $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$, порождаемой из заданного множества примеров $\boldsymbol{\Omega}$ над алфавитом $\boldsymbol{U}$. Процедура позволяет применять ее сперва к исходным данным, затем (по ходу их обработки) к определенным подмножествам исходных данных, и далее (продолжая такие рекурсивные «вложения») - до анализа соответствующих простейших конфигураций (объединение которых и дает исходные анализируемые данные). Именно таким путем формируется естественный вариант декомпозиции исходной решаемой задачи в набор подзадач меньшей сложности. Используемая технология «фокусировки» на менее сложных подзадачах позволяет также переходить от исходного множества примеров $\boldsymbol{\Omega}$ к его специальным – характеризующим соответствующие гиперкубы – подмножествам. Наконец, дополнительно в рассматриваемой процедуре может быть реализована *целенаправленная навигация* по формирующим анализируемую диаграмму псевдо-деревьям, например, в первую очередь по тем, которые «интересны» с точки зрения тех или иных дополнительных условий (в частности - релевантны некоторым заданным подмножествам исходных примеров, содержат или же, наоборот, не содержат некоторые заданные подмножества образующих из $\boldsymbol{U}$ и т.п.).

Рассмотрим предоставляемые предлагаемой процедурой возможности целенаправленной навигации в восстанавливаемой диаграмме сходств $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$ более подробно. Прежде всего, отметим, что здесь могут комбинироваться два типа «продвижений» по восстанавливаемой диаграмме сходств:

- движения «вширь» (в частности, переход с одного псевдо-дерева в исходной анализируемой диаграмме сходств на другое, переход от одного подмножества исходного множества примеров $\boldsymbol{\Omega}$, раскрывающего определенные вершины гиперкуба, содержащегося в заданном псевдо-дереве с одним из примеров исходной выборки $\boldsymbol{\Omega}$ в корне, к другому такому подмножеству и т.п.); и
- движения «вглубь» (в частности, переход от одного гиперкуба, вложенного в заданное псевдо-дерево с одним из исходных примеров из $\boldsymbol{\Omega}$ в корне, к другому гиперкубу той же «глубины» вложения (в анализируемое псевдо-дерево) или же переход от гиперкуба одной «глубины» вложения в анализируемое псевдо-дерево к гиперкубу другой «глубины» вложения и т.п.. При этом сперва производится переход к соответствующему гиперкубу той же «глубины» вложения, далее в нем - к одному из порождаемых при его раскрытии псевдо-деревьев – поддиаграмм псевдо-дерева, содержащих гиперкуб – исходную точку рассматриваемого перехода, а затем уже в этой под-диаграмме – к одному из вложенных уже в нее гиперкубов).

Исчерпывающую процедуру последовательного обхода (при восстановлении) исходной диаграммы по схеме

сперва - вглубь на сколько это возможно,
затем – назад до последней точки рекурсивного вложения,
где, далее, – вширь на один шаг и, затем,
вновь - вглубь, на сколько это окажется возможным,
и так далее до возвращения к последнему примеру исходной выборки $\boldsymbol{\Omega}$ *в качестве корня только что восстановленного псевдо-дерева,*

предлагает

Алгоритм НПС (Направленного Перебора Сходств).

ВХОД (Входные данные): алфавит $\boldsymbol{U}$ и множество примеров для обучения $\boldsymbol{\Omega}$.
ВЫХОД (Выходной результат): диаграмма сходств $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$.

Описание алгоритма:

ШАГ 1. Нумеруем образующие алфавита $\boldsymbol{U}=\{\mathbf{a_1}, \dots, a_n\}$ и примеры из множества $\boldsymbol{\Omega}=\{O_1, \dots, O_m\}$.

Счетчику *ГЛУБИНА ВЛОЖЕНИЯ*, контролирующему глубину вложения текущей анализируемой поддиаграммы (диаграммы сходств $\mathbf{D_GC}_{\varphi,f}$ $(\boldsymbol{U},\boldsymbol{\Omega})$), присваиваем значение 0.

Параметру *МНОЖЕСТВО ПРИМЕРОВ* присваиваем значение <1,1, … , 1> (кортеж, сформированный из в точности m единиц), кодирующее включение в текущее анализируемое множество примеров всех m элементов исходного множества $\boldsymbol{\Omega}$ (фактически положив текущее анализируемое множество примеров $\boldsymbol{\Omega^*} = \boldsymbol{\Omega}$).

Счетчику *ТЕКУЩИЙ ПРИМЕР*, контролирующему перебор в исходной диаграмме сходств $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$ корней текущих анализируемых псевдо-деревьев, присваиваем значение 0.

Счетчику *ТЕКУЩИЙ ГИПЕРКУБ* , контролирующему перебор гиперкубов в текущем анализируемом псевдо-дереве, присваиваем значение 0.

Параметру *РАЗМЕР ТЕКУЩЕГО МНОЖЕСТВА ПРИМЕРОВ*, фиксирующему число примеров в текущем *МНОЖЕСТВЕ ПРИМЕРОВ* присваиваем значение m.

Параметру *РАЗМЕР ТЕКУЩЕГО МНОЖЕСТВА ГИПЕРКУБОВ*, фиксирующему число гиперкубов в текущем анализируемом псевдо-дереве, присваиваем значение 0.

В исходно пустой стек **TRACK** помещаем (как единую запись) шестерку <*ГЛУБИНА ВЛОЖЕНИЯ, МНОЖЕСТВО ПРИМЕРОВ, ТЕКУЩИЙ ПРИМЕР, ТЕКУЩИЙ ГИПЕРКУБ, РАЗМЕР ТЕКУЩЕГО МНОЖЕСТВА ПРИМЕРОВ, РАЗМЕР ТЕКУЩЕГО МНОЖЕСТВА ГИПЕРКУБОВ* >.

ШАГ 2. Значение $\boldsymbol{i}$ счетчика *ТЕКУЩИЙ ПРИМЕР* элементов из $\boldsymbol{\Omega^*}$ присваиваем значение $\boldsymbol{i}$+1. Если значение $\boldsymbol{i}$ после этого превосходит значение параметра *РАЗМЕР ТЕКУЩЕГО МНОЖЕСТВА ПРИМЕРОВ* (т.е. все элементы множества $\boldsymbol{\Omega^*}$ - как корни соответствующих псевдо-деревьев на текущем уровне вложения анализируемых поддиаграмм исходной диаграммы сходств $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$ – перебраны), переходим на **ШАГ 7**. В противном случае начинаем восстановление псевдо-дерева $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega^*})\big|_{\{O\mathrm{i}\}}$, переходя на

ШАГ 3. Строим каркас $GK_{\mathbf{D},O\mathbf{i}}$ псевдо-дерева $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega^*})\big|_{\{O\mathrm{i}\}}$.

Выделяем в этом псевдо-дереве фрагмент типа α и проверяем наличие архитектурных фрагментов типа β. В случае их отсутствия переходим к ШАГу 2. В противном случае значение $\boldsymbol{j}$ счетчика *ТЕКУЩИЙ ГИПЕРКУБ* таких фрагментов типа β обращаем в 0, одновременно вычисленное (при построении каркаса $GK_{\mathbf{D},O\mathbf{i}}$) число фрагментов типа β присваиваем в качестве текущего значения параметру *РАЗМЕР ТЕКУЩЕГО МНОЖЕСТВА ГИПЕРКУБОВ*, текущее значение d счетчика *ГЛУБИНА ВЛОЖЕНИЯ* заменяем на значение d+1, в стек **TRACK** в качестве верхнего элемента помещаем обновленную шестерку <*ГЛУБИНА ВЛОЖЕНИЯ, МНОЖЕСТВО ПРИМЕРОВ, ТЕКУЩИЙ ПРИМЕР, ТЕКУЩИЙ ГИПЕРКУБ, РАЗМЕР ТЕКУЩЕГО МНОЖЕСТВА ПРИМЕРОВ, РАЗМЕР ТЕКУЩЕГО МНОЖЕСТВА ГИПЕРКУБОВ*>, после чего переходим на

ШАГ 4. Значение $\boldsymbol{j}$ счетчика *ТЕКУЩИЙ ГИПЕРКУБ* соответствующих фрагментов типа β присваиваем значение $\boldsymbol{j}$+1. Если вновь полученное значение $\boldsymbol{j}$ превосходит текущее

значение параметра *РАЗМЕР ТЕКУЩЕГО МНОЖЕСТВА ГИПЕРКУБОВ* (т.е. уже перебраны все такие фрагменты для текущего анализируемого псевдо-дерева, причем *текущего* - т.е. не обязательно *верхнего* - уровня вложения поддиаграмм исходной диаграммы сходств $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$), удаляем из стека **TRACK** последний элемент и возвращаемся к ШАГу 2. (Тем самым «курсор» актуального состояния анализа исходной диаграммы сходств $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$ переводится к предпоследнему - с точки зрения движения *вширь* и *вглубь* по этой диаграмме – положению, которое фиксируется соответствующими значениями трех используемых счетчиков - глубины вложения, примеров из исходной выборки $\boldsymbol{\Omega}$ а также гиперкубов в текущей анализируемой поддиаграмме). В противном случае обращаемся к анализу той части псевдо-дерева $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega^*})\big|_{\{O_i\}}$, которая лежит на гиперкубе с вновь выбранным текущим номером ***j*** и переходим на

ШАГ 5. Выделяем то подмножество $\boldsymbol{\Omega^{**}}$ исходного множества примеров $\boldsymbol{\Omega}$, все элементы которого раскрываются в текущий анализируемый гиперкуб.
Проверяем выполнимость условия **Утверждения 5.2.15**, гарантирующего полное покрытие соответствующим фрагментом текущего анализируемого псевдо-дерева всего текущего анализируемого гиперкуба. В случае выполнимости **Утверждения**, восстанавливаем гиперкуб, соответствующий анализируемому фрагменту текущего анализируемого псевдо-дерева, затем переходим на **ШАГ 4**. В противном случае

ШАГ 6. Переходим к анализу подзадачи с текущими входными данными $<\boldsymbol{U},\boldsymbol{\Omega^{**}}>$ (рекурсивно вкладывая описанные **ШАГами 1-5** действия в описываемый исходный алгоритм: в качестве исходных данных на ВХОДЕ такого рекурсивного вложения текущая пара $<\boldsymbol{U},\boldsymbol{\Omega^{**}}>$ заменит предыдущую анализируемую пару $<\boldsymbol{U},\boldsymbol{\Omega^{*}}>$).

Параметру *МНОЖЕСТВО ПРИМЕРОВ* присваиваем значение $<x_1, x_2, \dots, x_m>$, где для каждого $i\in\{1,2, \dots, m\}$ значение $x_i = 1$, если O_i содержится в множестве $\boldsymbol{\Omega^{**}}$. В противном случае полагаем $x_i=0$.

Счетчику *ТЕКУЩИЙ ПРИМЕР*, контролирующему перебор в исходной диаграмме сходств $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega^{*}})$ корней текущих анализируемых псевдо-деревьев, присваиваем значение 0.
Счетчику *ТЕКУЩИЙ ГИПЕРКУБ*, контролирующему перебор гиперкубов в текущем анализируемом псевдо-дереве, присваиваем значение 0.

Параметру *РАЗМЕР ТЕКУЩЕГО МНОЖЕСТВА ПРИМЕРОВ*, фиксирующему число примеров в текущем *МНОЖЕСТВЕ ПРИМЕРОВ* присваиваем значение $\big|\boldsymbol{\Omega^{**}}\big|$ - число единиц в текущем виде кортежа $<x_1, x_2, \dots, x_m>$, отражающее число элементов в текущем множестве примеров $\boldsymbol{\Omega^{**}}$.

Параметру *РАЗМЕР ТЕКУЩЕГО МНОЖЕСТВА ГИПЕРКУБОВ*, фиксирующему число гиперкубов в текущем анализируемом псевдо-дереве, присваиваем значение 0.

В стек **TRACK** помещаем (как единую запись) шестерку текущих значений *<ГЛУБИНА ВЛОЖЕНИЯ, МНОЖЕСТВО ПРИМЕРОВ, ТЕКУЩИЙ ПРИМЕР, ТЕКУЩИЙ ГИПЕРКУБ, РАЗМЕР ТЕКУЩЕГО МНОЖЕСТВА ПРИМЕРОВ, РАЗМЕР ТЕКУЩЕГО МНОЖЕСТВА ГИПЕРКУБОВ>*. Далее переходим на **ШАГ 2** (имея ввиду при этом, что роль $\boldsymbol{\Omega^{*}}$ теперь будет играть $\boldsymbol{\Omega^{**}}$).

ШАГ *7*.СТОП

□

Свойства корректности **Алгоритма НПС** демонстрирует следующее

5.3.1. Утверждение.

Алгоритм НПС характеризуется полнотой и точностью при восстановлении диаграммы сходств $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$:

1) порождает все принадлежащие ей сходства (с указанием для каждого из них всех ближайших к нему по вложимости элементов восстанавливаемой диаграммы), при этом
2) не порождает никаких других (не являющихся элементами восстанавливаемого множества сходств) объектов.

Доказательство.

ПОЛНОТА. Полнота **Алгоритма НПС** – естественное следствие полноты организованного в нем последовательного перебора всех возможных (в рамках восстанавливаемой диаграммы сходств) псевдо-деревьев, их каркасов а также (вложенных в эти псевдо-деревья и диагностируемых при формировании соответствующих каркасов) гиперкубов.

ТОЧНОСТЬ. Точность **Алгоритма НПС** – очевидное следствие использования им для восстановления каждой рассматриваемой диаграммы сходств процедуры порождения каркасов (и только ее - !) соответствующих поддиаграмм: каждой вершине каждого порождаемого каркаса соответствует то или иное замкнутое подмножество образующих исходного алфавита $\boldsymbol{U}$.

□

Дополнительную полезную информацию о необходимых **Алгоритму НПС** ресурсах памяти дает

5.3.2. Утверждение.

Глубина заполнения стека **TRACK** при работе **Алгоритма НПС** ограничена сверху величиной $|\boldsymbol{U}|$ (числом образующих исходного алфавита $\boldsymbol{U}$).

Доказательство.

Глубина стека **TRACK** на каждом шаге **Алгоритма НПС** ограничена сверху длиной соответствующей ветки частичного порядка в каждой из анализируемых по ходу его работы псевдо-деревьев. Таким образом, доказываемое **Утверждение** есть очевидное следствие двух фактов:

1) длина каждой ветки частичного порядка (задаваемого исходной диаграммой вложимости сходств $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U},\boldsymbol{\Omega})$) не может превышать числа образующих, которыми сформирован лежащий в корне соответствующего псевдо-дерева пример из исходного множества $\boldsymbol{\Omega}$.

2) Любой из элементов исходной выборки $\boldsymbol{\Omega}$ может быть сформирован лишь образующими из исходного алфавита $\boldsymbol{U}$.

□

Следует обратить внимание на две особенности обсуждаемого **Алгоритма**. Во первых, при восстановлении целевой диаграммы взаимной вложимости сходств основным рабочим инструментом **Алгоритма** оказывается (полиномиально сложная - !) процедура порождения каркасов соответствующих псевдо-деревьев. Именно это обстоятельство оставляет возможности при организации целенаправленной навигации по восстанавливаемой диаграмме сходств, например: при проверке условий каузальной полноты или же при поиске ДСМ-гипотез, способных обеспечить доопределение новых τ–примеров и т.п., быстро проверять (восстанавливать) подходящие фрагменты целевой диаграммы всех сходств, получая (быстрыми алгоритмами) таким способом ответы на соответствующие вопросы (выполнения условий каузальной полноты, доопределимости рассматриваемых τ–примеров и т.п.).

Во вторых, если рассматривать транзитивные замыкания анализируемых нами здесь графов – диаграмм сходств, псевдо-деревьев и их каркасов, то лежащая в основе **Алгоритма НПС** техника порождения каркасов и достраивания гиперкубов позволяет, для каждого содержащего в своем корне один из примеров – элементов исходной выборки $\boldsymbol{\Omega}$ – псевдо-дерева, стартовав с порождения его каркаса (и, как мы только что условились, достроив его до соответствующего транзитивного замыкания), далее, порождая утоняющие

«картину» фрагменты – гиперкубы и их под-диаграммы, обеспечить процесс монотонно возрастающего пополнения «стартового» графа каркаса до соответствующего псевдо-дерева. (Разумеется, подобное пополнение может также выполняется на уровне транзитивного замыкания каждого порождаемого подграфа, и, как нетрудно убедиться, исходные графы и их транзитивные замыкания будут различаться лишь числом ребер, имея совпадающие множества вершин).

Таким образом, используя эти два свойства можно, соответствующим образом организовав последовательность перебора восстанавливаемых в каждом псевдо-дереве подграфов, наладить порождение в первую очередь полезных результатов (например, ДСМ-гипотез, обеспечивающих выполнение условий каузальной полноты или же доопределимость заданных τ–примеров), а уже далее, если это потребуется, восстановить целиком все множество ДСМ-гипотез, формируемых из заданных алфавита ***U*** и набора примеров **Ω**. Другими словами, это позволяет организовать процесс формирования ***управляемых приближений*** при порождении множества всех задаваемых каждой парой <***U***,**Ω**> ДСМ-гипотез. Чуть позднее мы подробнее рассмотрим ряд интересных особенностей порождения полезных гипотез (разумеется, уточнив в явном виде формальное значение этого термина) средствами обсуждаемой здесь техники.

Оценивая возможные объемы комбинаторного перебора, возникающего при восстановлении соответствующих диаграмм вложимости ДСМ-сходств, порождаемых заданными парами вида <***U***,**Ω**>, представляется естественным попытаться выделить наиболее сложные комбинации (в том числе – и случаи рекурсивной вложимости) расположения гиперкубов на каркасах соответствующих псевдо-деревьев.

Здесь, рассматривая различные комбинации архитектурных фрагментов типа α и типа β в структуре соответствующих псевдо-деревьев, прежде всего, заметим, что в части возможно большего числа вершин (и ребер) в таких архитектурах конструкции, где на одном уровне «глубины» находятся два и более гиперкубов сопоставимого размера, уступают конструкциям, содержащим лишь один гиперкуб (на соответствующем уровне глубины). Действительно, для каждого уровня «глубины» необходим соответствующий «запас» образующих из ***U***, позволяющий (за счет комбинирования подмножеств этого «запаса») обеспечить «строительным материалом» достаточное количество вершин рассматриваемого псевдо-дерева, лежащих на этих – разных – гиперкубах. Таким образом, наихудший (с точки зрения объемов вычислений) вариант архитектуры – чередование (при движении по вертикали от корня рассматриваемого псевдо-дерева к его листьям) линейных участков (фрагментов типа α) и единичных фрагментов типа β (гиперкубов). Попробуем аппроксимировать конфигурации такого вида, своего рода «предельной» конструкцией – иерархией лишь из заданного числа гиперкубов (пренебрегая в рассматриваемом «предельном» случае необходимостью зарезервировать хотя бы часть образующих из алфавита ***U*** на формирование архитектурных фрагментов типа α в анализируемом псевдо-дереве, и осознавая также, что возможное число вершин восстанавливаемого псевдо-дерева можно максимально увеличить за счет комбинирования всех различных подмножеств образующих, помещаемых на нижней границе соответствующего гиперкуба).

В рамках предлагаемой аппроксимации попробуем оценить, для какого количества гиперкубов в такой конфигурации число вершин в псевдо-дереве анализируемой архитектуры окажется максимально большим. Положим, что в алфавите ***U*** у нас (как и ранее) ровно n элементов. Рассмотрим выражение вида

$$N(n) = Const \times (k2^{n/k}) \le O(k2^{n/k}),$$

где k – параметр, характеризующий число гиперкубов в анализируемом («модельном») псевдо-дереве выбранной нами архитектуры, $O(2^{n/k})$ – дает оценку числа вершин гиперкуба, построенного на n/k образующих, а N(n) таким образом позволяет оценить общее число вершин этого псевдо-дерева. Переходя к функции вида

$$f(x) = Const \times (x2^{n/x}) \le O(x2^{n/x}),$$

поведение которой (в частности – участки роста и убывания, а также экстремумы) нас будет интересовать для целых и положительных значений переменной x, вычислим производную $f'(x)$ для $f(x)$ и приравняем её нулю:

$$f'(x) = Const \times (2^{n/x} - n\, ln2\, (2^{n/x}) /x) = Const \times (2^{n/x} (1- (n\, ln2)/x)) = 0$$

Разрешая полученное уравнение относительно x, получаем искомое экстремальное значение

$$x = n\, ln2 .$$

При этом вычисление второй производной для $f(x)$ показывает, что мы имеем дело с минимумом этой функции в найденной точке экстремума.

Принимая во внимание, что

$$ln2 \approx 0{,}69 > 0{,}5 ,$$

заключаем, что на интервале значений $x \in [1, n/2]$ рассматриваемая нами функция $f(x)$ монотонно убывает. Таким образом, при заданном размере алфавита ***U*** оцениваемая величина N(n) – число вершин восстанавливаемого псевдо-дерева – достигает максимальных возможных значений для простейшей конфигурации – с одним гиперкубом. А в другом предельном случае (для $x = n/2$) каждый из попадающих в рассматриваемое псевдо-дерево гиперкубов вырождается до трех (соответствующих непустым подмножествам множества из двух образующих) вершин, что в свою очередь минимизирует количество требующих рассмотрения вершин анализируемого псевдо-дерева, ограничивая его сверху величиной 3n/2.

Наглядное представление о характере поведения функции $f(x) = Const \times (x2^{n/x})$ на сегменте положительных значений переменной x дают диаграммы на Рис.5.3.1 и 5.3.2.

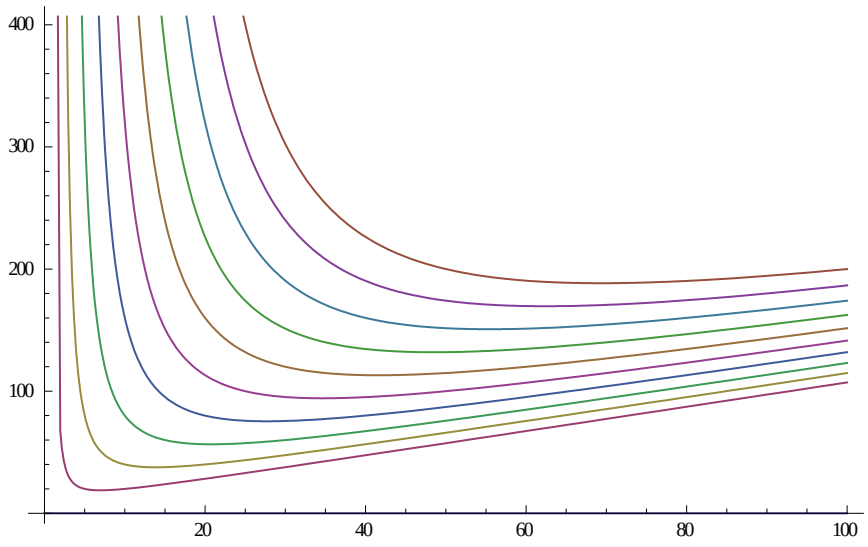


Рис. 5.3.1. Графики функции $f(x) = Const \times (x2^{n/x})$ для $n \in \{10, 20, \ldots 100\}$ и $x \in (0,100]$.

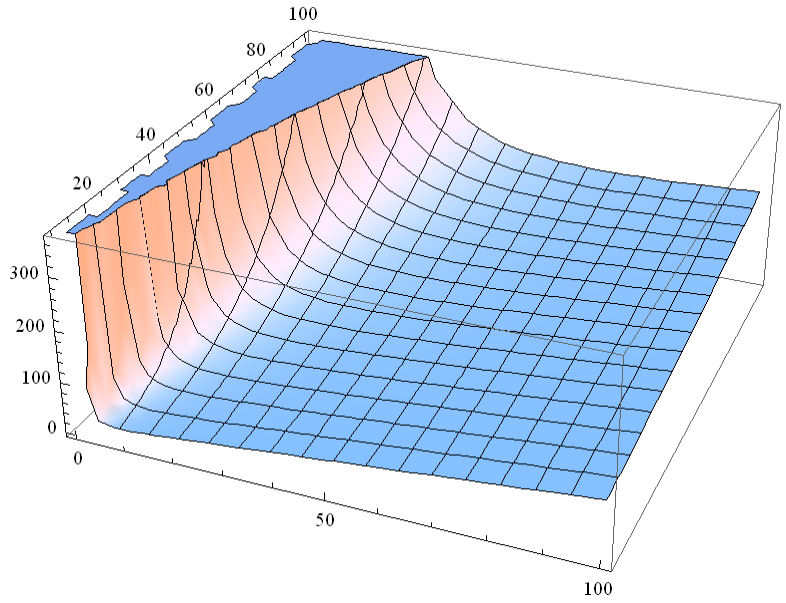


Рис. 5.3.2. Графики функции f(x) = Const x ($x2^{y/x}$) для $x \in (0, 100]$ и y∈[10,100] .

Полезные дополнительные (по отношению к традиционной алгоритмике ДСМ-метода – см. например, [*ДСМ АПГ*] и др.) возможности развиваемая здесь техника целенаправленного перебора вариантов предоставляет также и при порождении ДСМ-гипотез средствами более сильных, чем простое сходство (см. в частности, [*Финн 10, ДСМ АПГ*] и др.) стратегий рассуждений. Целенаправленная навигация вдоль соответствующих псевдо-деревьев помогает ускорить проверку соответствующих дополнительных условий в используемых решающих предикатах и правилах правдоподобного вывода. Наиболее существенные особенности анализа подобного рода дополнительных условий мы рассмотрим на примере проверки выполнимости условий так называемого запрета на контрпримеры (см. [*Финн 10*] и др.). Критически важной особенностью этого расширения стандартных предикатов и ППВ простого сходства является необходимость оперировать одновременно примерами обоих знаков из исходного множества ДСМ-примеров **Ω** (т.е. одновременно – и примерами, и контрпримерами).

Логическое условие запрета на контрпримеры (будем именовать его также как условие ***ЗКП***) выражается следующим соотношением[1]:

$$\forall X \forall Y \; (((v \subset X) \;\&\; (w \subseteq Y)) \rightarrow (J_{(1,n)}(X \Rightarrow_1 Y) \vee J_{(\tau,n)}(X \Rightarrow_1 Y))), \qquad (\textbf{\textit{ЗКП}})^+$$

и утверждает, что (для позитивного ***ЗКП***) искомая ДСМ-причина **v** ее следствие **w** не содержатся в (–)-примерах (т.е. в контрпримерах из исходного множества **Ω,** возможно, расширенного вновь доопределенными на предыдущих шагах ДСМ-рассуждения примерами) а также в примерах с оценкой ⟨0, n⟩ (фактическое противоречие на текущем n-ном шаге ДСМ-рассуждения).

Обсудим некоторые «технические» особенности рассматриваемой ситуации. Прежде всего, дадим соответствующие определения:

5.3.3. Определение.

1. Задачей *о существовании ДСМ-сходства, удовлетворяющего условию запрета на контрпримеры* (Задачей ***С-ЗКП***), будем называть следующую задачу:

---

[1] Здесь представлен случай позитивного условия ЗКП . Случай негативного условия формулируется аналогичным образом (см., например [*Финн 10*] и др.)

Дано: конечный алфавит $\boldsymbol{U} = \{a_1, a_2, \ldots, a_n\}$, конечные множества примеров $\boldsymbol{\Omega}^{\alpha} = \{O_1, O_2, \ldots, O_n\}$ и контрпримеров $\boldsymbol{\Omega}^{-\alpha} = \{O^-_1, O^-_2, \ldots, O^-_m\}$ для ДСМ-обучения.

Найти: существует ли в множестве $GC(\boldsymbol{U}, \boldsymbol{\Omega}^{\alpha})$ ДСМ-сходств, порожденных на примерах из множества $\boldsymbol{\Omega}^{\alpha}$, хотя бы один элемент, удовлетворяющий условию запрета на контрпримеры (т.е. не вкладывающееся ни в один из элементов множества $\boldsymbol{\Omega}^{-\alpha}$ – подробнее см. также [1] и др.).

2. Задачей *о перечислении всех ДСМ-сходств, удовлетворяющих условию запрета на контрпримеры* (Задачей ***Все_С-ЗКП***), будем называть следующую задачу:

Дано: конечный алфавит $\boldsymbol{U} = \{a_1, a_2, \ldots, a_n\}$, конечные множества примеров $\boldsymbol{\Omega}^{\alpha} = \{O_1, O_2, \ldots, O_n\}$ и контрпримеров $\boldsymbol{\Omega}^{-\alpha} = \{O^-_1, O^-_2, \ldots, O^-_m\}$ для ДСМ-обучения.

Найти: перечислить в множестве $GC(\boldsymbol{U}, \boldsymbol{\Omega}^{\alpha})$ ДСМ-сходств, порожденных на примерах из множества $\boldsymbol{\Omega}^{\alpha}$, все элементы, удовлетворяющие условию запрета на контрпримеры (т.е. не вкладывающиеся ни в один из элементов множества $\boldsymbol{\Omega}^{-\alpha}$).

□

5.3.4. Определение.

При заданных

- конечном алфавите $\boldsymbol{U} = \{a_1, a_2, \ldots, a_n\}$ а также
- конечных множествах примеров $\boldsymbol{\Omega}^{\alpha} = \{O_1, O_2, \ldots, O_n\}$
  и контрпримеров $\boldsymbol{\Omega}^{-\alpha} = \{O^-_1, O^-_2, \ldots, O^-_m\}$ для ДСМ-обучения

стоп-листом SL_j для каждого контрпримера O^-_j из множества $\boldsymbol{\Omega}^{-\alpha}$ будем называть соответствующее множество $\{ x^j_1, x^j_2, \ldots, x^j_{n(j)} \}$ образующих из $\boldsymbol{U}$, которые не входят в O^-_j.

□

Далее убедимся, что имеет место следующее полезное

5.3.5. Утверждение.

Элемент $\mathbf{x}$ множества $GC(\boldsymbol{U}, \boldsymbol{\Omega}^{\alpha})$ всех ДСМ-сходств, порожденных из множества примеров $\boldsymbol{\Omega}^{\alpha}$ удовлетворяет условию запрета на контрпримеры (условию ***ЗКП***) тогда и только тогда, когда его пересечение с каждым из стоп-листов $SL_1, SL_2, \ldots, SL_m$ не пусто:

$$\forall j \ [(j \in \{1,2, \ldots, m\}) \ \& \ (\mathbf{x} \cap SL_j \neq \varnothing)]$$

Доказательство.

=>. Если $\mathbf{x} \in GC(\boldsymbol{U}, \boldsymbol{\Omega}^{\alpha})$ и выполнено условие запрета на контрпримеры, то для каждого контрпримера O^-_j выполняется условие $\mathbf{x} \subseteq O^-_j$, т.е.

$$\mathbf{x} \cap O^-_j \neq \mathbf{x},$$

т.е. в множестве образующих $\mathbf{x}$ найдется по крайней мере один элемент из множества $\boldsymbol{U}$, который не входит в O^-_j, что означает выполнимость условия:

$$\mathbf{x} \cap SL_j \neq \varnothing .$$

<=. В случае, когда для каждого стоп-листа SL_j имеем

$$\mathbf{x} \cap SL_j \neq \varnothing,$$

следствием этого соотношения будет соотношение

$$\mathbf{x} \cap O^-_j \neq \mathbf{x},$$

которое и означает, что

$$\mathbf{x} \not\subseteq O^-_j \ .$$

□

Перейдем к обсуждению комбинаторных свойств рассматриваемых нами переборных задач. Имеет место

5.3.6. Утверждение.

Задача *о существовании ДСМ-сходства, удовлетворяющего условию запрета на контрпримеры* (задача ***С-ЗКП***), принадлежит к классу полиномиально разрешимых (см., например, [*Гэри 82*] и др.).

Доказательство.

1. Построим по заданным алфавиту $\boldsymbol{U}$ и множеству контрпримеров $\boldsymbol{\Omega}^{-\alpha}$ соответствующее множество стоп-листов SL_1, SL_2, … , SL_m. (Такое построение – полиномиально сложная процедура). Если среди сформированных стоп-листов имеется ∅, то множество решений задачи ***С-ЗКП*** пусто.

2. Если ни один из стоп-листов не оказался пустым множеством, выберем O_1 – первый из элементов множества примеров $\boldsymbol{\Omega}^{\alpha}$, чтобы затем (полиномиально быстро - !) построить каркас $GK_{\mathbf{D},O1}$ соответствующего ему псевдо-дерева $\mathbf{D_GC}(\boldsymbol{U}, \boldsymbol{\Omega}^{\alpha})\big|_{\{O1\}}$ в диаграмме $\mathbf{D_GC}(\boldsymbol{U},\boldsymbol{\Omega}^{\alpha})$ всех возможных ДСМ-сходств, порождаемых на множестве примеров $\boldsymbol{\Omega}^{\alpha}$. Далее последовательно рассмотрим вершины (которых не более чем n = $|\boldsymbol{U}|$) псевдо-дерева $\mathbf{D_GC}(\boldsymbol{U},\boldsymbol{\Omega}^{\alpha})\big|_{\{O1\}}$, непосредственно доминируемые его корнем – O_1: для каждой такой вершины непустота ее пересечений (как множества образующих из $\boldsymbol{U}$) со всеми стоп-листами SL_1,SL_2, … , SL_m является достаточным условием существования в $\mathbf{D_GC}(\boldsymbol{U},\boldsymbol{\Omega}^{\alpha})\big|_{\{O1\}}$ вершины, порождающей решение задачи ***С-ЗКП***. Действительно, выбрав из каждого такого пересечения по одной образующей и вычислив замыкание объединения порожденного таким образом множества образующих, мы получим сходство (как элемент множества всех сходств $\mathbf{D_GC}(\boldsymbol{U},\boldsymbol{\Omega}^{\alpha})$), удовлетворяющее (по построению – см. общие со стоп-листами образующие) условию запрета на контрпримеры.

При этом, в случае, когда корень A_1 рассматриваемого псевдо-дерева $\mathbf{D_GC}(\boldsymbol{U}, \boldsymbol{\Omega}^{\alpha})\big|_{\{O1\}}$ есть «наибольшая» вершина некоторого гиперкуба в этом псевдо-дереве (т.е. O_1 лежит на архитектурном фрагменте типа β), рассматриваемое выше множество непосредственно доминируемых им вершин анализируемого псевдо-дерева полиномиально быстро восстанавливается по (найденной при построении каркаса $GK_{\mathbf{D},O1}$) нижней границе этого гиперкуба и (также полиномиально быстро восстанавливаемому – см. **Утверждение 5.2.14** выше, в первой части настоящей работы) подмножеству $\mathbf{H}(\boldsymbol{U},\boldsymbol{\Omega}^{\alpha},O_1)$ всех примеров из $\boldsymbol{\Omega}^{\alpha}$, раскрывающих вершины рассматриваемого гиперкуба. (Действительно, для восстановления искомых непосредственно доминируемых O_1 вершин, лежащих на рассматриваемом гиперкубе, достаточно: сформировать попарные пересечения O_1 и элементов подмножества $\mathbf{H}(\boldsymbol{U},\boldsymbol{\Omega}^{\alpha},O_1)$, затем удалить из полученного перечня каждое такое пересечение, которое вкладывается хотя бы в один из других элементов этого перечня, не совпадая с ним. Предлагаемая процедура редактирования рассматриваемого перечня пересечений, как несложно убедиться, может быть реализована не более чем попарными сравнениями его элементов, т.е. - полиномиально быстро).

В случае принадлежности корня O_1 к архитектурному фрагменту типа α, непосредственно доминируемые этим корнем вершины анализируемого псевдо-дерева лежат на (только что построенном) каркасе $GK_{\mathbf{D},O1}$.

3. Если для текущего элемента O_i исходного множества примеров представленная выше в п.2 последовательность процедур (из-за пустоты пересечений сопоставленных доминируемым вершинам множеств образующих из $\boldsymbol{U}$ с какими-либо из рассматриваемых стоп-листов), переходим от примера O_i к примеру O_{i+1} , и так до завершения (полиномиально быстро реализуемого) перебора всех элементов множества $\boldsymbol{\Omega}^{\alpha}$.

□

5.3.7. Утверждение.

Задача *о числе ДСМ-сходств, удовлетворяющих условию запрета на контрпримеры* (***ВСЕ_ С-ЗКП***), принадлежит к классу ***#РС*** перечислительно полных (см. подробнее например, [*Гэри 82, Simon 77, Valiant 79, Valiant 79-a*] и др.) комбинаторных проблем.

Доказательство.
Предыдущим **Утверждением 5.3.6** показана полиномиально быстрая разрешимость задачи ***С-ЗКП***. Продемонстрируем возможности полиномиальной сводимости известной перечислительно полной задачи ***МОНОТОННАЯ ВЫПОЛНИМОСТЬ*** (см., например, [*Гэри 82, Simon 77, Valiant 79, Valiant 79-a*] и др.) к задаче ***ВСЕ_С-ЗКП.*** Для этого мы покажем, как по заданной булевской функции **Ф** в виде 2-КНФ будут порождены множества $\boldsymbol{U}(\boldsymbol{\Phi})$, $\boldsymbol{\Omega}^{\alpha}(\boldsymbol{\Phi})$ и $\boldsymbol{\Omega}^{-\alpha}(\boldsymbol{\Phi})$, а также будет показано, что множество таких сходств из $\mathbf{D_GC}(\boldsymbol{U}(\boldsymbol{\Phi}),\boldsymbol{\Omega}^{\alpha}(\boldsymbol{\Phi}))$, которые удовлетворяют также условию запрета на контрпримеры, изоморфно множеству единиц функции **Ф**.

Обратимся к более подробному описанию предлагаемого механизма сводимости рассматриваемых комбинаторных задач. Прежде всего, уточним, что при упоминании задачи о выполнимости монотонной булевской функции, представленной в виде 2-КНФ, речь идет о следующей комбинаторной проблеме:

Дано: монотонная булевская функция **Ф** в виде 2-КНФ

$$\boldsymbol{\Phi}(x_1,x_2, \dots ,x_n) = \boldsymbol{\Phi_1} \,\&\, \boldsymbol{\Phi_2} \,\&\, \dots \,\&\, \boldsymbol{\Phi_m} = (x_{11} \vee x_{12}) \,\&\, (x_{21} \vee x_{22}) \,\&\, \dots \,\&\, (x_{m1} \vee x_{m2}) \quad \textbf{(!)}$$

Найти: число наборов значений переменных $x_1,x_2, \dots ,x_n$, выполняющих функцию **Ф**.

Пусть функция **Ф** имеет представленный выше вид - (**!**). Определим:

- алфавит $\boldsymbol{U}$ для формирования примеров и контрпримеров в соответствующей задаче ***ВСЕ_С-ЗКП*** (**Ф**) как
$$\boldsymbol{U} = \{ x_1,x_2, \dots , x_n\},$$
- множество $\boldsymbol{\Omega}^{\alpha}$ примеров для ДСМ-обучения в соответствующей задаче ***ВСЕ_С-ЗКП*** (**Ф**) как
$$\boldsymbol{\Omega}^{\alpha} = \{O_1, O_2, \dots , O_{n+2}\},$$
- множество $\boldsymbol{\Omega}^{-\alpha}$ контрпримеров для ДСМ-обучения в соответствующей задаче ***ВСЕ_С-ЗКП*** (**Ф**) как
$$\boldsymbol{\Omega}^{-\alpha} = \{O^-_1, O^-_2, \dots , O^-_m\},$$

где

- $O^-_1 = \boldsymbol{U} \setminus \{ x_{11} , x_{12}\}$
- $O^-_2 = \boldsymbol{U} \setminus \{ x_{21} , x_{22}\}$
- …
- $O^-_m = \boldsymbol{U} \setminus \{ x_{m1} , x_{m2}\}$

и также

- $O_1 = \{ \quad x_2,x_3, \dots , x_n \}$
- $O_2 = \{ x_1, \quad x_3, \dots , x_n \}$
- ...
- $O_n = \{ x_1,x_2, \dots , x_{n-1} \}$
- $O_{n+1} = O_{n+2} = \{ x_1,x_2, \dots , x_n\}$,

и, наконец,

- $SL_1 = \{ x_{11} , x_{12}\}$
- $SL_2 = \{ x_{21} , x_{22}\}$
- …
- $SL_m = \{ x_{m1} , x_{m2}\}$.

Таким образом:

1) множество всех ДСМ-сходств, порождаемых из примеров $\boldsymbol{\Omega}^{\alpha} = \{O_1, O_2, \dots , O_{n+2}\}$ есть

$$\{\varnothing; \{x_1\},\{x_2\}, \dots , \{x_n\}; \{ x_1,x_2\}, \{ x_1,x_3\}, \dots , \{ x_{n-1},x_n \}; \{ x_1,x_2,x_3 \}, \dots , \{ x_1,x_2, \dots , x_n\}\},$$

что позволяет перечислить (кодируя единицей на позиции i вхождение соответствующей переменной x_i в порождаемое ДСМ-сходство) все 2^n булевских наборов значений переменных $x_1,x_2, \dots , x_n$:

<00...00>, <00...01>, ... , <11...10>, <11...11>;

2) В соответствии с **Утверждением 5.3.5** условие ***ЗКП*** выполняется для тех и лишь для тех наборов значений переменных $x_1,x_2, \dots , x_n$, которые имеют непустое пересечение (общие переменные) с каждым из стоп-листов $SL_1, SL_2, \dots, SL_m$. Но это, в свою очередь, означает, что для каждой из подформул

$$\mathbf{\Phi}_i = (x_{i1} \vee x_{i2}), \quad \text{где } i \in \{1,2, \dots , m\},$$

выражения

$$\mathbf{\Phi}(x_1,x_2, \dots , x_n) = \mathbf{\Phi}_1 \,\&\, \mathbf{\Phi}_2 \,\&\, \dots \,\&\, \mathbf{\Phi}_m = (x_{11} \vee x_{12}) \,\&\, (x_{21} \vee x_{22}) \,\&\, \dots \,\&\, (x_{m1} \vee x_{m2})$$

имеется по крайней мере одна (общая с соответствующим стоп-листом) из пары переменных $<x_{i1}, x_{i2}>$, определяющая также и выполнимость подформулы $\mathbf{\Phi}_i$. Таким образом, выполнимость функции **Ф** имеет место тогда и только тогда, когда в соответствующем «контексте» $<\boldsymbol{U}(\mathbf{\Phi}), \mathbf{\Omega}^{\alpha}(\mathbf{\Phi})$ и $\mathbf{\Omega}^{-\alpha}(\mathbf{\Phi})>$ выполняется условие ***ЗКП***.

Предлагаемый «контекст» $<\boldsymbol{U}(\mathbf{\Phi}), \mathbf{\Omega}^{\alpha}(\mathbf{\Phi})$ и $\mathbf{\Omega}^{-\alpha}(\mathbf{\Phi})>$ строится по соответствующей формуле **Ф** полиномиально быстро, что и определяет полиномиальную сводимость задачи ***МОНОТОННАЯ ВЫПОЛНИМОСТЬ*** к задаче ***ВСЕ_С-ЗКП***.

□

Уже подробно рассмотренная нами для случая работы с простыми ДСМ-сходствами техника управления перебором *восстановление псевдо-деревьев по их каркасам* (сокращенно: *ПД + Каркасы*) также может быть полезна и при обработке дополнительных логических условий – запрета на контрпримеры, единственности причины и др. (см., например, [*Финн 10*] и др.) обоих знаков. Особенностью применения подобной техники оптимизации перебора в данном случае являются как (уже названная выше) необходимость оперировать примерами обоих знаков (примерами и контрпримерами), так и дополнительные ограничения на допустимость порождаемых гипотез, обусловленные структурными особенностями τ-примеров, доопределяемых в каждом конкретном случае (например, при проверке выполнимости Условия Каузальной Полноты - см., например, [*Финн 10*] и др.).

Конкретные алгоритмы перечисления всех сходств, удовлетворяющих условию запрета на контрпримеры легко сформировать, например, дополнив уже подробно представленную выше технику *ПД + Каркасы* процедурами направленного перебора стоп-листов, рассмотренную при доказательстве **Утверждения 5.3.6**.

Итак, переформулируем условие принадлежности пары <v,w> множеству всех сходств из $\mathbf{D_GC}(\boldsymbol{U}, \mathbf{\Omega})$, дополнительно удовлетворяющих (α)-условию ***ЗКП*** (т.е. условию принадлежности множеству ***С-ЗКП***$^{\alpha}$ – позитивных, когда α есть +, и негативных, когда α есть -):

$$[(<v,w> \in (\boldsymbol{C\text{-}ЗКП}^{\alpha})] \quad \textit{тогда и только тогда, колгда}$$

$$\textit{iff} \;\; \forall \omega \; \big((\omega \in \mathbf{\Omega}^{-\alpha}) \rightarrow \exists v_\omega \, [(v_\omega \subseteq SL_\omega) \,\&\, (v_\omega \subseteq v)] \,\&\, (v \in \mathbf{D_GC}(\boldsymbol{U}, \mathbf{\Omega})) \big)$$

т.е. пара <v,w> соответствует условию ***ЗКП*** тогда и только тогда, когда для каждого контрпримера ω из множества $\mathbf{\Omega}^{-\alpha}$ найдется некоторое непустое подмножество v_ω образующих из $\boldsymbol{U}$, которое содержится одновременно в стоп-листе SL_ω для контрпримера ω и в сходстве v из пары <v,w>.

Алгоритм перечисления всех элементов соответствующего множества ***С-ЗКП***$^{\alpha}$ (**Алгоритм *С-ЗКП***$^{\alpha}$) определяет следующая

Схема **Алгоритма *С-ЗКП***$^{\alpha}$

1. Рассмотрим все непосредственно доминируемые корнем O_i вершины текущего псевдо-дерева номер i. (Их не более чем $|O_i| = \{a_1, a_2, \dots, a_r\} = r \le n = |\boldsymbol{U}|$).

2. Для каждой из таких вершин с текущим номером j рассмотрим кортеж ее пересечений (как множества образующих из ***U***) со всеми элементами имеющегося множества стоп-листов $\boldsymbol{SL}(\boldsymbol{\Omega}^{-\alpha})$:

$$\boldsymbol{U_SL} = < U_SL_1, U_SL_2 , \dots , U_SL_s >.$$

3. Если кортеж $\boldsymbol{U_SL}(O_{ij})$ не содержит пустых элементов, то искомая ДСМ-гипотеза из $\boldsymbol{C\text{-}3КП}^{\alpha}$ существует (т.е. множество $\boldsymbol{C\text{-}3КП}^{\alpha}$ в этом случае не пусто).

4. Возьмем в каждом из элементов кортежа $\boldsymbol{U_SL}(O_{ij})$ непустое подмножество $U_SL_k\,(O_{ij})^*$ и рассмотрим (как соответствующее замыкание)

$$v^*(O_{ij}) = [\ \bigcup_{k=1}^{s} U_SL_k\,(O_{ij})^*\ ]_{(U,\boldsymbol{\Omega})} .$$

5. Нетрудно убедиться, что по построению $v^*(O_{ij})$ есть элемент множества $\boldsymbol{C\text{-}3КП}^{\alpha}$.

□

Дополнительный комментарий (ДК).

При этом все множество $\boldsymbol{C\text{-}3КП}^{\alpha}$ будет располагаться в данном случае между

- верхней по вложению границей, определяемой замыканием объединения всех элементов кортежа $\boldsymbol{U_SL}(O_{ij})$ и
- нижней границей, определяемой всеми замыканиями объединений всех одноэлементных подмножеств всех компонентов кортежа $\boldsymbol{U_SL}(O_{ij})$ (когда из каждого такого компонента берется ровно одна образующая).

Опираясь на использованную при доказательстве **Утверждение 5.3.7** «техническую» конструкцию, можно показать, что размеры нижней границы множества $\boldsymbol{C\text{-}3КП}^{\alpha}$ определяет

5.3.8. Следствие.

Задача о числе элементов нижней границы (см. *ДК* выше) множества $\boldsymbol{C\text{-}3КП}^{\alpha}$ перечислительно полна.

□

Таким образом, для «быстрого» поиска подходящих элементов множества вновь можно воспользоваться техникой *ПД+Каркасы*:

- построить необходимые стоп-листы и
- перейти к систематическому обзору вершин, непосредственно доминируемых корнями соответствующих псевдо-деревьев (например, воспользовавшись возможностями полиномиально быстрого порождения каркасов или же - в случае попадания в соответствующие гиперкубы - возможностями быстрого порождения «верхней границы» - т.е. множества вершин вида *единственный 0 в кортеже 1* – в каждом из таких гиперкубов);
- далее, используя явный вид уточнения условий «полезности» соответствующих элементов $\boldsymbol{C\text{-}3КП}^{\alpha}$ (например, перечень «обязательных» образующих и т.п.) сперва породить «полезные» элементы множества $\boldsymbol{C\text{-}3КП}^{\alpha}$, а затем, если потребуется, - и все остальные.

Перейдем к обсуждению возможностей предлагаемой техники сокращений перебора при реализации ДСМ-правил правдоподобного вывода второго рода (ППВ-II - см., например, [*Финн 10*] и др.). Предметом анализа здесь будет техника поиска причинных зависимостей, позволяющих формировать заключения (ДСМ-прогноз) о свойствах новых объектов (исходно недоопределенных примеров).

Пусть заданы:

$\boldsymbol{Ex} = \{Ex_1, Ex_2, \dots , Ex_m\}$ – исходное множество примеров (в том числе – примеров и контрпримеров для ДСМ-обучения. При этом, если не оговорено противное, будем считать, что в множестве ***Ex*** перечислены примеры одного «знака». Для разделения

примеров разных знаков будем использовать дополнительный верхний индекс - $\boldsymbol{Ex}^{+}$ и $\boldsymbol{Ex}^{-}$). Будем также считать, что каждый пример

Ex_i из $\boldsymbol{Ex}$ характеризуется описанием соответствующего объекта O_i (формирующего его множества образующих исходного алфавита объектов U_L) а также перечнем его свойств P_i (представляемым соответствующим множеством образующих алфавита свойств U_R), связанными отношением $=>_1$ (отношением «объект обладает свойствами»)

$$(O_i =>_1 P_i), \quad i = 1,2, \ldots ,m;$$

$U = U_L \cup U_R$ - объединенный алфавит, а

$\Omega_L = \{O_1, O_2, \ldots , O_m\}$ - множество описаний объектов из $\boldsymbol{Ex}$: $O_i \in 2^{U_L}$,

$\Omega_R = \{P_1, P_2, \ldots , P_m\}$ - множество описаний свойств этих объектов: $P_i \in 2^{U_R}$, кроме того:

$\boldsymbol{Ex}^{\tau}= \{Ex^{\tau}_1, Ex^{\tau}_2, \ldots , Ex^{\tau}_s\}$ – исходное множество примеров, предложенных для ДСМ-доопределения (т.е. множество примеров Ex^{τ}_i вида $O^{\tau}_i =>_1 P^{\tau}_i$, где i = 1,2, ... ,s).

По определению ППВ-II для каждого Ex^{τ}_r из $\boldsymbol{Ex}^{\tau}$ необходимо найти соответствующий набор $\boldsymbol{H}$ гипотез h_j вида

$$(v_j =>_2 w_j), \quad j = 1,2, \ldots , q(r)$$

таких, что

$$\bigcup_{j=1}^{q(r)} w_j = P^{\tau}_r .$$

(При этом последнее равенство – ключевое условие в определении решающих предикатов и соответствующих правил правдоподобного вывода по аналогии – см. подробнее [*Финн 10*] и др., - где, в частности, для позитивной аналогии

$$\begin{aligned}\boldsymbol{\Pi}^{+}_{\boldsymbol{n}}(O^{\tau}_r, P^{\tau}_r, k) = \exists w_1 \ldots \exists w_k \ \{ \boldsymbol{(} & [\exists v_1 (\mathbf{J}^{<1,r>}(v_1 ==>_2 w_1) \ \& \ (v_1 \subset O^{\tau}_r)] \ \& \\ & \& \ [\exists v_2 \ [\mathbf{J}^{<1,r>}(v_2 ==>_2 w_2) \ \& \ (v_2 \subset O^{\tau}_r)] \ \& \ \ldots \\ & \ldots \ \& \ [\exists v_k \ [\mathbf{J}^{<1,r>}(v_k ==>_2 w_r) \ \& \ (v_k \subset O)] \ \& \\ & \& \ (w_1 \cup w_2 \cup \ldots \cup w_r = P^{\tau}_r \ )\boldsymbol{)} \ \& \\ & \& \ \forall U \ (\boldsymbol{(} \ (U \subseteq P^{\tau}_r) \ \& \ (U \neq \varnothing)\boldsymbol{)} \rightarrow \\ & \qquad \rightarrow \ \neg \exists Z \ \boldsymbol{(} \ [\mathbf{J}^{<-1,r>}(Z ==>_2 U) \vee \mathbf{J}^{<0,r>}(Z ==>_2 U) \ ] \ \& \ (Z \subseteq O^{\tau}_r)\boldsymbol{)} \ \}, \end{aligned}$$

а предикаты $\boldsymbol{\Pi}^{-}_{\boldsymbol{n}}$ и $\boldsymbol{\Pi}^{0}_{\boldsymbol{n}}$ формулируются путем перехода, соответственно, к учету негативных зависимостей и зависимостей одновременно обоих знаков по отношению к доопределяемому примеру Ex^{τ}_r).

Искать необходимые для проверки целевого (напомним, речь идет о покрытии P^{τ}_r множествами w_j , где j пробегает значения от 1 до q(r)) соотношения w_j удобно в псевдо-деревьях $\mathbf{D_GC}(\boldsymbol{U}_R, \boldsymbol{\Omega}_R)\big|_{Pi}$, построенных на диаграмме сходств $\mathbf{D_GC}(\boldsymbol{U}_R, \boldsymbol{\Omega}_R)$ примеров (в данном случае – свойств изучаемых объектов) из Ω_R.

Для этого: последовательно для каждого C_i из Ω_R строим каркас $GK_{\mathbf{D},\mathbf{C}i}$ соответствующего псевдо-дерева $\mathbf{D_GC}(\boldsymbol{U}_R, \boldsymbol{\Omega}_R)\big|_{Pi}$. На каждом таком каркасе выбираем все те $w_{j(i)}$, для которых

$$w_j \subseteq P^{\tau}_r ,$$

затем вычисляем $\bigcup_j w_j$.

Если это объединение строго вкладывается в P^{τ}_r, переходим к следующему элементу P_{i+1} из Ω_R а также следующему псевдо-дереву $\mathbf{D_GC}(\boldsymbol{U}_R, \boldsymbol{\Omega}_R)\big|_{Pi+1}$ (т.к. для текущих множества свойств P_{i+1} и псевдо-дерева $\mathbf{D_GC}(\boldsymbol{U}_R, \boldsymbol{\Omega}_R)\big|_{Pi}$ имеющиеся ППВ-II не работают).

Если же

$$\bigcup_j w_j = P^{\tau}_r ,$$

обращаемся к перечислению всех таких подпокрытий найденного нами покрытия $\boldsymbol{w}(r)$ для анализируемого множества свойств P^{τ}_r, которые, как и прежде, сохраняют найденное качество - покрывать P^{τ}_r.

Для этого можно использовать, например, следующий

5.3.9. Алгоритм.

Вход: множество P^{τ}_r (свойств примеров из $\boldsymbol{Ex}^{\tau}$) и покрывающее в точности его множество

$$\boldsymbol{w}(r) = \{w_{j1(r)}, w_{j2(r)} , \ldots, w_{jp(r)}\}$$

наборов свойств в ДСМ-гипотезах h_j вида

$$v_j =>_2 w_j, \quad j = 1,2, \ldots , q(r) ,$$

порожденных применением ППВ-I из примеров множества $\boldsymbol{Ex}$.

Выход: множество всех подмножеств $\boldsymbol{w'}(r)$ множества $\boldsymbol{w}(r)$, таких, что объединение элементов каждого из них по-прежнему покрывает целевое множество свойств P^{τ}_r.

*Описание **Алгоритма**:*

Комментарий: Перенумеруем элементы исходного покрытия $\boldsymbol{w}$ и поместим их в исходный перечень **ТЕКУЩЕЕ ПОКРЫТИЕ**. Стек **ИЗБЫТОЧНЫЕ ЭЛЕМЕНТЫ** используем для перечисления (в лексикографическом порядке - !) комбинаций избыточных (т.е. таких, удаление которых не нарушает покрытие множества свойств C^{τ}_r элементами перечня **ТЕКУЩЕЕ ПОКРЫТИЕ**) элементов исходного покрытия.

ШАГ 1. В список **ТЕКУЩЕЕ ПОКРЫТИЕ** заносим все r элементов исходного покрытия $\boldsymbol{w}(r)$. Переменной **КОРЕНЬ** присваиваем значение w[1].

ШАГ 2. Переменной **ТЕКУЩИЙ ЭЛЕМЕНТ ПОКРЫТИЯ** присваиваем текущее значение переменной **КОРЕНЬ**. Стек **ИЗБЫТОЧНЫЕ ЭЛЕМЕНТЫ** назначаем пустым.

ШАГ 3. Переносим элемент **ТЕКУЩИЙ ЭЛЕМЕНТ ПОКРЫТИЯ** из списка **ТЕКУЩЕЕ ПОКРЫТИЕ** в стек **ИЗБЫТОЧНЫЕ ЭЛЕМЕНТЫ**.

ШАГ 4. Проверяем условие «*входящие в* **ТЕКУЩЕЕ ПОКРЫТИЕ** *элементы исходного перечня* $\boldsymbol{w}(r)$ *покрывают целевое множество свойств* P». Если это условие выполнено, отправляем текущий вариант списка **ТЕКУЩЕЕ ПОКРЫТИЕ** на **ВЫХОД Алгоритма** (например, на печать) и переходим на ШАГ 6. В противном случае

ШАГ 5. Возвращаем текущее значение **ТЕКУЩИЙ ЭЛЕМЕНТ ПОКРЫТИЯ** из стека **ИЗБЫТОЧНЫЕ ЭЛЕМЕНТЫ** в список **ТЕКУЩЕЕ ПОКРЫТИЕ**, затем переходим на ШАГ 7.

ШАГ 6. Переменной **ТЕКУЩИЙ ЭЛЕМЕНТ ПОКРЫТИЯ** присваиваем следующее за текущим значение (заносим сюда следующий элемент исходного покрытия $\boldsymbol{w}(r)$). Если это невозможно (т.е. на данном шаге Алгоритма **ТЕКУЩИЙ ЭЛЕМЕНТ ПОКРЫТИЯ** содержит последний элемент исходного покрытия $\boldsymbol{w}(r)$),

ШАГ 7. Откатываемся в списке **ИЗБЫТОЧНЫЕ ЭЛЕМЕНТЫ** к предпоследнему элементу w[i] (удаляем из стека **ИЗБЫТОЧНЫЕ ЭЛЕМЕНТЫ** в его текущем состоянии последний элемент - w[j]). Если это невозможно (т.е. w[j] – корень этого стека: его значение совпадает с текущим значением переменной **КОРЕНЬ**), переходим на ШАГ 9, в противном случае

ШАГ 8. Переходим к следующему (за помещенным в **ТЕКУЩИЙ ЭЛЕМЕНТ ПОКРЫТИЯ** на данный момент w[j] в исходном перечне $\boldsymbol{w}(r)$) элементу w[j+1]:

помещаем в **ТЕКУЩИЙ ЭЛЕМЕНТ ПОКРЫТИЯ** этот элемент w[j+1] и возвращаемся на ШАГ 3.

ШАГ 9. Переменной **КОРЕНЬ** присваиваем следующее за текущим значение (помещаем в нее следующий элемент исходного покрытия ***w***(r)). Если это невозможно (т.е. все элементы исходного покрытия уже были рассмотрены в качестве корня стека избыточных элементов покрытия), переходим на ШАГ 10. В противном случае переходим на ШАГ 2.

ШАГ 10. **СТОП**

□

Корректность предложенного **Алгоритма** определяет

5.3.10. Утверждение.

В части перечисления всех удовлетворяющих условию «*входящие в* **ТЕКУЩЕЕ ПОКРЫТИЕ** *элементы исходного перечня* ***w***(r) *покрывают целевое множество свойств P*» подпокрытий для заданного исходного покрытия ***w***(r) **Алгоритм 5.3.9** характеризуют свойства полноты и точности.

Доказательство.

Реализованная в **Алгоритме 5.3.9** тактика перебора подпокрытий исходного покрытия ***w***(r) в *лексикографическом порядке* (т.е. начиная с самых «длинных» стеков избыточных элементов, и далее - ко все более и более коротким – с выбором в качестве корня каждого такого стека избыточных элементов только таких элементов исходного покрытия ***w***(r), которые имеют все более и более «старшие» номера; при этом переход от текущего корня стека к следующему организован так, что после каждого такого перехода рассматриваются лишь те элементы исходного покрытия ***w***(r), номера которых строго больше, чем номер элемента, стоящего в корне на текущем шаге) гарантирует, что каждая целевая последовательность отбрасываемых элементов исходного покрытия ***w***(r) будет сформирована лишь один раз. В дополнение к этому, постоянный контроль выполнимости условия «*входящие в* **ТЕКУЩЕЕ ПОКРЫТИЕ** *элементы исходного перечня* ***w***(r) *покрывают целевое множество свойств C*» гарантирует, что ни одно из подходящих подпокрытий не будет пропущено. Наконец, никаких других (отличных от перечисленных) объектов **Алгоритм 5.3.9** не строит.

□

Будем называть покрытия, порождаемые **Алгоритмом 5.3.9**, *полезными*. Для каждого такого покрытия ***w'***(r) выделим из множества ***Ex*** подмножество ***Ex***(r) всех формирующих ***w'***(r) примеров:

$$\boldsymbol{Ex}(r) = \{ Ex_{i1(j)}, Ex_{i2(j)}, \dots , Ex_{im(j)} \} .$$

Далее на всех элементах $\Omega_{L(r)} = \{ O_{i1(j)}, O_{i2(j)}, \dots , O_{im(j)} \}$, т.е. на левых частях этих примеров будем перечислять соответствующие ДСМ-сходства (замыкания Галуа). Для этого: для каждого порождаемого на элементах множества $\Omega_{L(r)}$ псевдо-дерева **D_GC**($\boldsymbol{U}_L$, $\boldsymbol{\Omega}_L(r)$) строим каркас $GK_{\mathbf{D},O}$ и для каждого его элемента проверяем, лежит ли он в порождаемом всеми примерами из Ω_L множестве ДСМ-сходств (т.е., сперва: лежит ли он на обоих каркасах, далее, - шаг за шагом по расширениям этих каркасов – возможно, вплоть до проверки его принадлежности одновременно обоим соответствующим псевдо-деревьям, – последовательно перебирая, если потребуется их «промежуточные» приближения). При этом для каждого элемента w_j текущего (анализируемого) покрытия для P^{τ}_r достаточно найти в обсуждаемом здесь процессе анализа хотя бы одну использующую его ДСМ-гипотезу. Любой набор таких гипотез (для каждого элемента покрытия – хотя бы по одной), очевидно, удовлетворяет условиям выполнимости ППВ-II.

При проверке попадания текущего анализируемого сходства v_j из текущей диаграммы **D_GC**($\boldsymbol{U}_L$, $\boldsymbol{\Omega}_L(r)$), построенной на левых частях отобранных в ***Ex***(r) из ***Ex*** примеров, в

диаграмму $\mathbf{D_GC}(\boldsymbol{U}_{\mathrm{L}},\boldsymbol{\Omega}_{\mathrm{L}})$, построенную на всех левых частях примеров из ***Ex***, возможны две ситуации:

a) орбиты обоих сходств совпали (искомый нами случай);
b) орбита в диаграмме $\mathbf{D_GC}(\boldsymbol{U}_{\mathrm{L}},\boldsymbol{\Omega}_{\mathrm{L}})$ больше, чем орбита в диаграмме $\mathbf{D_GC}(\boldsymbol{U}_{\mathrm{L}},\boldsymbol{\Omega}_{\mathrm{L}}(\mathrm{r}))$ (меньше она быть не может, т.к. ***Ex***(r) отбирают из ***Ex***). В этом случае следует проверить, как дополнительные (к выделенным в ***Ex***(r)) примеры из ***Ex*** сокращают (ведь их правые части – множества свойств, в отличие от правых частей примеров из ***Ex***(r), не входят в рассматриваемое *полезное* покрытие для P^{τ}_{r} - !) размеры текущего w_{j0}: удастся ли оставшимися зависимостями (более точно – множествами свойств w_j в их правых частях) все же «покрыть» удаляемые из w_{j0} (учетом дополнительных к ***Ex***(r) примеров из ***Ex***) образующие. В случае положительного ответа на этот вопрос – на текущем шаге перебора получено искомое решение, в противном случае следует переходить к следующему полезному покрытию для P^{τ}_{r}.

Основные особенности предлагаемой процедуры:

- размер $\boldsymbol{U}_{\mathrm{R}}$ - алфавита правых частей (анализируемых примеров и каузальных зависимостей) обычно существенно меньше, чем размер алфавита $\boldsymbol{U}_{\mathrm{L}}$ левых частей. Таким образом, характеризующая объемы перебора при поиске ДСМ-сходств «комбинаторика» правых частей, вообще говоря, менее «тяжела», чем «комбинаторика» анализа их левых частей. Как следствие:
- возникновение гиперкубов в псевдо-деревьях на правых частях исходных примеров и вновь порождаемых зависимостей есть существенно менее частое явление, чем возникновение аналогичных «тяжелых» комбинаторных объектов на левых частях изучаемых пар <*объект (подобъект), свойства*>;
- открывающиеся возможности сокращения «комбинаторики» перебора при работе с правыми частями определяют роль каркасов псевдо-деревьев (как инструмента быстрого порождения приближенных описаний соответствующих псевдо-деревьев): они (такие каркасы) на правых частях примеров и зависимостей более точно аппроксимируют соответствующие псевдо-деревья. Действительно, гиперкубов на правых частях (описаниях свойств) – существенно меньше, чем на левых (описаниях объектов). Да и размер их (если такие гиперкубы все же возникают), соответственно, существенно меньше, чем в случае операций с левыми частями;
- при переносе рассмотрения на левые части $=>_1$ (после выделения соответствующих покрытий и соотнесения их с каркасами и др. на правых частях) последовательный перебор по представленной схеме

  от примеров обучающей ДСМ-выборки - к каркасам соответствующих псевдо-деревьев, и далее (если потребуется) – ко все более и более точным приближениям соответствующих псевдо-деревьев и всей диаграммы ДСМ-сходств

  позволяет, стартовав с быстро порождаемых объектов – каркасов, - вести предлагаемое сравнение на все более точных приближениях для восстанавливаемых псевдо-деревьев так, чтобы завершить его (это сравнение) при первом же получении искомой причинной зависимости вида $v =>_2 w$.

Говоря о возможностях и ограничениях рассматриваемой техники управления перебором, дополнительно также следует заметить, что задействованное в формулировке **Утверждения 5.2.12** требование об отсутствии дубликатов в исходном множестве примеров Ω не накладывает каких-либо существенных ограничений на общность рассматриваемой здесь конструкции. Действительно, при наличии в Ω пары одинаковых примеров $O_1 = O = O_2$ достаточно удалить их из обработки и добавить к диаграмме $\mathbf{D_GC}_{\varphi,\mathrm{f}}(\boldsymbol{U},\boldsymbol{\Omega}\backslash\{O\})$ дополнительную вершину O, чтобы, сократив объем вычислений, необходимый в случае работы с исходным множеством примеров Ω, получить тот же самый финальный результат.

5.4. Корректность процедурной конструкции приближенного ДСМ-метода

Суммируем предложенные в предыдущих **Разделах 5.2-3** возможности управления перебором в рамках приближенного ДСМ-метода. Итак, в нашем распоряжении имеются:

- процедура разделения диаграммы взаимной вложимости неподвижных точек используемого нами замыкания Галуа на самостоятельно обрабатываемые под-диаграммы (так называемые псевдо-деревья);
- специальные (быстро формируемые) структурные конструкции - базис замыканий Галуа и каркас псевдо-дерева;
- процедуры идентификации и формирования (в анализируемых диаграммах и псевдо-деревьях) архитектурных фрагментов трех типов а также способы быстрой диагностики их позиционирования в восстанавливаемых псевдо-деревьях (при формировании каркасов таких псевдо-деревьев);
- технологии направленного восстановления гиперкубов (фрагментов типа β) на каркасах псевдо-деревьев;
- возможности оптимизации порождения ДСМ-зависимостей за счет «запуска» перебора сперва на правых частях[1] пар вида **ОБЪЕКТ=>СВОЙСТВА**;
- две главные «рабочие идеи»:

  (a) последовательный переход от анализа исходного множества примеров **Ω** к обработке его подмножеств (характеризующих требующие восстановления фрагменты типа β) и далее, направленным образом, к соответствующим подмножествам таких подмножеств, порождаемым рекурсивной процедурой последовательного восстановления всех необходимых фрагментов каждого псевдо-дерева по его каркасу (т.е. – последовательно «***сжимающая»*** исходно заданное множество примеров **Ω** до подмножеств все меньшего и меньшего размера ***рекурсивная процедура***, которая и позволяет за конечное число шагов *направленным образом* сформировать целевое псевдо-дерево);

  (b) последовательный переход от каркаса восстанавливаемого псевдо-дерева к «охватывающим» его диаграммам, получаемым последовательным присоединение к исходному каркасу (а затем – и к соответствующим его расширениям) тех вновь восстанавливаемых частей архитектурных фрагментов типа β, которые шаг за шагом порождаются уже представленными выше процедурами типа (a). (Таким образом формируется процедура, ***монотонно расширяющая*** структурные «дополнения» исходного каркаса и ***ограниченная сверху*** (в формировании таких расширений) конфигурацией восстанавливаемого по этому каркасу псевдо-дерева, которая собственно и позволяет за конечное число шагов *направленным образом* сформировать целевое псевдо-дерево).

Завершая описание предложенной нами процедурной конструкции приближенного ДСМ-метода представляется естественным убедиться в корректности предлагаемой схемы управления перебором и прежде всего - в том,

- что она действительно позволяет строить *приближенные* описания обсуждаемых диаграмм частичного порядка (и формирующих их псевдо-деревьев);
- что такие приближения можно *целенаправленно формировать* направленным образом (в частности, руководствуясь потребностями
  - скорейшего доопределения - прогнозирования свойств - новых, ранее не изученных, объектов,
  - проверки условий каузальной полноты для исходной выборки примеров и контрпримеров,

[1] Как правило, имеющих существенно более простую комбинаторную структур, чем описания левых частей (анализируемых объектов).

- скорейшего порождения зависимостей с заранее заданными «структурными» характеристиками – наличием образующих из определенного подмножества исходного алфавита **U**, вхождением в соответствующие им классы эквивалентности примеров из заданного подмножества исходного множества примеров **Ω** и т.п.),

- что, не будучи прерванной на промежуточных шагах ее исполнения, предлагаемая процедура приближенных ДСМ-вычислений в результате сформирует нам *всю целевую диаграмму* частичного порядка (со всеми образующими ее псевдо-деревьями).

Пусть заданы исходный алфавит **U** и множество объектов (примеров) **Ω**. Сформированную на них диаграмму взаимных вложений подмножеств образующих из **U**, характеризующих неподвижные точки замыкания[2] Галуа s^*g, будем обозначать как D = $\mathbf{D}_GC_{s^*g}(\mathbf{U},\Omega)$. Базисом замыканий Галуа s^*g мы будем (как и ранее) называть множество $B_GC_{s^*g}(\mathbf{U},\Omega)$ замыканий

$$B_GC_{s^*g}(\mathbf{U},\Omega) = \{s^*g\ (\{a_i\}) \mid i \in \{1,2, \dots ,n\}\}$$

одно-элементных подмножеств алфавита U= {a_1, a_2, ... , a_n} на множестве примеров **Ω**.

5.4.1. Определение

(1) *Каркасом* GK_D *ранга 1* (или просто - *каркасом*) мы будем называть диаграмму взаимной вложимости элементов Базиса замыканий Галуа s^*g .

(2) *Каркасом* $GK^{(2)}_D$ *ранга 2* мы будем называть диаграмму взаимной вложимости элементов множества замыканий s^*g всех двух-элементных подмножеств алфавита U, которое, в свою очередь, мы будем называть базисом замыканий Галуа ранга 2 и обозначать как

$$B^{(2)}_GC_{s^*g}(\mathbf{U},\Omega) = \{s^*g\ (\{a_{i_1},a_{i_2}\}) \mid i_1,i_2 \in \{1,2, \dots ,n\}\}.$$

…

(n) *Каркасом* $GK^{(n)}_D$ *ранга n* мы будем называть диаграмму, образованную одним элементом – замыканием s^*g (**U**) множества всех элементов алфавита **U**, который будем называть базисом замыканий Галуа ранга n и обозначать как $B^{(n)}_GC_{s^*g}(\mathbf{U},\Omega)$.

□

Взаимосвязи между структурой диаграммы $\mathbf{D}_GC_{s^*g}(\mathbf{U},\Omega)$ и множеством каркасов $GK^{(i)}_D$ всех возможных рангов, порождаемыми на паре <**U**,**Ω**> , характеризует

5.4.2. Утверждение (*Теорема о представлении*[3])

Для заданных алфавита **U** и множества примеров **Ω** диаграмма $\mathbf{D}_GC_{s^*g}(\mathbf{U},\Omega)$ представима (наследующей непосредственную вложимость в формирующих ее каркасах) объединенной диаграммой

$$\mathbf{D}_GC_{s^*g}(\mathbf{U},\Omega) = < (\bigcup_{i=1}^{n} B^{(i)}_GC_{s^*g}(\mathbf{U},\Omega)), \subseteq >,$$

сформированной[4] из каркасов $GK^{(i)}_{\mathbf{D}}$ замыканий всех возможных рангов i на алфавите **U**.

Доказательство.

=>. Пусть *O'* = { a_{j1}, a_{j2}, ... , a_{ji} } есть непустое подмножество элементов исходного алфавита **U**, представляющее некоторую вершину диаграммы $\mathbf{D}_GC_{s^*g}(\mathbf{U},\Omega)$. Рассмотрим каркас

---

[2] См. подробнее выше – **Раздел 5.3**.

[3] Диаграммы $\mathbf{D}_GC_{s^*g}(\mathbf{U},\Omega)$ объединением всех каркасов $GK^{(i)}_D$ по множеству всех допустимых для заданной пары <**U**,**Ω**> рангов i.

[4] С учетом используемого при формировании каркасов отношения *непосредственной вложимости* соответствующих множеств образующих из алфавита **U**, принадлежащих при этом каркасам разных рангов.

$GK^{(i)}_{\boldsymbol{D}}$ ранга $\boldsymbol{i}$, представляющий собою диаграмму взаимной вложимости определенных нами ранее замыканий Галуа $\boldsymbol{s}^*\boldsymbol{g}$ для всех содержащих ровно $\boldsymbol{i}$ элементов подмножеств исходного алфавита **U**. По определению $GK^{(i)}_{\mathbf{D}}$ содержит и замыкание

$$\boldsymbol{s}^*\boldsymbol{g}\ (\{\ a_{j1}, a_{j2}, \dots , a_{ji}\ \}) = \boldsymbol{s}(\boldsymbol{g}\ (\{\ a_{j1}, a_{j2}, \dots , a_{ji}\ \}))$$

множества $O' = \{\ a_{j1}, a_{j2}, \dots , a_{ji}\ \}$. Т.к. множество $O' = \{\ a_{j1}, a_{j2}, \dots , a_{ji}\ \}$ соответствует одной из вершин диаграммы $\mathbf{D_}\boldsymbol{GC}_{\boldsymbol{s}^*\boldsymbol{g}}(\mathbf{U},\boldsymbol{\Omega})$, то O' представляет собою неподвижную точку рассматриваемого замыкания Галуа, т.е.

$$O' = \{\ a_{j1}, a_{j2}, \dots , a_{ji}\ \} = \boldsymbol{s}^*\boldsymbol{g}\ (\{\ a_{j1}, a_{j2}, \dots , a_{ji}\ \}) = \boldsymbol{s}(\boldsymbol{g}\ (\{\ a_{j1}, a_{j2}, \dots , a_{ji}\ \})).$$

Таким образом, O' содержится в качестве вершины в $GK^{(i)}_{\boldsymbol{D}}$ и, как следствие – в объединении

$$\bigcup_{i=1}^{n} \boldsymbol{B}^{(i)}_\boldsymbol{GC}_{\boldsymbol{s}^*\boldsymbol{g}}(\mathbf{U},\boldsymbol{\Omega}).$$

<=. Пусть $O' = \{\ a_{j1}, a_{j2}, \dots , a_{ji}\ \}$ содержится в объединении каркасов всех возможных рангов, порожденных на заданной паре **U** и $\boldsymbol{\Omega}$. Предположим, что этого множества O' нет среди вершин диаграммы $\mathbf{D_}\boldsymbol{GC}_{\boldsymbol{s}^*\boldsymbol{g}}(\mathbf{U},\boldsymbol{\Omega})$, т.е. O' не является неподвижной точкой рассматриваемого нами замыкания Галуа, порождаемых на этой паре **U** и $\boldsymbol{\Omega}$. Тогда по *Теореме о представлении* составных замыканий Галуа через замыкания одноэлементных подмножеств исходного алфавита **U** (см. **Утверждение 5.2.3**) замыкание

$$\boldsymbol{s}^*\boldsymbol{g}\ (\{\ a_{j1}, a_{j2}, \dots , a_{ji}\ \}) = \boldsymbol{s}(\boldsymbol{g}\ (\{\ a_{j1}, a_{j2}, \dots , a_{ji}\ \}))$$

содержит по крайней мере одну дополнительную образующую $a_{\mathbf{0}}$ из текущего алфавита **U**, которая входит в замыкание

$$\boldsymbol{s}^*\boldsymbol{g}\ (\{\ a_{j0}\ \}) = \boldsymbol{s}(\boldsymbol{g}\ (\{\ a_{j0}\ \}))$$

по крайней мере одного из одноэлементных подмножеств $\{a_{j0}\}$ множества O':

$$a_{j0} \in \{\ a_{j1}, a_{j2}, \dots , a_{ji}\ \} = O'$$

и

$$\{\ a_{j0}, a_0\} \subseteq \boldsymbol{s}^*\boldsymbol{g}\ (\{\ a_{j0}\ \}) = \boldsymbol{s}(\boldsymbol{g}\ (\{\ a_{j0}\ \})).$$

В таком случае образующая a_{j0} из O' входит в (порождаемые *замкнутыми* множествами образующих из **U**) вершины каркасов $GK^{(i)}_{\boldsymbol{D}}$ (для всех возможных $\boldsymbol{i} \in \{1,2, \dots ,n\}$) *только вместе* с образующей $a_{\mathbf{0}}$. Таким образом, наше исходное предположение о принадлежности множества образующих $O' = \{\ a_{j1}, a_{j2}, \dots , a_{ji}\ \}$ множеству вершин объединения каркасов $GK^{(i)}_{\boldsymbol{D}}$ по всем возможных $\boldsymbol{i} \in \{1,2, \dots ,n\}$ оказывается неверным.

□

Таким образом, представление диаграммы $\boldsymbol{D} = \mathbf{D_}\boldsymbol{GC}_{\boldsymbol{s}^*\boldsymbol{g}}(\mathbf{U},\boldsymbol{\Omega})$ объединением каркасов вида $GK^{(i)}_{\boldsymbol{D}}$ обеспечивает возможности приближенного ее описания:

- ограничиваясь тем или иным набором *первых членов в разложении* диаграммы $\boldsymbol{D}$ в объединение каркасов $GK^{(i)}_{\boldsymbol{D}}$ рангов i для $i \in \{1,2,\dots, i_e\}$ (где $i_e < n$), мы будем получать *приближенное* описание для $\boldsymbol{D} = \mathbf{D_}\boldsymbol{GC}_{\boldsymbol{s}^*\boldsymbol{g}}(\mathbf{U},\boldsymbol{\Omega})$,
- выбрав после построения каркаса $GK_{\boldsymbol{D}} = GK^{(1)}_{\boldsymbol{D}}$ для первоочередного восстановления определенный архитектурный фрагмент типа β («накладываемый» на булевский гиперкуб $\mathbf{HC}_{\boldsymbol{j}}$ уже диагностированной при формировании каркаса $GK_{\boldsymbol{D}}$ размерности $\boldsymbol{j}$), далее можно оперировать лишь теми образующими из исходного алфавита **U**, которые «участвуют» в формировании именно этого архитектурного фрагмента типа β; причем – оперировать на соответствующем *подмножестве* исходного множества образующих из **U** теми же конструкциями каркасов различных рангов[5] (однако, сформированными уже

[5] А также и другими представленными в **Разделах 5.2-3** процедурами сокращения перебора.

лишь на соответствующем подмножестве алфавита **U**). Таким путем, переходя от одного архитектурного фрагмента типа β к другому, а также регулируя «детальность»[6] восстановления каждого соответствующего фрагмента типа β, можно выстраивать те или иные *целенаправленно формируемые* приближенные описания целевой диаграммы ***D***.

Именно так и могут быть продемонстрированы имеющиеся возможности управления перебором (путем выбора и последующего приоритетного восстановления необходимых архитектурных фрагментов типа β) и сокращения его объемов (за счет работы лишь с некоторыми – целенаправленно выделяемыми[7] - архитектурными фрагментами типа β).

На **Примере 5.4.3** представлена иллюстрация обсуждаемой схемы управления перебором в приближенном ДСМ-методе.

5.4.3. Пример.

Пусть $\boldsymbol{U} = \{a_1, a_2, \ldots, a_{28}\}$, $\boldsymbol{\Omega} = \{ h_1, h_2, \ldots, h_{18}\}$ и

$\mathbf{h_1} = \{a_1, a_2, a_3, a_5, a_{10}, a_{11}\}$, $\mathbf{h_2} = \{ a_1, a_2, a_4, a_{10}, a_{11}, a_{13}\}$,
$\mathbf{h_3} = \{ a_1, a_3, a_4, a_{10}, a_{11}, a_{13}\}$, $\mathbf{h_4} = \{ a_1, a_2, a_3, a_4, a_5, a_9, a_{10}, a_{11}, a_{13}, a_{26}\}$,
$\mathbf{h_5} = \{ a_2, a_3, a_4, a_5, a_6, a_{10}, a_{11}, a_{13}, a_{26}, a_{27}\}$,
$\mathbf{h_6} = \{ a_2, a_3, a_4, a_5, a_7, a_{10}, a_{11}, a_{13}, a_{26}\}$,
$\mathbf{h_7} = \{ a_{10}, a_{11} \}$, $\mathbf{h_8} = \{ a_{11}, a_{12}\}$, $\mathbf{h_9} = \{ a_{14}, a_{15}\}$, $\mathbf{h_{10}} = \{ a_{14}, a_{16}, a_{17}\}$,
$\mathbf{h_{11}} = \{ a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{18}, a_{19}, a_{26}\}$,
$\mathbf{h_{12}} = \{ a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{18}, a_{20}, a_{26}\}$,
$\mathbf{h_{13}} = \{ a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{22}, a_{26}\}$,
$\mathbf{h_{14}} = \{ a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{21}, a_{26}\}$,
$\mathbf{h_{15}} = \{ a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{23}, a_{26}\}$,
$\mathbf{h_{16}} = \{ a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{21}, a_{22}, a_{23}, a_{24}, a_{26}\}$,
$\mathbf{h_{17}} = \{ a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{21}, a_{22}, a_{23}, a_{25}, a_{26}\}$,
$\mathbf{h_{18}} = \{ a_2, a_3, a_4, a_5, a_6, a_{10}, a_{11}, a_{13}, a_{26}, a_{28}\}$.

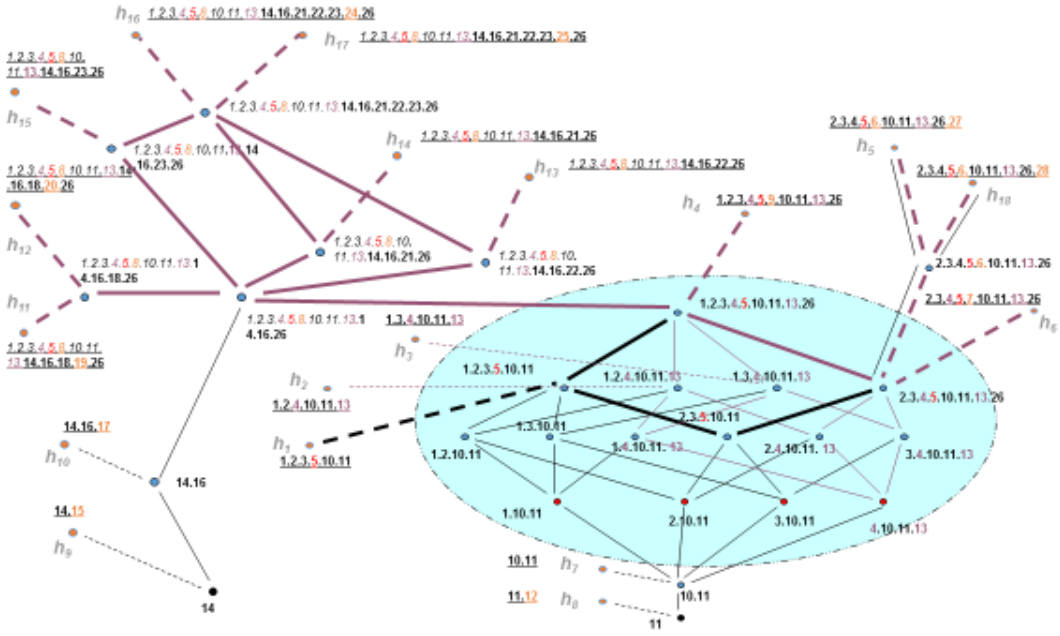


Рис. 5.4.1. Диаграмма $\boldsymbol{D} = \mathbf{D_GC}_{s*g}(\mathbf{U},\boldsymbol{\Omega})$ для пары <**U**,**Ω**> из **Примера 5.4.3**.

---

[6] Например, ограничиваясь определенным набором первых членов в представлении целевой диаграммы ***D*** объединением каркасов соответствующих рангов.

[7] С учетом уже называвшихся выше приоритетов - доопределения конкретных новых объектов, проверки специальных условий корректности реализуемого в данном случае ДСМ-рассуждения и т.п. .

Диаграмма $\mathbf{D_GC}_{\varphi,f}(\boldsymbol{U}, \boldsymbol{\Omega})$ в данном случае имеет представленный на Рис.5.4.1 вид.

На Рис. 5.4.2 – пример входящего в эту диаграмму псевдо-дерева (которое содержит в корне объект h_{17} из исходного множества $\boldsymbol{\Omega}$):

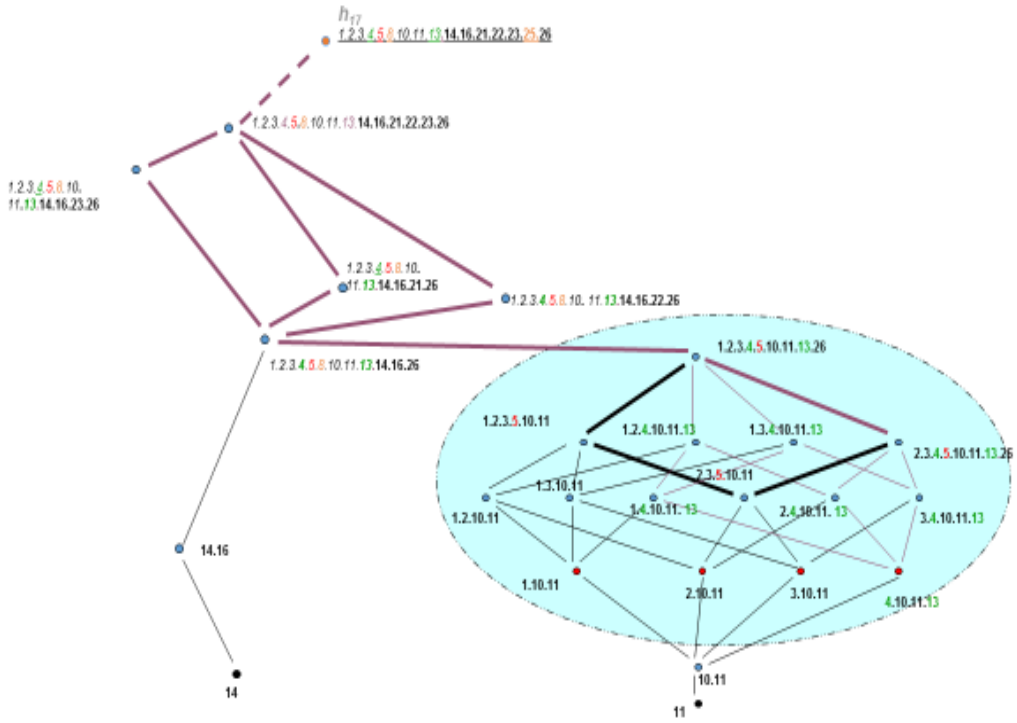


Рис. 5.4.2. Псевдо-дерево из диаграммы $\boldsymbol{D}$ с объектом h_{17} в корне.

Наконец, на Рис. 5.4.3 выделенная темным фоном в овале область – пример подструктуры (фрагмента) псевдо-дерева $\mathbf{D_}\boldsymbol{GC}_{\boldsymbol{s}*\boldsymbol{g}}(\mathbf{U},\boldsymbol{\Omega})\big|_{\{h_{17}\}}$ с Рис.5.4.2, имеющий архитектуру типа β:

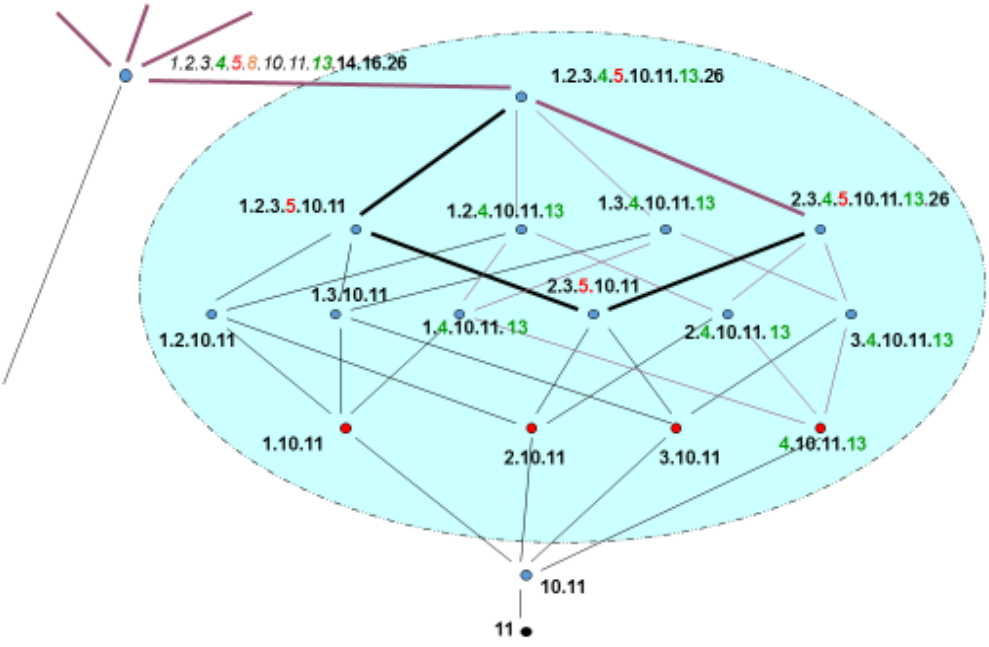


Рис. 5.4.3. Фрагмент архитектуры типа β в псевдо-дереве с Рис. 5.4.2.

При восстановлении псевдо_дерева $\mathbf{D_GC}_{s^*g}(\mathbf{U},\boldsymbol{\Omega})\big|_{\{h_{17}\}}$ процесс вычислений начинается с построения $\boldsymbol{B_GC}_{s^*g}(\mathbf{U},\boldsymbol{\Omega})\big|_{\{h_{17}\}}$ - сопоставляемого его корню (т.е. $\mathbf{h_{17}} = \{ a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{21}, a_{22}, a_{23}, a_{25}, a_{26}\}$) фрагмента соответствующего *базиса замыканий Галуа* $\boldsymbol{B_GC}_{s^*g}(\mathbf{U},\boldsymbol{\Omega})$:

$[\{a_1\}]_{(U,\Omega)} = s^*g(\{a_1\}) = \{a_1, a_{10}, a_{11}\}$
$[\{a_2\}]_{(U,\Omega)} = s^*g(\{a_2\}) = \{a_2, a_{10}, a_{11}\}$
$[\{a_3\}]_{(U,\Omega)} = s^*g(\{a_3\}) = \{a_3, a_{10}, a_{11}\}$
$[\{a_4\}]_{(U,\Omega)} = s^*g(\{a_4\}) = \{a_4, a_{10}, a_{11}, a_{13}\}$
$[\{a_5\}]_{(U,\Omega)} = s^*g(\{a_5\}) = \{a_2, a_3, a_5, a_{10}, a_{11}\}$
$[\{a_8\}]_{(U,\Omega)} = s^*g(\{a_8\}) = \{a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{26}\}$
$[\{a_{10}\}]_{(U,\Omega)} = = s^*g(\{a_{10}\}) = \{a_{10}, a_{11}\}$
$[\{a_{11}\}]_{(U,\Omega)} = s^*g(\{a_{11}\}) = \{a_{11}\}$
$[\{a_{13}\}]_{(U,\Omega)} = s^*g(\{a_{13}\}) = \{a_4, a_{10}, a_{11}, a_{13}\}$
$[\{a_{14}\}]_{(U,\Omega)} = s^*g(\{a_{14}\}) = \{a_{14}\}$
$[\{a_{16}\}]_{(U,\Omega)} = s^*g(\{a_{16}\}) = \{a_{14}, a_{16}\}$
$[\{a_{21}\}]_{(U,\Omega)} = s^*g(\{a_{21}\}) = \{a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{21}, a_{26}\}$
$[\{a_{22}\}]_{(U,\Omega)} = s^*g(\{a_{22}\}) = \{a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{22}, a_{26}\}$
$[\{a_{23}\}]_{(U,\Omega)} = s^*g(\{a_{23}\}) = \{a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{23}, a_{26}\}$
$[\{a_{25}\}]_{(U,\Omega)} = s^*g(\{a_{25}\}) = \{a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{21}, a_{22}, a_{23}, a_{25}, a_{26}\}$
$[\{a_{26}\}]_{(U,\Omega)} = s^*g(\{a_{26}\}) = \{a_2, a_3, a_4, a_5, a_{10}, a_{11}, a_{13}, a_{26}\}$

Далее, анализом взаимной вложимости составляющих $\boldsymbol{B_GC}_{s^*g}(\mathbf{U},\boldsymbol{\Omega})\big|_{\{h_{17}\}}$ множеств образующих из алфавита $\mathbf{U}$, несложно убедиться, что

$$[\{a_{11}\}]_{(U,\Omega)} = \{a_{11}\} \subset$$
$$\subset \{a_{10}, a_{11}\} = [\{a_{10}\}]_{(U,\Omega)} \subset$$
$$\subset \{a_1, a_{10}, a_{11}\} = [\{a_1\}]_{(U,\Omega)} \subset$$
$$\subset [\{a_8\}]_{(U,\Omega)} = \{ a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{26}\}$$

и

$$[\{a_{11}\}]_{(U,\Omega)} = \{a_{11}\} \subset$$
$$\subset \{a_{10}, a_{11}\} = [\{a_{10}\}]_{(U,\Omega)} \subset$$
$$\subset \{a_2, a_{10}, a_{11}\} = [\{a_2\}]_{(U,\Omega)} \subset$$
$$\subset [\{a_8\}]_{(U,\Omega)} = \{ a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{26}\},$$

а также

$$[\{a_{11}\}]_{(U,\Omega)} = \{a_{11}\} \subset$$
$$\subset \{a_{10}, a_{11}\} = [\{a_{10}\}]_{(U,\Omega)} \subset$$
$$\subset \{a_3, a_{10}, a_{11}\} = [\{a_3\}]_{(U,\Omega)} \subset$$
$$\subset [\{a_8\}]_{(U,\Omega)} = \{ a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{26}\}$$

и, наконец,

$$[\{a_{11}\}]_{(U,\Omega)} = \{a_{11}\} \subset$$
$$\subset \{a_{10}, a_{11}\} = [\{a_{10}\}]_{(U,\Omega)} \subset$$
$$\subset \{a_4, a_{10}, a_{11}, a_{13}\} = [\{a_4\}]_{(U,\Omega)} \subset$$
$$\subset [\{a_8\}]_{(U,\Omega)} = \{ a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{26}\}$$

вместе с

$$[\{a_{11}\}]_{(U,\Omega)} = \{a_{11}\} \subset$$
$$\subset \{a_{10}, a_{11}\} = [\{a_{10}\}]_{(U,\Omega)} \subset$$
$$\subset \{a_4, a_{10}, a_{11}, a_{13}\} = [\{a_{13}\}]_{(U,\Omega)} \subset$$
$$\subset [\{a_8\}]_{(U,\Omega)} = \{ a_1, a_2, a_3, a_4, a_5, a_8, a_{10}, a_{11}, a_{13}, a_{14}, a_{16}, a_{26}\}.$$

Принимая во внимание

$$\mathbf{[\{}a_4\mathbf{\}]}_{(U,\Omega)} = \{a_4, a_{10}, a_{11}, a_{13}\} = \mathbf{[\{}a_{13}\mathbf{\}]}_{(U,\Omega)},$$

заключаем, что a_4 и a_{13} - так называемые «склеенные» образующие (см. выше – **Раздел 5.2**) и для удобства последующих расчетов их совместные вхождения в диаграмму и ее фрагменты можно рассматривать как одну «новую образующую» $a_{4,13}$. Кроме того, одновременная вложимость $\mathbf{[\{}a_1\mathbf{\}]}_{(U,\Omega)}$, $\mathbf{[\{}a_2\mathbf{\}]}_{(U,\Omega)}$, $\mathbf{[\{}a_3\mathbf{\}]}_{(U,\Omega)}$ и $\mathbf{[\{}a_{4,13}\mathbf{\}]}_{(U,\Omega)}$ в $\mathbf{[\{}a_8\mathbf{\}]}_{(U,\Omega)}$ указывают нам на наличие в $\boldsymbol{B_GC_{s*g}}(\mathbf{U},\boldsymbol{\Omega})\big|_{\{h_{17}\}}$ архитектурного фрагмента типа β (по крайней мере – части булевского гиперкуба $\mathbf{HC_4}$ размерности 4), т.к. в $\mathbf{[\{}a_8\mathbf{\}]}_{(U,\Omega)}$ могут быть также вложены по-парные, по-троечные и т.д. объединения соответствующих замыканиям $\mathbf{[\{}a_1\mathbf{\}]}_{(U,\Omega)}$, $\mathbf{[\{}a_2\mathbf{\}]}_{(U,\Omega)}$, $\mathbf{[\{}a_3\mathbf{\}]}_{(U,\Omega)}$ и $\mathbf{[\{}a_{4,13}\mathbf{\}]}_{(U,\Omega)}$ множеств образующих из исходного алфавита **U**. (При этом, наличие в $\boldsymbol{D_GC_{s*g}}(\mathbf{U},\boldsymbol{\Omega})\big|_{\{h_{17}\}}$ соответствующих вершин будет, как было показано выше в **Разделе 5.2**, определяться наличием в **Ω** необходимого подмножества так называемых «раскрывающихся» в этот гиперкуб $\mathbf{HC_4}$ примеров).

В такой ситуации, выбрав в качестве приоритетной цели для восстановления ту часть псевдо_дерева $\boldsymbol{D_GC_{s*g}}(\mathbf{U},\boldsymbol{\Omega})\big|_{\{h_{17}\}}$, которая «наложена» на идентифицированный нами только что булевский гиперкуб $\mathbf{HC_4}$, нам следует обратиться к формированию каркаса ранга 2 (а далее – ранга 3 и т.п.) на образующих из подмножества $\{a_1, a_2, a_3, a_{4,13}\}$[8] исходного алфавита **U**.

Таким образом, на первом шаге приближений (т.е. - при использовании[9] лишь каркасов ранга 2) будут сформированы новые (расширяющие каркас $GK_D = GK^{(1)}_D$) и непосредственно доминирующие соответствующие множества из четверки $\mathbf{[\{}a_1\mathbf{\}]}_{(U,\Omega)}$, $\mathbf{[\{}a_2\mathbf{\}]}_{(U,\Omega)}$, $\mathbf{[\{}a_3\mathbf{\}]}_{(U,\Omega)}$ и $\mathbf{[\{}a_{4,13}\mathbf{\}]}_{(U,\Omega)}$ вершины $\{a_1, a_2, a_{10}, a_{11}\}$, $\{a_1, a_3, a_{10}, a_{11}\}$, $\{a_1, a_{4,13}, a_{10}, a_{11}\}$, $\{a_2, a_3, \mathbf{a_5}, a_{10}, a_{11}\}$, $\{a_2, a_{4,13}, a_{10}, a_{11}\}$, $\{a_3, a_{4,13}, a_{10}\}$. (При этом, следуя ранее предложенному в **Разделе 5.2** порядку классификации образующих из исходного алфавита **U**, $\mathbf{a_5}$ есть «наклеенная[10]» на пару $<a_2,a_3>$ образующая).

Затем (при необходимости) переходом к каркасам ранга 3 и далее в процессе формирования (последовательными *монотонно расширяющими* дополнениями от соответствующего каркаса GK_D ко всей восстанавливаемой поддиаграмме для **D**) будут порождаться следующие приближенные структурные описания текущего псевдо-дерева $\boldsymbol{D_GC_{s*g}}(\mathbf{U},\boldsymbol{\Omega})\big|_{\{h_{17}\}}$. Рассматриваемый здесь процесс, не будучи остановлен на каком-либо промежуточном шаге, восстановит весь «наложенный» на выбранный гиперкуб $\mathbf{HC_4}$ фрагмент псевдо-дерева $\boldsymbol{D_GC_{s*g}}(\mathbf{U},\boldsymbol{\Omega})\big|_{\{h_{17}\}}$. Очевидно, что таким образом, шаг за шагом, можно целенаправленно строить приближенные описания всей диаграммы $\boldsymbol{D_GC_{s*g}}(\mathbf{U},\boldsymbol{\Omega})$.

Более того, приближенные описания такого типа для каждого из формирующих диаграмму $\boldsymbol{D_GC_{s*g}}(\mathbf{U},\boldsymbol{\Omega})$ псевдо-деревьев $\boldsymbol{D_GC_{s*g}}(\mathbf{U},\boldsymbol{\Omega})\big|_{\{h_i\}}$ можно строить независимо одно от другого, в том числе – в режиме *параллельных* вычислений. Соответствующие алгоритмические и программные построения характеризуют *параллельную* версию *приближенного* ДСМ-метода.

□

Таким образом, нами представлены дополнительные средства целенаправленного управления перебором в (том числе – в приближенном) ДСМ-методе:

- декомпозиция исходной задачи (о реконструкции целевой диаграммы частичного порядка) на независимо друг от друга решаемые подзадачи восстановления псевдо-деревьев, формирующих исходную диаграмму;

а также вытекающие из этого

---

[8] Разумеется – с учетом выполненной нами « склейки» образующих a_4 и a_{13}.

[9] Заметим – *только* на восстанавливаемом гиперкубе $\mathbf{HC_4}$ (а *не на всем* псевдо-дереве $\boldsymbol{D_GC_{s*g}}(\mathbf{U},\boldsymbol{\Omega})\big|_{\{h_{17}\}}$)!

[10] Встречающаяся в $\boldsymbol{D_GC_{s*g}}(\mathbf{U},\boldsymbol{\Omega})\big|_{\{h_{17}\}}$) *только вместе* с ними.

- очевидные возможности использования параллельных вычислений. (Один из перспективных вариантов организации таких вычислений – работа в *облачных средах* типа SDI[11])

Возможности использования развитой для теоретико-множественного случая техники приближенного ДСМ-метода при работе с другими типами данных можно суммировать к следующим двум группам процедур:

- сперва структурные описания исходных объектов (примеров и контрпримеров), формируемые в языке типа *множества+отношения на них* (например, графы как пары множеств – *вершин* и *ребер* (т.е. пар вершин), цепочки как множества символов, где игнорируется порядок их повторных вхождений, и т.п.), можно приближенным[12] образом «обсчитать» в теоретико-множественном представлении,

а далее

- полученные такими приближенными вычислениями сужения исходно обрабатываемых выборок примеров и контрпримеров – удобная возможность сократить объемы вычислений, требуемых для последующей обработки таких подвыборок *точными* (т.е учитывающими все особенности используемого структурного описания) средствами.

Приближенные вычисления при работе с данными, которые представлены числовыми векторами значений параметров могут быть организованы на базе анализа трехстолбцовых миноров – подматриц исходно заданной матрицы-универсума.

* * *

Подведем итоги предпринятого нами рассмотрения. Прежде всего отметим, что представленная выше техника декомпозиции множества всех ДСМ-зависимостей, порождаемых на заданных алфавите $\boldsymbol{U}$ и множестве примеров $\boldsymbol{\Omega}$, дает возможности последовательно формировать (быстрой процедурой - !) каркасы соответствующих деревьев отдельно для каждого - 1-го, 2-го, …, последнего объектов из исходного множества примеров $\boldsymbol{\Omega}$, выбирая (надежно управляемым образом - !) на каждом из каркасов те веточки частичного порядка и те конкретные ДСМ-зависимости (с учетом входящих в них образующих, с учетом порождающих их примеров – «родителей», …), которые являются «интересными» в смысле тех или иных «внешних» (по отношению к организации процедуры перебора вариантов, в том числе – содержательных, отражающих специфику предметного знания об исследуемой средствами ДСМ-метода прикладной предметной области) соображений. (Например, это могут быть ДСМ-зависимости, которые непременно содержат образующие из некоторого специально выделенного подмножества, или же зависимости, удовлетворяющие заданным ограничениям на число входящих в них образующих, и т.п.).

Не менее интересные возможности предлагаемая техника декомпозиции предоставляет в части проверки выполнимости весьма важного в ДСМ-рассуждении условия так называемой каузальной полноты (которое предполагает для всех элементов заданного конкретного множества примеров наличие ДСМ-«объяснения» порождаемыми из него зависимостями,). В частности, здесь появляются возможности для *ускоренной* проверки выполнимости условия каузальной полноты, причем, если потребуется, - это можно сделать ДСМ-зависимостями необходимого формата (например, - заданного размера, или же – содержащими заданные образующие и т.п.). Аналогичным образом может быть

---

[11] Software Defined Infrastructure - см., например, [*SDI, SDN&NFV-2013, SDN&OF-2014*] и др. .

[12] Где, вообще говоря, могут порождаться в том числе и неинтерпретируемые в исходном структурном описании результаты (например, порожденные в теоретико-множественном описании зависимости типа <*подмножество вершин, подмножество ребер*> могут оказаться рассогласованными: в выделенном *подмножестве ребер* могут оказаться встречающимися также и вершины, которые оказались *невыделенными* в соответствующее *подмножество вершин*).

организована ускоренная (т.е. не требующая порождения всех возможных в каждом рассматриваемом случае ДСМ-зависимостей) проверка ДСМ-доопределимости заданного нового объекта (примера). Таким образом, можно организовать *приближенные* (однако, снабженные надежным механизмом *управления* порождением рассматриваемых приближений) ДСМ-рассуждения: в первую очередь порождать те или иные «полезные» результаты за счет первоочередного формирования «полезных» ДСМ-зависимостей. Далее, перебрав последовательно все ветки частичного порядка во всех рассматриваемых деревьях частичного порядка, (порождаемых в свою очередь в результате декомпозиции описанными выше средствами всего множества ДСМ-зависимостей на отдельные подструктуры), сформировать «предельный случай» такого сорта приближений – целиком *все множество* ДСМ-зависимостей.

Наконец, предоставляемая обсуждаемым подходом возможность анализировать отдельно каждое из соответствующих деревьев рассмотренного выше типа, имеющих в корне примеры из исходной выборки **Ω**, позволяет организовать *параллельную обработку* имеющихся ДСМ-данных.

Итак, предлагаемая техника управления перебором (техника *ПД+Каркасы*) позволяет организовать управляемую «навигацию» в множествах соответствующих ДСМ-сходств. Таким образом появляется возможность

- сперва, искать «полезные» (в уже оговоренном выше смысле – т.е. ДСМ-сходства, содержащие, например, некоторые наперед заданные образующие и т.п.) результаты, а далее, если потребуется,
- последовательно достраивать их множество до множества всех соответствующих результатов (простых ДСМ-сходств; гипотез, удовлетворяющих заданной системе дополнительных ограничений ***ЗКП*** и др.; гипотез, полезных для проверки ДСМ-доопределимости элементов заданного множества τ-примеров или же проверки выполнимости Условия Каузальной Полноты).

При этом в каждом из представленных выше направлений организации перебора:

- при поиске простых сходств,
- при анализе дополнительных логических условий (ЗКП, единственность сходства и др.) а также
- при проверке доопределимости τ-примеров и\или проверке Условия Каузальной Полноты

общая «стратегия» последовательного восстановления соответствующих псевдо-деревьев по их каркасам (с порождением в первую очередь «полезных» ДСМ-гипотез) оказывается эффективным инструментом оптимизации объемов необходимых вычислений.

Дополнительные возможности для обсуждаемой оптимизации дает распараллеливание реализуемого ДСМ-поиска: как уже было показано выше, каждое из псевдо-деревьев, составляющих восстанавливаемую диаграмму сходств, может быть «обсчитываемо» в независимом от других оставшихся псевдо-деревьев режиме. Более того, каждый из соответствующих каркасов (а также его последовательные расширения до соответствующего псевдо-дерева) может также быть восстанавливаем независимо от других (т.е., в том числе - и в параллельном режиме вычислений).

Одним из вариантов организации подобного распараллеленного «счета» может быть, например, следующий:

- на имеющемся «физическом» системно-техническом ландшафте (вычислительных мощностях, коммуникационном оборудовании и устройствах хранения данных) формируем виртуальную (логическую) «облачную» вычислительную среду;
- в сформированной логической среде, исходя из размеров исходного множества примеров для ДСМ-обучения и уточнения понятие «полезные» зависимости а также выбора стратегии порождения ДСМ-результатов (например, восстановления лишь «полезных» зависимостей, или же восстановления сперва лишь «полезных», а затем

достраивания их множества до полного множества ДСМ-зависимостей и т.п.), «поднимаем» каждый раз *необходимое* (т.е. определяемое заданными параметрами исходных данных для ДСМ-обучения а также текущими комбинаторными характеристиками анализируемых псевдо-деревьев - их архитектурой, глубиной рекурсии при раскрытии встречающихся в них гиперкубов и т.п.) число параллельно работающих копий «типовой» виртуальной машины (исполнимого «типового» программного модуля ДСМ-анализа каждого такого дерева), реализующей алгоритмику ДСМ-вычислений по представленной выше схеме *ПД+Каркасы* (обеспечивая при этом существенное ускорение ДСМ-вычислений за счет параллельной обработки всех псевдо-деревьев рассмотренного нами выше типа);

- при этом при вычислении базиса замыканий Галуа для рассматриваемых подмножеств образующих определенную пользу (в части сокращения времени вычислений) может оказать использование специализированных аппаратно-схемных технологических комплексов (например, TCAM-решений[13] и т.п.).

Наконец, отметим, что предложенная техника управления перебором при порождении ДСМ-зависимостей – возможный вариант решения проблемы эффективной применимости ДСМ-метода при работе с большими выборками исходных данных. Действительно: описанные нами возможности

- декомпозиции исходной задачи на подзадачи, дополненные очерченными выше перспективами
- организации параллельных вычислений при компьютерной реализации ДСМ-рассуждений, а также возможности
- вести приближенные ДСМ-вычисления (при которых целенаправленным образом порождается лишь некоторое «полезное» подмножество ДСМ-гипотез)

позволяют надеяться при решении прикладных задач на *существенное расширение* сложившихся к настоящему времени рамок использования техники интеллектуального анализа данных средствами ДСМ-метода.

Итак, при формировании детального описания диаграммы $\mathbf{D_}\boldsymbol{GC}_{s*g}(\mathbf{U},\boldsymbol{\Omega})$ в ряде случаев идентифицируются *классы переборных задач*, которые для различных типов данных (т.е. для различных способов описания исходной обучающей выборки прецедентов) *доказуемо* принадлежат к некоторым известным классам *трудно-разрешимых комбинаторных проблем* (в частности – к хорошо известным классам ***NPC*** и ***#PC***, соответственно, ***NP***–полных и перечислительно полных комбинаторных проблем). Таким образом, при ориентации на реальные приложения развиваемого подхода приходится столкнуться с актуальной экспоненциальной сложностью реализуемых вычислений и быстро растущими объемами перебора при поиске решений.

Принимая во внимание *неустранимость* в общем случае экспоненциально быстро растущих оценок сложности вычислений, представляется естественным искать такие *приближенные описания* диаграмм $\mathbf{D_}\boldsymbol{GC}_{s*g}(\mathbf{U},\boldsymbol{\Omega})$, которые позволяли бы решать в каждом конкретном случае задачу прогнозирования свойств новых прецедентов (элементов множества $\boldsymbol{\Omega}^{\tau}$). Для этого организуется *целенаправленное* (*управляемое*) порождение релевантных цели прогноза (т.е. цели формализованного ДСМ-рассуждения) вершин формируемой диаграммы $\mathbf{D_}\boldsymbol{GC}_{s*g}$ и делается это средствами специально разработанного проблемно-ориентированного *метода последовательных приближений*.

Процедуры управления целенаправленностью реализуемого поиска эмпирических зависимостей в исходных данных опираются на *мета-характеристики* анализируемого объекта $\mathbf{D_}\boldsymbol{GC}_{s*g}(\mathbf{U},\boldsymbol{\Omega})$ и формирующих его алгоритмов. Так в ситуации, когда для заданных исходной выборки прецедентов $\boldsymbol{\Omega}$ и операции сходства $\otimes$ оказывается, что порождаемых

[13] TCAM (ternary content addressable memory) – специализированные схемные решения реализации ассоциативной памяти.

сходств (а с ними – соответствующих эмпирических зависимостей) «*слишком много*», представляется естественным задаться вопросом: «хорошо ли это»? (Ведь экспоненциально быстрый рост числа зависимостей, формируемых из текущей выборки, сигнализирует о возможностях комбинирования любых входящих в описания прецедентов характеристик с любыми, что вряд ил интересно с точки зрения поиска практически значимых для решения прикладных задач зависимостей). Это позволяет сформировать *мета-средства управления ДСМ-рассуждением*.

Дополнительно в подобных случаях представляется естественным задуматься об адекватности используемого инструментария анализа данных «природе» скрытых в этих данных эмпирических зависимостей, в частности проверить, может быть

- исходная обучающая выборка **Ω** – «плохая»? (Например, не содержит «значимых[14]» прецедентов), или же
- используемый язык описания прецедентов неадекватен поиску характерных для анализируемой предметной области эмпирических зависимостей?

Для идентификации подобных (комбинаторно «плохих») ситуаций представляется уместным искать удобные средства диагностики «качества» исходной обучающей выборки **Ω**, которые позволяли бы обойтись без порождения множества всех скрытых в ней эмпирических зависимостей при принятии решения о пополнении исходной выборки дополнительными («информативными» прецедентами) или же о модификации используемого языка описания прецедентов (в частности, модификации операции сходства ⊗ и др.). Выше в **Главе 5** представлены «диагностические» инструменты «быстрого» управления поиском зависимостей, позволяющие полиномиально сложными процедурами диагностировать наличие в формируемой диаграмме $\mathbf{D_GC}_{s*g}(\mathbf{U},\mathbf{\Omega})$ экспоненциально сложных фрагментов типа β. Тем самым появляется возможность как целенаправленно обходить ряд комбинаторно «плохих» переборных конфигураций, оперируя с другими фрагментами текущей диаграммы **D**, так и (при необходимости) переходить к расширениям исходной выборки **Ω**, пополняя ее новыми прецедентами и фактически оперируя направленно формируемыми *последовательностями расширяющихся обучающих выборок*.

Переход к операциям с последовательностями расширяющихся обучающих выборок имеет принципиально важной значение с точки зрения выделения так называемых *эмпирических закономерностей* – таких эмпирических зависимостей, которые сохраняют типы своих истинностных значений при расширениях исходной обучающей выборки. (Эмпирические закономерности имеют две существенные характеристики: каждая из них не только остается выводимой средствами ДСМ-рассуждения при расширении исходной выборки **Ω** прецедентов, но и сохраняет тип приписываемого ей истинностного значения – эмпирическую истину, эмпирическую ложь или эмпирическое противоречие).

Фокусировка на поиске именно эмпирических закономерностей дает дополнительные возможности управления комбинаторным перебором в рассматриваемом нами случае: поиск сходств, которые были бы инвариантны (т.е. порождались бы также и) в рассматриваемых направленных[15] расширениях исходной выборки **Ω**, позволяет существенным образом сузить рамки реализуемого при их формировании перебора (по сравнению с анализом лишь исходной обучающей выборки прецедентов **Ω**).

---

[14] Позволяющих «сузить» комбинаторику вариантов до значимых зависимостей.

[15] В реальных прикладных исследованиях такие расширения отражают историю процесса включения новых экспериментальных данных в выполняемое исследование (и, как правило, не носят случайного характера).

Глава 6

Некоторые функциональные особенности ДСМ-ИАД

6.1. Эффекты немонотонности ДСМ-рассуждений.

Обсуждение наиболее интересных функциональных особенностей ДСМ-ИАД мы начнем с рассмотрения эффектов немонотонности ДСМ-рассуждений. Традиционные формальные системы, изучающие проблемы выводимости формул из фиксированных множеств аксиом посредством использования корректных правил – правил *достоверного вывода* в принятой нами ранее терминологии, - обладают свойством монотонности (по длине вывода) в том смысле, что все теоремы какой-либо теории такого типа содержатся в множестве теорем всякого расширения этой теории. При таком подходе, фактически, аксиомы и правила вывода – "компактная" форма представления полной информации о теоремах соответствующей формальной системы.

В свою очередь, исследования в области интеллектуальных систем со времени возникновения этого направления связаны с проблемами анализа рассуждений по "*здравому смыслу*" (ЗС-рассуждений – см., например, [*Winograd 80*] и др.). Значительная часть этих работ посвящена попыткам формализовать ЗС-рассуждения в рамках традиционных логических систем. Однако при построении прикладных интеллектуальных систем для открытых предметных областей формализации подобного рода оказались неадекватными: решающим обстоятельством оказалась необходимость принимать решения и действовать в условиях ограниченности знаний о "реальном мире" - т.е. в условиях *неполноты* информации. Практические приемы, используемые в такого рода решениях в рамках прикладных компьютерных систем, как правило, основаны на аналогии с действиями человека в подобных ситуациях: допустимыми часто провозглашаются "следствия", для которых нет точных доказательств, но которые, тем не менее, получены "*естественным путем*" и выглядят достаточно "*правдоподобными*". Вполне естественно, что эти "следствия", полученные на основе *неполной* информации, могут впоследствии оказаться "*невыводимыми*" на базе более полных данных. Необходимость учитывать сформулированный принцип *немонотонности* вывода в рамках ЗС-рассуждений привела, в частности, к развитию соответствующих нестандартных логических средств (см., например, обзор работ по немонотонным рассуждениям, выполненным группой Д.Маккарти в специальном номере журнала ***Artificial Intelligence*** – [*AI 80*]). Аналогичные эффекты имеют место и в системах реализации правдоподобных рассуждений в стиле ДСМ-метода: при пополнении множества Σ' фактов квази-аксиоматической теории

$$\Gamma = <\Sigma,\Sigma',\Re>$$

новыми элементами, вообще говоря, возможны и ситуации, когда ряд зависимостей в множестве Σ, выполнявшихся до такого пополнения, оказывается невыполнимыми после его выполнения. Ниже мы приведем пример названного эффекта немонотонности ДСМ-рассуждения.

6.1.1. Пример

Пусть

$\mathbf{U^{(1)}} = \{A_1, A_2, B_1, B_2, C_1, C_2, D_1, D_2, E\}$, a

$\mathbf{U^{(2)}} = \{a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2, e\}$.

Будем считать при этом, что все попарные сходства (пересечения) элементов в каждом из множеств $\mathbf{U^{(1)}}$ и $\mathbf{U^{(2)}}$ пусты.

Пусть далее

$\Omega_1 = \{\omega \mid \omega_i = (X_i ==> x_i)\}$, где i = 1,2,3,4;

$\Omega_{-1} = \{\omega \mid \omega_i = (X_i ==> x_i)\}$, где i = 5,6,7,8;

$\Omega_\tau = \{\omega \mid \omega_9 = (X_9 ==> x_9)\}$,

при этом

$X_1 = \{ A_1, A_2, B_1\}$	и	$x_1 = \{ a_1, a_2, b_1\}$
$X_2 = \{ A_1, A_2, B_2\}$		$x_2 = \{ a_1, a_2, b_2\}$
$X_3 = \{ A_1, C_1, E\}$		$x_3 = \{ a_1, c_1, e\}$
$X_4 = \{ A_1, C_2, E\}$		$x_4 = \{ a_1, c_2, e\}$
$X_5 = \{ A_1, A_2, D_1\}$		$x_5 = \{ a_1, a_2, d_1\}$
$X_6 = \{ A_1, A_2, D_2\}$		$x_6 = \{ a_1, a_2, d_2\}$
$X_7 = \{ C_1\}$		$x_7 = \{ a_1, c_1, e\}$
$X_8 = \{ C_2\}$		$x_8 = \{ a_1, c_2, e\}$
$X_9 = \{ A_1, A_2, E\}$		$x_9 = \{ a_1, e\}$

Тогда, если

$X_0 = \{ A_1, A_2\}$

$x_0 = \{ a_1, a_2\}$

и

$\omega_0 = (X_0 ==> x_0)$,

то одновременно

- ω_1 и ω_2 дают позитивное, а
- ω_3 и ω_4 - негативное сходство,

представляемое ω_0. Т.е.

$\omega_1 , \omega_2 ; \omega_3 , \omega_4 \ \vdash\!\!\!-\!\!\!-_{\mathbf{I}^0_a} \omega_0$.

(Здесь $\vdash\!\!\!-\!\!\!-_{\mathbf{I}^0_a}$ означает выводимость с помощью правила правдоподобного вывода $\mathbf{I}^0_a$)

Таким образом, множеству гипотез, порождаемых за один шаг ДСМ-рассуждения, в данном случае принадлежит зависимость ω_0, порождающее *эмпирическое противоречие*. Итак, имеет место

$$\mathbf{J}_{<0,1>}(\omega_0) .$$

Кроме того, для $\omega_{10} = (X_{10} ==> x_{10})$, где

$X_{10} = \{ A_1, E\}$

$x_{10} = \{ a_1, e\}$,

имеем:

$$\omega_3 , \omega_4 \ \vdash\!\!\!-\!\!\!-_{\mathbf{I}^+_a} \omega_{10}$$

и

$\neg(\Omega_{-1} \vdash\!\!\!-\!\!\!-_{\mathbf{I}^-_s} \omega_0)$

для любых элементов $\mathbf{I}^-_s$ из множества отрицательных правил правдоподобного вывода типа **I**.

Далее:

$$\omega_1 , \omega_2 , \omega_7 , \omega_8 , \omega_{10} \ \vdash\!\!\!-\!\!\!-_{\mathbf{II}^+_a} \omega_9$$

и при этом для ДСМ-вывода зависимости ω_9 нет возможности использовать ни $\mathbf{II}^-_a$, ни $\mathbf{II}^0_a$ (т.к. на это невозможно сделать ни на зависимости ω_0, ни на его производных – отрицательных гипотезах, получаемых взятием непустых подмножеств соответствующего x_0 при неизменной левой части стрелки $X_{10} ==> x_{10}$).

Теперь мы имеем уже

$$\mathbf{J}_{<1,2>}(\omega_0) .$$

Тогда для

$$\omega_{11} = (X_{11} ==> x_{11}) ,$$

где

$X_{11} = \{ A_1, A_2\}$

$x_{11} = \{ a_1\}$

имеем

$$\omega_1 , \omega_2 , \omega_9 \vdash\!\!\!- \mathbf{I}^{+}_{\mathbf{a}} \omega_{11}$$

но

$$\neg(\Omega_1 \cup \Omega_{<1,2>} \vdash\!\!\!- \mathbf{I}^{+}_{\mathbf{a}} \omega_0),$$

т.е. мы получили

$$\mathbf{J}_{<1,3>}(\omega_{11})$$

и

$$\mathbf{J}_{<-1,3>}(\omega_0).$$

Таким образом, полученное за один шаг эмпирическое противоречие ω_0, т.е.

$$\omega_0 \in \Omega_{<0,1>,}$$

отсутствует в множестве гипотез $\Omega_{<0,3>}$, полученных за три шага ДСМ-рассуждения (стартовавшего с представленных исходных данных). При этом зависимость ω_0 лежит в множестве $\Omega_{<-1,3>}$, и не представляется возможным получить ω_0 применением какого-либо позитивного ППВ.

o

Этот и некоторые другие примеры немонотонности ДСМ-рассуждения можно найти в работе [*Забежайло 84*].

Предъявленные соображения позволяют вести речь по крайней мере о двух принципиально различных версиях ДСМ-рассуждения:

- *монотонной* (когда каждое из ранее порожденных, либо имевшихся в исходном состоянии процесса рассуждений истинностных значений) не пересматривается (не ревизуется) на последующих шагах ДСМ-рассуждения. При этом если одна и та же пара <C,A> может выводиться (доопределяться) на разных шагах рассуждения, то это не приводит к противоречию, т.к. считается, что разное число шагов ДСМ-вывода порождает разные *степени правдоподобия*; и
- *немонотонной*, когда из всех истинностных значений, приписываемых каждой паре вида <C,A> на различных шагах ДСМ-рассуждения, *актуальным* (и, возможно, пересматриваемым на последующих шагах[1]) считается лишь значение, порождаемое на *текущем* шаге.

[1] Разумеется, до достижения состояния стабилизации ДСМ-рассуждения.

6.2. О функциональности отношения причинности в ДСМ-методе.

Одной из часто используемых (востребованных) характеристик бинарных отношений является свойство так называемой функциональности, когда в принадлежащих соответствующему отношению *xRy* парах вида < *x,y* > каждому конкретному значению *x* сопоставляется не более чем единственное значение *y*. С «технической» точки зрения проверка выполнимости (т.е. подтверждение или же, наоборот, - опровержение) этого свойства может носить весьма непростой характер. Такая ситуация характерна, например, для случаев, когда изучаемое отношение *R* задается теми или иными нетривиальными условиями – комбинациями различных ограничений, алгоритмическими процедурами и др. .

Таким образом, пусть заданы два множества *X* и *Y*. Для бинарного отношения

$$R \subseteq (X \text{ x } Y),$$

как подмножества декартова произведения *X* и *Y*, в каждой паре < *x,y* > такой, что имеет место *xRy* , сопоставим каждому x_0 из *X* множество Y_{x0}:

$$Y_{x0} = \{ y \mid (xRy) \,\&\, (x = x_0) \,\&\, (x_0 \in X) \},$$

т.е. для каждого значения *x* соответствующее Y_x - это множество всех *y* в множестве всех принадлежащих отношению *R* пар < *x,y* > с этим *x*.

6.2.1. Определение.

Отношение будем называть функциональным тогда и только тогда, когда выполняется условие

$$\forall x \, [\exists y \, (xRy) \rightarrow \{ \, \forall z \, [(xRz) \rightarrow \, (z = y)] \, \}],$$

□

В рамках только что обозначенного контекста представляется естественным обратиться к диагностике обсуждаемого свойства функциональности, в частности, у базового отношения причинности (т.е. отношения *под-объект V⊂X есть причина наличия у объекта X множества фиксированных свойств W:* $V =>_2 W$) в ДСМ-методе автоматического порождения гипотез (см., например, [*Финн 10, ДСМ АПГ*] и др.). На важность и в то же время нетривиальный характер этой проблемы обратил внимание В.К.Финн[1].

Переходя к детальному анализу проблемы функциональности отношения причинности в ДСМ-методе, обратим внимание на следующие существенные обстоятельства:

- исходными для реализации ДСМ-рассуждения являются три множества примеров:

  - множество Ω^+ позитивных примеров – множество пар вида

    <объект - X, множество свойств - Y>,

    для которых известно, что объект X обладает множеством свойств Y. В терминах частично определенного отношения $=>_1$ (*объект обладает свойствами*) это представляется как

    $$J_{<1,0>}(X =>_1 Y);$$

  - множество Ω^- негативных примеров – множество пар вида <X,Y>, таких, что известно, что объект X не обладает множеством свойств Y. В терминах частично определенного отношения $=>_1$ это представляется как

    $$J_{<-1,0>}(X =>_1 Y) \text{ и}$$

  - множество Ω^τ недоопределенных (предлагаемых для ДСМ-прогноза) примеров – множество пар вида <X,Y>, таких, что неизвестно, обладает ли объект X множеством свойств Y. В терминах частично определенного отношения $=>_1$ это представляется как

    $$J_{<\tau,0>}(X =>_1 Y).$$

---

[1] В свою очередь, обсуждение этой проблемы с В.К.Финном и М.А.Михеенковой послужило отправной точкой для написания автором настоящей заметки.

Здесь также истинностное значение формулы $J_{<\alpha,k>}(\varphi)$ есть 1 (*логическая истина*), если истинностное значение формулы φ есть α, и 0 (*логическая ложь*) в противном случае;

а второе значение – k – в угловых скобках в нижнем индексе при J-функции представляет номер шага ДСМ-рассуждения (в данном случае – нулевого, т.е. это – исходные данные на «входе» ДСМ-рассуждения);

- частично определенное отношение $=>_2$ (*данный под-объект является причиной наличия\отсутствия у соответствующих объектов заданного множества свойств*) последовательно формируется (см. например [*Финн 10, ДСМ АПГ*] и др.) по шагам ДСМ-рассуждения применением соответствующих правил правдоподобного вывода. При этом:
- на нулевом шаге отношение $=>_2$ представляет собою пустое множество пар вида <V,W>;
- на первом шаге из исходных примеров (элементов множеств Ω^+ и Ω^-) порождаются первые пары вида <V,W>, для которых

$$J_{<\alpha,1>}(V =>_2 W),$$

а α может принимать значения 1 (*позитивная зависимость*), -1 (*негативная зависимость*) или 0 (случай *эмпирического противоречия*), затем

- на следующем шаге ДСМ-рассуждения найденные зависимости используются для ДСМ-доопределния (там, где это разрешают соответствующие ДСМ-правила правдоподобного вывода) примеров из множества Ω^τ,
- вновь доопределенные примеры из Ω^τ пополняют соответствующие множества позитивных и негативных примеров Ω^+ и Ω^-, после чего
- процесс ДСМ-рассуждения итерируется до состояния стабилизации (см. [*Финн 10, ДСМ АПГ*] и др.).

Теперь перейдем собственно к вопросу о функциональности отношения $=>_2$. Прежде всего отметим, что так называемое условие[2] исчерпываемости

$$\forall X\ \forall Y\ \mathbf{[}\ [\mathbf{J}_{<1,r>}(X==>_1 Y)\ \&$$
$$\&\ \forall A\ (\ \mathbf{J}_{<1,r>}(X==>_1 A) \rightarrow (A \subseteq Y)\ )\ \&\ (V \subset X)] \rightarrow$$
$$\rightarrow [\ (W \neq \varnothing)\ \&\ (W \subseteq Y)\ \&\ ((X = C_1) \vee \ldots \vee (X = C_k))\ ]\ \mathbf{]}$$

в формулировке (см. [*Финн 10, ДСМ АПГ*] и др.) решающего предиката Простого сходства

$$\boldsymbol{M}^+_{\boldsymbol{a}}(V,W) = \exists k\ \underline{\boldsymbol{M}^+_{\boldsymbol{a}}}(V,W,k) =$$
$$= \exists k\ \exists C_1 \ldots \exists C_k\ \exists A_1 \ldots \exists A_k\ \mathbf{(}\ [\ \mathbf{J}_{<1,r>}(C_1==>_1 A_1)\ \&$$
$$\&\ \forall A\ (\ \mathbf{J}_{<1,r>}(C_1==>_1 A) \rightarrow (A \subseteq A_1))\ ]\ \& \ldots$$
$$\ldots\ \&\ [\ \mathbf{J}_{<1,r>}(C_k==>_1 A_k)\ \&$$
$$\&\ \forall A\ (\ \mathbf{J}_{<1,r>}(C_k==>_1 A) \rightarrow (A \subseteq A_k))\ ]\ \&$$
$$\&\ [\ (\bigcap_{i=1}^{k} C_i = V)\ \&\ (V \neq \varnothing)\ ]\ \&$$
$$\&\ [\ \forall i \forall j\ (\ [(i \neq j) \& (1 \leq i,j \leq k)] \rightarrow (C_i \neq C_j)\ )\ ]\ \&$$
$$\&\ \forall X\ \forall Y\ \mathbf{[}\ [\mathbf{J}_{<1,r>}(X==>_1 Y)\ \&$$
$$\&\ \forall A\ (\ \mathbf{J}_{<1,r>}(X==>_1 A) \rightarrow (A \subseteq Y)\ )\ \&\ (V \subset X)] \rightarrow$$
$$\rightarrow [\ (W \neq \varnothing)\ \&\ (W \subseteq Y)\ \&\ ((X = C_1) \vee \ldots \vee (X = C_k))\ ]\ \mathbf{]}\ \&$$
$$\&\ (k \geq 2)\ \mathbf{)}$$

[2] В данном случае - для позитивных примеров, анализируемых на r–том шаге ДСМ-рассуждения.

на каждом фиксированном шаге ДСМ-рассуждения при вычислении соответствующего *y* по рассчитанному (как результат вычисления сходства соответствующих примеров) *x* не оставляет иных возможностей кроме вычисления максимальной общей части множеств свойств A_i всех k штук задействованных при вычислении *x* примеров:

$$y = \bigcap_{i=1}^{k} A_i$$

Таким образом, имеет место

6.2.2. Утверждение

На каждом фиксированном шаге ДСМ-рассуждения отношение $=>_2$ функционально.

Доказательство.

Построим доказательство индукцией по номеру шага ДСМ-рассуждения (рассматривая лишь шаги с нечетными номерами: на четных шагах отношение $=>_2$ лишь используется для доопределения новых примеров из множества Ω^{τ}).

1. На шаге 1 каждая из зависимостей со $=>_2$ порождается из исходно заданных примеров – элементов множеств Ω^+ и Ω^-. Предположим, что найдутся V, W_1 и W_2 такие, что

$$J_{<\alpha,1>}(V =>_2 W_1) \text{ и } J_{<\alpha,1>}(V =>_2 W_2),$$

при этом $\alpha \in \{1,-1\}$ и

$$\varnothing \neq W_1 \neq W_2 \neq \varnothing .$$

Тогда для $<V,W_1>$ и $<V,W_2>$ в множестве Ω^{α} найдутся соответственно по крайней мере по два примера

$$J_{<\alpha,0>}(X_1 =>_1 Y_1), J_{<\alpha,0>}(X_2 =>_1 Y_2)$$

и

$$J_{<\alpha,0>}(X_3 =>_1 Y_3), J_{<\alpha,0>}(X_4 =>_1 Y_4),$$

из которых и порождаются две рассматриваемые нами зависимости. Но в этом случае, в соответствии с условием исчерпываемости (см. выше) необходимо рассмотреть сходство всех имеющихся в нашем распоряжении примеров вида $J_{<\alpha,0>}(X_i =>_1 Y_i)$, для которых

$$(V \subset X_i),$$

т.е. перебрать все i из соответствующего множества индексов (в рассматриваемом примере содержащего по крайней мере 4 элемента: $i \in \{1,2,3,4, \ldots\}$). При этом у нас останется лишь единственная возможность для вычисления W, соответствующего этому множеству всех выделенных условием исчерпываемости примеров:

$$W = (\bigcap_{i=1}^{k} Y_i)$$

где k пробегает все уже перечисленные в условии исчерпываемости примеры. Таким образом, мы пришли к противоречию (с исходным предположением):

$$W_1 = W = W_2.$$

2. Положим, что рассматриваемое нами утверждение верно на шаге 2n+1. Покажем, что оно также справедливо и на шаге 2n+3 = 2(n+1) + 1.

Предположим противное. Тогда, также как и в только что рассмотренном случае для n = 0, при порождении двух зависимостей вида

$$J_{<\alpha,2n+3>}(V =>_2 W_1) \text{ и } J_{<\alpha,2n+3>}(V =>_2 W_2),$$

должны быть задействованы по крайней мере 2 пары примеров вида

$$J_{<\alpha,s>}(X_1 =>_1 Y_1), J_{<\alpha,s>}(X_2 =>_1 Y_2)$$

и

$$J_{<\alpha,s>}(X_3 =>_1 Y_3), J_{<\alpha,s>}(X_4 =>_1 Y_4),$$

где s будет пробегать либо соответствующее множество индексов примеров знака α, порожденных на 2n+2 шаге ДСМ-рассуждения, либо множество примеров того же знака,

используемых для порождения причинных зависимостей на шаге 2n (т.е. в ситуации, когда отношение $=>_2$ функционально).

Применяя тот же - что и для случая n=0 - прием (с проверкой условия исчерпываемости на объединении двух множеств примеров, используемых при порождении причинных зависимостей на данном шаге ДСМ-рассуждения), также приходим здесь к единственному варианту вычисления соответствующего W:

$$W = (\bigcap_{i=1}^{k} Y_i).$$

Полученное таким образом противоречие (с исходным допущением о существовании двух различных W_1 и W_2) завершает доказательство рассматриваемого **Утверждения**.

□

Переходя к обсуждению исследуемых свойств отношения $=>_2$ на *последовательностях* шагов (применения Правил правдоподобного вывода ДСМ-метода), следует прежде всего иметь ввиду (уже достаточно давно описанные - см., например, [*Забежайло 84, ДСМ АПГ*] и др.) эффекты немонотонности ДСМ-рассуждений.

В свою очередь, говоря о немонотонности в обсуждаемом контексте, представляется целесообразным различать два существенно отличающихся друг от друга (процедурных) режима организации ДСМ-рассуждений. В первом из них (будем называть его режимом *слабой* немонотонности) эффекты невыводимости (см., например [*ДСМ АПГ*] и др.) охватывают лишь порождаемые причинные зависимости (формулы, содержащие соответствующий отношению $=>_2$ предикат). Все результаты ДСМ-доопределений примеров из исходного множества Ω^τ на всех последующих шагах ДСМ-рассуждения объявляются неизменными (имеющими статус, который не подлежит пересмотру на всех последующих шагах ДСМ-рассуждения). Другими словами, в этом режиме множество примеров, используемых последовательно на соответствующих шагах ДСМ-рассуждения по крайней мере не уменьшается (может либо расти, либо оставаться неизменным[3]).

Во втором процедурном режиме (будем называть его режимом *сильной* немонотонности) вместе с невыводимостью тех или иных причинных зависимостей, содержащих $=>_2$, регулярному анализу[4] дополнительно подвергаются также и зависимости, содержащие $=>_1$ (на старте ДСМ-рассуждения пребывавшие в множестве Ω^τ): представляется естественным здесь вместе с невыводимостью (на определенном шаге ДСМ-рассуждения) тех или иных причинных зависимостей, содержащих $=>_2$, считать невыводимыми также и те исходные τ-примеры, содержащие $=>_1$, которые переводились из Ω^τ в текущее множество позитивных (или же негативных) примеров с помощью теперь (т.е. на данном шаге ДСМ-рассуждения) уже утративших выводимость причинных зависимостей, содержащих $=>_2$.

Одним из доводов в пользу разделения режимов сильной и слабой немонотонности ДСМ-рассуждения может служить, например, ситуация с использованием в качестве компонентов задействованной стратегии ДСМ-рассуждения лишь Правил правдоподобного вывода с запретом на контрпримеры (см. [*Финн 10, ДСМ АПГ*] и др.). Как нетрудно убедиться, в этом случае порождение в рамках ДСМ-рассуждения на каком-либо его шаге номер k формул вида

$$J_{<0,k>}(V =>_2 W)$$

оказывается попросту невозможным. Тем не менее, здесь все же можно сконструировать ситуации, когда доопределение новых примеров (элементов исходного множества Ω^τ) оказывается противоречивым. Простейшую иллюстрацию алгоритмических конструкций такого типа дает

6.2.3. Пример.

Исходные данные (ШАГ 0)

---

[3] В последнем случае достигнуто состояние стабилизации ДСМ-рассуждения.

[4] Например, на каждой последовательной паре шагов (т.е. так называемом *такте*) ДСМ-рассуждения.

Множество позитивных примеров Ω^+ :

$J_{<1,0>}$(**A_1 A** $=>_1$ **ac**)
$J_{<1,0>}$(**A_2 A** $=>_1$ **ac**)

Множество негативных примеров Ω^- :

$J_{<-1,0>}$(**B_1 B** $=>_1$ **bc**)
$J_{<-1,0>}$(**B_2 B** $=>_1$ **bc**)

Множество недоопределенных примеров Ω^τ :
$J_{<\tau,0>}$ (**AB** $=>_1$ **abc**)

ШАГ 1.

Порождаем (правилами правдоподобного вывода по Методу позитивного и негативного сходства с запретом на контрпримеры) возможные причинные зависимости:

$J_{<1,1>}$ (**A** $=>_2$ **ac**)
$J_{<-1,1>}$ (**B** $=>_2$ **bc**)

ШАГ 2:

Доопределяем найденными причинными зависимостями соответствующий элемент множества Ω^τ :

$J_{<0,2>}$ (**AB** $=>_1$ **abc**)

Таким образом, трехзначный случай (т.е. ситуация, когда для означивания порождаемых зависимостей задействованы лишь истинностные значения типов $\{1,-1, \tau\}$) уже в простейшем примере расширяется также и на четвертый тип истинностных значений – 0 (*эмпирическое противоречие*).

Принимая во внимание тот факт, что в рамках слабо-немонотонной версии ДСМ-рассуждения исходное множество примеров, используемых для обучения, может лишь расширяться (за счет новых элементов, переводимых – доопределяемых - на соответствующих шагах ДСМ-вывода из исходного множества Ω^τ), несложно убедиться, что в силу условия исчерпываемости (также как и в случае **Утверждения 6.2.2** – см. выше) при порождении новых причинных зависимостей справедливо

6.2.4. Утверждение

На каждом фиксированном шаге слабо-немонотонного ДСМ-рассуждения отношение $=>_2$ функционально, при этом на более поздних шагах множество свойств (правая часть) в соответствующих причинных зависимостях может либо оставаться неизменной, либо только сокращаться[5].

Теперь обратимся к случаю сильной немонотонности ДСМ-рассуждения. Здесь возможны ситуации, в которых промежуточные результаты - конкретные причинные зависимости, формулируемые как принадлежащие отношению $=>_2$ пары <*под-объект*, *множество свойств*>, а также вновь доопределенные примеры, ранее находившиеся в множестве Ω^τ, - на последующих (более поздних) шагах ДСМ-рассуждения могут оказаться невыводимыми. При этом место соответствующих причинных зависимостей, принадлежащих отношению $=>_2$, займут новые пары вида <*под-объект, множество свойств*>, в которых ранее уже найденные *под-объекты* окажутся связанными этим же отношением причинности с новыми (в т.ч. – **большими**, **меньшими** или же **несравнимыми** в смысле взаимной вложимости множеств) наборами *свойств*. Таким образом, при

[5] Т.е. запрещены случаи порождения (при неизменной левой части $=>_2$) правой части, несравнимой с ранее порожденной или же объемлющей ранее порожденную.

сохранении функциональности отношения $=>_2$ «внутри» каждого шага ДСМ-рассуждений, на некоторых последовательностях таких шагов могут возникать ситуации, когда:

- выводившаяся на некотором шаге причинная зависимость вида $J_{<\alpha,k1>}(V =>_2 W)$, на некотором более позднем шаге k_2 (где $k_2 > k_1$) перестанет выводиться (например, за счет доопределения новых элементов из множества Ω^{τ} , присоединение которых к множеству примеров, используемых при порождении новых причинных зависимостей, за счет исполнения условия исчерпываемости сокращает ранее порождавшееся с этой «*причиной*» множество *свойств*), и кроме того
- на некотором еще более позднем шаге k_3 (где $k_3 > k_2 > k_1$) вновь выводящейся окажется новая зависимость вида $J_{<\alpha,k3>}(V =>_2 W_0)$, при этом возможны все три варианта: $W_0 \subset W$, $W \subset W_0$ и $[W \supset (W \cap W_0) \subset W_0]$.

Детальную иллюстрацию только сформулированной схемы дает следующий

6.2.5. Пример

Исходные данные (ШАГ 0)

Множество позитивных примеров:
Ω^{+} = { (1.0.1), (1.0.2), … , (1.0.24) }

Множество негативных примеров Ω^{-} – пусто.

Множество недоопределенных примеров:
Ω^{τ} = { (τ.0.1), (τ.0.2), … , (τ.0.20) }

При этом:

(1.0.1) $J_{<1,0>}(P_{111}\, Q_{1111} =>_1 \mathbf{abc})$
(1.0.2) $J_{<1,0>}(P_{111}\, Q_{1112} =>_1 \mathbf{abc})$
(1.0.3) $J_{<1,0>}(P_{112}\, Q_{1121} =>_1 \mathbf{abc})$
(1.0.4) $J_{<1,0>}(P_{112}\, Q_{1122} =>_1 \mathbf{abc})$
(1.0.5) $J_{<1,0>}(P_{121}\, Q_{1211} =>_1 \mathbf{abc})$
(1.0.6) $J_{<1,0>}(P_{121}\, Q_{1212} =>_1 \mathbf{abc})$
(1.0.7) $J_{<1,0>}(P_{122}\, Q_{1221} =>_1 \mathbf{abc})$
(1.0.8) $J_{<1,0>}(P_{122}\, Q_{1222} =>_1 \mathbf{abc})$
(1.0.9) $J_{<1,0>}(P_{211}\, Q_{2111} =>_1 \mathbf{abc})$
(1.0.10) $J_{<1,0>}(P_{211}\, Q_{2112} =>_1 \mathbf{abc})$
(1.0.11) $J_{<1,0>}(P_{212}\, Q_{2121} =>_1 \mathbf{abc})$
(1.0.12) $J_{<1,0>}(P_{212}\, Q_{2122} =>_1 \mathbf{abc})$
(1.0.13) $J_{<1,0>}(P_{221}\, Q_{2211} =>_1 \mathbf{abc})$
(1.0.14) $J_{<1,0>}(P_{221}\, Q_{2212} =>_1 \mathbf{abc})$
(1.0.15) $J_{<1,0>}(P_{222}\, Q_{2221} =>_1 \mathbf{abc})$
(1.0.16) $J_{<1,0>}(P_{222}\, Q_{2222} =>_1 \mathbf{abc})$
(1.0.17) $J_{<1,0>}(D_1\, X_1 =>_1 \mathbf{ab})$
(1.0.18) $J_{<1,0>}(D_1\, X_2 =>_1 \mathbf{ab})$
(1.0.19) $J_{<1,0>}(D_2\, X_3 =>_1 \mathbf{ab})$
(1.0.20) $J_{<1,0>}(D_2\, X_4 =>_1 \mathbf{ab})$
(1.0.21) $J_{<1,0>}(Y_1\, Z_1 =>_1 \mathbf{a}d_1)$
(1.0.22) $J_{<1,0>}(Y_1\, Z_2 =>_1 \mathbf{a}d_1)$
(1.0.23) $J_{<1,0>}(Y_2\, Z_3 =>_1 \mathbf{a}d_2)$
(1.0.24) $J_{<1,0>}(Y_2\, Z_4 =>_1 \mathbf{a}d_2)$

(τ.0.1) $J_{<\tau,0>}(Q_{11}\, P_{111} =>_1 \mathbf{abc})$
(τ.0.2) $J_{<\tau,0>}(Q_{11}\, P_{112} =>_1 \mathbf{abc})$
(τ.0.3) $J_{<\tau,0>}(Q_{12}\, P_{121} =>_1 \mathbf{abc})$
(τ.0.4) $J_{<\tau,0>}(Q_{12}\, P_{122} =>_1 \mathbf{abc})$
(τ.0.5) $J_{<\tau,0>}(Q_{21}\, P_{211} =>_1 \mathbf{abc})$
(τ.0.6) $J_{<\tau,0>}(Q_{21}\, P_{212} =>_1 \mathbf{abc})$
(τ.0.7) $J_{<\tau,0>}(Q_{22}\, P_{221} =>_1 \mathbf{abc})$
(τ.0.8) $J_{<\tau,0>}(Q_{22}\, P_{222} =>_1 \mathbf{abc})$
(τ.0.9) $J_{<\tau,0>}(P_1 Q_{11} =>_1 \mathbf{abc})$
(τ.0.10) $J_{<\tau,0>}(P_1 Q_{12} =>_1 \mathbf{abc})$
(τ.0.11) $J_{<\tau,0>}(P_2 Q_{21} =>_1 \mathbf{abc})$
(τ.0.12) $J_{<\tau,0>}(P_2 Q_{22} =>_1 \mathbf{abc})$
(τ.0.13) $J_{<\tau,0>}(D_1 X_5 =>_1 \mathbf{a})$
(τ.0.14) $J_{<\tau,0>}(D_2 X_5 =>_1 \mathbf{a})$
(τ.0.15) $J_{<\tau,0>}(\mathbf{AB}P_1 =>_1 \mathbf{abc})$
(τ.0.16) $J_{<\tau,0>}(\mathbf{AB}P_2 =>_1 \mathbf{abc})$
(τ.0.17) $J_{<\tau,0>}(\mathbf{AB}D_1 =>_1 \mathbf{ab})$
(τ.0.18) $J_{<\tau,0>}(\mathbf{AB}D_2 =>_1 \mathbf{ab})$
(τ.0.19) $J_{<\tau,0>}(Y_1\, X_5 =>_1 \mathbf{a}d_1)$
(τ.0.20) $J_{<\tau,0>}(Y_2\, X_5 =>_1 \mathbf{a}d_2)$

ШАГ 1.

Порождаем (правилом правдоподобного вывода по методу позитивного сходства) причинные зависимости:

(1.1.1) $J_{<1,1>}(P_{111} =>_2 \mathbf{abc})$ из (1.0.1) и (1.0.2)
(1.1.2) $J_{<1,1>}(P_{112} =>_2 \mathbf{abc})$ из (1.0.3) и (1.0.4)
(1.1.3) $J_{<1,1>}(P_{121} =>_2 \mathbf{abc})$ из (1.0.5) и (1.0.6)
(1.1.4) $J_{<1,1>}(P_{122} =>_2 \mathbf{abc})$ из (1.0.7) и (1.0.8)
(1.1.5) $J_{<1,1>}(P_{211} =>_2 \mathbf{abc})$ из (1.0.9) и (1.0.10)
(1.1.6) $J_{<1,1>}(P_{212} =>_2 \mathbf{abc})$ из (1.0.11) и (1.0.1)
(1.1.7) $J_{<1,1>}(P_{221} =>_2 \mathbf{abc})$ из (1.0.13) и (1.0.14)
(1.1.8) $J_{<1,1>}(P_{222} =>_2 \mathbf{abc})$ из (1.0.15) и (1.0.16)

(1.1.9) $J_{<1,1>}(D_1 =>_2 \mathbf{ab})$ из (1.0.17) и (1.0.18)
(1.1.10) $J_{<1,1>}(D_2 =>_2 \mathbf{ab})$ из (1.0.19) и (1.0.20)

(1.1.11) $J_{<1,1>}(Y_1 =>_2 \mathbf{a}d_1)$ из (1.0.21) и (1.0.22)
(1.1.12) $J_{<1,1>}(Y_2 =>_2 \mathbf{a}d_2)$ из (1.0.23) и (1.0.24)

ШАГ 2:

Доопределяем найденными причинными зависимостями соответствующие элементы множества Ω^{τ} :

(1.2.1) $J_{<1,2>}(Q_{11}\,P_{111} =>_1 \mathbf{abc})$ из (τ.0.1) и (1.1.1)
(1.2.2) $J_{<1,2>}(Q_{11}\,P_{112} =>_1 \mathbf{abc})$ из (τ.0.2) и (1.1.2)
(1.2.3) $J_{<1,2>}(Q_{12}\,P_{121} =>_1 \mathbf{abc})$ из (τ.0.3) и (1.1.3)
(1.2.4) $J_{<1,2>}(Q_{12}\,P_{122} =>_1 \mathbf{abc})$ из (τ.0.4) и (1.1.4)
(1.2.5) $J_{<1,2>}(Q_{21}\,P_{211} =>_1 \mathbf{abc})$ из (τ.0.5) и (1.1.5)
(1.2.6) $J_{<1,2>}(Q_{21}\,P_{212} =>_1 \mathbf{abc})$ из (τ.0.6) и (1.1.6)
(1.2.7) $J_{<1,2>}(Q_{22}\,P_{221} =>_1 \mathbf{abc})$ из (τ.0.7) и (1.1.7)
(1.2.8) $J_{<1,2>}(Q_{22}\,P_{222} =>_1 \mathbf{abc})$ из (τ.0.8) и (1.1.8)

(1.2.9) $J_{<1,2>}(\mathbf{AB}D_1 =>_1 \mathbf{ab})$ из (τ.0.17) и (1.1.9)
(1.2.10) $J_{<1,2>}(\mathbf{AB}D_2 =>_1 \mathbf{ab})$ из (τ.0.18) и (1.1.10)

(1.2.11) $J_{<1,2>}(Y_1X_5 =>_1 \mathbf{a}d_1)$ из (τ.0.19) и (1.1.11)
(1.2.12) $J_{<1,2>}(Y_2X_5 =>_1 \mathbf{a}d_2)$ из (τ.0.20) и (1.1.12)

ШАГ 3.

Расширяем множество позитивных примеров вновь доопределенными примерами и вновь порождаем причинные зависимости:

(1.3.1) $J_{<1,3>}(Q_{11} =>_2 \mathbf{abc})$ из (1.2.1) и (1.2.2)
(1.3.2) $J_{<1,3>}(Q_{12} =>_2 \mathbf{abc})$ из (1.2.3) и (1.2.4)
(1.3.3) $J_{<1,3>}(Q_{21} =>_2 \mathbf{abc})$ из (1.2.5) и (1.2.6)
(1.3.4) $J_{<1,3>}(Q_{22} =>_2 \mathbf{abc})$ из (1.2.7) и (1.2.8)

(1.3.5) $J_{<1,3>}(\boldsymbol{AB} =>_2 \boldsymbol{ab})$ из (1.2.9) и (1.2.10)

(1.3.6) $J_{<1,3>}(X_5 =>_2 \mathbf{a})$ из (1.2.11) и (1.2.12)

Заметим также, что на данном шаге сохраняются (вновь выводятся) соответствующие причинные зависимости, формируемые из позитивных примеров исходного (см. ШАГ 0) множества Ω^{+} .

ШАГ 4.

Переносим порождаемые зависимости на новые примеры, требующие доопределения:

(1.4.1) $J_{<1,4>}$ ($P_1Q_{11} =>_1$ **abc**) из (τ.0.9) и (1.3.1)
(1.4.2) $J_{<1,4>}$ ($P_1Q_{12} =>_1$ **abc**) из (τ.0.10) и (1.3.2)
(1.4.3) $J_{<1,4>}$ ($P_2Q_{21} =>_1$ **abc**) из (τ.0.11) и (1.3.3)
(1.4.4) $J_{<1,4>}$ ($P_2Q_{22} =>_1$ **abc**) из (τ.0.12) и (1.3.4)

(1.4.5) $J_{<1,4>}$ ($D_1X_5 =>_1$ **a**) из (τ.0.13) и (1.3.6)
(1.4.6) $J_{<1,4>}$ ($D_2X_5 =>_1$ **a**) из (τ.0.13) и (1.3.6)

ШАГ 5.

Вновь расширяем множество позитивных примеров доопределяемыми примерами и вновь порождаем причинные зависимости:

(1.5.1) $J_{<1,5>}$ ($P_1 =>_2$ **abc**) из (1.4.1) и (1.4.2)
(1.5.2) $J_{<1,5>}$ ($P_2 =>_2$ **abc**) из (1.4.3) и (1.4.4)

(1.5.3) $J_{<1,5>}$ ($D_1 =>_2$ **a**) из (1.0.17), (1.0.18) и (1.4.5)
(1.5.4) $J_{<1,5>}$ ($D_2 =>_2$ **a**) из (1.0.19), (1.0.20) и (1.4.6)

Дополнительно заметим, что на данном шаге также сохраняются, т.е. вновь выводятся - за исключением пары перестающих выводиться:

(1.1.9) $J_{<1,1>}$ ($D_1 =>_2$ **ab**)
(1.1.10) $J_{<1,1>}$ ($D_2 =>_2$ **ab**),

соответствующие причинные зависимости, формируемые из позитивных примеров исходного множества Ω^+. Кроме того, сохраняются (вновь выводятся) соответствующие причинные зависимости, порожденные и на ШАГе 3.

ШАГ 6.

Переносим порождаемые зависимости на новые примеры, требующие доопределения:

(1.6.1) $J_{<1,6>}$ (**AB**$P_1 =>_1$ **abc**) из (τ.0.15) и (1.5.1)
(1.6.2) $J_{<1,6>}$ (**AB**$P_2 =>_1$ **abc**) из (τ.0.16) и (1.5.2)

При этом перестают доопределяться

(τ.0.17) $J_{<\tau,0>}$ (**AB**$D_1 =>_1$ **ab**)
(τ.0.18) $J_{<\tau,0>}$ (**AB**$D_2 =>_1$ **ab**)

т.к. больше - из-за (1.5.3) и (1.5.4) - не выводятся

(1.1.9) $J_{<1,1>}$ ($D_1 =>_2$ **ab**)
(1.1.10) $J_{<1,1>}$ ($D_2 =>_2$ **ab**)

ШАГ 7.

Расширяем множество позитивных примеров вновь доопределенными примерами и вновь порождаем причинные зависимости:

(1.7.1) $J_{<1,7>}$ (***AB*** $=>_2$ ***abc***) из (1.6.1) и (1.6.2)

При этом перестает выводиться (1.3.5), т.к. перестали выводиться (1.2.9) и (1.2.10).

□

6.2.6. Замечание.

Несложно убедиться, что заменой ***abc*** на ***ac*** приведенная выше в **Примере 6.2.5** техническая конструкция может быть воспроизведена также и при выводе зависимости вида

$$J_{<1,7>} (\boldsymbol{AB} =>_2 \boldsymbol{ac}),$$

о порождаемости которой мы говорили выше, рассматривая случай [$W \supset (W \cap W_0) \subset W_0$].

□

6.2.7. Замечание.

Также несложно убедиться, что случай $W_0 \subset W$ (в текущем контексте, например, положив $W_0 = \{\boldsymbol{a}\}$) можно описать, добавив (через доопределение соответствующим образом внесенных в исходное множество Ω^{τ}) к ранее уже получаемым позитивным примерам с множеством свойств $\{\boldsymbol{ab}\}$ дополнительные примеры с той же левой частью $\{\boldsymbol{AB}\}$ стрелки $=>_2$ и сокращенной правой частью - $\{\boldsymbol{a}\}$.

□

В заключение представляется важным обратить внимание на особый статус причинных зависимостей, порождаемых из исходного (доступного на нулевом шаге ДСМ-рассуждения) множества примеров. Их декларируемый здесь особый статус выражается в том, что в порождаемых из них причинных зависимостях нельзя (используя те или иные «трюки» с доопределением в ходе ДСМ-рассуждения тех или иных ранее недоопределенных примеров из множества Ω^{τ}) *расширить* соответствующее (т.е. формируемое на ШАГе 1) *множество свойств*. В ходе ДСМ-рассуждения эти причинные зависимости могут – за счет расширения[6] используемых при их порождении исходных данных – либо «*усекаться*» по множеству участвующих в соответствующей причинной зависимости *свойств*, либо – за счет порождения аналогичных зависимостей на примерах противоположного знака – переходить в статус *эмпирически противоречивых*. Этот особый статус исходных примеров а также порождаемых из них на первом ШАГе причинных зависимостей, по-видимому, может рассматриваться в качестве дополнительного аргумента как в защиту важности используемого в ДСМ-методе (см., например, [*Финн 10, ДСМ АПГ*] и др.) *критерия каузальной полноты*, так и в пользу особого статуса доопределений новых (исходно находящихся в множестве Ω^{τ}) примеров причинными зависимостями ШАГа 1.

Итак, исходные примеры и порождаемые из них зависимости имеют в ДСМ-рассуждении особый статус.

Однако, не меньшего внимания, по-видимому, заслуживают также и результаты, порождаемые на последующих (втором и далее) шагах ДСМ-рассуждения. Действительно, представленный в **Примере 6.2.3** эмпирический эффект *«неустранимости» четырехзначности* множества типов задействованных в ДСМ-рассуждении истинностных значений (т.е. эффект существования особых ситуаций, когда недостаточно трех типов истинностных значений при работе даже в особом – исключающем *противоречивые* причинные зависимости - случае «*запрета на контрпримеры*»), демонстрирует принципиальную важность учета *числа шагов* ДСМ-рассуждения при порождении тех или иных ДСМ-зависимостей. Здесь, использовав уже представленную в **Примере 6.2.5** технику конструирования и синхронизации (по длине вывода) «конкурирующих» цепочек рассуждений, можно моделировать эффект порождения на *любом наперед заданном* шаге ДСМ-рассуждения эмпирического противоречия в рамках ДСМ-стратегии с запретом на контрпримеры.

[6] Дополнения обучающей выборки вновь доопределенными примерами.

6.3. «Наследуемость» эмпирических зависимостей и порождение закономерностей в рамках ДСМ-ИАД.

Обнаружив эффект «*наследуемости*» (*сохранения «порождаемости»* эмпирических зависимостей - ЭЗ) при расширении исходно заданного для ДСМ-обучения множества примеров и контрпримеров, мы имеем очевидны сигнал о том, что в процессе выполняемой экспертизы удалось обнаружить (путем расширения исходной выборки данных) пример перехода от *эмпирических зависимостей* к *эмпирическим закономерностям*.

В свою очередь возможные причины «ненаследуемости» (т.е. несохранения эффекта «порождаемости») ЭЗ - это:

a. появление в расширенном множестве примеров новых объектов со сходной (с некоторыми уже построенными ЭЗ) «структурой» и с такими множествами свойств, что на расширенном множестве примеров сходство на правых частях выражений вида **СТРУКТУРА ОБЪЕКТА** $=>_1$ **СВОЙСТВА** перестает порождаться из-за «сужения» (или же, вообще, пустоты) сходства свойств. (Легко видеть, что такая ситуация возможна лишь в случае работы с неодноэлементным множеством свойств);
b. появление новых контрпримеров, которые опровергают (на расширенной выборке исходных ДСМ-данных) условие *запрета на контрпримеры* (*ЗКП*).

Процедура «диагностики» наследуемости (или же, наоборот, ненаследуемости) порождаемых ЭЗ в данном случае может выглядеть следующим образом:

c. каждой ЭЗ, порождаемой из исходного множества примеров, может быть сопоставлен некоторый класс эквивалентности объектов[1] из исходного множества $\mathbf{\Omega_1}$;
d. при расширении $\mathbf{\Omega_1}$ добавлением $\mathbf{\Omega_{12}}$ до некоторого $\mathbf{\Omega_2}$ каждый новый пример из $\mathbf{\Omega_{12}}$ проверим на попадание (его «структурой») в имеющиеся классы эквивалентности, т.е.
    I. будем перечислять все такие классы эквивалентности и для каждого E_i из них
    II. выделим те элементы $\mathbf{\Omega_{12}}(E_i)$ из $\mathbf{\Omega_{12}}$, которые «попадают» в E_i , чтобы далее
    III. проверить вкладываются ли общие свойства $W_1(E_i)$ всех объектов из E_i в общую часть $W_{12}(\mathbf{\Omega_{12}}(E_i))$ свойств новых объектов из множества $\mathbf{\Omega_{12}}(E_i)$,
    IV. если такое вложение (в том числе - совпадение) имеет место, то соответствующая классу E_i эмпирическая зависимость наследуется с $\mathbf{\Omega_1}$ на $\mathbf{\Omega_2}$;
    V. в противном случае мы имеем пример ненаследуемости соответствующей ЭЗ с $\mathbf{\Omega_1}$ на $\mathbf{\Omega_2}$;
    VI. рассмотрев все порожденные на $\mathbf{\Omega_1}$ классы E_i , мы получим явный вид множества всех наследуемых и всех ненаследуемых с $\mathbf{\Omega_1}$ на $\mathbf{\Omega_2}$ ЭЗ;
e. при заданном множестве $\mathbf{\Omega_\tau}$ требующих прогноза свойств новых примеров можно сформировать также (задействовав процедуру из п.п. 3.b.I-VI) соответствующие карты наследуемых и ненаследуемых ЭЗ, полезных при доопределении свойств новых объектов из заданного множества $\mathbf{\Omega_\tau}$.

В случае операций с *одноэлементным* множеством свойств (см. ситуацию ***b*** выше) рассмотрим исходное множество примеров и контрпримеров

$$\mathbf{\Omega_1} = \mathbf{\Omega^+_1} \cup \mathbf{\Omega^-_1}$$

а также его расширение

[1] См. подробнее [*Забежайло 15*] а также **Раздел 3.1**.

$$\mathbf{\Omega_2} = \mathbf{\Omega^+_2} \cup \mathbf{\Omega^-_2}.$$

Положим

$$\mathbf{\Omega^i_2} = \mathbf{\Omega^i_1} \cup \mathbf{\Omega^i_{12}}$$

для $i \in \{+,-\}$. Будем считать, что

$$\mathbf{\Omega_1} \subset \mathbf{\Omega_2} \subseteq 2^{\mathbf{U}}$$

для некоторого конечного алфавита

$$\mathbf{U} = \{a_1, a_2, \dots, a_n\}.$$

Далее:

f. Для каждого объекта A_j из $\mathbf{\Omega^-_1}$ и $\mathbf{\Omega^-_{12}}$ построим множество $SL(A_j)$ не входящих в него образующих из алфавита **U** (которое будем называть **СТОП-ЛИСТОМ**).

g. Для каждого объекта $A_k = \{a^k_1, a^k_2, \dots, a^k_s\}$из множества $\mathbf{\Omega^+_1}$ множество $GC_{\text{ЗКП}}(\mathbf{\Omega^+_1})$ всех ДСМ-ЭЗ, удовлетворяющих условию *ЗКП* запрета на контрпримеры, будем формировать следующим алгоритмом:

   I. для каждого **СТОП-ЛИСТА** $SL(A_j)$ для $\mathbf{\Omega^-_1}$ найдем его общую часть $SL(A_j)\big|_{A_k}$ со «структурой» $A_k = \{a^k_1, a^k_2, \dots, a^k_s\}$ соответствующего примера из $\mathbf{\Omega^+_1}$;

   II. если какой-то из усеченных таким образом **СТОП-ЛИСТОВ** $SL(A_j)\big|_{A_k}$ окажется пустым, то переходим к работе со следующим примером A_{k+1} из $\mathbf{\Omega^+_1}$;

   III. если же все (т.е. если j пробегает по всему множеству $\mathbf{\Omega^-_1}$) усеченные предложенным способом **СТОП-ЛИСТЫ** $SL(A_j)\big|_{A_k}$ окажутся непустыми, то выбирая последовательно из каждого из них одноэлементные, двухэлементные и т.п. подмножества, затем собирая их (по схеме: каждое непустое подмножество для каждого непустого сужения **СТОП-ЛИСТА** с каждым непустым подмножеством для каждого из оставшихся непустых сужений **СТОП-ЛИСТОВ**) по всему множеству рассматриваемых **СТОП-ЛИСТОВ**, и, наконец, вычисляя замыкание для каждого объединения сформированных таким способом подмножеств образующих из алфавита **U**, мы последовательно построим все вкладывающиеся в текущий объект A_k **ЗКП**-ЭЗ, которые могут быть восстановлены ДСМ-методом из $\mathbf{\Omega_1}$. При этом следует учитывать, что для целей ДСМ-прогноза (т.е. для доопределения новых объектов) достаточно ограничиться комбинациями лишь одноэлементных подмножеств рассматриваемых сужений **СТОП-ЛИСТОВ**;

   IV. рассмотрев все варианты выбора A_k в множестве $\mathbf{\Omega^+_1}$, мы получим все множество **ЗКП**-ЭЗ, порождаемых из $\mathbf{\Omega_1}$.

h. Параллельно с формированием **СТОП-ЛИСТОВ** для множества $\mathbf{\Omega^-_1}$ (см. п.п. g.I-III выше) следует организовать абсолютно аналогичную процедуру и для множества $\mathbf{\Omega^-_{12}}$. В этом случае пустота соответствующего усечения **СТОП-ЛИСТА** для какого-либо элемента (расширяющего $\mathbf{\Omega^-_1}$) множества контрпримеров $\mathbf{\Omega^-_{12}}$ будет сигнализировать о *ненаследуемости* соответствующей (позитивной) **ЗКП**-ЭЗ с $\mathbf{\Omega_1}$ на $\mathbf{\Omega_2}$. Аналогичный сигнал о ненаследуемости ЭЗ формируется и при наличии (см. **Пример 6.3.1**) в соответствующих множеству $\mathbf{\Omega^-_{12}}$ усеченных **СТОП-ЛИСТАХ** по крайней мере одного такого SL_0, что в нем нет ранее задействованных для усеченных **СТОП-ЛИСТОВ** для множества $\mathbf{\Omega^-_1}$ образующих.

6.3.1. Пример

Пусть:

$\mathbf{U} = \{a_1, a_2, \dots, a_5\}$

$\mathbf{\Omega_1} = \mathbf{\Omega^+_1} \cup \mathbf{\Omega^-_1}$

$\mathbf{\Omega^+_1} = \{A^+_1, A^+_2, A^+_3, A^+_4\}$,

где $A^+_1 = \{a_1, a_2, a_3, a_4\}$, $A^+_2 = \{a_1, a_2, a_3, a_4\}$, $A^+_3 = \{a_1, a_2\}$ и $A^+_4 = \{a_1, a_2\}$

$\mathbf{\Omega^-_1} = \{A^-_1, A^-_2\}$, где $A^-_1 = \{a_2, a_5\}$ и $A^-_2 = \{a_1, a_3, a_4, a_5\}$

$\mathbf{\Omega_2} = \mathbf{\Omega^+_2} \cup \mathbf{\Omega^-_2}$
$\mathbf{\Omega^+_2} = \varnothing$
$\mathbf{\Omega^-_2} = \{A^-_1, A^-_2, A^-_3, A^-_4\}$, где $A^-_3 = \{a_2, a_3, a_5\}$ и $A^-_4 = \{ a_1, a_2, a_3, a_5\}$

Тогда:

$SL(A^-_1)\,|_{A^+_1} = \{a_1, a_3, a_4\}$
$SL(A^-_2)\,|_{A^+_1} = \{a_2\}$
$SL(A^-_3)\,|_{A^+_1} = \{a_1, a_3\}$
$SL(A^-_4)\,|_{A^+_1} = \{a_4\}$

Таблица 6.3.2

Структура **СТОП-ЛИСТОВ**:

	a_1	a_2	a_3	a_4	a_5
$SL(A^-_1)\,\|_{A^+_1}$	■		■	■	
$SL(A^-_2)\,\|_{A^+_1}$		■			
$SL(A^-_3)\,\|_{A^+_1}$	■		■		
$SL(A^-_4)\,\|_{A^+_1}$				■	

При этом:

из множества примеров $\mathbf{\Omega^+_1}$ с учетом *запрета на контрпримеры* из множеств $\mathbf{\Omega^-_1}$ и $\mathbf{\Omega^-_2}$ соответственно порождаются (см. **Таблицу 1**)

- две ЭЗ со «структурами» $\{a_1, a_2\}$ и $\{a_1, a_2, a_3, a_4\}$, полученными соответственно как замыкания Галуа на исходном множестве примеров $\mathbf{\Omega_1}$

  $[\{a_1, a_2\}]_{\mathbf{\Omega}_1} = \{a_1, a_2\}$ и
  $[\{a_3, a_2\}]_{\mathbf{\Omega}_1} = [\{a_4, a_2\}]_{\mathbf{\Omega}_1} = [\{a_1, a_2, a_3, a_4\}]_{\mathbf{\Omega}_1} = \{a_1, a_2, a_3, a_4\}$,

  в первом случае, и лишь
- одна ЭЗ со «структурой» $\{a_1, a_2, a_3, a_4\}$, полученной как замыкание Галуа на расширенном множестве примеров $\mathbf{\Omega_2}$

  $[\{a_1, a_2,, a_4\}]_{\mathbf{\Omega}_2} = [\{a_3, a_2, a_4\}]_{\mathbf{\Omega}_2} = [\{a_3, a_2, a_1, a_4\}]_{\mathbf{\Omega}_2} = \{a_1, a_2, a_3, a_4\}$,

  во втором. ЭЗ со структурой $\{a_1, a_2\}$ «проваливается» на присутствующем в расширении $\mathbf{\Omega^-_{12}}$ множества $\mathbf{\Omega^-_1}$ до множества $\mathbf{\Omega^-_2}$ контрпримере A^-_4.

□

Дополнительные (*быстро* вычислимые) *индикаторы ненаследуемости* ЭЗ в рассматриваемом нами случае можно получить, например, опираясь на метрические свойства порождаемых предлагаемым выше способом усечений соответствующих **СТОП-ЛИСТОВ**. (Так, если число усеченных **СТОП-ЛИСТОВ** больше текущего анализируемого $SL(A_j)\,|_{A_k}$ и в них нет повторяющихся – т.е. входящих в усечения разных **СТОП-ЛИСТОВ** – образующих, то для порождения соответствующей ЭЗ в подобной ситуации просто не хватит «строительного материала» - элементов исходного алфавита **U**).

Предлагаемая техника порождения и контроля наследуемости ЭЗ при расширении исходно заданного множества примеров новыми элементами позволяет «экономным» образом получать ответы на следующие вопросы:

Дано:

- множество $\mathbf{\Omega_1}$ примеров и контрпримеров для ДСМ-обучения, а также
- его расширение $\mathbf{\Omega_{12}}$ до множества $\mathbf{\Omega_2}$, а также
- множество $\mathbf{\Omega_\tau}$, предложенных для ДСМ-прогноза новых примеров.

Требуется:

- найти (построить) множество **зкп-ЭЗ$_{н}$** всех наследуемых при переходе от $\mathbf{\Omega_1}$ к $\mathbf{\Omega_2}$ эмпирических ДСМ-зависимостей **зкп**-ЭЗ, удовлетворяющих условию *запрета на контрпримеры* (*ЗКП*);
- проверить, совпадает ли **зкп-ЭЗ$_{н}$** с множеством всех порождаемых из $\mathbf{\Omega_1}$ ДСМ-зависимостей **зкп**-ЭЗ;
- проверить, совпадает ли подмножество **зкп-ЭЗ$_{н}$**$\big|_{\Omega\tau}$ множества **зкп-ЭЗ$_{н}$** , сформированное всеми теми ЭЗ, которые полезны для прогноза свойств новых объектов из множества $\mathbf{\Omega_\tau}$ (т.е. подмножество всех наследуемых ЭЗ, полезных для прогноза свойств примеров из $\mathbf{\Omega_\tau}$), с множеством всех порожденных из исходного множества $\mathbf{\Omega_1}$ ЭЗ,
- найти (построить) множества **зкп-ЭЗ$_{\neg н}$** и **зкп-ЭЗ$_{\neg н}$**$\big|_{\Omega\tau}$ соответственно всех ненаследуемых ЭЗ и всех полезных для прогноза свойств примеров из $\mathbf{\Omega_\tau}$ ненаследуемых ЭЗ.

Наследуемость соответствующих ЭЗ при последовательном расширении исходной выборки примеров для ДСМ-обучения новыми примерами и контрпримерами – очевидный аргумент, чтобы утверждать, что в рамках такого «рекурсивного» ДСМ-ИАД[2] найдены (восстановлены из данных) не только некоторые эмпирические зависимости, но и соответствующие ***эмпирические закономерности***.

[2] ИАД – интеллектуальный анализ данных.

6.4. Некоторые комментарии к проблеме эпистемологических оснований ИАД.

Представляя ДСМ-метод как платформу для проведения интеллектуального анализа данных (ИАД), мы условились формализовать рассматриваемый нами особый класс некорректных вычислительных проблем[1] как задачи восстановления эмпирических зависимостей (ЭЗ) из данных (фактов). Здесь нам пришлось столкнуться с четырьмя критически важными проблемами, умение «разумным образом» решать которые и определяет качество порождаемых в процессе ИАД результатов:

(I) как, располагая целым рядом возможностей, выбрать *адекватный* язык описания изучаемых явлений,

(II) как, располагая набором инструментов экспертизы, выбрать из них те, которые позволяют *учесть* (при формировании заключения) критически важные *характеристические особенности* изучаемого явления;

(III) как сформировать адекватные подходы и «инструменты» для выбора *наилучшего результата* из множества тех, которые получены разными средствами экспертизы (причем – в процессе обработки данных, представленных разными средствами – языками разной выразительной силы и вычислительной сложности);

(IV) как *оценивать* получаемые результаты в случае *ненаследуемости* порождаемых эмпирических зависимостей при *расширении* (*пополнении новыми объектами*) исходной обучающей выборки прецедентов (примеров и контрпримеров)?

Обозначенные в перечне (I) – (IV) вопросы, как несложно убедиться, в той или иной форме уже известны исследовательскому сообществу, интересующемуся проблематикой теории познания, логики открытия и роста научного знания.

В предлагаемом перечнем (I) – (IV) контексте представляется полезным учесть основные специфические особенности рассматриваемого круга проблем, а также опыт решения ряда задач этого типа, ранее уже идентифицированных и осознанных исследовательским сообществом, в частности, в свете классических результатов, которые были получены в рамках работ по

- формализации *индукции* (индуктивных выводов как заключений универсального характера, выводимых из заданного набора примеров-прецедентов) – см. например, [*Carnap 45, Carnap 50*] и др.,
- изучению структуры научных теорий и логики научного открытия – см., например, [*Popper 59, Поппер 02, Поппер 04, Кун 02, Лакатос 67 Фейерабенд 07*] и др., а также
- изучению и формализации *абдуктивных выводов* (процедур порождения частных зависимостей *от заключения к посылкам*, логическим следствием которых и является это заключение) – см., например, [*Пирс 00, Пирс 05*] и др. .

Ключевая роль здесь отводится уже хорошо знакомому перечню вопросов:

(i) какими рациональными средствами (формальными «инструментами») следует формировать (описывать и аргументировать приемлемость) утверждения вида «*зависимость* $\boldsymbol{d}$ *объясняет факты* $\boldsymbol{f_1, f_2, \ldots, f_n}$»;

(ii) как понимать утверждение о том, что «*зависимость* $\boldsymbol{d}$ *объясняет факты* $\boldsymbol{f_1, f_2, \ldots, f_n}$»;

(iii) каковы необходимые дескриптивные характеристики (доступные для использования выразительные возможности) языка представления данных, который можно считать *приемлемым* для описания *объясняющей факты* $\boldsymbol{f_1, f_2, \ldots, f_n}$ эмпирической *зависимости* $\boldsymbol{d}$;

---

[1] Задач о *доопределении частично-определенного отношения* в смысле В.К.Финна (см. например, [*Финн 76, Финн 81*] и др.) или же задач *phenomenal data mining* в смысле Дж.Маккарти (см., например, [*McCarthy 00, McCarthy*] и др.)

(iv) как оценить *достаточность оснований* для принятия утверждения «*зависимость* ***d*** *объясняет факты* $\boldsymbol{f_1, f_2,\ldots, f_n}$» (в частности, для приписывания ему *выделенной истинностной* оценки);

(v) существуют ли (проверяемые[2]) условия, выполнимость которых обеспечивала бы устойчивость[3] выделяемых из фактов эмпирических зависимостей при пополнении имеющихся исходных данных новыми прецедентами?

Переходя к комментариям опыта, к настоящему моменту уже накопленного при поиске ответов на эти вопросы, представляется уместным обратить внимание на следующие критичные для эффективной организации ИАД обстоятельства:

(1) оказалось, что эмпирические закономерности в естественных и социальных науках могут иметь существенно различающуюся природу (см., например, [*Поппер 93, Popper 59, Hempel 66, Feyerabend 75, Поппер 00*] и др.). Так называемые «законы» природы ориентированы на зависимости универсального[4] характера, в то время как в ряде наук об обществе востребованы суждения относительно конкретного единичного факта (или же определенной коллекции таких фактов).

(2) Указанные различия требуют уточнения понятия *истинность утверждения*[5]. Классическое понимание (содержание) этого понятия, опирающееся на представления об истинности как выводимости в дедуктивных теориях, оказывается недостаточным при работе с открытыми (незамкнутыми) предметными областями. Классической схемы *объяснения* по К.Гемпелю[6] (см., например, [*Hempel 45, Hempel 66*] и др.) оказывается недостаточно для ведения анализа данных уже не только в области социальных наук, но и в таких традиционных естественно-научных областях как, например, техническая[7] или же медицинская диагностика.

(3) Востребованность задач восстановления зависимостей из данных в открытых предметных областях формирует потребность в соответствующих – *открытых – теориях*[8]. Продуктивное использование частичных (неполных, фрагментарных) аксиоматизаций анализируемой предметной области (например, постулатов значения в смысле Р.Карнапа [*Carnap 36, Carnap 56* и др.]), тем не менее, не снимает вопроса о выборе приемлемых процедур *пересмотра*[9] (обновления) таких аксиоматизаций при расширении описания изучаемой предметной области (получении новых фактов и\или эмпирических зависимостей).

(4) Более того, упомянутые расширения описаний открытых предметных областей, на что впервые обратили внимание исследовательского сообщества Дж.Маккарти и

---

[2] «Разумным» образом.

[3] Т.е. «*наследуемость*» (сохранение *порождаемости*) при расширении используемой для их формирования базы фактов (прецедентов).

[4] Особенности употребления этого термина достаточно подробно комментируется в профессиональной литературе, где, в частности, демонстрируется, что в рамках соответствующих контекстов он имеет вполне определенные смысловые границы (см. например, понимание универсальности законов в ньютоновской и релятивистской механике).

[5] В частности утверждения вида «*зависимость* ***d*** *объясняет факты* $\boldsymbol{f_1, f_2, \ldots, f_n}$» а также подобных ему.

[6] Предполагающей использование «универсальных» утверждений, специфичных для соответствующей предметной области.

[7] Понимаемая достаточно широко: включая в том числе, например, задачи *идентификации* явлений заданного типа (например, технических сбоев в нормальной работе компьютерного оборудования, вызванных целенаправленным внешним вредоносным воздействием – компьютерной атакой) или же их атрибуции (например, выявления источников компьютерных атак и др.).

[8] См., например, предложенный Д.А.Поспеловым [*Поспелов 81, Поспелов 86* и др.] аппарат *семиотических моделей* или же предложенный В.К.Финном [*Финн 88, Финн 91* и др.] формализм *квази-аксиоматических теорий*.

[9] Причем, не только о собственно технологиях реинжиниринга подобных систем постулатов значения, но и о сложности вычислений, характеризующих этот процесс.

его коллеги (см. например, [*AI 80*] и др.), характеризуются эффектом *немонотонности* (когда некоторая часть ранее выводившихся утверждений[10] может оказаться *невыводимыми* при расширении текущего описания открытой предметной области новыми элементами – фактами и\или зависимостями).

(5) В таких условиях для анализа данных приходится наряду с классическим достоверным[11] выводом использовать также более широкую конструкцию – *формализованное рассуждение*[12] (где дедуктивные выводы можно комбинировать с различными видами правдоподобных рассуждений, в том числе – так называемыми рассуждениями в духе здравого смысла[13] и др.), интегрируя их с традиционными методами численного анализа данных.

(6) Использование правил *правдоподобного* вывода в цепочке формализованных рассуждений требует привлечения специальных инструментов для *оценивания* порождаемых результатов[14]. Этот путь естественным образом ведет к выбору специальных классов многозначных логик[15], позволяющих организовать адекватным образом процедуры многозначного оценивания результатов (формализованного) рассуждения. При этом несложно привести примеры эффекта *немонотонности* рассуждений (см. выше – п.4), выражающегося в *ненаследуемости* (несохранении в прежнем виде) ранее зафиксированных *типов истинностных оценок*[16] при расширении исходных данных.

(7) Особую роль при пополнении накопленного массива сведений об изучаемом явлении в открытой предметной области новой информацией приобретает попперовский *принцип фальсифицируемости* порождаемых результатов (см., например, [*Popper 59, Поппер 04, Поппер 02, Kuhn 62, Kuhn 63*] и др.). Учитывая, как правило, существенным образом *ограниченный*[17] объем фактов - примеров и контрпримеров, доступных для «первичного» восстановления скрытых в них эмпирических зависимостей, представляется вполне естественным использовать в процессе оценки *достаточности оснований*[18] для принятия результатов анализа данных процедурную «технику» проверки на *фальсифицируемость* эмпирических зависимостей одного знака, порождаемых из имеющихся фактов, зависимостями противоположного знака (например, организовать проверку фальсифицируемости *позитивных* эмпирических зависимостей результатами порождения, соответственно, *негативных* зависимостей и наоборот). В такой ситуации нефальсифицируемость зависимостей, порождаемых из имеющихся эмпирических данных, может сигнализировать о достаточности оснований для принятия этих результатов (эмпирических зависимостей). Однако, в случае

---

[10] В отличие от соответствующей ситуации в дедуктивных теориях.

[11] Дедуктивным выводом, формализованным для «инструментального» использования известными логическими средствами (см., например, [*Маслов 68, Маслов 68, Робинсон 70*] и др.).

[12] См., например, [*Финн 04*] и др. .

[13] *Common sense reasoning* – см., например, сборник [*AI 80*] и др. .

[14] Несложно представить себе ситуацию, когда некоторое утверждение может быть порождено парой цепочек правдоподобных рассуждений, имеющих несовпадающие длины. Представляется естественным иметь в этом случае возможности приписать каждому из использованных «путей» формализованного рассуждения свою *истинностную оценку* (характеризующую, в том числе, как соответствующую «*длину*» *вывода*, так и соответствующую ему *степень правдоподобия* полученного заключения).

[15] Так называемых логик ***BM***-типа (см., например, работы [*Финн 76, Аншаков 86, Аншаков 87, Anshakov 89, Аншаков 93*] и др.), характеризуемых 6 типами истинностных значений: ***t*** - логическая истина, ***f*** - логическая ложь, <+**1,n**> - эмпирическая истина на n–ном «шаге» формализованного рассуждения, <-**1,n**> - эмпирическая ложь на n–ном «шаге» формализованного рассуждения, <**0,n**> - эмпирическое противоречие на n–ном «шаге» формализованного рассуждения и ***τ*** -недоопределенность.

[16] См., например, работы [*Забежайло 84, Забежайло 13*] о немонотонности ДСМ-рассуждений и др. .

[17] В целом ряде практически значимых случаев исследователю здесь доступны лишь так называемые *статистически незначимые* выборки прецедентов.

[18] См., например, [*Есенин-Вольпин 93, Финн 83, Финн 88, Финн 91*] и др. .

расширения массива исходной информации (примеров и контрпримеров) новыми фактами такая нефальсифицируемость является, вообще говоря, ненаследуемой: несложно привести примеры *ненаследуемости* (*немонтонности* - см. п.6 выше) ранее уже зафиксированных *типов истинностных оценок* при расширении исходных данных новыми фактами (прецедентами).

(8) Критически важным шагом при организации ИАД оказывается выбор языка представления данных (и знаний) об изучаемом явлении. Здесь важно понимать, что, следуя сформулированной, в частности, П.Бернайсом [*Бернайс 00, Поппер 00-а*] и П.Фейерабендом [*Feyerabend 75* и др.] классической точке зрения на эту проблему, обсуждать и анализировать удается лишь то, что *выражается в языке*. Таким образом, адекватность выразительных (дескриптивных) возможностей используемого в рамках ИАД языка формализованного представления сведений об изучаемом явлении[19] самым непосредственным образом влияет на качество порождаемых результатов экспертизы (причем не только в части типов соответствующих истинных оценок для порождаемых ЭЗ или их наследуемости при расширении исходного «корпуса» фактов и зависимостей новыми элементами, но и в части собственно выводимости тех или иных ЭЗ при соответствующих расширениях исходных данных новыми прецедентами).

(9) В части выбора собственно *средств формирования зависимостей* из данных (средств экспертизы, задействованных в процессе ИАД) традиционный[20] взгляд на проблему *метода* научного исследования требует особого внимания к проблеме «параллельного» применения (использования) различных средств экспертизы. В ситуации, когда в арсенале исследователя могут находиться одновременно несколько, вообще говоря, несравнимых (по используемым в них средствам представления и математическим моделям анализа данных) «инструментов» экспертизы, отсутствие *аргументированных* возможностей предпочесть один из этих «инструментов» другому (в полном соответствии с предложенным П.Фейерабендом «постулатом значения» «***mach, was Du willst***» [21]) ведет к необходимости с одинаковым вниманием относиться к каждому из получаемых результатов экспертизы. Таким образом, возникает проблема сравнения (например, в стремлении выбрать *наилучший*[22]) порождаемых результатов ИАД, которая разумными средствами может быть формализована и решена, по-видимому, лишь при использовании «*однородных*» средств анализа данных[23].

(10) Понимая[24] науку как *систему ответов* на *вопросы*, формируемых с помощью *метода*, который характеризуется как определенными *дескриптивными* (описательными), так и *аргументационными* (обеспечивающими *достаточные*

---

[19] Т.е. выразимость средствами этого языка всех критически важных характеристик и особенностей, присущих изучаемому явлению (например, каузальных факторов влияния, определяющих наличие или же, наоборот, отсутствие анализируемого явления).

[20] Формировавшийся еще со времен Аристотеля ([*Аристотель 81*] и др.), а в рассматриваемой нами области существенным образом опирающийся на каноническую точку зрения, предложенную Ф.Бэконом [*Бэкон 02*] и Д.С.Миллем (см., например, [*Милль 00, Милль 64*] и др.), которая в дальнейшем развивалась и дорабатывалась целым рядом блестящих исследователей (см., например, [*ЭЭ 00*] и др.).

[21] Т.е. «*делай, что хочешь*» (в немецком варианте) или же «*anything goes*» - «*все подойдет*» (в английском варианте авторского текста П.Фейерабенда [*Feyerabend 75*]).

[22] Например, с такой точки зрения, которая учитывала бы как *описательную*, так и *предсказательную* силу соответствующих формализованный моделей.

[23] В случае использования «*неоднородных*» средств ИАД-экспертизы отдельных усилий требует собственно корректное «погружение» используемых «инструментов» в *единую* среду оценивания получаемых результатов. В отсутствие такой среды (а также корректно определенного в ней отношения предпочтения) следует быть готовым оказаться в ситуации, когда потребуется сравнивать, например, «*мокрое*» с «*зеленым*» или же «*горячее*» с «*длинным*».

[24] Вслед за А.С.Есениным-Вольпиным и В.К.Финном.

основания для принятия порождаемого результата) возможностями, при решении прикладных задач ИАД представляется естественным особое внимание уделить подбору эффективных процедур собственно формирования зависимостей из фактов. При этом используемый в таких процедурах механизм формирования и оценивания порождаемых результатов не должен зависеть от конкретного способа (языка) представления исходных данных. Обсуждаемый механизм (представления данных) должен быть «встроен» в соответствующие процедуры ИАД: этот путь позволяет сформировать условия для сравнимости результатов экспертизы при варьировании[25] выбираемых средств представления данных. Таким образом можно способствовать выбору *рациональных* в смысле П.Бернайса (который считал, что «*рациональность в угадывании*» и есть «*предмет эвристики*» - см. [*Бернайс 00* и *Поппер 00-а*]) понятий для формирования языка представления данных а также релевантных этим дескриптивным «инструментам» средств экспертизы (процедур формирования эмпирических зависимостей из данных).

(11) Рациональности выбора обсуждаемых дескриптивных инструментов ИАД может способствовать использование проблемно-ориентированной системы языков описания изучаемого явления, формируемой, например, по следующему принципу:

- первичный (простейший) уровень этой системы формируется языками, построенными как множества соответствующих признаков (теоретико-множественное описание[26]);
- затем первичное теоретико-множественное описание может быть расширено использованием также и отношений (причем, как статических[27], так и динамических[28]) между (подмножествами) используемых признаков;
- далее, для детализации описаний «поведения» соответствующих признаков можно использовать те или иные шкалы[29] их изменения.

При этом переход от более простого к более сложным вариантам описания может быть использован, например, для устранения (за счет более детального отражения «природы» изучаемого явления соответствующими дескриптивными средствами) возникающих в ходе экспертизы «коллизий» (в частности – для разрешения эмпирических противоречий и т.п.).

(12) *Наследуемость* тех или иных коллекций ЭЗ, порождаемых в процессе ИАД, при варьировании как средств экспертизы, так и задействованных в ее рамках языков представления данных, будучи дополненной *устойчивостью* (наследуемостью) соответствующих эмпирических зависимостей и при расширении исходных данных новыми фактами, может служить естественным сигналом о том, что в процессе выполняемого интеллектуального анализа данных удалось «нащупать» переход от эмпирических зависимостей к *эмпирическим закономерностям* (см., например, [*Финн 04*] и др.).

(13) В свою очередь, ненаследуемость соответствующих коллекций ЭЗ при расширении исходных данных новыми примерами и контрпримерами – явный сигнал о необходимости продолжить «настройку» используемых средств экспертизы в поиске такой их конфигурации, которая позволяла бы обеспечить достаточные основания для принятия соответствующих результатов. Ненаследуемость тех или иных коллекций ЭЗ – это также повод задействовать

---

[25] Например, с целью выбора наиболее подходящих.

[26] Возможно – с учетом кратности вхождения соответствующих дескрипторов в описания анализируемых объектов (язык так называемых нумерованных множеств).

[27] Совместно встречающиеся признаки.

[28] Последовательности (цепочки, направленные сети) встречаемости признаков (и\или же их комбинаций).

[29] Наименований, порядковые, метрические.

«интеллектуальные возможности» выполняющего ИАД эксперта[30] с целью проанализировать выявленные коллизии (в частности – попытаться локализовать, если это возможно, те факторы, которые определяют такую ненаследуемость, реконструировать соответствующие абдуктивные объяснения в смысле Ч.С.Пирса, сформировать на базе подобных факторов и объяснений подходящие *контриндуктивные* заключения в смысле П.Фейерабенда и т.п.).

(14) Наконец, ненаследуемость эмпирических зависимостей, порождаемых в рамках ИАД на конкретном фактическом материале (в конкретной исследовательской ситуации), также может быть сигналом о проблемах следующих трех типов:

(14.1) мало данных для формирования обоснованных заключений о всех имеющихся (в этих данных) ЭЗ[31];

(14.2) обнаружены критически важные «дефекты» в используемых представлениях имеющихся исходных данных (дефекты выразительных возможностей выбранного языка). Необходимо менять задействованный дескриптивный «инструментарий» ИАД;

(14.3) используемые средства осуществления ИАД неадекватны «природе» анализируемых данных, например, мы пытаемся изучать *другую* по своей физической природе *реальность* (например, каузальные зависимости, характеризующие броуновское движение частиц[32] и т.п.) нерелевантными ей инструментами познания (характерными для мира «*структурируемой* каузальности[33]»).

В заключение – несколько соображений о соотносимости сформулированных нами выше комментариев со свойства и функциональными характеристиками ДСМ-метода восстановления зависимостей из данных как платформы интеллектуального анализа данных.

Развитый в рамках ДСМ-метода инструментарий средств поддержки ИАД позволяет сформировать собственный конструктивный вариант действий в контексте каждого из 14 представленных выше аспектов обсуждаемой нами группы проблем, предлагая при этом свою версию ответов на каждый из вопросов (i) – (v). Более того:

- используя формируемую по описанной выше, в п.(11), схеме иерархию ДСМ-языков представления данных, затем
- подключая сюда дополнительно возможности варьировать выбор соответствующего языка с целью проверить наследуемость порождаемых на его основе эмпирических зависимостей (в том числе – порождаемых различными средствами ДСМ-экспертизы, варьируемыми в рамках выбора возможных стратегий ДСМ-рассуждения),
- а затем в случае выявления коллизий (ненаследуемости ЭЗ, возникновения эмпирических противоречий и т.п.), используя[34] контриндуктивный (в смысле П.Фейерабенда) анализ факторов, «ответственных» за соответствующие конкретным ситуациям эффекты ненаследуемости ЭЗ,

можно сформировать эффективный «процедурный механизм» целенаправленного управляемого *подбора подходящего* (т.е. дающего результаты, которые признаются

---

[30] «Усиливаемые» ассистирующей ему интеллектуальной компьютерной системой.

[31] Накопленные данные пока что не содержат сведений (фактов – примеров и контрпримеров), которые были бы *достаточны* для восстановления из них всех зависимостей, представленных в этих данных неявным образом.

[32] Где предметом рассмотрения может быть лишь соответствующая *совокупность* объектов (см., например, модели статистической физики [*Einstein 05 , Ланду 76*] и т.п.).

[33] Где анализ причинности может быть подвергнут *декомпозиции* от *комплексов* описаний объектов (наборов факторов) к их *значимым коллекциям* (*определенным совокупностям* тех и только тех каузальных факторов, которые и определяют наличие или же, наоборот, - отсутствие изучаемого явления) и далее – к *элементарным* каузальным «*компонентам*» (факторам, оказывающим «самостоятельное» влияние на изучаемое явление).

[34] Разумеется, уже в «ручном» режиме, где ведущая роль – у эксперта, выполняющего ИАД.

экспертами приемлемыми для описания изучаемого явления в конкретной прикладной ситуации) *приближенного описания* «природы» анализируемого средствами ДСМ-ИАД отношения **ОБЪЕКТ => СВОЙСТВА**[35].

Если же путем предлагаемого варьирования и оптимизации выбора средств ДСМ-ИАД (средств представления данных а также «инструментов» ДСМ-экспертизы) искомые области «устойчивости» восстанавливаемых из фактов ЭЗ - т.е. области *Эмпирических Закономерностей* - идентифицировать все же не удается, следует проанализировать две возможности принятия решений о дальнейших действиях:

а) в выполняемом процессе ИАД обработан все еще недостаточный (для обоснованности порождаемых заключений) объем данных, таким образом, процесс ДСМ-экспертизы следует продолжить (последовательно расширяя используемую выборку исходных примеров и контрпримеров новыми фактами);

в) возможно, мы пытаемся анализировать (релевантными миру *детерминистских каузальных влияний* средствами) явления, которые характерны для мира (реальности) совсем иного типа: по-видимому, это - не мир направленных на конкретный объект («персонифицированных») детерминистских причинных зависимостей и обусловленных ими каузальных влияний. (Примером «другой реальности» подобного типа может служить, в частности, уже упоминавшийся нами мир реальностей статистической физики). В этом случае, продолжать процесс ИАД прежними средствами, по всей видимости, не следует.

[35] Ср., например, с выбором «*объяснительных*» - а, вообще говоря, вовсе не *каузальных* – переменных при построении аппроксимаций числовых функций регрессионными зависимостями.

Заключение

Полученные в рассматриваемой диссертационной работе результаты позволяют при создании компьютерных систем анализа данных использовать современные методы искусственного интеллекта и корректные формализации исследовательских эвристик, формируемые средствами алгоритмических конструкций в том числе и высокой вычислительной сложности. Таким образом, имеются основания говорить о реализации идей, моделей и методов теоретической информатики в компьютерном инструментарии решения сложных задач анализа данных в прикладных предметных областях.

Результаты, выносимые на защиту:

- Разработка (обеспечивающей унифицированные возможности оперировать различными типами данных) ***алгебраической формализации*** для используемой процедурной конструкции извлечения зависимостей из структурных описаний прецедентов, исходно заданных для обучения.
- Идентификация и детальный анализ ***специального класса комбинаторных объектов*** (диаграмм взаимной вложимости классов эквивалентности описаний объектов-прецедентов), ***определяющих особенности алгоритмики*** порождения эмпирических зависимостей предлагаемым способом.
- Расширение процедурной конструкции ДСМ-метода на ***новые типы данных*** (структурные описания ***нечисловых объектов***, дополненные ***числовыми значениями*** существенных параметров).
- Доказательство ***оценок сложности вычислений***, требуемых для решения основных переборных задач, которые возникают в процессе порождения диаграмм вложимости классов эквивалентности при обработке ***различных типов данных*** (вариантов описаний исходно заданных прецедентов).
- Разработка метода ***дискретной оптимизации*** и ***управления*** комбинаторным ***перебором*** вариантов при формировании рассматриваемых диаграмм и анализе переносимости порождаемых эмпирических зависимостей на описания новых - ранее еще не изученных - прецедентов (***метода последовательных приближений*** при формировании диаграмм вложимости, реализуемого в том числе и в режиме ***параллельных вычислений***, что в совокупности позволяет ***снять ограничения*** на ***размеры*** эффективно обрабатываемых ***исходных коллекций прецедентов***).
- ***Обоснование корректности*** процедурной конструкции метода последовательных приближений, предложенного для формирования диаграмм вложимости классов эквивалентности прецедентов.
- ***Демонстрация работоспособности*** предложенных математических моделей, методов и алгоритмов при решении прикладных задач в различных предметных областях (решение прикладных задач интеллектуального анализа данных средствами ДСМ-метода автоматического порождения зависимостей).

Перспективы дальнейших разработок:

1. Развитие моделей и средств анализа устойчивости порождаемых эмпирических зависимостей при расширении исходной выборки прецедентов и варьировании средств их описания (развитие моделей и средств «опознания» *эмпирических закономерностей*).
2. Учет динамики (возможных изменений в описаниях анализируемых объектов или в структуре заданных для обучения коллекций прецедентов).
3. Распространение развитой техники ДСМ-ИАД на коллекции последовательно расширяемых новыми элементами наборов прецедентов (учет явного вида последовательностей пополнения базы фактов новыми прецедентами. Анализ возможностей организации эффективного мониторинга соответствующих изменений в исходной структуре классов эквивалентности).
4. Развитие средств ранжирования зависимостей, порождаемых из исходно заданных эмпирических данных. Учет имеющихся сведений о решетках ДСМ-правил правдоподобного вывода и стратегиях (комбинациях используемых правил) правдоподобного вывода при формировании оценок качества и упорядочениях по «силе» порождаемых эмпирических зависимостей.
5. Развитие взаимодействия с прикладными областями. Адаптация предложенной техники приближенных ДСМ-вычислений для эффективного использования в некоторых специальных прикладных предметных областях (медицинская и техническая диагностика, анализ социологических данных и др.).

Список рисунков и таблиц

Рис. 3.1.2.1. Уточнение понятия *сопутствующие изменения*……..	…...77
Рис. 3.1.2.2. Сопутствующие изменения и свойства новых объектов…….	…...78
Рис. 3.1.2.3. Отношение **ОБЪЕКТ=>$_1$СВОЙСТВА** для числовых данных…….	…...78
Рис. 3.1.2.4. Матрица-универсум **X**, монотонные подматрицы и м-башни в **X**.	……79
Рис. 4.2.1. Диаграмма **D_*GC$_{s*g}$*(U,Ω)** в **Примере 4.2.38**……...	..…158
Рис. 4.2.2. Диаграмма **D_*GC$_{s*g}$*(U,Ω)** в **Примере 4.2.39**……...	..…..159
Рис. 5.2.1. Пример архитектур типов α и β …….	……185
Рис. 5.2.2. Пример архитектуры типа γ ……	…...186
Рис. 5.2.3. Примеры встречаемости образующих…..……..	…...189
Рис. 5.2.4. Варианты раскрытия вершин гиперкуба примерами из исходного множества прецедентов……..……	…...194
Рис. 5.3.1. Графики функции f(x) = Const x ($x2^{n/x}$) для n ∈ {10, 20, … 100} и x∈(0,100] ……	…...201
Рис. 5.3.2. Графики функции f(x) = Const x ($x2^{y/x}$) для x ∈ (0, 100] и y∈[10,100]	…...202
Рис. 5.4.1. Диаграмма ***D*** = **D_*GC$_{s*g}$*(U,Ω)** для пары <**U**,**Ω**> из **Примера 5.4.3**…	…...215
Рис. 5.4.2. Псевдо-дерево из диаграммы ***D*** с объектом h_{17} в корне……	…...216
Рис. 5.4.3. Фрагмент архитектуры типа β в псевдо-дереве с Рис. 5.4.2…….	…...216
Таблица 6.3.2. Структура **СТОП-ЛИСТОВ** ..……...	…...237
Приложения	
Таблица П 2.1.1. Обзор результатов фармакологических приложений…….	……288
Таблица П 2.2.1 Анализ химической канцерогенности (пример исходных…… данных)	……292
Таблица П.2.3.1 Анализ токсичности противосудорожных соединений…… (исходные данные)	……294
Таблица П 2.3.2 Токсичность противосудорожных соединений (гипотезы)	……296
Таблица П 2.3.3 Токсичность противосудорожных соединений. Зависимости, порождаемые только на графах исходных соединений	……297
Таблица П.3.3.1. Тематическое индексирование системой ***ORACLE ConText*** рассматриваемого множества документов. (Векторы тем)……	…...317
Таблица П.3.3.2. Полный каталог тем, выделенных системой ***ORACLE ConText*** в рассматриваемом множестве документов……..…	……318
Таблица П.3.3.3 Полная система классов тематического сходства рассматриваемых документов……	…...319
Таблица П 4.2.1. Список анализируемых параметров……	…...334
Таблица П.4.2.2 Значения параметров для позитивных примеров……..……..	…...334
Таблица П.4.2.3 Значения параметров для негативных примеров…….	…...335
Таблица П.4.2.4 Примеры ДСМ-гипотез, порожденных по методу геологических аналогий…….	…...335
Рис. П 5.1.1. Схема управления, основанная на «балансовых» моделях…….	..…..337
Рис. П 5.1.2. Схема управления, основанная на семиотических моделях …..	..…..338
Таблица П 8.1.1 Исходная выборка прецедентов ……	…….372
Таблица П 8.1.2 Позитивные гипотезы ……	……372
Таблица П 8.1.3 Негативные гипотезы …….	…...373
Таблица П 8.1.4 Пополнение стартовой выборки прецедентов ……	…...374
Таблица П 8.1.5 Расширенная выборка прецедентов……..	...…374
Таблица П 8.1.6 Новые гипотезы, полученные на расширенной выборке……	…….375

Список литературы.

1. [АД]
   Анализ данных - ***Википедия***. - https://ru.wikipedia.org/wiki/Анализ_данных.
2. [Авидон 74]
   Авидон В.В., Лексина Л.А. Дескрипторный язык для анализа сходства химических структур органических соединений //***НТИ***, Сер.2. - 1974. - N3. - С.22-25.
3. [Аншаков 86]
   Аншаков О.М., Скворцов Д.П., Финн В.К. Логические средства экспертных систем типа ДСМ // ***Семиотика и информатика***. - Вып.28. - 1986. - С.65-102.
4. [Аншаков 87]
   Аншаков О.М., Скворцов Д.П., Финн В.К., Ивашко В.Г. Логические средства ДСМ-метода автоматического порождения гипотез: основные понятия и система правил вывода // ***НТИ***. Сер.2. - 1987. - N9. - С.10-18.
5. [Аншаков 93]
   Аншаков О.М., Скворцов Д.П., Финн В.К. О дедуктивной имитации некоторых вариантов ДСМ-метода автоматического порождения гипотез // ***Семиотика и информатика***. - Вып.33. - 1993. - С.164-233.
6. [Апресян 74]
   Апресян Ю.Д. *Лексическая семантика. Синонимические средства языка*. - М., 1974.
7. [Апресян 88]
   Апресян Ю.Д. и др. *Лингвистическое обеспечение системы ЭТАП-2*. - М.: Наука. - 1988.
8. [Апресян 92]
   Апресян Ю. Д. и др. Лингвистический процессор для сложных информационных систем. М.: Наука, 1992. - 256 С.
9. [Аристотель 81]
   Аристотель. *Сочинения в четырех томах*. - М: Мысль, 1981.
10. [Ахо 80]
    Ахо А., Хопкрофт Дж., Ульман Дж. *Построение и анализ вычислительных алгоритмов.* - М.: Мир. - 1979. - 536 С.
11. [Белоногов 10]
    Белоногов Г.Г., Хорошилов Александр А., Хорошилов Алексей А. Автоматизация составления англо-русских двуязычных фразеологических словарей по массивам двуязычных текстов. - ***Научно-техническая информация***. Сер. 2, Информационные процессы и системы. - 2010. - N 5. - Сс. 1-8.
12. [Баргесян 09]
    Баргесян А.А., М.С. Куприянов М.С., Холод И.И., Тесс М.Д., Елизаров С.И. Анализ данных и процессов. - СПб.: БХВ – Петербург, 2009. - 512 с.
13. [Бернайс 00]
    Бернайс П. О рациональности. - В кн.: Эволюционная эпистемология и логика социальных наук. Карл Поппер и его критики. – М.: Эдиториал УРСС. – 2000. – Сс. 154 – 162.
14. [Биркоф 76]
    Биркоф Г., Барти Т. *Современная прикладная алгебра*. - М: Мир. - 1976. - 400 с.
15. [Блинова 95]
    Блинова В.Г. О результатах и применении ДСМ-метода порождения гипотез к задачам анализа связи “структура химических соединений ==> биологическая активность” // ***НТИ***. - Сер.2. - 1995. - N5. - С.14-17.
16. [Бобылева 88]
    Бобылева Н.В., Ивашко В.Г., Краснова В.М., Финн В.К. Применение ДСМ-систем в задачах технической диагностики - ***Семиотические аспекты формализации интеллектуальной деятельности***: Тез. докл. школы-семинара "Боржоми-88". - М.: ВИНИТИ. 1988.- С.178-182.

17. [Боголюбов 83]
Боголюбов А. Н. Математики. Механики. Биографический справочник. — Киев: Наукова думка, 1983. — 639 С.
18. [Большаков 00]
Большаков И.А., Гельбух А.Ф. Модель «Смысл <=> Текст»: тридцать лет спустя. - ***J. Internat. Forum on Information and Documentation.*** - FID 519. - 2000, N1.- Сс. 10-16.
19. [Бочвар 38]
Бочвар Д.А. Об одном трехзначном исчислении и его применении к анализу парадоксов классического расширенного функционального исчисления // **Математический сборник**. - 1938. - Т.**4**. - N2. С.287-308.
20. [Будунова 93]
Будунова А.П., Поройков В.В., Блинова В.Г., Финн В.К. ДСМ-метод формирования гипотез: применение для анализа связи "структура - гепатопротекторная детоксицирующая активность". - ***НТИ***, Сер.2. - 1993. - N7. - С. 12-15.
21. [Бунге 68]
Бунге Н. Джон Стюарт Милль как экономист // **Журнал Министерства Народного Просвещения**. - 1868, N10, С.1-100.
22. [Бутлеров 53]
Бутлеров А.М. Сочинения. – М: Изд. АН СССР. –
    a. Т.1 Теоретические и экспериментальные работы по химии. - 1953г. -639 С.
    b. Т.2. Введение к полному изучению органической химии. - 1953г. - 624 С.
    c. Т.3 Научно-популярные, исторические, критико-библиографические и другие работы по химии. Путешествия. -1958г. - 429 С.
23. [Бэкон 35]
Бэкон Ф. *Новый органон.* - Л.: Соцэкгиз, 1935. - 382 С.
Бэкон Ф. Новый органон. – М.:Директ-Медиа. – 2002. – 384 С.
24. [Вапник 74]
Вапник В. Н., ***Червоненкис А. Я***. Теория распознавания образов. - М.: Наука, 1974. – 416 С.
25. [Вапник 79]
Вапник В. Н. Восстановление зависимостей по эмпирическим данным. - М.: Наука, 1979. – 448 С.
26. [Вапник 84]
Вапник В.Н., Глазкова Т.Г., Кощеев В.А., Михальский А.И., Червоненкис А.Я. Алгоритмы и программы восстановления зависимостей. - М.: Наука, 1984. - 815 С.
27. [ВВГУ]
Строев А.А., Забежайло И.М. К проблеме прогнозирования объемов остатков средств на счетах коммерческих банков. – ***Вестник Волгоградского Гос. Унив***. (в печати)
28. [Вершинин 91]
Вершинин К.П., Романенко И.Б. Распространение метода резолюций на некоторые недвузначные логики // *Семиотика и информатика*. - Вып.31. - 1991. - С.104-122.
29. [Виноградов 97]
Виноградов Д.В., Гиляревский Р.С., Забежайло М.И., Мельников Н.И., Столярова О.В. Об экспериментах по машинному обучению релевантности текста и темы // **Труды 3-й Международной конференции “НТИ-’97: Информационные ресурсы, интеграция, технологии**” (Москва, 26-28 ноября 1997 г.), М.: ВИНИТИ, Часть 1, С.59-61.
30. [Волкова 11]
Волкова А.Ю. Алгоритмизация процедур ДСМ–метода автоматического порождения гипотез. - ***НТИ***. Сер. 2. – 2011 – № 5 – Сс. 6–12;
31. [Волкова 11-а]
Волкова А.Ю. Анализ данных различных предметных областей с помощью процедур ДСМ–метода автомат. порождения гипотез. - ***НТИ***, сер. 2. – 2011. – № 6. – Сс. 9–18.

32. [Волкова 13]
Волкова А.Ю. Опыт создания интеллектуальной ДСМ-системы для исследования данных различных предметных областей. - ***НТИ***. Сер. 2. - 2013. – № 11. – Сс.12-26.
33. [Воронин 83]
Воронин Ю.А. Исследование операций при поисках и разведке месторождений полезных ископаемых. - Новосибирск: Наука (Сиб. отд.). -1983. - 291 С.
34. [Воронин 91]
Воронин Ю.А. *Начала теории сходства*. - Новосиб.: Наука (Сиб. отд.). -1991. - 128 С.
35. [Воронов 96]
Воронов Ю.С., Забежайло М.И., Смирных Е.А. О технологии и моделях процесса оптимизации оплаты документов при проведении клиринга // ***Открытые системы***. - 1996. - N5.
36. [Воронцов 04]
Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов. - ***Математические вопросы кибернетики*** (Под ред. О. Б. Лупанова).- М.: Физматлит, 2004. - Т. 13. – Сс. 5–36.
37. [Воронцов 04-а]
Воронцов К. В. Обзор современных исследований по проблеме качества обучения алгоритмов. - ***Таврический вестник информатики и математики***._ 2004, № 1. - Сс. 5–24.
38. [ГА]
Генетический алгоритм - ***Википедия***. – Генетический алгоритм. - https://ru.wikipedia.org/wiki/Генетический_алгоритм
39. [Гаджиев 12]
Гаджиев А.М. - Управление рентабельностью коммерческого предприятия в условиях неполноты информации. – Магистерская диссертация. – Факультет ВМК МГУ им. М.В.Ломоносова. – 2012. – 64 С.
40. [Гиляревский 88]
Гиляревский Р.С., **Субботин М.М.** О возможностях оценки перспективных новых информационных технологий// ***НТИ***, Сер.2, 1988, N12, С.2-5.
41. [Гиляревский 88-а]
Гиляревский Р.С., Калошин В.В. Тенденции развития информатики (по отечественным диссертациям от 1965 до 1980 гг.// ***НТИ***, Сер.2, 1988, N4, С.2-8.
42. [Гладков 06]
Гладков Л. А., Курейчик В. В., Курейчик В. М. Генетические алгоритмы: Учебное пособие. — М: Физматлит, 2006. — 320 С.
43. [Голендер 78]
Голендер В.Е., Розенблит А.Б. Вычислительные методы конструирования лекарств. - Рига: Зинатне. - 1978. - 232 С.
44. [Гретцер 82]
Гретцер Г. Общая теория решеток. – М.: Мир, 1982. – 456 С.
45. [Гусакова 87]
Гусакова С.М., Финн В.К. Сходства и правдоподобный вывод //***Изв. АН СССР***. Сер. Тех. Кибернет.-1987. - N 5. - С.42-63.
46. [Гусакова 04]).
Гусакова С.М. Подход к решению задачи атрибуции исторических источников с помощью ДСМ-метода. - ***Новости искусственного интеллекта***. – 2004, № 3. – С. 42-49.
47. [Гэри 82]
Гэри М., Джонсон Д.С. *Вычислительные машины и трудно-решаемые задачи.* - М.: Мир. - 1982. - 416 С.
48. [Долягин 86]
Долягин А.Б., Блинова В.Г., Забежайло М.И., Финн В.К., Хемпсон М.С., Воловик А.С., Никитина А.И., Капустина В.М., Писаренко Э.К. Исследование рационального пути

подбора химических соединений для испытания их на возбудителе рака картофеля. // ***Микология и фитопатология***. - 1986. - N6.

49. [ДСМ АПГ]
Автоматическое порождение гипотез в интеллектуальных системах. (Ред. - Финн В.К.). – М.: Либроком, 2009. – 528 С.
50. [ДИ]
Диаграммы Исикавы – ***Википедия***. - https://ru.wikipedia.org/wiki/Диаграмма_Исикавы
51. [Дьячков 90]
Дьячков П.Н. Квантово-химические расчеты в исследовании механизмов активности и токсичности химических соединений. - ***Итоги науки и техники. - Сер. Токсикология***. - Т.**16**. - ВИНИТИ. - 1990.
52. [Дюк 04]
Дюк В.А. Технологии Data Mining в медико-биологических исследованиях. - - ***Новости искусственного интеллекта***. – 2004, № 3. – С. 49-57.
53. [Дюк 01]
Дюк В.А., Самойленко А.П. Data Mining: учебный курс. – СПб: Питер, 2001. – 368 С.
54. [Емельянов 03]
Емельянов В. В., Курейчик В. В., Курейчик В. М. Теория и практика эволюционного моделирования. — М: Физматлит, 2003. — 432 С.
55. [Есенин 93]
Есенин-Вольпин А.С. Об антитрадиционной (ультраинтуиционистской) программе оснований математики и естественнонаучном мышлении // **Семиотика и информатика**. - Вып.**33**. - М.: ВИНИТИ. - 1993. - С.3-67.
56. [Есенин-Вольпин 99]
А.С. Есенин-Вольпин. Философия. Логика. Поэзия. Защита прав человека: Избранное. — М.: РГГУ, 1999. — 450 С.
57. [Журавлев 74-а]
Журавлев Ю.И. Алгоритмы построения минимальных ДНФ для функций алгебры логики // В кн.: *Дискретная математика и математические вопросы кибернетики.* - М.: Наука, 1974. - С. 67-98.
58. [Журавлев 74-б]
Журавлев Ю.И. Метрические свойства ДНФ // В кн.: *Дискретная математика и математические вопросы кибернетики.* - М.: Наука, 1974. - С. 99-148.
59. [Журавлев 77]
Журавлев Ю.И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов//***Кибернетика***. - 1977, N4, С.5-17.
60. [Журавлев 77а]
Журавлев Ю.И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов II// ***Кибернетика***. - 1977, N6, С.21-27.
61. [Журавлев 78]
Журавлев Ю.И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов III// ***Кибернетика***. - 1978, N2, С. 35-43.
62. [Журавлев 80]
Журавлев Ю.И. и др. Задачи распознавания и классификации со стандартной обучающей информацией // ***ЖВМиМФ***.-1980.-Т20,N5.-С.1294-1309.
63. [Журавлев 06]
Журавлев Ю.И., ***Рязанов В. В., Сенько О. В***. «Распознавание». Математические методы. Программная система. Практические применения. - М.: Фазис, 2006. – 176 С.
64. [Забежайло 82]
Забежайло М.И., Финн В.К., Козлова С.П., Катамадзе Т.Г., Авидон В.В., Рабинков А.А. Об одном методе автоматического формирования гипотез и его программной реализации // ***НТИ***, Сер.2. - 1982. - N4. - С.20-26.

65. [Забежайло 83]
Забежайло М.И., Финн В.К., Авидон В.В., Катамадзе Т.Г., Блинова В.Г., Бодягин Д.А., Рабинков А.А. Об экспериментах с базой данных с неполной информацией посредством ДСМ-метода автоматического порождения гипотез // ***НТИ***, Сер.2. - 1983. - N2. - С.28-32.
66. [Забежайло 84]
Забежайло М.И. Ненаследуемость эмпирического противоречия в ДСМ-методе и немонотонные рассуждения // ***НТИ***, Сер.2. - 1984. - N11. - С.14-17.
67. [Забежайло 87]
Забежайло М.И., Ивашко В.Г., Кузнецов С.О., Михеенкова М.А., Хазановский К.П., Аншаков О.М. Алгоритмические и программные средства ДСМ-метода автоматического порождения гипотез. // ***НТИ***. Сер.2. - 1987. - N10. - с.1-13.
68. [Забежайло 90]
Забежайло М. И. Новые информационные технологии и системы НТИ // Науч.-техн. информация. Сер. 2, Информационные процессы и системы. — 1990. — №5. — С. 2—9.
69. [Забежайло 90-а]
Забежайло М.И. Некоторые тенденции в развитии интеллектуальных систем // ***Программные продукты и системы***. - 1990. - N 4. - С.86-94.
70. [Забежайло 91]
Забежайло М.И. Интеллектуальные системы и задача восстановления эмпирических зависимостей структурно-числового характера // **Итоги науки и техники.** Сер." Информатика". Т.**15** "*Интеллектуальные информационные системы*" - М: ВИНИТИ. - 1991. - С. 102-114.
71. [Забежайло 92]
Забежайло М.И К проблеме расширения ДСМ-метода автоматического порождения гипотез на данные с числовыми параметрами // ***НТИ***. Сер.2. - 1992. - N11. - с.11-21
72. [Забежайло 92-а]
Забежайло М.И. Новые информационные технологии в научных исследованиях и технологических разработках // ***НТИ***. Сер. 2.- 1992.- № 6.-С.1-11.
73. [Забежайло 92-б]
Забежайло М.И. Интеллектуальные системы: на пути к новым поколениям. // ***Новости* искусственного *интеллекта***. N1, 1992. -Сс.8-24.
74. [Забежайло 93]
Забежайло М.И. К проблеме расширения ДСМ-метода автоматического порождения гипотез на данные с числовыми параметрами II // ***НТИ.*** Сер.2. - 1993. - N2. - с.6-16.
75. [Забежайло 93-а]
Забежайло М.И. (***Zabezhailo M.I.***) The extension of the JSM-method: plausible reasoning dealing with numeric data // ***Новости Искусственного Интеллекта***. Специальный выпуск: *Международная конференция по искусственному интеллекту* ***"Восток-Запад - 93"*** (**Artificial Intelligence News**. Special issue*: **EWAIC-93***, Moscow, 7-9 Sept. 1993). 1993.
76. [Забежайло 96]
Забежайло М.И. Формальные модели рассуждений в принятии решений: приложения ДСМ-метода в системах интеллектуального управления и автоматизации научных исследований // ***НТИ***, Сер.2. - 1996. - N5-6. - С.20-32.
77. [Забежайло 96-а]
Забежайло М.И. О комбинаторной природе одной задачи оптимизации // ***НТИ***, Сер.2. - 1996. - N1. - С.19-27.
78. [Забежайло 97]
Забежайло М.И. К проблеме формализации метода геологических аналогий. **// *Известия РАН. Сер. "Теория и системы управления*".** - 1997. - N2.- С.151-164.
79. [Забежайло 97-а]

Забежайло М.И., Емеленко М.Н., Липчинский Е.А., Максин М.В. К разработке платформенно-независимой версии программной системы, реализующей ДСМ-метод автоматического порождения гипотез// **Труды 3-й Международной конференции "НТИ-'97: Информационные ресурсы, интеграция, технологии"** (Москва, 26-28 ноября 1997 г.), М.: ВИНИТИ, Часть 1, С.91-93.

80. [Забежайло 98]
Забежайло М.И., Емеленко М.Н., Липчинский Е.А., Максин М.В. К пониманию термина "прикладная семиотика" // ***НТИ***, Сер.2 - 1998. - N1.- С.11-18.
81. [Забежайло 98-а]
Забежайло М.И. К проблеме автоматического понимания полнотекстовых документов в информационном поиске. **// *Известия РАН. Сер. "Теория и системы управления*".** - 1998. - N5. - С.167-176.
82. [Забежайло 98-б]
Забежайло М.И. Информационный поиск в полнотекстовых документах: некоторые проблемы и перспективы. **// *Новости искусственного интеллекта.*** - 1998. - N.1.- С. 24-59
83. [Забежайло 98-в]
Забежайло М.И. Интеллектуальный анализ данных – новое направление развития информационных технологий //***НТИ***. Сер. 2.- 1998. - №8. - С. 6-17.
84. [Забежайло 98-г]
Забежайло М.И. Data Mining & Knowledge Discovery in Data Bases: предметная область, задачи, методы и инструменты // *Труды 6-ой национальной конференции по искусственному интеллекту с международным участием*. - Пущино. - 1998. - Том 1. - С. 592-600.
85. [Забежайло 09]
Забежайло М.И. Банковский бизнес в России: индустрия или искусство? – Волгоград: Волгоградское научное издательство. – 2009. – 331 С.
86. [Забежайло 12]
Забежайло М.И. К проблеме быстрого анализа заголовков при адресации пакетов данных в компьютерных сетях. - ***НТИ***. Сер.2. – 2012. – №12, С.8-15.
87. [Забежайло 13]
Забежайло М.И. О функциональности отношения причинности, используемого в ДСМ-рассуждениях. – ***НТИ***. Сер.2. - 2013. – N7. - Сс.1-7.
88. [Забежайло 13-а]
Забежайло М.И. К задаче анализа вложимости подслов в заголовки пакетов данных. - ***Системы и средства информатики***. – 2013, Т.23, №1. – С. 58-68.
89. [Забежайло 14]
Забежайло М.И. О некоторых возможностях управления перебором в ДСМ-методе. - ***Искусственный интеллект и принятие решений***. – 2014. – Часть I: № 1, С.95 -110. – Часть II: № 3, С.3 - 21.
90. [Забежайло 14-а]
Забежайло М.И., Синякова Е.В. К вопросу об «интеллектуальности» интеллектуального анализа данных. – ***НТИ***. Сер. 2, Информационные процессы и системы.- 2014.- № 3.-С.1-9.
91. [Забежайло 14-б]
Забежайло М.И. Приближенный ДСМ-метод на примерах. – ***НТИ.*** Сер.2. – 2014. – №10, С. 1-12.
92. [Забежайло 15]
Забежайло М.И. К вопросу о достаточности оснований для принятия результатов интеллектуального анализа данных средствами ДСМ-метода. – ***НТИ.*** Сер.2.– 2015.– №1. – С.1-9.
93. [Забежайло 15-а]

Забежайло М.И. О некоторых оценках сложности вычислений в ДСМ-рассуждениях. - ***Искусственный интеллект и принятие решений***. - 2015. – Часть I: №1, С.3-17. Часть II: №2. С.3-17.

94. [Загоруйко 99]
Загоруйко Н. Г. Прикладные методы анализа данных и знаний. - Новосибирск: ИМ СО РАН, 1999.- 270 С.
95. [Зализняк 00].
Зализняк А.А., Янин В.Л. Новгородские грамоты на бересте (из раскопок 1990-1996 гг.) – М: Русские словари, 2000. – 430 С.
96. [Зенкин 91]
Зенкин А.А. *Когнитивная компьютерная графика*. - М.: Наука. - 1991. - 188 С.
97. [ИАД]
Интеллектуальный анализ данных. – ***Википедия***. http://ru.wikipedia.org/wiki/Data_mining
98. [Ивашко 84]
Ивашко В.Г. Об одной формализации индуктивных рассуждений для экспертных информационных систем // ***НТИ***. Сер.2. - 1984. - N8. - С.28-32.
99. [Ивашко 86]
Ивашко В.Г., Финн В.К. Экспертные системы и некоторые проблемы их интеллектуализации// В сб.: ***Семиотика и информатика***. Вып. **27** (1986). - М.: ВИНИТИ. - С.25-61.
100. [ИНС]
Искусственная нейронная сеть. – ***Википедия***. – Искусственная нейронная сеть. - https://ru.wikipedia.org/wiki/Искусственная_нейронная_сеть
101. [ИНТ 91]
Интеллектуальные информационные системы/ **Итоги науки и техники.** Сер. Информатика - Т. **15** (ред. - ***Финн В.К., Черный А.И***.) - М: ВИНИТИ. - 1991.
102. [ИИС 90]
Искусственный интеллект: справочник. (Под редакцией ***Э.В.Попова***). М.: Радио и связь. - 1990. - Т.**1**: -464 С., т.**2**: - 304 С., т.**3**: - 363 С.
103. [Кайберг 78]
Кайберг Г. *Вероятность и индуктивная логика***.** - М.: Прогресс, 1978. - 374 С**.**
104. [Кант 66]
Кант И. Сочинения в шести томах. - М: Мысль, 1966.
105. [Кендал 76]
Кендал Дж.М., Стюарт А. Многомерный статистический анализ.-М.:Наука, 1976. – 736С.
106. [Клайн 84]
Клайн М. Математика. Утрата определенности. - М.: Мир. - 1984. - 446 С.
107. [Клайн 88]
Клайн М. Математика. Поиск истины. - М.: Мир. - 1988. - 295 С.
108. [Кон 68]
Кон П. Универсальная алгебра. - М.: Мир, 1968. – 359 С.
109. [Кохонен 80]
Кохонен Т. Ассоциативная память - М.: Мир. - 1980. - 239 С.
110. [Кочетков 89]
Кочетков Д.В. Распознающие алгоритмы, инвариантные относительно преобразований пространства признаков. - ***Распознавание, классификация, прогноз: Математические методы и их применение***. - М.: Наука. 1989. - Вып. 11. - Сс. 178-206.
111. [Круглов 01]
Круглов В. В., Борисов В. В. Искусственные нейронные сети. Теория и практика. - М.: Горячая линия – Телеком, 2001. – 382 С.
112. [Кузнецов 86]
Кузнецов А.М. *От компонентного анализа к компонентному синтезу*.- М., 1986.– 126 С.

113.[Кузнецов 97]
Кузнецов Б.А., Солнцева Е.К., Закамская Д.В., Леонтьев А.А., Деревянкин М.В., Быховский Д.В., Ашкинадзе Б.Л. Интеллектуальный поиск в текстовых БД с помощью системы ДИАНА // **Труды 3-й Международной конференции "НТИ-'97: Информационные ресурсы, интеграция, технологии"** (Москва, 26-28 ноября 1997 г.), М.: ВИНИТИ, Часть 1, С.131-135.

114.[Кузнецов 91]
Кузнецов С.О. ДСМ-метод как система аксиоматического обучения// **Итоги науки и техники.** Сер." Информатика". Т.**15** "*Интеллектуальные информационные системы"* - М: ВИНИТИ - 1991. - С. 17 - 53.

115.[Кузнецов 91-а]
Кузнецов С.О. Модели и методы автоматического обучения // ***Итоги науки и техники.*** Сер. Вычислительные науки. Т.**7**. - М.: ВИНИТИ. - 1991. - С. 89 - 139.

116.[Кузнецов 93]
Кузнецов С.О. Быстрый алгоритм построения всех пересечений объектов из конечной полурешётки. - ***НТИ***. Сер.2. 1993. - №1. Сс. 17-20.

117.[Лакатос 67]
Лакатос И. Доказательства и опровержения: Как доказываются теоремы. – М.: Наука. - 1967. - 152 С.

118.[Ландау 76]
Ландау Л.Д., Лифшиц Е.М. Теоретическая физики.Т.5: Статистическая физика (Часть I). - М.; Наука. – 1976. – 584 С.

119.[Ланкастер 72]
Ланкастер Ф. Информационно-поисковые системы / М.: Мир, 1972. – 308 С.

120.[Ларичев 87]
Ларичев О.И. *Объективные модели и субъективные решения* / М.: Наука, 1987.

121.[Ларичев 96]
Ларичев О.И., Мошкович Е.М. *Качественные методы принятия решений* / М.: Наука, 1996.

122.[Лаудон 81]
Лаудон Г. ЭВМ и машинные методы в геологии. - М: Мир.- 1981.- 352 С.

123.[Лахути 97]
Лахути Д.Г., Рубашкин В.Ш. Семантический словарь как инструмент информационных технологий // **Труды 3-й Международной конференции "НТИ-'97: Информационные ресурсы, интеграция, технологии"** (Москва, 26-28 ноября 1997 г.), М.: ВИНИТИ, Часть 2, С.7-12.

124.[Лейбов 90]
Лейбов А.Е. Алгоритмы преобразования графов, используемые для правдоподобных выводов и интеллектуального интерфейса в экспертных системах // **II Всесоюзная конференция** "*Искусственный интеллект 90*", Минск 1990. - Т.1. - С.141-143.

125.[Лейбов 91]
Лейбов А.Е. Автоматическое кодирование химических структур кодом ФКСП.// **Итоги науки и техники.** Сер. "Информатика" - Т.**15** "*Интеллектуальные информационные системы*" - М: ВИНИТИ. - 1991.- С.141-158.

126.[Лейбов 94]
Лейбов А.Е. Программные инструменты представления знаний для решателей типа "правдоподобный вывод + достоверный вывод" в предметных областях со сложно структурированными объектами // **IY Национальная конференция** с международным участием "*Искусственный интеллект - 94*", Рыбинск, 1994, Т.2. - С.278-281.

127.[Лотман 92]
Лотман Ю.М. *Культура и взрыв*. - М.: "Гнозис"(Издательская группа "Прогресс"). - 1992. - 272 С.

128.[Мазур 01]
Мазур И.И., Шапиро В.Д. Управление проектами. Справочник для профессионалов. – М.: Высшая школа, 2001. – 875 С.
129.[Магинскас 63]
Магинскас Ю.А. Об одном подходе к анализу причинности // В кн.: **Проблемы логики.** - М.: АН СССР. - 1963. - С. 144-150.
130.[Маренков 06]
Маренков Н.Л. Бухгалтерская отчетность и международные стандарты. – М.: Флинта: МПСИ , 2006. – 272 с.
131.[Маслов 64]
Маслов С.Ю. Обратный метод установления выводимости в классическом исчислении // **ДАН СССР**, 1964, **159**, N1, с.17-20.
132.[Маслов 68]
Маслов С.Ю. Обратный метод установления выводимости для логических исчислений // **Труды МИ АН СССР им. В.А.Стеклова**. - **98**. - М.: Наука, 1968. - С.26-87.
133.[Мельников 91]
Мельников Н.И. О новой версии СУБД для систем, основанных на знаниях // **Итоги науки и техники.** Сер. "Информатика" - Т.**15** "*Интеллектуальные информационные системы*" - М: ВИНИТИ. - 1991.- С.159-169.
134.[Мельчук 95]
Мельчук И.А. Русский язык в модели «Смысл ⇔ Текст». - "Языки русской культуры". Москва - Вена. 1995. - 682 С.
135.[Мельчук 99]
Мельчук И.А. Опыт теории лингвистических моделей «Смысл ⇔ Текст». М., 1974 (2-е изд., 1999). М.: Школа «Языки русской культуры», 1999 - 346 С.
136.[Меркурьева 91]
Меркурьева Г.В., Меркурьев Ю.А. Экспертные системы имитационного моделирования // ***Изв. АН СССР***. Сер. Техн. Кибернет. - 1991. - N3. - С. 156 - 173.
137.[Милль 00]
Милль Д.С. *Система логики силлогистической и индуктивной. Изложение принципов доказательства в связи с методами научного познания.* - Пер. С англ. под ред. и с предисл. приват-доц. Имп. Московского Унив. В.Н.Ивановского. - М.: Изд. маг. “Книжное Дело”, 1900. - XXX, 781С.
138.[Милль 07]
Милль Д.С. Об определении предмета политической экономии; и о методе исследования, свойственном ей. - В книге: Основы политической экономии. – М.: Эксмо. – 2007. – Сс. 985-1023.
139.[Милль 64]
Милль Дж.С. *Рассуждения и исследования*. (В трех частях). - С.-Пб.: Типография А.С.Голицына, 1864-65.
140.[Михеенкова 86]
Михеенкова М.А., Финн В.К. Об одном классе экспертных систем с неполной информацией // ***Известия АН СССР,*** сер. **Техническая кибернетика**. - 1986. -N5. - С.82-103.
141.[Михеенкова 87]
Михеенкова М.А. Об экспертных системах типа ДСМ // ***НТИ.*** Сер.2. - 1987. - N9. - С.23-31.
142.[Михеенкова 97]
Михеенкова М.А. ДСМ-метод правдоподобного рассуждения как средство анализа социального поведения // ***Известия РАН.*** Сер. Теория и системы управления. - 1997. - N5. - С.62-70.
143.[Михеенкова 98]

Михеенкова М.А. *Развитие ДСМ-метода автоматического порождения гипотез для его применения при анализе социологических данных типа "СУБЪЕКТ==>ПОВЕДЕНИЕ".* - Автореферат диссертации на соискание ученой степени кандидата технических наук. - М.: ВИНИТИ, 1998. - 28 С.

144.[Михеенкова 13]
Михеенкова М.А., Волкова А.Ю. Спецификация интеллектуальной системы типа ДСМ/ - ***НТИ***. Сер. 2. – 2013. – № 7. – Сс. 5-19.

145.[МЛиП]
Многозначные логики и их применения. (Ред. – Финн В.К.). – М: Издательство ЛКИ, 2008. – Т.1 (Логические исчисления, алгебры и функциональные свойства). – 502 С., Т.2 (Логики в системах искусственного интеллекта). - 240 С.

146.[Моисеев 93]
Моисеев Н.Н. *Восхождение к Разуму*. - М.:ИздАТ. - 1993.

147.[Моисеев 98]
Моисеев Н.Н. *Расставание с простотой*. - М.:Аграф. - 1998. - 474 С.

148.[Мюллер 87]
Мюллер В.К. *Англо-Русский Словарь*.Изд.21. - М.: "Русский язык". - 1987. - 843С.

149.[НИИ 04]
Новости искусственного интеллекта. – 2004, №3. – М.: Российская Ассоциация Искусственного Интеллекта.

150.[Нильсон 73]
Нильсон Н. *Искусственный интеллект*. - М.: Мир. - 1973. - 270 С.

151.[НТИ 95]
Науч.-Тех. Информация. - Сер2. - 1995. - N.5 (Спец. вып.: ***Интеллектуальные системы***).

152.[Ньюэлл 67]
Ньюэлл А., Саймон Г. GPS-программа, моделирующая процесс человеческого мышления // В кн.: Вычислительные машины и мышление. - М.: Мир. - 1967. - С.283 - 300.

153.[Падучева 05]
Падучева Е.В. Соответствие «Смысл ⇔ Текст» в исторической перспективе. // Восток - Запад: Вторая международная конференция по модели «Смысл ⇔ Текст». - М.: Языки славянской культуры, 2005. - С. 330-349

154.[Паклин 12]
Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям. — СПб: Изд. Питер, 2009. — 624 с.

155.[ПЗ 84]
Представление знаний в человеко-машинных и робототехнических системах. - М.: ВЦ и ВИНИТИ АН СССР. - 1984. - Т.**1**: - 261 С., т.**2**: - 236 С., т.**3**: - 380 С., т.**4**: - 291 С.

156.[Пирс 00]
Пирс Ч.С. Избранные философские произведения. – М.: Логос. – 2000. – 448 С.

157.[Пирс 05]
Пирс Ч.С. Рассуждение и логика вещей. Лекции для Кембриджских конференций 1898 года. – М.: РГГУ. – 2005. – 371 С.

158.[Панкратова 95]
Панкратова Е.С. Задача прогнозирования результирующего эффекта от введения в организм двух химических веществ и ее решение средствами ДСМ-системы // ***НТИ***. - Сер.2. -1995. - N5. - С. 24-29.

159.[Пойа 70]
Пойа Д. *Математическое открытие*. - М.: Наука, 1970. - 452С.

160.[Пойа 75]
Пойа Д. *Математика и правдоподобные рассуждения*. - М.: Наука, 1975. - 463С.

161.[Полак 74]

Полак Э. *Численные методы оптимизации.* - М.: Мир. - 1974. - 376 С.

162.[Попов 87]
Попов Э.В. Экспертные системы, решение неформализованных задач в диалоге с ЭВМ. - М: «Наука». - 1987. - 284 С.

163.[Поппер 83]
Поппер К. *Логика и рост научного знания* - М., 1983.

164.[Поппер 93]
Поппер К. Нищета историцизма. - М.: Путь. – 1993. – 188 С.

165.[Поппер 04]
Поппер К. Предположения и опровержения. Рост научного знания**.**– М.:Аст.- 2004.- 638С.

166.[Поппер 02]
Поппер К. Нормальная наука и опасности, связанные с ней. - В кн.: Структура научных революций. – М.: Аст. – 2002. – Сс. 525- 538.

167.[Поппер 00]
Поппер К. Логика социальных наук. - В кн.: Эволюционная эпистемология и логика социальных наук. Карл Поппер и его критики. – М.: Эдиториал УРСС. – 2000. – Сс. 298 – 313.

168.[Поппер 00-а]
Поппер К. Призыв Бернайса к более широкому пониманию рациональности. - В кн.: Эволюционная эпистемология и логика социальных наук. Карл Поппер и его критики. – М.: Эдиториал УРСС. – 2000. – Сс. 163 – 175.

169.[Поспелов 81]
Поспелов Д.А. *Логико-лингвистические модели в системах управления*. - М.: Энергоиздат. - 1981. - 232 С.

170.[Поспелов 86]
Поспелов Д.А. *Ситуационное управление - теория и практика*. - М.: Наука. - 1986.

171.[Пролог]
Пролог (язык программирования). – ***Википедия***. - https://ru.wikipedia.org/wiki/Пролог_(язык_программирования)

172.[Раевский 88]
Раевский О.А., Сапегин А.М. Возможности и перспективы конструирования биологически активных веществ// ***Успехи химии***. - 1988. - **57**, N9. - С. 1565-1586.

173.[РКП 89]
Распознавание, классификация, прогноз: математические методы и их применение. / Вып. 1, 2. (под. ред. ***Ю.И.Журавлева***). - М.: Наука, 1989.

174.[Робинсон 70]
Робинсон Дж. Машинно-ориентированная логика, основанная на принципе резолюции // **Кибернетический сборник** (нов. сер.). - **7**. - М.: Мир. - 1970.

175.[Рождественский 67]
Рождественский Н.Н. *О значении Милля в ряду современных экономистов*. - С.-Пб.: Типография Э.Праца, 1867.

176.[Рутковская 06]
Рутковская Д., Пилиньский М., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы. – М.: Горячая линия – Телеком, 2006. – 452 С.

177.[Салле 78]
Салле К., Дебизер Ж. Формирование нефтяных залежей. - М.: Недра, 1978. - 245 С.

178.[САФИД 85]
Труды Школы-семинара "Семиотические аспекты формализации интеллектуальной деятельности" (Кутаиси, 1985). - М.: ВИНИТИ, 1985.

179.[САФИД 88]

Труды Школы-семинара "Семиотические аспекты формализации интеллектуальной деятельности" (Боржоми, 1988). - М.: ВИНИТИ, 1988. - 420 С.

180.[Сапегин 92]
Сапегин А.М., Раевский О.А., Зефиров Н.С., Финн В.К., Скитер П.В. Гибридная интеллектуальная система молекулярного дизайна// Тез. докл. **9 Всес. конф. по химической информатике** (г.Черноголовка, Московская обл., 11-15 янв. 1992).

181.[Себер 80]
Себер Дж. *Линейный регрессионный анализ*. - М.: Мир, 1980.

182.[Семенихин 14]
Семенихин А.С. IBM Watson: выявление скрытых взаимосвязей. – **AINL-2014** (Moscow, Skolkovo). – 2014, Sept. 13. - http://www.slideshare.net/AINL/ibm-watson-39104971

183.[Симонс 85]
Симонс Дж. *ЭВМ пятого поколения: компьютеры 90-х годов.* - М.: Фин. и стат., 1985.

184.[Синякова 13]
Синякова Е.В. Использование методов интеллектуального анализа данных для решения задачи прогнозирования показателей производственных процессов. – Магистерская диссертация. – Факультет ВМК МГУ им. М.В.Ломоносова. – 2013. – 135 С.

185.[Скворцов 83]
Скворцов Д.П. О некоторых способах построения логических языков с кванторами по кортежам \\ ***Семиотика и информатика***. - Вып.**20**. - М.:ВИНИТИ, 1983. - С.102-126.

186.[Слэйгл 73]
Слэйгл Дж. *Искусственный интеллект*. - М.: Мир. - 1973. - 319 с.

187.[Смелянский 11]
Смелянский Р.Л. Компьютерные сети (в 2-х томах). - М: Издательский центр "Академия", 2011. (Том 1. ***Системы передачи данных***. - 304 С. . Том 2. ***Сети ЭВМ***. - 240 С.)

188.[Сокал 80]
Сокал Р.Р. Кластер-анализ и классификация. Предпосылки и новые направления. - В сб.: **Классификация и кластер**. - М: Мир. - 1980. - С. 7-19.

189.[СТ 12]
Смыслы, тексты и другие захватывающие сюжеты (Сборник статей в честь 80-летия Игоря Александровича Мельчука). - Москва: Языки славянской культуры, 2012. - 703 С.

190.[Степанов 75]
Степанов Ю.С. *Методы и принципы современной лингвистики*. - М., 1975.

191.[Субботин 94]
Субботин М.М. Гипертекст. Новая форма письменной коммуникации//**Итоги науки и техники**. Сер. Информатика/ ВИНИТИ - 1994. - Т.18. - 158 С.

192.[Тихонов 43]
Тихонов А. Н. «Об устойчивости обратных задач». - ***Докл. АН СССР***. 1943. Т. 39. № 5. С. 195—198.

193.[Тихонов 69]
Тихонов А.Н. Корректные и некорректные задачи. - Большая советская энциклопедия (В 30 т. / гл. ред. А. М. Прохоров. — 3-е изд.). — М.: Советская энциклопедия, 1969—1978.

194.[Фабрикантова 91]
Фабрикантова Е.Ф. Проблемы компьютерного моделирования метаболизма чужеродных веществ в организме человека// **Итоги науки и техники.** Сер. "Информатика" - Т.**15** "*Интеллектуальные информационные системы"* - М: ВИНИТИ. - 1991. - С. 115 - 135.

195.[Фабрикантова 93]

Фабрикантова Е.Ф. Описание моделей метаболизма веществ в виде квази-аксиоматических теорий // ***Новости Искусственного Интеллекта***. Специальный выпуск: *Международная конференция по искусственному интеллекту* ***"Восток-Запад - 93"*** (**Artificial Intelligence News**. Special issue*:* ***EWAIC-93***, Moscow, 7-9 Sept. 1993). 1993.

196.[Филимонов 06]
Филимонов Д.А., В. В. Поройков В.В. Прогноз спектра биологической активности органических соединений. - ***Российский химический журнал*** (***Журнал Росссийского хим. об-ва им. Д.И.Менделеева***), 2006, L (2). - Сс. 66-75.

197.[Финн 76]
Финн В.К. О возможностях формализации правдоподобных рассуждений средствами многозначных логик // В кн.: *VII Всес. симпозиум по логике и методологии науки.* - Киев: Наукова Думка, 1976. - С. 82-83.

198.[Финн 80]
Финн В.К., Аншаков О.М., Григолия Р.Ш., Забежайло М.И. Многозначные логики как фрагменты формализованной семантики. – ***Семиотика и информатика***. Вып. 15. 1980. М.: ВИНИТИ. – Сс. 27-60.

199.[Финн 81]
Финн В.К. К формальному определению информационно-поисковой системы //***НТИ***, Сер.2, 1981, N5, С.1-14.

200.[Финн 81-а]
Финн В.К. Базы данных с неполной информацией и новый метод автоматического порождения гипотез // В кн.: *Диалоговые и фактографические системы информационного обеспечения.* - М., 1981. - С.153-156.

201.[Финн 83]
Финн В.К. О машинно-ориентированной формализации правдоподобных рассуждений в стиле Ф.Бэкона - Д.С.Милля // ***Семиотика и информатика***. Вып.**20**. - 1983. - С.35-101.

202.[Финн 84]
Финн В.К. Информационные системы и проблемы их интеллектуализации // ***НТИ***, Сер.2, 1984, N1, С.5-15.

203.[Финн 85]
Финн В.К. Правдоподобные вывода и проблемы автоматического порождения теорий из фактов //В кн. ***Интенсиональные логики и логическая структура теорий***. Тез. докл. советско-финского коллоквиума по логике. - Тбилиси: Мецниереба.- 1985.- С.156-158.

204.[Финн 88]
Финн В.К. Правдоподобные выводы и правдоподобные рассуждения// **Итоги науки и техники, сер. "Теория вероятностей. Математическая статистика. Теоретическая кибернетика".** Том **28**. - М.: ВИНИТИ,1988. - С. 2 - 84.

205.[Финн 91]
Финн В.К. Правдоподобные рассуждения в интеллектуальных системах типа ДСМ// **Итоги науки и техники**. Сер. "**Информатика**". - Т. 15 "Интеллектуальные информационные системы" - М: ВИНИТИ. - 1991. - С.54 - 101.

206.[Финн 91-а]
Финн В.К., Черный А.И. Интеллектуальные системы и новые информационные технологии в ВИНИТИ // **Итоги науки и техники.** Сер. "Информатика" - Т.**15** "*Интеллектуальные информационные системы*" - М: ВИНИТИ. - 1991. С. 3 - 15.

207.[Финн 93]
Финн В.К. О проблемах логики естественных наук // ***Семиотика и информатика***. - Вып.**33**. - М.: ВИНИТИ. - 1993. - С.68-70.

208.[Финн 93-а]

Финн В.К., Михеенкова М.А. Некоторые проблемы обобщенного ДСМ-метода автоматического порождения гипотез // ***Семиотика и информатика***. - Вып.**33**. - М.: ВИНИТИ. - 1993. - С.136-163.

209.[Финн 96]

Финн В.К. Об одном варианте логики аргументации // ***НТИ***. - Сер.2.- 1996. N5-6. - С.13-19.

210.[Финн 99]

Финн В.К. Синтез познавательных процедур и проблема индукции // ***НТИ***. - Сер.2. - 1999. N1-2. - С.8-45.

211.[Финн 99а]

Финн В.К., Гусакова С.М. Новые версии отрицательного предиката сходства в ДСМ-методе // ***НТИ***. - Сер.2. - 1999. N1-2. - С.53-56.

212.[Финн 99-б]

Финн В.К. О декларативном, процедурном и понятийном знании в интеллектуальных системах типа ДСМ. // **Труды 4-й международной конференции "Интеграция. Информационные технологии. Телекоммуникации"** (Москва, 17-19 марта 1999г.). - Секция "Интеллектуальные системы автоматизированной поддержки научных исследований". - М.:ВИНИТИ. - 1999. - С.3-4.

213.[Финн 99-в]

Финн В.К. Интеллектуальные системы и общество: идеи и понятия // ***НТИ***. - Сер.2. - 1999. - N10. - С.6-20.

214.[Финн 04]

Финн В.К. Об интеллектуальном анализе данных. – ***Новости искусственного интеллекта***. – 2004, №3. – Сс.3-18.

215.[Финн 10]

Финн В.К. Индуктивные методы Д.С. Милля в системах искусственного интеллекта // ***Искусственный интеллект и принятие решений***. - 2010. – Часть I: №3, С.3 -21, Часть II: №4, С. 14 -40.

216.[Финн 10-а]

Финн В.К. Об определении эмпирических закономерностей посредством ДСМ - метода автоматического порождения гипотез. - ***Искусственный интеллект и принятие решений***. – 2010. №4. – С. 41-48.

217.[Финн 11]

Финн В.К. Искусственный интеллект: методология, применения, философия. – М.: Красанд, 2011.– 448 С. Приложение 1. Основные публикации по ДСМ-методу.

218.[Хвостова 97]

Хвостова К.В., Финн В.К. Проблемы исторического познания в свете современных междисциплинарных исследований, М.: РГГУ. - 1997. - 256 С.

219.[Хомски 72]

Хомский Н. Аспекты теории синтаксиса. - М.: Изд-во МГУ. - 1972. – 259 С.

220.[Хомски 72-а]

Хомский Н. Язык и мышление. - М.: Изд-во МГУ, - 1972. - 121 С.

221.[Хохлов 90]

Хохолов А.П., Савченко Ю.Н., Мартынов И.В. *Миелинопатии и демиелинизирующие заболевания.* - М: Медицина, 1990. – 207 С.

222.[Цаленко 85]

Цаленко М.Ш. Канонические представления неприводимых систем и схемы классификации. // ***НТИ***, Сер.2. - 1985. - N1.

223.[Чень 83]

Чень Ч., Ли Р. *Математическая логика и автоматическое доказательство теорем.* - М: Наука. - 1983 - 358 С.

224.[Шашкин 10]

Шашкин Л.О. Применение методов эволюционного моделирования для оптимизации множества ДСМ-гипотез - ***Искусственный интеллект и принятие решений***, N 3, 2010. - Сс. 33-39.

225.[Шашкин 10-а]
Шашкин Л.О. Сравнение приближенных способов поиска сходств для ДСМ-метода // ***НТИ***. Сер. 2. 2010. N 12. - Сс. 9-13.

226.[Шрейдер 71]
Шрейдер Ю.А. Математическая модель теории классификации. - М.:ВИНИТИ, 1971.

227.[Шейдер 71-а]
Шейдер Ю.А. Равенство, сходство, порядок. – М: Наука, 1971. – 255 С.

228.[Эшби 59]
Эшби У.Р. Введение в кибернетику. - М.: Иностранная литература, 1959. – 432 С. (http://pcp.vub.ac.be/ASHBBOOK.html)

229.[ЭЭ 00]
Эволюционная эпистемология и логика социальных наук. Карл Поппер и его критики. – М.: Эдиториал УРСС. – 2000. – 464 С.

230.[Якубович 71]
Якубович С.М. Аксиоматическая теория сходства // ***НТИ***, Сер.2. - 1968. - N10.

231.[Acorn 92]
Acorn T., Walden S. SMART: Support Management Automated Reasoning Technology for Compaq Customer Service // **Innovative Applications of Artificial Intelligence** 4. - Menlo Park (CA): AAAI, 1992.

232.[ADed]
Automated deduction. – ***Wikipedia*** - http://en.wikipedia.org/wiki/Automated_theorem_proving

233.[Agrawal 93]
Agrawal R, ***Imielinski T***., ***Swami A***. Mining Associations between Sets of Items in Massive Databases. ***Proc. of the 1993 ACM-SIGMOD Int. Conf. on Managem. of Data***.- Pp. 207-216.

234.[AI]
Искусственный интеллект. – ***Википедия***. - https://ru.wikipedia.org/wiki/Искусственный_интеллект

235.[AI 80]
Artificial Intelligence. - V.**13**. - N1-2. - 1980.

236.[Aldridge 90]
Aldridge W.N. An assessment of the toxicological properties of pyre-throids and their neurotoxicity // **Crit. Rev. Toxicol**. - 1990. - V.21. - N2. - P.89-104.

237.[Anshakov 89]
Anshakov, O.M., Skvortsov, D.P. **and** ***Finn, V.K***. On axiomatization of many-valued logics associated with formalization of plausible reasonings // ***Studia Logica***. - 1989, **48**, N4, Pp. 423-447.

238.[AS'95]
Architectures for Semiotic Modeling and Situation Analysis in Large Complex Systems/ Proc. of the Workshop of 10th (1995) IEEE Symp. on Intelligent Control (Eds.: ***J.Albus***, ***A.Meystel***, ***D.Pospelov***, ***T.Reader***). - 27 -29 August 1995, Monterey, CA. - AdRem, Inc., 1995.

239.[Ausari 89]
Ausari A., Modaress Batoul. Commercial use of expert systems in the USA// ***ASM J. of Systems Management***. - 1990. - **41**, N 12. - C 10-13, 32.

240.[Bacon 60]
Bacon F. *The New Organon and Related Writings*. - Liberal Arts Press, N.-Y. - 1960.

241.[Barletta 91]
Barletta R. An Introduction to Case-Based Reasoning // ***AI Expert***. - August 1991. - Pp. 43-49.

242.[BD]
Большие данные – ***Википедия***. - Большие данные. - http://ru.wikipedia.org/wiki/Большие_данные
243.[Beall et al. 91]
Beall G., Dashiell F., Bratt S., Swanger H. SAIC - A Knowledge Based System to Support Nuclear Test Ban Treaty Verification // **Innovative Applications of Artificial Intelligence** 3. - Menlo Park (CA): AAAI, 1991.
244.[Benson 86]
Benson M., Benson E., Ilson R. *The BBI Combinatory Dictionary of English*. - John Benjamin Publ. Co., Amsterdam - Philadelphia, 1986.
245.[Bishop 06]
Bishop C.M. Pattern Recognition and Machine Learning. – Berlin: Springer Verlag, 2006. – XX+738 P.
246.[Bisson 90]
Bisson, G. A knowledge based generalizer. - *Proc. 7-th* **ICML**, Austin, 1990, pp. 9-15.
247.[Bisson 91]
Bisson, G. KBG: a Generator of Knowledge Bases. (In: ***Kodratoff, Y***. -ed. - *Machine Learning:* Proc. **EWSL 91**, Springer-Verlag, 1991, p. 137.) An extended version: Learning of Rule Systems by Combining Clustering and Generalization. *Proc. of the International Conference Symbolic-Numeric, Data analysis and Learning,* Paris 17-20 Sept. 1991, pp.399-415.
248.[Bradley 91]
Bradley S. AAAI-91 Conference Highlights // **Software Engineering Strategies**. Gartner Group, Inc. - August 21, 1991.
249.[Bradshow 97]
Bradshow J. *Introduction to Tversky similarity measure*. - **www.daylight.com/.../MUG'97**
250.[Braspenning 89]
Braspenning P.J., Uiterwijk J.W.H., Bakker H., van Leeuwen L.C.J. Conceptual tools for modelling complex objects// *Simulat. Methodol., Languag. and Architectures and AI and Griphics for Simulat.*. Proc. SCS Int. Conf. (Rome, June 7-9, 1989). - ***Iazeolla G., Lehmann A., Van Den Herik H.J***. - eds. - SCS Intern. - 1989. - C. 256 - 262.
251.[Breiman 84]
Breiman L., Fredman J., Olshen R.A., Stone C.J. Classification and Regression Trees. / Monterey, CA: Wadsworth& Brooks, 1984.
252.[Bondy 86]
Bondy S.C. The biochemical evaluation of the neurotoxic damage // ***Fundamental and Applied Toxicology***. - 1986. - V6. - N2. - P.208-217.
253.[Buchanan 78]
Buchanan B.G., Feigenbaum E.A. DENDRAL and MetaDENDRAL: Their applications dimension//***Artificial Inteligence***. - V.11. - P.5-24.
254.[Burch 91]
Burch R.W. *A Piercean Reduction Thesis: the Foundations of Topological Logic*. - Lubbock, TX: Texas Tech Univ. Press. - 1991. - 152 P.
255.[Cabena et al. 97]
Cabena P., Hadjinian P., Stadler R, Verhees J., Zanasi A. *Discovering Data Mining from Concept to Implementation* / Prentice Hall, 1997.
256.[Carnap 36]
Carnap R. Testability and Meaning. – ***Philosophy of Science***. – V.3 (1936). – Pp.419-471; V.4 (1937). – Pp.1-40.
257.[Carnap 45]
Carnap R. On Inductive Logic. - ***Philosophy of Science***. – V.12 (1945). – Pp.72-97.
258.[Carnap 50]

Carnap R. Logical Foundation of Probability. – Chicago: Univ. of Chicago Press. – 1950. – xvii+607 P.

259.[Carnap 56]

Carnap R. The Methodological Character of Theoretical Concepts. – ***Minnesota Studies in the Philosophy of Science***. – V.1 (Eds.: Feigl H., Scriven M.). – Minneapolis. – 1956.- Pp.38-76.

260. [CBR 94]

Topics in Case-Based Reasoning // **Lecture Notes In Artificial Intelligence**. N 837 (Eds.: ***Wess S., Althoff K.-D., Richter M***.)./ - Berlin, etc.: Springer Verlag. - 1994. - 471 P.

261.[CBR Express]

Designing CBR Express Case Base. // **Technical Paper**. - El Segundo, CA: Inference Corporation. - 1993. - Pp. 1-56.

262.[CBR-KTN 91]

Case-Based Reasoning: Relying on Experience // ***Knowledge Technology News***. - October 1991. - Vol. 3, Issue 2.

263.[CBR-PROC]

CBR Workshop Proceedings/ Menlo Park, CA: Morgan Kaufmann. - 1988-1991.

264.[Chandrasekaran 86]

Chandrasekaran B. Generic tasks in knowledge-based reasoning // **IEEE Expert** - 1986 - (Fall 1986) - C.23 - 30.

265.[Chandrasekaran 82]

Chandrasekaran B., Mittal S. Deep versus compiled knowledge approaches to diagnostic problem solving// Proc. Nat. Conf. AI (**AAAI-82**). Los Altos, CA. - 1982. - C. 349-354.

266.[Chester 93]

Chester M. *Neural Networks: a Tutorial*/ Englewood Cliffs, NJ: Prentice Hall, 1993.

267.[Chomsky 82]

Chomsky N. Some concepts and consequences of the theory of government and binding. - Cambridge, MA, 1982.

268.[Churbuck 92]

Churbuck D.C. Learning by Example // ***Forbes***. - June 8, 1992.

269.[Clancey 85]

Clancey W.J. Heuristic classification// ***Artif. Intell.*** - 1985. - **27**. - C. 289 - 350.

270.[Clocksin 03]

Clocksin W., Mellish C.S. Programming in Prolog. – Berlin et al.: Springer Science & Business Media. – 2003. – 299 P.

271.[CO-CO 87]

COLLINS COBUILD English Language Dictionary. - London: HarperCollins Publ. - 1987.

272.[Cox 86]

Cox B.J. *Object-Oriented Programming: An Evolutionary Approach*. - Addison-Wesley, Reading, MA. - 1986.

273.[Czerwinski 60]

Czerwinski Z. O pojeciu przyecyny I kanonach Milla // Studia Logica (Warsawa). 1960. - V.9.

274.[C&T]

Computers and Thought. A collection of articles. - (Eds.: ***Feigenbaum E.*** and ***Feldman J***.).- McGraw-Hill Book Co., New York - San Francisco - Toronto - London, 1963.

Вычислительные машины и мышление. - Под ред. ***Э.Фейгенбаума*** и ***Дж.Фельдмана***. - М: Мир, 1967. - 552 С.

275.[Darvas 87]

Darvas F. METABOLEXPERT: An expert system for predicting metabolism of substances// *QSAR in Environmental Toxicology.* - Vol.II/ Ed. K.L.E.Kaiser. - Dordrecht, Holland: Reidel Publishing Co., 1987. - pp. 71-81.

276.[Darvas 88]

Darvas F. Predicting metabolic pathways by logic programming// **J. Mol. Graph**. - 1988. - **6**, N6. - pp.80-86.

277.[Davis 83]
Davis R. Reasoning from first principles in electronic troubleshooting// **Int. J. Man-Mach. Studies.** - 1983. - 19. - C. 403 -423.

278.[Decarlo 04]
Decarlo D. Extreme Project Management. – San Francisco, CA. - (USA): JOSSEY-BASS, 2004.
Декарло Д. Экстремальное управление проектами. – М.: Комп. P.M. Office, 2005. – 588 С.

279.[DLI]
Digital Libraries Initiative - www.nsf.gov/pubs/1998/nsf9863/nsf9863.htm

280.[DMC]
DataMind Corporation / www.datamindcorp.com

281.[DM KDR]
Data Mining and Knowledge Discovery Resources / www.kdnuggets.com

282.[DM-IBM]
Data Mining - An IBM Overview / www.almaden.ibm.com

283.[DWHI]
Data WareHouse Institute / www.dw-institute.com.

284.[Duval 89]
Duval, B. and ***Kodratoff, Y***. A Tool for the Management of Incomplete Theories: Reasoning about Explanations. (In ***Brazdil, P***., and ***Konolige, K***. - eds. - *Machine Learning, Meta-Reasoning and Logics.* Kluwer, 1989, pp. 135-158.

285.[Dzwinel 83]
Dzwinel J. The structure and algorithm of the cybernetic exploration system for mineral deposits. - *Acta Geophysica Polonica*. - V. XXXI, N3 (1983). - Pp. 307-315.

286.[Dzwinel 81]
Dzwinel J., Cisek B. Prognostycze aspekty zastosowania sistemu bezposrednich poszukiwan weglowodorow WEGA-D w poludniowej Polsce (Prognostic aspects of the WEGA-D system applicat. for direct hydrocarbon exploration in southern Poland).- *Nafta*.- 9 (1981). - Pp. 265-269.

287.[EAI]
Enterprise application integration. – ***Wikipedia***. - http://en.wikipedia.org/wiki/Enterprise_application_integration

288.[EAI 87]
Encyclopedia of Artificial Intelligence (***Shapiro S.C.*** - ed.) - N.-Y. e.a.: J.Wiley & Sons, 1987. - Vol.**1**: - XXIV, 680 pp., vol.**2**: - 681-1220 pp.

289.[*Einstein* 05]
Einstein A. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten - *Annalen der Physik*.***-322****(8)– May 1905. – Pp. 549–560.*
Эйнштейн А. О движении взвешенных в покоящейся жидкости частиц, требуемом молекулярно-кинетической теорией теплоты. - Собр. науч. трудов. - М.: Наука, 1966. - Т. 3. - Сс. 108-117.

290.[ES]
Expert Systems. - ***Wikipedia***. - http://en.wikipedia.org/wiki/Expert_system

291.[EXCALIBUR]
EXCALIBUR Technologies Corp. (San Diego, CA, USA) - www.excalib.com **.**

292.[EXCALIBUR PS]
EXCALIBUR Products & Solutions. - www.excalib.com/rev3/products/products.html

293.[Falkenhainer 85]

Falkenhainer B. Propositionality graphs, units analysis and domain constraints: improving power and efficiency of the scientific discovery process // Proc. **9th IJCAI** Los Angeles, CA, 1985.-P.552-556.

294.[Falkenhainer 86]
Falkenhainer B., Michalski R.S. Integrating quantitative and qualitative discovery: the ABACUS system// ***Machine Learning***. - 1986.- **1**.- C.367-401.

295.[Falkenhainer 87]
Falkenhainer B.C. Scientific theory formation through analogical inference// In: **Proc. of the Fourth Intern. Workshop on Machine Learning.** - Irvine, CA. - 1987. - Pp. 218-229.

296.[Falkenhainer 88]
Falkenhainer B.C., Rajamoney S. The Independencies of Theory Formation, Revision and Experimentation // In: **Proc. of the Fifth International Conference on Machine Learning.** - Morgan Kaufmann, Los Altos, CA. - 1988. - Pp. 353-366.

297.[Fayyad et al. 96]
Fayyad U.M., Piatestsky-Shapiro G., Smith P., Uthurusam R. Advances in Knowledge Discovery and Data Mining. - Cambridge, MA: AAAI Press/MIT Press, 1996, -550 P.

298.[FEDERICA]
Исследовательский проект 7-й Рамочной Программы Европейского Союза «**FEDERICA» -** http://dana.i2cat.net/the-end-of-federica-project/platform/

299.[Feyerabend 75]
Feyerabend P. Against Method. Outline of an anarchistic theory of knowledge.–London.–1975.
Фейерабенд П. Против метода. Очерк анархистской теории познания. – М.: Хранитель. – 2007. – 413 С.

300.[FGCS 82]
Outline of research and development plans for Fifth Generation Computer Systems. - Tokyo, ICOT, 1982.

301.[FGCS 84]
ЭВМ пятого поколения. - (Под ред. Мото-ока). - М., 1984.- 104 С.

302.[Finn 93]
Finn V.K., Mikheyenkova M.A. Logical frameworks for intellectual systems of public opinion poll analysis. - **EWAIC'93: Proc**..(Moscow, 1993). - Pp.319-323.

303.[Finn 93-a]
Finn V.K., Mikheyenkova M.A. On the application of JSM-method of automatic hypotheses generation in sociological investigations. - ***Новости Искусственного Интеллекта***. Специальный выпуск: *Международная конференция по искусственному интеллекту* ***"Восток-Запад - 93"*** (***Artificial Intelligence News***. Special issue*:* ***EWAIC-93***, Moscow, 7-9 Sept. 1993). 1993. - Pp.91-98.

304.[Fischer 90]
Fischer P., Zytkow J.M. Discovering Quarks and Hidden Structure// *Methodologies for Intelligent Systems* **5**. (**Ras Z., Zemankova M., Emrich M.L.** - eds.) - Elsevir Science Publ. Co. Inc., N.-Y. - 1990. - Pp. 362-370.

305.[FPGA]
Программируемые Логические Интегральные Схемы. – ***Википедия***. – https://ru.wikipedia.org/wiki/%CF%CB%C8%D1

306.[FT 99]
Mastering Information Management. Part 6: Knowledge Management.// **Financial Times**. - 1999. - March, 8 .

307.[FULCRUM]
FULCRUM Technologies Inc. (Ottawa, Canada). - www.fulcrum.com **.**

308.[Gabbay 96]

Gabbay D.M., Ohlbach H.J. (Ed.s) *Practical Reasoning.* - **Lecture Notes in Artificial Intelligence**.- N1085.- Berlin et al.: Springer Verlag. - 1996.

309.[GBG HT]
Glass Bead Game (Highland Technologies Inc., MD, USA). - *www.htech.com* .

310.[GENI]
Национальная Программа США «***Global Environment for Network Innovations*** (**GENI**)» - http://www.geni.net

311.[Gergely 84]
Gergely T., Finn V.K. On solver "plausible inference + deduction type in intellectual informational computing systems// *Artif. Intell. IFAC Proc. Ser.* - 1984. - N9.

312.[Gitis 95]
Gitis V. GIS Technology for the Design of Computer-Based Models in Seismic Hazard Assessment.- *Geographical Information Systems in Assessing Natural Hazards*, (***A.Carrara*** and ***F.Guzzetti*** - eds.), 1995, 219-233, Kluver Academic Publishers.

313.[Gitis 95a]
Gitis V., Gergely T. Intelligence Assistant Systems for Geophysics and Geology. - IUGG General Assembly, Boulder, Colorado, July 2-14, 1995.

314.[Gotoda 96]
Gotoda H., An J., Fujiwara Y. Analogical Reasoning of Organic Reactions Based on the Structurized Compound-Reaction Diagram // **Practical Reasoning** (Gabbay D.M. and Ohlbach H.J. - eds.). **Lecture Notes in Artificial Intelligence**. - N1085. - Berlin et al.: Springer Verlag. - 1996. - Pp.276-290.

315.[Gotthard 97]
Gotthard W., Marwick A., Seiffert R. Mining Text Data // **DB2 Magazine**. - Winter 1997, V.2, N.3. - Pp. 45-48.

316.[Groth 98]
Groth R. *Data Mining. A Hands-On Approach for Business Professionals* / Prentice Hall PTR, Upper Saddle River, NJ 07458, 1998, - 264 P.

317.[Hadamard 02]
Hadamard J. Sur les problèmes aux dérivées partielles et leur signification physique. — 1902. — P. 49–52.

318.[Hajek 78]
Hajek P., Havranek T. *Mechanizing Hypothesis Formation.* – Berlin: Springer . - 1978. - 396 p.

319.[Hall 89]
Hall R.P. Computational approaches to analogical reasoning: A comparative analysis // **Artificial Intelligence** - N39. - 1989.-Pp.39-120.

320.[Hamsher 91]
Hamsher, W. Principles of diagnosis: current trends and a report on the First international workshop // **AI Mag.** -1991. - **12**, N 4. - pp.15 - 18, 20 - 23, 37.

321.[Harmon 91]
Harmon P. Case-Based Reasoning // ***Intell. Softw. Strat.*** - November 1991. - Vol. VII, N 11.

322.[Harmon 91-a]
Harmon P. Case-Based Reasoning II // ***Intell. Softw. Strat.*** - December 1991. - Vol. VII, N 12.

323.[Harmon 92]
Harmon P. Case-Based Reasoning III // ***Intel. Softw. Strat.*** - January 1992. - Vol. VIII, N 1.

324.[Hart 82]
Hart P.E. Direction for AI in the eighties// ***SIGART Newslett.*** - 1982. - **79**, N 11.

325.[Hastie 09]
Hastie, T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. — Berlin et al.: Springer-Verlag, 2009. — 746 P.

326.[Hearst 93]
Hearst M.A. *TextTiling: A quantitative approach to discourse segmentation* - http://sun-site.berkeley.edu
327.[Hearst 94]
Hearst M.A. *Multi-paragraph segmentation of expository texts* // **Report No. UCB/CSD 94/790**. - Berkeley, CA: Comp. Sci. Division, Univ. of California. - January 1994.
328.[Helton 91]
Helton T. The Hottest New AI Technology - Case-Based Reasoning // **The Spang Robinson Report on Artificial Intelligence.** - August 1991. - Vol.7, N 8.
329.[Hempel 45]
Hempel C.G. Studies in the Logic of Confirmation. – ***Mind***. – V.54 (1945). – Part I: Pp. 1-26; Part II: Pp. .97-121.
330.[Hempel 48]
Hempel C.G., Oppenheim P. Studies in the logic of explanation// ***Philos. of Science***. - 1948. - **15.**
331.[Hempel 66]
Hempel C.G. Philos. of Natur. Sciences. – Englewood Cliffs, N. J.: Prentice-Hall.- 1966**. -** 116 P.
332.[Hennesey 92]
Hennesey D., Hinkle D. Applying Case-Based Reasoning to Autoclave Loading // ***IEEE Expert***, October 1992.
333.[Holland 95]
Holland J. *Hidden Order* / New York: Addison Wesley, 1995.
334.[Hollingshaus 82]
Hollingshaus J.G., Fukuto T.R. Effect of chronic exposure to pesticides on delayed neurotoxicity // In: **Effect of chronic exposures to pesticides on animal systems**. - N.-Y.: Raven Press. - 1982. - P.85-120.
335.[Hume 88]
Hume D. *Treatise on human nature*. - (***Selby-Bigge*** - ed.). - 1888.
336.[Hyafil 76]
Hyafil L., Rivest R.L. Constructing Optimal Binary Decision Trees is NP-complete. - ***Information Processing Letters***. - 5, N1 (1976), pp.15-17.
337.[IBM-TM]
IBM Text Mining Technology. Turning Information Into Knowledge. A White paper from IBM/ (Ed. - ***D. Tkach***). - www.software.ibm.com/data/mediaminer/IMText/
338.[ID]
Information Discovery, Inc. / www.datamining.com
339.[IEEE Exp]
IEEE Expert. **Intelligent Systems & Their Applications** // V.11, N5, October , 1996.
340.[IDM]
Implementing Document Management // The Delphi Group: Boston, MA. – July, 1998. – http://www.delphigroup.com
341.[IFRS]
International Financial Reporting Standards. – ***Wikipedia***. - http://en.wikipedia.org/wiki/International_Financial_Reporting_Standards
342.[Inmon 94]
Inmon, W.H. *Building the Data WareHouse* / John Wiley, NY, NY, 1994.
343.[Inmon 95]
Inmon, W.H. Using the Data WareHouse / John Wiley, NY NY, 1995.
344.[ISAS'97]

Intelligent Systems and Semiotics: A Learning Perspective// Proc. Intern. Conf. ISAS'97 (Eds. - ***J.Albus***, ***T.Reader***, et al.). Sept. 22-25, 1997, NIST, Gaithersburg, MD, USA - см. также *http://isd.cme.nist.gov/proj/semiotics97/*.

345.[Iwata 91]
Iwata Y., Obama T. QDES: Quality Design Expert System for Steel Product.// **Innovative Applications of Artificial Intelligence** 3. (***Smith*** & ***Scott*** - eds.). - Menlo Park (CA): AAAI, 1991.

346.[Johnson 69]
Johnson M.K. The delayed neurotoxic effects of some organo-phophorus compounds: Identification of phosphorylation site as an esterase // ***Biochem. J.*** - V.114. - 1969. - P.711-717.

347.[Josephson 94]
Josephson J.R. (Ed.) *Abductive Inference: Computation, Philosophy, Technology.* - N.-Y..: Cambridge Univ. Press. - 1994.

348.[Keravnou 89]
Keravnou E.T., Washbrook J. What is a deep expert system? An analysis of the architectural requirements of second-generation expert systems//***Knowl.Eng.Rev***.-1989.- **4**, No 3. - P.205-233.

349.[Keravnou 90]
Keravnou E.T. What is a deep expert system? An analysis of first-generation limitations and a review of second-generation architectures// ***Comput.Phys.Commun***.-1990.-**61.** -C.3-12.

350.[Kimball 96]
Kimball R. The Data Warehouse Toolkit, Practical Techniques for Building Dimensional Data Warehouses / New York, NY: John Wiley & Sons, Inc. .- 1996.

351.[Kitova 94]
Kitova I.I., Raevsky O.A., Blinova V.G., Zefirov N.S. Statistical and logico-structural approaches in QSAR analysis of anti-ACHE activity of aryl-N-methylcarbamates. - ***Proc. of the 10th European Symposium on Structure-Activity Relationship: QSAR and Molecular Modelling*** (Barcelona, Spain), Sept. 4-9, 1994. A 184.

352.[Klavans 96]
The balancing act: combining symbolic and statistical approaches to language (***Klavans J.L***., ***Resnik P***. - eds.). - Cambridge, MA: MIT Press. - 1996.

353.[Klavans 98]
McKeown K.R., Klavans J.L. Generating coherent summaries of On-Line documents: combining statistical and symbolic techniques. - www.cs.columbia.edu/~klavans/home.html

354.[KnA]
Advances in Knowledge Acquisition. // **Lecture Notes in Artificial Intelligence**. - N 1076. / Berlin, etc.: Springer Verlag. - 1996.

355.[Kocabas 91]
Kocabas S. Conflict Resolution as Discovery in Particle Physics. // ***Machine Learning***. - **6**. - 1991.

356.[Kodratoff 92]
Kodratoff,Y. and ***Bisson,G***., The Epistemology of Conceptual Clustering: KBG, an Implementation // ***Journal of Intell. Information Systems***, 1992, N1, pp.57-84.

357.[Kodratoff 92-a]
Kodratoff, Y. Recent advances in machine learning - ***Intern. Journ. of Patt. Recogn. and Artif. Intell.*** - **6,** N4 (1992), pp. 469-511.

358.[Kohavi 96]
Kohavi R., Sommerfield D., Dougherty J. Data Mining using MLC++: A Machine Learning Library in C++. - http://www.sgi.com/Technology/mlc

359.[Kohonen 01]
Kohonen T. Self-Organizing Maps. - Berlin et al: Springer-Verlag, 2001. – 501 P.

360.[Kristensen 89]

Kristensen B.B. Multiple specialization and generalization concept modelling// *Simulat. Methodol., Languag. and Architectures and AI and Graphics for Simulat.*. Proc. SCS Int. Conf. (Rome, June 7-9, 1989). - ***Iazeolla G., Lehmann A., Van Den Herik H.J.*** - eds. - SCS Intern. - 1989. - C. 250 - 255.

361.[KT 98]
Knowledge Technology: Approaches and Innovations // The Delphi Group: Boston, MA. – July 1998. – http://www.delphigroup.com

362.[Kuhn 62]
Kuhn T.S. The Structure of Scientific Revolutions. – Chicago. – 1962. - 264 P.
Кун Т. Структура научных революций. – М.: Аст. – 2002. - 608 С.

363.[Kuhn 63]
Kuhn T.S. The Logic of Discovery or Psychology of Research? – In: Scientific Change (Ed.: Crombie A.C.). – London: Heinemann. – 1963. – Pp. 347-395.
Кун Т. Логика открытия или психология исследования. - В кн.: Структура научных революций. – М.: Аст. – 2002. – Сс. 539- 578.

364.[Kulkarni 87]
Kulkarni D., Simon H. The Processes of Scientific Discovery: The Strategy of Experimentation // ***Cognitive Science.*** - **12** (1987). - Pp. 139-175.

365.[Kumar 98]
Kumar V.P., Lakshman T.V. , Stiliadis D. Beyond Best Effort: Router Architectures for the Differentiated Services of Tomorrow's Internet. – In: ***IEEE Communications***, vol. 36, no. 5, May 1998. - Pp. 152-163.

366.[Lakatos 68]
Lakatos I. Changes in the Problem of Inductive Logic. – In: The Problem of Inductive Logic (Ed.: Lakatos I.). – London. – 1968.

367.[Langley 87]
Langley P., Simon H.A., Bradshaw G.L., Zytkov J.M. *Scientific Discovery: Computational Explorations of the Creative Processes*// MIT Press, Cambridge, MA. - 1987.

368.[Langley 88]
Langley P., Jones R. A computational model of scientific insight// In: *The Nature of Creativity* (**Sternberg R.** - ed.) - Cambridge Univ. Press., Cambridge. - 1988.

369.[Langley 89]
Langley P., Zytkov J.M. Data-Driven Approaches to Empirical Discovery// ***Artificial Intelligence.*** - 1989. - **40**. - Pp. 283-312.

370.[Langley 11]
Langley P. The changing science of machine learning – ***Machine Learning***. - **82** (3), 2011. – Pp. 275–279

371.[Larichev 92]
Larichev O. Cognitive validity in design of decision-aiding techniques // ***Journal of Multicriteria Decision Analysis***. - 1992. - V.1, N.3. - Pp. 127-138.

372.[Lenat 83]
Lenat D.B. The role of heuristics in learning by discovery: three case studies // *Machine Learning: an AI Approach*. - Tioga, Palo Alto, CA, 1983.

373.[Lenat 87]
Lenat D.B. Automated theory formation in mathematics // Proc. 11th *Int. Joint Conf. on AI*, Cambridge, MA, 1987. -. Pp.833-842.

374.[Lenat 90]
Lenat, D. B., Guha R.V. Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project. – Reading, MA: Addison-Wesley. -1990. – 372 P.

375.[Li-Min Fu 90]
Li-Min Fu. A rule-based causal simulation system // ***Trans. The Soc. for Comp. Simulat.*** - 1990. - **7**, N 3. - C. 251-264.

376.[Longman PD]
Language Activator. The World's First Production Dictionary, Longman Group UK Ltd., 1993.
377.[Los 77]
Los J. The Foundations of a Methodological Analysis of Mill's Methods / In: *25 Years of Logical Methodology in Poland*. - Warsawa, 1977. - Pp. 291-325.
378.[Luhn 58]
Luhn H.P. The automatic creation of literature abstracts // ***IBM Journal***. - 1958. - Pp. 159-165.
379.[Lyons 98]
Lyons, D., Tseytin G.S. 1998. Phenomenal data mining and link analysis. - In D. Jensen and C. Henry Goldberg (Eds.). - ***Artificial Intelligence and Link Analysis*** (1998 Fall Symposium). - AAAI, AAAI Press. - Pp. 68-75.
380.[McCarthy 00]
McCarthy J. Phenomenal Data Mining: from data to phenomena. - Stanford, CA: Computer Science Department, Stanford University. - Professor John McCarthy's page. – 2000, February 23. -http://jmc.stanford.edu/articles/phenomenal.html
381.[McCarthy]
McCarthy J. John McCarthy's Home Page. - http://www-formal.stanford.edu/jmc/
382.[McKeown 12]
McKeown N., Kazemian P., Varghese G. Header Space Analysis: Static Checking for Networks. – In: **9th USENIX Symposium on Networked Systems Design and Implementation. NSDI '12,** April 2012, San Jose, CA.
383.[MCS]
Microsoft Corporation/ www.microsoft.com
384.[Merkuryeva 90]
Merkuryeva G.V.,Merkuryev Yu.A., Toivonen H.T. Knowledge based simulation systems - a survey// ABO Academy, Finland - Report 90-12. - 1990. - 48C.
385.[Michalewicz 94]
Michalewicz Z. Genetic Algorithms + Data Structures = Evolution Programs / New York: Springer-Verlag, 1994.
386.[Michalski 78]
Michalski R.S., Larson J.B. *Selection of most representative training examples and incremental generation of VL1 hypotheses. The underlying methodology and the description of programs ESEL and AQ11* (Techn. Rep. UIUCDCS-R-78-867) - Dept. of Comp. Sci., Univ. of Illinois, Urbana,IL. - 1978
387.[Michalski 90]
Michalski, R.S. and Kodratoff,Y., Research in Machine Learning: Recent Progress, Classification of Methods, and Future Direction. (In: ***Kodratoff,Y. and Michalski,R.S.*** (Eds.) *Machine Learning: An Artificial Intelligence Approach*, Vol. III. Morgan Kaufmann, San Mateo, pp.3-30, 1990.)
388.[Michalski 98]
Machine Learning and Data Mining: Methods and Applications. / Eds. - ***Michalski, R.S. Bratko I., Kubat M.*** - John Wiley, 1998.
389.[Michie 82]
Michie D. High-road and low-road programs // AI Magazine - 1982 - **3**, N 21.
390.[Michie 94]
Michie D., Spiegelhalter D.J., Taylor C.C. Machine Learning, Neural and Statistical Classification. – NY: Ellis Horwood. – 1994. - 290 P.
391.[Mikheyenkova 95]
Mikheyenkova M.A. Application of JSM-Reasoning to Problems of Sociology. - *Architectures for Semiotic Modelling and Situation Analysis in Large Complex Systems*. - **Proc. of the**

Workshop of 10th (1995) IEEE Sysmposium on Intelligent Control (Eds.: J.Albus, A.Meystel, D.Pospelov, T.Reader). - 27-29 August 1995, Monterey, CA. - AdRem, Inc., 1995. - Pp. 429-433.

392.[Mill 66]
Mill J.S. *System of Logic. Стюарт Милль и Тэн.* - Перев. Н.Хмелевского, С.-Пб., 1866.

393.[Mill 74]
Mill J.S. Philosophy of Scientific Method. - Hafner Press, N.-Y. - 1974.

394.[Mill 89]
Mill J.S. On Liberty – In**:** «On Liberty» and Other Writings. – Cambridge: Cambridge University Press. – 1989. - 289 P.

395.[Minsky 61a]
Minsky M. Steps Toward Artificial Intelligence// ***Proc. IRE*****.** - V.**49** (January 1961). - Pp. 8-30.

396.[Minsky 61b]
Minsky M. A Selected Descriptor-Indexed Bibliography to the Literature on Artificial Intelligence // ***IRE Trans. Human Factors Electron*****.** - HFE-2. (March 1961). - Pp. 39-55.

397.[Minsky 68]
Semantic Information Processing - (***Minsky M***. - ed.). The M.I.T. Press, Cambridge, Mass..

398.[Mitchell 97]
Mitchell T. Machine Learning. - NY: McGraw-Hill Science/Engineering/Math, 1997. – 432 P.

399.[MLPT 89]
Machine Learning: Principles and Techniques. (***Forsuth, R***. - ed.) - London, Chapman and Hall. 1989. - xvi, 255 pp.

400.[Mohri 12]
Mohri M., Rostamizadeh A., Ameet Talwalkar A. Foundations of Machine Learning. - MIT Press, 2012. – 432 P.

401.[Morik 89]
Morik K. Sloppy Modelling // In: Knowledge Representation and Organization in Machine Learning (***Morik K.*** - ed.). - Berlin et al.: Springer Verlag, 1989. - Pp. 107-134.

402.[Morris 91]
Morris J., Hirst G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text // ***Computational Linguistics***. - 1991. - V.17, N1. - Pp.21-48.

403.[Mucoz 97]
Mucoz A. Compound Key Word Generation from Document Databases Using a Hierarchical Clustering ART Model// ***Intelligent Data Analysis***. - 1997, V**1**, N1.

404.[Murthy 98]
Murthy S. Automatic construction of decision trees from data: A multidisciplinary survey. - ***Data Mining and Knowledge Discovery***. - V.2, Issue 4, (December 1998). - Pp.345 - 389.

405.[Narain 90]
Narain S., Rothenberg J. Qualitative modelling using the causality relation // ***Trans. The Soc. for Comp. Simulat***. - 1990. - **7**, N 3. - C. 265-289.

406.[Nicolas 91]
Nicolas, J., Lebbe, J., and ***Vignes, R.*** From knowledge to similarity. - *Proc. of the International Conference Symbolic-Numeric, Data analysis and Learning,* Paris 17-20 Sept. 1991, pp.585-597

407.[Nida 75]
Nida E.A. Componential analysis of meaning: an introduction to semantic structures. - Hague-Paris, 1975.

408.[Nilsson 71]
Nilsson N.J. Problem-Solving Methods in Artificial Intelligence. - N.-Y.: McGraw-Hill. - 1971.

409.[NL-Lrn]
Connectionist, Statistical and Symbolic Approaches to Learning in Natural Language Processing.// (Eds.: **Wermter S., Riloff E., Scheler C**.)./ Berlin: Springer Verlag. - 1996.

410.[Nordhausen 90]
Nordhausen B., Langley P. A Robast Approach to Numeric Discovery// In: *Proc. of the Seventh Intern. Conf. on Machine Learning.* - Morgan Kaufmann Publ. Inc., Palo Alto, CA. - 1990. - Pp. 411-418.

411.[Nordhausen 90-a]
Nordhausen B., Langley P. An Integrated Approach to Empirical Discovery // In: *Computational Models of Scientific Discovery and Theory Formation* (***Shrager J., Langley P.*** - eds.). - Morgan Kaufmann Publ. Inc., Palo Alto, CA. - 1990. - Pp. 97-128.

412.[NWI]
NeuralWare Inc./ www.NeuralWare.com

413.[N-YA 88].
***Living in a Chemical World: Occupational and Environmental Significance of Industrial Carcinogens*. - C.Maltoni, I.J.Selikoff (eds).** *Annals of the New-York Academy of Sciences*, V.534. - 1988. - 1045 p.

414.[OCT]
Context: Introduction to Oracle ConText. **ORACLE White Paper.-** ORACLE Corp., Sept., 1993.

415.[OFELIA]
Исследовательский проект 7-й Рамочной Программы Европейского Союза «**OFELIA**» – http://www.fp7-ofelia.eu/

416.[Ohlsson 87]
Ohlsson S. Transfer of training in procedural learning: a matter of conjectures and refutetions? // *Computat. Models of Learning* (ed.***L.Bolc***). Springer-Verl. - 1987. - C. 55-88.

417.[ONF]
Ассоциация «**Open Networking Foundation**» - https://www.opennetworking.org/, http://www.openflow.org

418.[ON.Lab]
Open Networking Laboratory\ Open Networking Research Center – http://onrc.stanford.edu

419.[PA & DM]
Predictive Analytics and Data Mining. - https://www.sas.com/technologies/analytics/datamining/

420.[PAKDD-98]
Proceedings of the PAKDD-98: Second Pacific-Asia Conf. on Knowledge Discovery and Data Mining / **Lecture Notes in Artif. Intell**. (Eds.: Wu X., Katagiri R., Korb K.), 1394. - Heidelberg et al.: Springer Verlag, 1998.

421.[Patil 81]
Patil R.S. Causal representation of patient illness for electrolyte and acid-base diagnosis// **MIT/LCS/TR-267**. - - 1981.

422.[PATHFINDER]
PATHFINDER (Version 7.0) // **User's Guide** / Presearch Inc., Fairfax, VA. - 1994.

423.[PATHFINDER-s]
Time Inc.' New Media division (PATHFINDER Web site). - www.pathfinder.com .

424.[Pearl 87]
Pearl, J. Distributed revision of composite beliefs // **Artif. Intell.**, 1987, **33**, pp. 173-215.

425.[Peirce 67]
Peirce C.S. On the natural classification of arguments // ***Proc. of the American Academy of Arts and Sciences***. - 1867. - **7**. - Pp.261-287.

426.[Peirce 67-79]
Peirce C.S. Writings. - www.peirce.org/ writings.html

427.[Peirce 57-78]
Writings of Charles S. Peirce. (***Kloesel C.J.*** et al. - eds.). Vol.**1**, *1857-1866.* Vol.**2**, *1867-1871*. Vol.**3**, *1872-1878.* - Bloomington: Indiana University Press. - 1982-86.
428.[Peterson 61]
Peterson W.W. *Error-correcting Codes*. - MIT Press, 1961.
429.[Plotkin 70]
Plotkin G.D. A Note on Inductive Generalization.- ***Machine Intelligence***, **5.**- 1970.- Pp. 153-164.
430.[Plotkin 71]
Plotkin G.D. A Further Note on Inductive Generalization. - ***Machine Intelligence***, **6.** - 1971. - Pp. 101-124.
431.[Plotkin 72]
Plotkin G.D. Automatic methods of inductive inference. – Ph.D. Thesis, The University of Edinburgh, 1971. - 222 P.
432.[Polya 54]
Polya G. *Mathematics and Plausible Reasoning*. - Princeton Univ. Press. – Princeton, NJ.- 1954.
Пойа Д. *Математика и правдоподобные рассуждения*. - М.: Наука, 1975. - 463С.
433.[Popper 52]
Popper K.R. The nature of philosophical problems and their roots in science // ***Brit. J. Philos. Sci***. - 1952. - V.**3**. - Pp.124-156.
434.[Popper 58]
Popper K.R. *The logic of scientific discovery*. - London, 1958.
435.[Popper 59]
Popper K.R. The Logic of Scientific Discovery. – London – Ney York: Routledge Classics. – 1959. – xxvii+513 P.
436.[QSAR]
QSAR - Википедия. - https://ru.wikipedia.org/wiki/QSAR
437.[Quinlan 79]
Quinlan J.R. Discovery roles from large collections of examples: a case study // **Expert Systems in the Micro Electronic Age**. - Edinburgh: Edinburgh Univ. Press. - 1979.
438.[Quinlan 83]
Quinlan J.R. Learning efficient classification procedures // *Machine Learning: An Artificial Intelligence Approach*. - Palo Alto (CA): Morgan Kaufmann. - 1983.
439.[Quinlan 85]
Quinlan J.R. Internal consistency in plausible reasoning systems // ***New Generation Computing***. - 1985. - V.**3**, N.2. - Pp. 157-180.
440.[Quinlan 86]
Quinlan J.R. Induction of Decision Trees // ***Journal of Machine Learning***, 1986, N1.
441.[Quinlan 88]
Quinlan J.R. C4.5: Programs for Machine Learning - Redwood City, CA: Morgan Kaufmann, 1988.
442.[Radiya 92]
Radiya A., Zytkov J.M. A Framework for Discovering Discrete Event Models// In: *Machine Learning (Proc. of the Ninth International Conference, July 1992, Aberdin, U.K.).* - Morgan Kaufmann Publ. Inc. - 1992.
443.[RADWL 98]
Russian - American Digital Libraries Workshop (Moscow, April 16-17, 1998). - http://koi.risi.ru/radwl98/
444.[Ramsey 50]
Ramsey F.P. The foundations of mathematics. - London: Routledge & Kegan Paul, 1950.
445.[RCA]

Root Cause Analysis – ***Википедия*** - http://en.wikipedia.org/wiki/Root_cause_analysis

446.[Refenes 95]
Refenes A. Neural Networks in the Capital Markets/ New York, NY: John Wiley & Sons, Inc., 1995.

447.[Reichenbach 49]
Reichenbach H. The theory of probability. - Berkeley, 1949.- xvi+492 Pp.

448.[ReMind 1.1]
Functional Specifications for the ReMind Development System. (Version 1.1) // **Documentation ReMind 1.1.** - Issy-Les-Moulineaux: Cognitive Systems, Inc. (USA) & Ingenia (France). - Mardi, 20 Avril, 1993.

449.[Reyes 14]
Reyes A., Rosso P. On the Difficulty of Automatically Detecting Irony: Beyond a Simple Case of Negation. - ***Knowledge and Information Systems***. -V.40, N3 (Sept., 2014), Pp. 595-614. - http://link.springer.com/article/10.1007%2Fs10115-013-0652-8.

450.[RDOOP 87]
Research Directions in Object-Oriented Programming (***Shriver B.D., Wegner P.*** - eds.) - MIT Press. - 1987.

451.[Robinson 65]
Robinson J.A. Automatic deduction with hyper-resolution // ***Internat. Journ. Comput. Math***.. - 1965, Vol.**1**. - Pp. 227-234.

452.[Robinson 65-a]
Robinson J.A. A Machine-Oriented Logic Based on the Resolution Principle. – ***J.ACM***, 1965, V.12, N1, Pp.23-41.

453.[Robinson 67]
Robinson J.A. Heuristic and Complete Processes in the Mechanization of Theorem Proving. – In: ***Systems and Computer Science*** (Hart J.T. and Takasu S. – eds.). – Toronto: Univ. of Toronto Press. – 1967. – Pp. 116-124.

454.[Rose 89]
Rose D. Using Domain Knowledge to Aid Scientific Theory Revision // In: **Proc. of the Sixth Intern. Workshop on Machine Learning.** - Morgan Kaufmann, San Mateo, CA. - 1987. - Pp. 218-229.

455.[Rosso 14]
Rosso P. On irony detection in social media. - **AINL-2014** (Moscow, Skolkovo). – 2014, Sept. 12. - http://www.slideshare.net/AINL/1-paolo-rosso-ainl-irony

456.[Salton 88]
Salton G. Automatic text processing: the transformation, analysis, and retrieval of information by computer. - Reading, MA: Addison-Wesley. - 1988.

457.[Sanger 12].
Sanger D.E. Obama Order Sped Up Wave of Cyberattacks Against Iran. - ***The New York Times***: June 1, 2012. – http://www.nytimes.com/2012/06/01/world/middleeast/obama-ordered-wave-of-cyberattacks-against-iran.html?_r=1&

458.[Sanger 13].
Sanger D.E. Confront and Conceal. Obama's Secret Wars and Surprising Use of American Power. – NY.: Broadway Books, 2013.- 512 P.

459.[SAS DT]
Decision Trees for Analytics Using SAS® Enterprise Miner. - http://www.sas.com/storefront/aux/en/spdectree/63319_excerpt.pdf

460.[SASI]
SAS Institute. - www.sas.com.

461.[SAS-EMS]
SAS Enterprise Miner Software - www.sas.com.

462.[SA'96]

Issues of Measurements and Semiotic Analysis/- Proc. of the Workshop on Control Mechanisms for Complex Systems. (Eds. - ***T.Reader***, ***M.Coombs***, et al.) - Las Cruces (NM) 8-12 Dec. 1996.

463.[Schank 89]
Schank R., Riesbeck C. Inside Case-Based Reasoning / Hillsdale, NJ: Lawrence Erlbaum Assoc.. - 1989.

464.[Schatz 84]
Schatz W., Verity G.W. DARPA's big Puch in AI // **Datamation**. - 1984. - **30**, N2. - pp. 48-50.

465.[Schwartz 92]
Schwartz E., Treece J. Smart Programs Go to Work // **Business Week**. - March 2, 1992.

466.[SCI 85]
Strategic computing: first annual report. - DARPA, Feb.1985.

467.[SDI]
Software Defined Infrastructure. – ***Википедия***. - http://en.wikipedia.org/wiki/Software_Defined_Infrastructure

468.[Sells 85]
Sells P. *Lectures on Contemporary Syntactic Theories*. – Stanford, CA: Center for the Study of Language and Information (CSLI). - 1985. - viii+214 Pp.

469.[SDN&NFV-2013]
SDN and NFV – 2013 (2nd Annual Software Defined Networking Summit). – Praha, CZ, 2013, Sept. 30-Oct. 02. - https://www.sdncentral.com/event/sdn-conference-2013-prague/

470.[SDN&OF-2014]
SDN & OpenFlow World Congress - 2014. - Düsseldorf, Germany. – 2014, October 14-17. - http://www.layer123.com/sdn

471.[SIGCOM- 2014]
Annual conference of the ACM Special Interest Group on Data Communication (SIGCOM) - Chicago, Illinois, USA, 2014, August 17-22. - http://conferences.sigcomm.org/sigcomm/2014/

472.[Simon 91]
Simon H.A. Artificial Intelligence: Where Has It Been, and Where is it Going? // ***IEEE Transactions on Knowledge and Data Engineering***. - V.3. - N.2. - June 1991. - Pp.128-136.

473.[Simon 77]
Simon J. On the difference between one and many. - **Lect. Not. Comp. Sci.** -V.52 (1977). - Pp. 480-491.

474.[SIP 68]
Semantic Information Processing - (***M.Minsky*** - ed.). MIT Press, Cambridge, MA. - 1968.

475.[Skorochod'ko 72]
Skorochod'ko E.F. Adaptive method of automatic abstracting and indexing // In: **Information Processing 71**. Proc. of the IFIP Congress 71 (***Freiman C.V***. - ed.). - North Holland Publ. Comp. - 1972. - Pp. 1179-1182.

476.[SR]
Summarization bibliography - www.dcs.shef.ac.uk/~gael/alphalist.html

477.[SRCU]
Summarization - http://www.cs.columbia.edu/~radev/summarization

478.[SRCU-1]
Summarization Resources - http://www.cs.columbia.edu/~hjing/summarization.html

479.[Steels 86]
Steels L. Components of expertise// *AI Memo 88-16*. Artif. Intell. Labor., Vrije Universiteit, Brussel, Belgium. - 1986.

480.[Stefik 86]
Stefik M., Bobrow D.G. Object-oriented programming: themes and variations// **AI Magazine** - 1986 - **6**, N 4. - C. 40 - 62.

481.[SUN]
SUN Microsystems Computer Company / www.sun.com
482.[Suppes 84]
Suppes,P., Conflicting Intuitions about Causality.(In: *Midwest Studies in Philosophy IX, Causation and Causal Theories*. - Univ. of Minnesota Press, Minneapolis, 1984, pp.151-168.)
483.[Tanimoto 87]
Tanimoto S.L. *The Elements of Artificial Intelligence*. - Computer Science Press. - 1987.
484.[TCAM]
Ассоциативная память – ***Википедия***. – http://en.wikipedia.org/wiki/Content-addressable_memory
485.[Testa 76]
Testa B., Jenner P. Drug metabolism: Chemical and Biochemical Aspects. - N.-Y. - Basel: Marcel Dekker Inc., 1976. - 500 p.
486.[TMC]
Thinking Machine Corporation / www.think.com
487.[Thomsett 02]
Thomsett R. Radical Project Management. – Upper Saddle River, NJ: Prentice Hall, 2002.
Томсетт Р. Радикальное управление ИТ-проектами. – М.: Лори, 2005. – 292 С.
488.[Tversky 77]
Tversky A. **Psychological Reviews**. - 1977, **V.84**, N 4.-Pp.327-352.
489.[Tukey 77]
Tukey, John W. Exploratory Data Analysis. - Reading, Mass.: Addison-Wesley Publishing Company, 1977. - XVI+688 P
490.[Turing 50]
Turing A.M. Computing Machinery and Intelligence// ***Mind.*** - **59** (October 1950). - Pp. 433-460.
491.[Valiant 79]
Valiant L.G. The complexity of enumeration and reliability problems //***SIAM J. Comput.*** - 1979. - **8**. - pp. 410-421.
492.[Valiant 79a]
Valiant L.G. The complexity of computing the permanent. - ***Theoretical Computer Science.*** - 8 (1979). - Pp. 189-201.
493.[Valiant 84]
Valiant L.G. A theory of learnable// ***Commun. ACM***. - 1984. - **27**, N11. - pp. 1134 - 1142.
494.[Vapnik 98]
Vapnik V.N. Statistical Learning Theory. - NY, NY: John Wiley & Sons, 1998. - 732 P.
495.[Vapnik 00]
Vapnik V.N. The Nature of Statistical Learning Theory. – Berlin: Springer Verlag, 2000. – XX+314 P.
496.[Weiss 89]
Weiss S.M., Kapouleas I. An Empirical Comparison of Pattern Recognition, Neural Nets, and Machine Learning Classification Methods// Proc. 11th IJCAI, 1989.
497.[Weiss 97]
Weiss S., Indurkhya N. Predictive Data Mining: a Practical Guide/ Morgan Kaufmann Publ., 1997.
498.[Welstead 94]
Welstead S.T. Neural Network and Fuzzy Logic Applications in C\C++. / John Wiley & Sons, Inc., 1994.
499.[Winograd 80]
Winograd T. Extended inference modes in reasoning by computer systems // ***Artificial Intelligence***, V.**13** (1980). – N12. – Pp.5-26.
500.[Winston 80]

Winston P.H. Learning and reasoning by analogy// ***Comm. ACM***. - 23(12). - 1980. - Pp.689-703.

501.[von Wright 60]
von Wright G.H. *A* Treatise on Induction and Probability. - Paterson, NJ: Littlefield, Adams & Co, 1960. - 310 P.

502.[Witten 05]
Witten I.H., Frank E. Data Mining: Practical Machine Learning Tools and Techniques. – Amsterdam et al.: Morgan Kaufmann, 2005. – xxxi + 524 P.

503.[Wrobel 89]
Wrobel S. Demand-driven Concept Formation // In: *Knowledge Representation and Organization in Machine Learning* (***Morik K.*** - ed.). - Berlin et al.: Springer Verlag, 1989. - Pp. 289-319.

504.[Wu 89]
Wu Y., Wang S. Discovering Knowledge from Observational Data// In: *Knowledge Discovery in Databases*, IJCAI-89 Proc. *(**Piatetsky-Shapiro G.*** - ed.) - Detroit, MI. - Pp. 369-377.

505.[Zabezhailo 94]
Zabezhailo M.I., Finn V.K., Leybov A.E.,Melnikov N.I., Pankratova E.S. CBR-technology in the chemical safety control // **Proc. EWCBR'94** (France, November 7-11, 1994).- 1994.

506.[Zabezhailo M.I. et al. 95]
Zabezhailo M.I. et al. Reasoning Models for Decision Making: Applications of JSM-Method for Intelligent Control Systems//*Architectures for Semiotic Modeling and Situation Analysis in Large Complex Systems.* - **Proc. of the Workshop of 10th (1995) IEEE Symp. on Intelligent Control** (Eds.: ***J.Albus, A.Meystel, D.Pospelov, T.Reader***). - 27 -29 August 1995, Monterey, CA../ AdRem, Inc., 1995. - Pp. 99-108.

507.[Zabezhailo 96]
Zabezhailo M.I., Finn V.K. Intelligent Information Systems. - ***International Forum on Information and Documentation*** (FID, The Hague, Netherlands). - 1996. - V21. - N2. - Pp.21-31.

508.[Zabezhailo 96-a]
Zabezhailo M.I. On the Application of the Semiotic Modelling Technique in the Problem of Clearing Payments Control.// **Proc. 12th European Conference on Artificial Intelligence** (ECAI'96, Budapest, 11-16 August, 1996). - Workshop N30 "Applied Semiotics". Pp.17-21.

509.[Zabezhailo 96-b]
Zabezhailo M.I., Finn V.K., Gergely T. JSM-method as an Instrument for Semiotic Modelling: Application to Geological Forecasting // **Proc. 12th European Conference on Artificial Intelligence** (ECAI'96, Budapest, 11-16 Aug., 1996). - Workshop N30 "Applied Semiotics". Pp.13-16.

510.[Zabezhailo 97]
Zabezhailo M.I. To the Scaleable Technology of Automated Document Understanding Based on Quasi-Axiomatic Theories.// Proc. 1997-th **International Conference on Intelligent Systems and Semiotics "A Learning Perspective" - ISAS'97** (NIST, Gaithersburg, MD, September 22-25, 1997). - NIST Special Publications. - N918. - 1997. - Pp.100-102.

511.[Zabezhailo 13]
Zabezhailo M. I. On the Functionality of the Causal Relationship Used in JSM Reasoning. - ***Automatic Documentat. and Mathemat. Linguistics***. - 2013, Vol. 47, No. 4. - Pp. 162–167

512.[Zembowicz 92]
Zembowicz R., Zytkov J.M. Discovery of Equations: Experimental Evaluation of Convergence// In: *Proc. of AAAI-92.* - AAAI Press, Menlo Park, CA. - 1992. - Pp. 70-75.

513.[Zenkin 90]
Zenkin A.A. Waring's problem from the standpoint of the cognitive interactive computer graphics // ***Mathem. & Computer Modelling***. - 1990. - V.13, N.11. - Pp.9-25.

514.[Zenkin 95]

Zenkin A.A. Intelligent control and cognitive computer graphics// *Architectures for Semiotic Modeling and Situation Analysis in Large Complex Systems.* - **Proc. of the Workshop of 10th (1995) IEEE Symp. on Intelligent Control** (Eds.: J.Albus, A.Meystel, D.Pospelov, T.Reader). - 27 -29 August 1995, Monterey, CA../ AdRem, Inc., 1995.. - Pp. 366-371.

515.[Zytkow 86]
Zytkow J.M., Simon H.A. A Theory of Historical Discovery: The Construction of Componential Models// ***Machine Learning***. - 1986. - **1.** - Pp. 107-136.

516.[Zytkow 87]
Zytkow J.M. Combining Many Searches in the FAHRENHEIT Discovery System // **Proc. of the Fourth Intern. Workshop on Machine Learning**. - Las Altos, CA: Morgan Kaufmann, 1987. - Pp. 281-287.

517.[Zytkow 90-a]
Zytkow J.M., Lewenstam A. Analytical Chemistry; the science of many models// **Fresenius Journal of Analytical Chemistry**. - Springer-Verlag. - 1990. - **338.** - Pp. 225-233.

518.[Zytkow 90-b]
Zytkow J.M. Deriving Laws Through Analysis of Processes and Equations// In: *Computational Models of Discovery and Theory Formation* (**Langley P.** and **Shrager J.** - eds.). - Morgan Kaufmann Publ. Inc., San Mateo, CA, 1990. - Pp. 129-156.

519.[Zytkow 90-c]
Zytkov J.M., Zhu J., Hussam A. Automated Discovery in a Chemistry Laboratory// In: *Proc. of the Eight National Conference on Artificial Intelligence.* - AAAI Press. - 1990. - Pp. 889 - 894.

520.[Zytkow 90-d]
Zytkow J.M., Baker J. Interactive Mining of Regularities in Databases// In: *Knowledge Discovery in Databases* (**Piatetsky-Shapiro G., Frawley W.** - eds.). - AAAI Press, Menlo Park, CA. - 1990.

521.[Zytkow 91-a]
Zytkow J.M., Zhu J. Automated Empirical Discovery in a Numerical Space// In: *Proc. of the Third Annual Chinese Machine Learning Workshop '91.* - Harbin, Peoples' Republic of China, 1991. - Pp.1-11.

522.[Zytkov 91-b]
Zytkow J.M., Zytkow A.N. Human discovery of laws and concepts; An experiment// In: *Proc. of the Thirteenth Annual Conference of the Cognitive Science Society (Hillsdale, New Jersey).* - Lawrence Erlbaum Associates Inc. - 1991. - Pp. 617-622.

523.[Zytkow 91-c]
Zytkow J.M. Integration of Knowledge and Method in Real-world Discovery// In: *Proc. SIGART-91.* - 1991.

524.[Zytkow 92]
Zytkow J.M., Zhu J., Zembowicz R. Operational definition refinement: a discovery process// In: *The Proc. of the Tenth National Conference on Artificial Intelligence (AAAI-92: July, 1992. San Jose, CA.).* - AAAI Press, Menlo Park, CA. - Pp. 76-81.

525.[Zytkow 96]
Zytkow J.M. Automated Discovery of Empirical Laws // ***Fundamenta Informaticae*****.** - 27 (1996). - Pp.299-318.

Приложение 1

Общая характеристика приложений ДСМ-ИАД.

Все рассматриваемые далее в данной **Главе** примеры приложений уже подробно обсуждавшегося нами в предыдущих **Главах** ДСМ-инструментария автоматического порождения зависимостей из данных в той или иной мере опираются на нижеследующую единую полуформальную схему описания (см. также ранее - **Раздел 2.3 Главы 2**) задачи о доопределении частично-определенного отношения **ОБЪЕКТ=>СВОЙСТВА**.

Пусть заданы два конечных множества:

C = {C_1, C_2,..., C_n} - объектов, характеризуемых наличием определенной "внутренней структуры" таким образом, что есть возможность выделять "подобъекты" в объектах из множества ***C*** (при этом мы будем считать что множество всех возможных подобъектов для объектов из ***C*** есть некоторое множеств ***C'***), и

A = {A_1, A_2,..., A_m} - анализируемых свойств у объектов из множества ***C***.

На декартовом произведении ***C***x***A*** задается частично определенное отношение ==>$_1$ *обладать свойством*, интерпретируемое как *объект* C_i *обладает свойствами* A_j и образованное парами <C_i,A_j> из множеств следующих трех типов

Ω^+ - множество пар <C_i,A_j> (их мы будем также называть *позитивными примерами* или просто *примерами*), таких что объект C_i действительно обладает свойствами A_j;

Ω^- - множество пар <C_i,A_j> (их мы будем также называть *негативными примерами* или *контрпримерами*), таких что объект C_i не обладает свойствами A_j;

Ω^τ - множество пар <C_i,A_j> (их мы будем также называть *недоопределенными примерами*), таких что о наличии у объекта C_i свойств A_j нет сведений (эти знания и предполагается породить в ходе ДСМ-рассуждения).

В свою очередь на декартовом произведении ***C'***x***A*** задается частично определенное отношение ==>$_2$ *быть причиной*, интерпретируемое как *подобъект* C'_i *является причиной наличия\отсутствия свойств* A_j и образованное парами <C'_i,A_j> из множеств следующих четырех типов

ω^+ - множество пар <C'_i,A_j> (их мы будем также называть *позитивными гипотезами – позитивными эмпирическими зависимостями - первого рода*), таких что подобъект C'_i действительно обладает свойствами A_j (т.е. C'_i считается "причиной" присутствия для A_j);

ω^- - множество пар <C'_i,A_j> (их мы будем также называть *негативными гипотезами – негативными эмпирическими зависимостями - первого рода*), таких что объект C'_i не обладает свойствами A_j (т.е. C'_i считается "причиной" отсутствия для A_j или, другими словами, "антипричиной, контрпричиной" для присутствия A_j);

ω^0 - множество пар <C'_i,A_j> (их мы будем также называть *противоречивыми*[1] *гипотезами первого рода*), таких что информация о наличии\отсутствии у подобъекта C'_i свойств A_j противоречива;

ω^τ - множество пар <C'_i,A_j> (их мы будем также называть *недоопределенными гипотезами первого рода*), таких что о наличии у подобъекта C'_i свойств A_j нет сведений (эти знания и предполагается породить в ходе ДСМ-рассуждения).

Для порождения новых эмпирических зависимостей из данных используется следующая процедурная схема:

---

[1] Точный смысл этого термина определен в **Главе 3** уточнением следующей идеи: противоречивыми оказываются те гипотезы первого рода, которые могут быть порождены одновременно применением и *позитивных*, и *негативных* ППВ-I.

- стартуя с элементов множеств $\boldsymbol{\Omega^+}$ и $\boldsymbol{\Omega^-}$, применением правил правдоподобного вывода первого рода (ППВ-I), мы будем стремиться породить все возможные элементы множеств $\boldsymbol{\omega^+}$, $\boldsymbol{\omega^-}$ и $\boldsymbol{\omega^0}$; после чего
- использовать порожденные гипотезы из множеств $\boldsymbol{\omega^+}$ и $\boldsymbol{\omega^-}$ для доопределения (там, где это возможно) применением правил правдоподобного вывода второго рода (ППВ-II) элементов множества $\boldsymbol{\Omega^\tau}$, затем
- пополнить вновь доопределенными[2] элементами множества $\boldsymbol{\Omega^\tau}$ соответствующие текущие множества примеров и контрпримеров, после чего итерировать процесс ДСМ-вывода до ситуации, когда соответствующее множество $\boldsymbol{\Omega^\tau} = \varnothing$, либо же до наступления состояния стабилизации (когда новые зависимости из уже имеющихся данных больше породить не представляется возможным).

(Таким образом, речь идет о максимально возможном доопределении использованием правил правдоподобного вывода в рамках ДСМ-метода - т.е. в процессе ДСМ-рассуждения - частично определенных отношений $==>_1$ и $==>_2$.)

Задавая те или иные конкретные спецификации понятия *внутренняя структура* объектов из множества ***C*** (и, соответственно, *структура подобъектов* из множества ***C'***), мы будем получать те или иные варианты формализации ДСМ-метода.

В подавляющем большинстве случаев представленные далее эксперименты были ориентированы на порождение эмпирических зависимостей, релевантных прогнозу определенных свойств конкретных новых объектов (т.е. прикладная задача, как правило, была ориентирована на «*точечную*» экспертизу свойств конкретных объектов, а не на порождение «*универсальных*» зависимостей). При этом доступные для выполнения ДСМ-ИАД исходные данные характеризовались необходимостью получать результаты на малых (статистически незначимых) выборках прецедентов – т.е. в ситуациях (см. подробнее **Главу 1**), «предшествующих» формированию таких коллекций данных, которые достаточны по своим объемам для корректного применения к ним статистических моделей и «инструментов» восстановления эмпирических зависимостей.

В части характеристики использовавшихся средств представления данных следует отметить, что в большинстве случаев были задействованы те или иные версии теоретико-множественных (дескрипторных) языков.

Однако, в ряде экспериментов (например, с использованием кода ФКСП, позволяющего формировать покрытие графа химической структуры анализируемого объекта – физиологически-активного соединения – подграфами определенного вида[3], а также при изучении некоторых патологий картофеля[4]) был задействован механизм работы с так называемыми нумерованными множествами (допускающими многократное вхождение в соответствующее множество дескрипторов, характеризующих изучаемый объект, того или иного элемента - образующей). Значительная часть экспериментов в части анализа и прогноза свойств физиологически активных выполнялась также при их представлении в виде плоских симметризованных графов.

Попытки использовать цепочки образующих (как средство описания динамики изменений состояния анализируемых объектов во времени) были предприняты при анализе крупных технологических проектов в банковских приложениях (проектов модернизации ИТ-систем в ряде крупных российских банков). Здесь объектами исследования оказались технические карты описания проектов, формируемые за счет выбора тех или иных конкретных (управленческих и системно-технических) решений вдоль «траектории» реализации

---

[2] При этом результаты противоречивого доопределения (т.е. те, которые получены применением равных по силе ППВ-II, утверждающих *наличие* искомых свойств, с одной стороны, и их *отсутствие* - с другой) помещаются в множество $\boldsymbol{\Omega^0}$ и исключаются из соответствующего пополнения множеств примеров и контрпримеров.

[3] См. подробнее **Главу 1**, **Приложение 8.2** а также работы [*Авидон 74, Лейбов 91*] и др. .

[4] См. [*Долягин 86*].

проекта, определяемой его «типовыми» промежуточными этапами. Перечень таких этапов дает (типовая) Дорожная Карта проекта[5].

К сожалению, объективный анализ данных в этой интересной и весьма важной области приложений существенным образом ограничен характерными для нее требованиями конфиденциальности и коммерческой тайны. Даже в рамках общения в специальной профессиональной среде здесь можно получить доступ лишь к очень малым объемам соответствующих данных[6].

Однако, в рамках доступных автору настоящей работы[7] фактических данных о модернизационных ИТ-проектах в нескольких крупных российских коммерческих банках удалось показать, что описывая динамику того или иного варианта развития проекта как последовательность выбора решений (организационных, управленческих и системно-технических) из заданного множества вариантов, можно формировать, вообще говоря, не любые, а лишь определенным образом согласованные варианты таких «проектных траекторий» (если, конечно, руководствоваться принципом *минимизации* или же, по крайней мере, не порождения новых *рисков* на промежуточных этапах реализации проекта). Так, например, можно показать, что:

(1) в условиях *неполноты* исходных функциональных *требований* к финальному целевому решению (характеристикам создаваемой банковской информационной системы) выбор в качестве технологии *управления проектами* так называемой *классической* модели[8] вместе со *стандартной* технологией разработки и внедрения[9] формируемых ИТ-решений, как следствие (этого выбора), как правило, порождает ситуации, когда при *дополнении* исходных *требований* новыми элементами существенная часть ранее уже выполненных работ оказывается несовместимой со вновь зафиксированными параметрами целевого решения и необходим *реинжиниринг* части уже сформированных компонентов (приходится вновь возвращаться к стадии проектирования, выполнять перепроектирование соответствующих архитектурных компонентов и далее вновь двигаться по цепочке *проектирование => разработка => внедрение*. Однако, накапливаемые таким образом проблемы со сроками, необходимыми ресурсами и, разумеется, постоянно растущим бюджетом проекта могут достаточно быстро стать *критически* значимыми *факторами риска* и воспрепятствовать успешному завершению проекта в запланированные сроки, с заданными параметрами качества создаваемых решений и, конечно же, в рамках заранее согласованного бюджета). Избежать представленных рисков позволяет, например, выбор технологии *масштабирования прототипов*[10] в качестве инструментария разработки и внедрения создаваемых решений вместе с так называемой технологией *экстремального управления проектами* (как релевантного[11] project_management-инструментария);

---

[5] Вариант подобной *Дорожной Карты*, при необходимости пополняемый более детальным описанием промежуточных стадий модернизационного проекта, можно найти, например, в монографии [*Забежайло 09*].

[6] Подобные проекты имеют, как правило, весьма существенные бюджеты. Как следствие приоритетами при принятии соответствующих решений в таких проектах часто оказываются вовсе не «*поиск истины*» или же стремление к *непременно успешному* завершению работ и достижению исходно сформулированных целей, а индивидуальные *интересы* конкретных персонажей (как на стороне исполнителя, так и, как это не странно, и на стороне заказчика), отражаемые известным «олимпийским» принципом: главное не выигрывать, главное – участвовать (решать, распределять, …).

[7] В силу занимаемых в тот период позиций одного из руководителей входящего в первую тройку российских коммерческих банков (а также руководителя Проекта модернизации ИТ-систем этого банка).

[8] См., например, [*Мазур 01*]и др. .

[9] Имеется ввиду технология последовательной смены стадий проекта: *проектирование => разработка => внедрение*.

[10] См., например, [*Забежайло 09*] и др.

[11] Ситуации с неполнотой исходных требований – см., например, [*Thomsett 02, Decarlo 04*] и др. .

(2) в условиях выбора в качестве модели *интеграции приложений* в целевом ИТ-решении так называемой технологии EAI[12] последующие усилия по доработке зарубежного прикладного программного обеспечения (так называемых *Core Banking Systems*[13]) под требования российских государственных регуляторов а также национального законодательства РФ[14] практически в 100% случаев натыкаются на проблемы с защитой *интеллектуальной собственности*[15]. Особые трудности здесь возникают также при интеграции в базовое ПО национальных средств защиты данных и обеспечения информационной безопасности (ведь как и в случае доработки базового ПО здесь необходимы соответствующие действия по интеграции программных компонентов, призванные обеспечить нормальное взаимодействие базового ПО и средств обеспечения информационной безопасности). Альтернативой, позволяющей в данном случае надежно *управлять* обозначенными проектными *рисками*, является, например, выбор так называемых *компонентных моделей* описания функциональных возможностей целевого решения вместе с *сервис-ориентированными архитектурами*[16] соответствующих программных приложений.

Наконец, характеризуя особенности конфигурирования собственно средств ДСМ-экспертизы, используемых в описываемых ниже экспериментах, отметим, что в большинстве представленных ситуаций базовыми инструментами анализа данных и восстановления соответствующих эмпирических зависимостей в части задействованных *правил правдоподобного вывода* первого рода (см. подробнее **Главу 3**) были ППВ, реализующие так называемые методы *простого сходства* и *запрета на контрпримеры*. Полный обзор экспериментально оценивавшихся возможностей а также особенностей использования ДСМ-ППВ в задачах восстановления зависимостей из эмпирических данных можно найти, например, в работах А.Ю.Волковой (см. [*Волкова 11, Волкова 11-а, Михеенкова 13, Волкова 13*] и др.).

---

[12] *Enterprise Application Integration*- интеграция приложений уровня предприятия (см., например, [*EAI*]), и др.).

[13] В русскоязычной специальной литературе для них используется (по-видимому, не самый точный) термин *Автоматизированная Банковская Система* (АБС).

[14] Ведь зарубежное банковское программное обеспечение ориентировано на модель финансового учета (как правило, это – IFRS [*IFRS*]), базирующуюся на отличных от российской (Российская системы бухгалтерского учета - РСБУ) фундаментальных принципах: в IFRS это - мониторинг текущей капитализации, а в РСБУ – фискальные интересы государства (см., в частности, [*Маренков 06*] и др.).

[15] Владелец прав на базовое программное обеспечение (ПО), как правило, отрицательно реагирует на предложения заказчика о внесении тех или иных изменений в исходные коды. (И это естественно: ведь в таком случае предстоит договориться о всех деталях разделения ответственности за возможные ошибки в коде интегрированного ПО. Как правило, владелец прав интеллектуальной собственности на ПО попросту *отказывается* его *сопровождать*, если какие-либо изменения внесены в это ПО без его санкции. А вместо внесения такого рода изменений часто предлагает заказать ему *отдельную новую разработку*, в которой были бы учтены все текущие требования заказчика. Заметим, умалчивая при этом, какие проблемы в части сопровождения и модификации этой специфической версии базового ПО ждут впоследствии заказчика, и какими бюджетными расходами может далее сопровождаться такой путь).

[16] См., подробнее, например - [*Забежайло 09*] и др..

Приложение 2.

Анализ свойств физиологически активных веществ (ФАВ): Приложения в задаче СТРУКТУРА ==> СВОЙСТВА физиологически активных химических соединений.

С точки зрения когнитивных наук современная химия - мир веществ и их свойств - представляется одной из наиболее интересных областей знания, где существует широкий спектр специальных проблемно-ориентированных средств представления знаний о структурах и свойствах изучаемых объектов - всевозможных химических соединений. При формировании такого сорта средств задействована обширная гамма принципов и "*физических механизмов*": от "*геометрического*" представления структурных графов химических веществ (например, "плоских" или трехмерных), до моделей, основанных, например, на учете *спектральных* характеристик или же на представлении распределений *электронных плотностей*. Проблема разумного баланса между "структурной простотой" языков представления знаний и адекватностью предоставляемых ими средств описания изучаемых явлений имеет долгую, весьма разнообразную и, на наш взгляд, весьма поучительную историю.

Действительно, практически вся используемая на сегодняшний день информация о мире веществ и их свойств опирается на известную гипотезу ***А.М.Бутлерова*** о том, что свойства химического соединения полностью определяются его химической структурой.

Важно заметить, что реально используемые для описания химических соединений языки представления знаний могут не только иметь совершенно разную "физическую" природу но и различные степени строгости (т.е. детальности) описания структур (а значит и свойств) изучаемых соединений.

Фактически, здесь перед исследователем стоит дилемма: выбирать между сложностью (а значит глубиной и адекватностью) языка представления знаний о структурах химических соединений, с одной стороны, и реальной осуществимостью их анализа (в смысле оценок сложности вычислений и, соответственно, необходимых вычислительных ресурсов - быстродействия процессора, объемов памяти и т.п.) с другой. Так, например, представление химических структур с помощью спектральных характеристик требует использования дескрипторного[1] языка, где каждой спектральной линии сопоставляется отдельный дескриптор. Таким образом, большая часть задач информационного поиска (например, анализ сходства соединений, проверка вложимости структурного фрагмента в соединение и т.п.) в данном случае сводится к простейшим вычислениям в булевой алгебре множеств (к вычислению пересечения множеств, проверке вложимости множества в множество и т.п.), реализуемым алгоритмами полиномиальной вычислительной сложности[2]. Использование же представлений химических структур в виде 2-D симметризованных (т.е. построенных без учета реальных углов и расстояний между атомами внутри молекулы) графов требует значительно больших вычислительных усилий. Так, например, проверка вложимости подграфа

---

[1] Таким образом, в данном случае используемый тип данных - множества.

[2] Своего рода "платой" за быстроту порождения результатов в данном случае являются некоторые проблемы, возникающие в процессе интерпретации получаемых результатов. Например, в ситуации, когда найден некоторый набор спектральных линий, характеризующих искомый фрагмент структуры химического соединения (в частности - так называемый фармакофор - т.е. структурный фрагмент, являющийся в смысле гипотезы Бутлерова "носителем" некоторого множества свойств), требуется предпринять специальные усилия для восстановления структуры этого фрагмента, представляемой соответствующей совокупностью элементов таблицы Менделеева а также валентных связей между ними. Нетрудно убедиться, что такое обратное преобразование (т.е. переход от дескрипторного кода - множества "смысловых" примитивов - к, например, графовому представлению, существенным образом использующему еще и отношения между элементами дескрипторного языка), вообще говоря, не является однозначным и требует привлечения дополнительных знаний для устранения возникающих неоднозначностей. Поучительный пример решения этой проблемы - принципы построения и практика использования знаменитой экспертной системы ***DENDRAL*** (см., например, [*Buchanan 78*] и др.).

в граф требует уже реализации алгоритма, вообще говоря, экспоненциальной вычислительной сложности.

Опыт реальных прикладных химических исследований демонстрирует полезность и эффективность совместного использования иерархических проблемно-ориентированных семейств[3] языков представления знаний о химических структурах, где есть возможность согласованного перехода от менее детального (но более простого в вычислениях) к более детальному (но требующему больших объемов вычислений) способу описания химических структур. Стандартный вариант такой иерархии может выглядеть следующим образом:

- дескрипторный язык (как правило, образованный той или иной системой "типовых" подграфов, покрытие которыми анализируемой структуры и порождает ее представление в выбранном языке);
- язык симметризованных (см. выше) 2-D графов - "плоских" симметризованных изображений анализируемой структуры;
- язык симметризованных (см. выше) 3-D графов - учитывающих пространственные конфигурации симметризованных изображений анализируемой структуры;
- язык 3-D графов, учитывающих реальные пространственные углы и расстояния в изображении анализируемой структуры.

Совместное использование иерархий представленного вида позволяет варьировать детальность анализа изучаемых структур по схеме:

- сперва быстрый, но приближенный поиск нужной химической информации в больших БД, затем
- уточнение его результатов на массивах значительно меньшей размерности с использованием более детальных языков представления знаний (ЯПЗ).

Рассмотрим практически используемый пример простейшего ЯПЗ в только что представленной иерархии. Речь пойдет о так называемом Фрагментарном Коде Суперпозиции Подструктур (ФКСП - см., например, [*Авидон 74*] и др.) - проблемно-ориентированном языке представления знаний о структурах химических соединений в задачах анализа физиологических активностей соединений (т.е. задачах анализа связи ***СТРУКТУРА ==> СВОЙСТВА*** химических соединений). Ключевая идея построения кода ФКСП - выделение в структуре анализируемого соединения так называемых *активных центров* (специфических групп химических соединений, сведенных в заранее построенный каталог[4]). При построении ФКСП-представления химической структуры

- выделяются входящие в нее активные центры (из заданного списка), затем
- для каждой пары активных центров прослеживается соединяющая их углеродная цепочка (если таковая существует[5]) минимальной длины;
- фиксируется характер связей в выделенной углеродной цепочке (простые или сопряженные, и т.п.).

Таким образом, при построении ФКСП-представления используется покрытие графа анализируемой химической структуры стандартными (по способу порождения) подграфами вида

{<***активный центр 1***>
<**углеродная цепочка**>
<***активный центр 2***>}.

Оказалось, что для кодирования известных химических соединений используется достаточно небольшой каталог стандартных ФКСП-дескрипторов (подграфов) - всего около 20000 штук.

---

[3] См. также соображения о вариативности языков представления данных для реализации их интеллектуального анализа (**Глава 1**).

[4] В этом каталоге - около 80 фиксированных подграфов (специфических групп элементов).

[5] Таковой может не быть если на любом пути между выделенными активными центрами находится какой-либо третий активный центр.

Легко видеть, что язык ФКСП ориентирован в первую очередь на описание простейших взаимосвязей между “смысловыми примитивами” химического “текста”, описывающего структуру химического соединения, - *активными центрами*. Не составляет труда расширять его (в случае необходимости, например, при обработке тех или иных конкретных проблемно-ориентированных массивов химических структур) специальными проблемно-ориентированными дескрипторами, характеризующими какие-либо специфические особенности изучаемой предметной области, и порождать более точные (но не универсальные, а проблемно-ориентированные) варианты ЯПЗ в духе обсуждавшейся нами выше иерархии.

Главным достоинством такого типа средств представления знаний является их относительная вычислительная простота: большинство основных процедур их структурного анализа имеют полиномиальную вычислительную сложность и не требует каких-либо особых вычислительных ресурсов. (Обзор ряда приложений в области биохимии, основанных на использовании языка ФКСП, и реализованных на персональных компьютерах, можно найти, например, в работах [*ИНТ 91, Zabezhailo 95, Забежайло 96*] и др.).

Далее каждый раз при обсуждении приложений ДСМ-рассуждений в задаче анализа связи ***СТРУКТУРА*** ==> ***АКТИВНОСТЬ*** везде, где специально не оговаривается способ представления данных, - это будет язык ФКСП. Во всех случаях использования иных средств (например, когда будут задействованы структуры собственно в виде 2-D или 3-D графов) мы будем объявлять об этом отдельно.

П.2.1 Фармакологические приложения.

Исторически сложилось так, что анализ связи ***СТРУКТУРА*** ==> ***АКТИВНОСТЬ*** физиологически активных веществ является одной из наиболее широких областей практических приложений технологий ДСМ-ИАД. Конструирование новых лекарственных препаратов, фармакология, анализ антипродуктивных свойств физиологически активных соединений, медицинская диагностика и рекомендации по выбору терапии - вот наиболее существенные направления приложений в этой области. Осуществленные здесь приложения ДСМ-метода выполнялись в кооперации с экспертами в соответствующих направлениях исследований.

Общая структура всех проводившихся здесь работ такова. Исследовались наличие или отсутствие заранее выбранных физиологических активностей у химических соединений. При этом предполагалось, что для каждой физиологической активности существует конкретный объект - соответствующий фрагмент химической структуры изучаемого соединения, который может рассматриваться как “*причина*” (т.е. *структурный носитель*) анализируемой активности изучаемого соединения[6]. Целью применения ДСМ-инструментария ИАД в приложениях обсуждаемого типа является прогнозирование наличия или отсутствия изучаемых свойств у новых соединений, имеющих подобную (“*сходную*”, “*аналогичную*”, “*похожую*”, ...) исходно заданным соединениям химическую структуру.

Грубая схема принятия решений в таком прогнозировании может быть представлена следующими шагами:

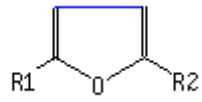
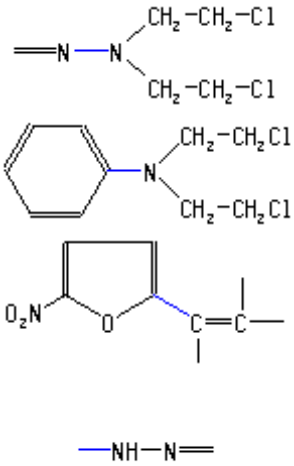
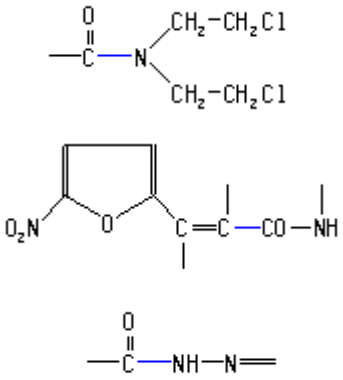
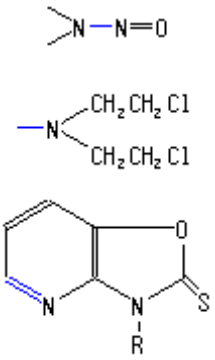
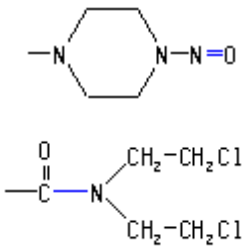
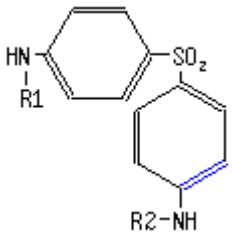
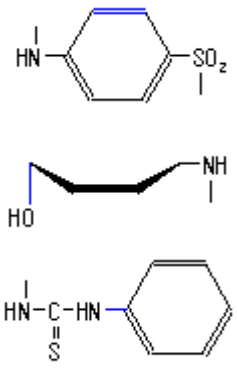
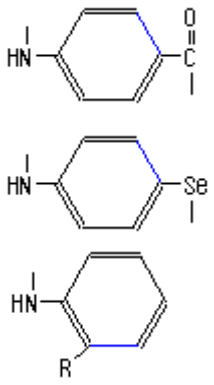
- по исходно заданному множеству примеров и контрпримеров (т.е. множеству соединений, обладающих исследуемой активностью, и множеству соединений, не обладающих исследуемой активностью) в процессе индуктивного обучения на примерах и контрпримерах выделяются соответствующие фармакофоры и антифармакофоры;
- для порожденных множеств фармакофоров и антифармакофоров строятся условия переносимости задаваемых ими “причинных” зависимостей на новые соединения;
- для каждого нового химического соединения строится (с использованием как фармакофоров и антифармакофоров, так и условий переносимости задаваемых ими “причинных” зависимостей) прогноз наличия или отсутствия у него изучаемых активностей.

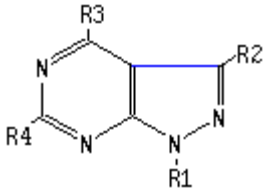
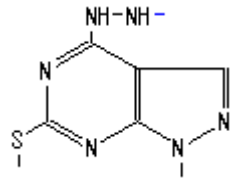
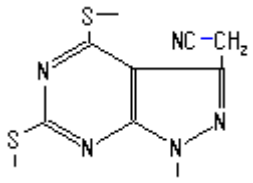
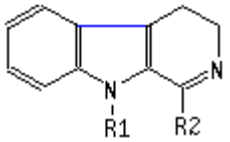
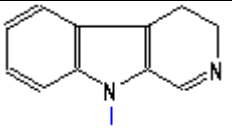
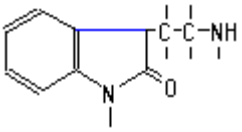
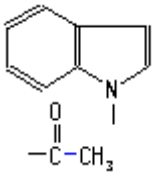
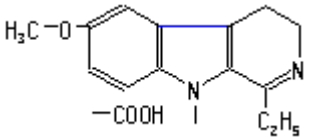
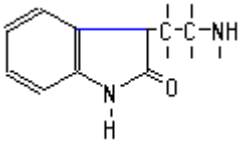
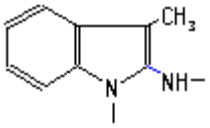
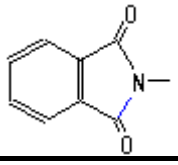
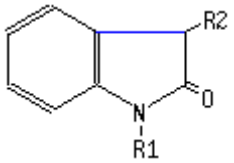
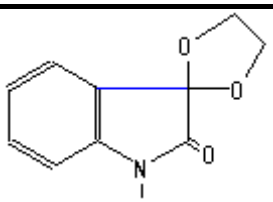
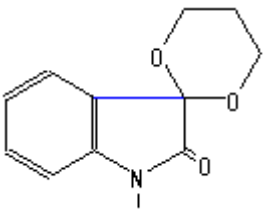
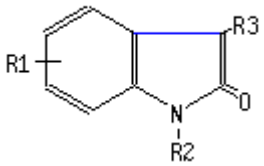
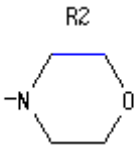
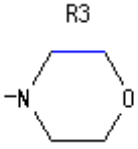
---

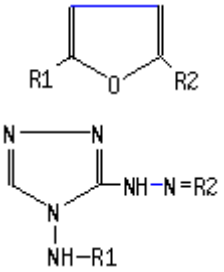
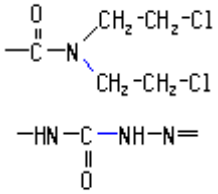
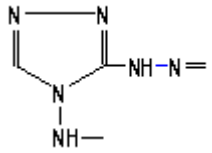
[6] Другими словами, предполагалось, что справедлива гипотеза о так называемых фармакофорах (структурных носителях биологических свойств химических соединений) и антифармакофорах (структурных фрагментах, препятствующих проявлению анализируемых свойств).

В соответствии с представленной схемой было осуществлено несколько десятков серий экспериментов, в которых были проанализированы (в качестве примеров и контрпримеров) по различным физиологическим активностям более 2000 химических соединений. На прогнозирование предлагались около 100 новых химических соединений, для более 90 из которых ДСМ-прогноз физиологических активностей был подтвержден результатами независимых биологических лабораторных тестов. С учетом порожденных фармакофоров и антифармакофоров были синтезированы 5 новых лекарственных соединений. “*In vivo*” и “*in vitro*” тесты этих соединений подтвердили наличие у них физиологических активностей, предсказанных средствами ДСМ-метода. Ниже представлена таблица, дающая обозрение некоторых из обсуждавшихся выше экспериментальных результатов [*Финн 91, Забежайло 83, Kitova 94, Будунова 93, Забежайло 96*].

Таблица П 2.1.1

Химические соединения, активности	**Фармакофоры**	**Антифармакофоры**	**Точный прогноз свойств**
Фураны  **противоопухолевая** 53 соединения			37 из 40
Соединения различных химических классов **противоопухолевая** 67 соединений			7 из 7
Производные дифенилсульфонов  **антилепрозная** 86 соединений			5 из 6

Пиразоло-пиримидины **противоопухолевая**  81 соединение			10 из 10
β-Карболины, Индолы  **антиалкогольная** **гепатопротекторная** 31 соединение	  	   	8 из 11
Изатины, 90 соединений  **антидепрессанты**			21 из 25
Изатины  **антидепрессант** **стимулирующая** **антиконвульсант** **седативная** 90 соединений	антидепрессант 	 антиконвульсант 	21 из 25

Фураны Аминотриазолы  **токсичность** 120 соединений	 

и ежедневные, способ введения и т.п.), сведения о наблюдаемых результатах (наличие или отсутствие канцерогенного эффекта, целевой орган, гистология, количественные характеристики - латентный период, частоты опухолей и т.п.).

В процессе ДСМ-мониторинга используется целый ряд моделей изучения эффекта канцерогенности. В простейшем случае эффект канцерогенности рассматривается с позиций анализа связи ***СТРУКТУРА==>АКТИВНОСТЬ*** химических соединений. В терминах ДСМ-метода изучаемые объекты (химические соединений) представлены описаниями их химических структур и анализируемой активности (свойством вызывать эффект канцерогенности). В соответствии со стандартной схемой принятия решений в рамках ДСМ-метода интеллектуальная система компьютерного мониторинга порождает структурные “причины” и “антипричины” проявления исследуемого свойства, используя их далее в прогнозировании свойств новых (ранее не изученных) соединений.

В более продвинутой модели описание объектов становится более сложным: используются кортежи, компонентами которых являются *тип животного, доза, способ введения* и *химическая структура* соединения. Таким образом появляется возможность прогнозировать эффекты канцерогенеза применительно к конкретным типам животных, способам введения вещества, суммарным дозам и т.п..

Модель следующего уровня детализации предполагает использование механизмов разделения так называемых прямых и непрямых канцерогенов. Интеллектуальная система мониторинга порождает соответствующие фармакофоры как в случае прямых, так и в случае непрямых канцерогенов. Далее для прямых потенциальных канцерогенов система использует модель рассуждений второго уровня (см. выше). Для анализа непрямых канцерогенов используется ДСМ-система анализа метаболических преобразований химических соединений. Для каждого структурного фрагмента, порождаемого в процессе метаболизма, система прогнозирует его потенциальную канцерогенность или неканцерогенность. Исходное изучаемое химическое соединение считается канцерогеном, если хотя бы один из его метаболитов прогнозируется как потенциальный канцероген[8]. Наконец, имеется возможность прогнозировать взаимные влияния химических соединений при выделении потенциальных канцерогенов. В этом случае объект представляется кортежем, элементы которого - пары химических соединений (так сказать “*эффектор*” + ”*катализатор*”), вместе с информацией о способах их введения, дозах и характеристиках экспериментальных животных. Интеллектуальная система мониторинга, фактически прогнозирует наличие или отсутствие эффектов синергизма, антагонизма или аддитивности (каждый из них представлен как соответствующее ДСМ-свойство) пар химических канцерогенов.

Ниже представлен иллюстративный пример рассуждений ДСМ-системы мониторинга потенциально канцерогенных химических соединений [*Панкратова 95, Zabezhailo 94, Zabezhailo 95*]. Цель ДСМ-рассуждения - прогноз наличия или отсутствия канцерогенных свойств у новых соединений на базе информации об известных канцерогенах. Библиотека прецедентов системы сформирована “записями”, характеризующими влияния известных канцерогенов на экспериментальных животных. Каждая “запись” содержит информацию о

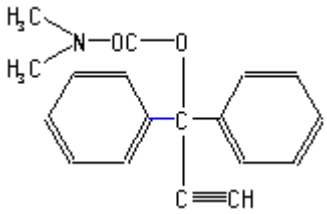
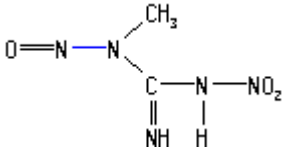
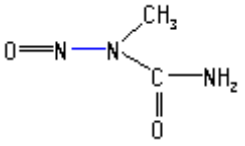
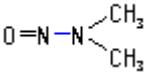
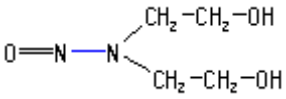
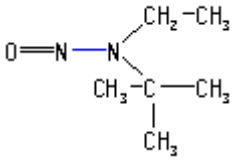
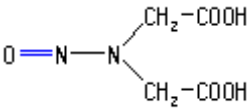
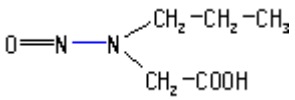
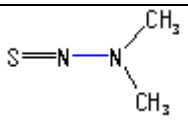
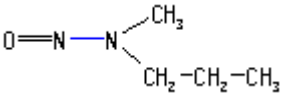
- химической структуре соответствующего соединения (т.е. некоторое изоморфное представление молекулярного графа соединения вместе с информацией о вхождении в него “стандартных” подграфов из наперед заданного списка - списка структурных индексов базы структурных данных;
- типе экспериментальных животных (крыса, хомяк, ...);
- способе введения химического вещества (орально, подкожно, ...);

[8] Представляемая модель может быть проиллюстрирована следующим результатом компьютерного эксперимента, полученным Е.С.Панкратовой и ее коллегами [*Панкратова 95, Zabezhailo 95*]: средствами модели второго типа интеллектуальная компьютерная система мониторинга не смогла прогнозировать наличие или отсутствие потенциальной канцерогенности для 3-гидроксибутил-нитрозамина; используя же сведения о возможных путях метаболических преобразований этого соединения, она построила корректный прогноз, аргументировав его опасность как непрямого канцерогена.

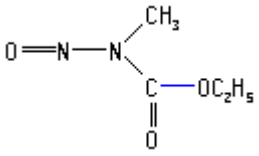
- наличии или отсутствии канцерогенных свойств в соответствующих экспериментах и т.д..

В **Таблице П 2.2.1** представлены 10 “записей” из Библиотеки прецедентов:

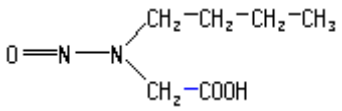
Таблица П 2.2.1

R1) КРЫСА ОРАЛЬНО КАНЦЕРОГЕН		**R2)** КРЫСА ОРАЛЬНО КАНЦЕРОГЕН	
R3) КРЫСА ОРАЛЬНО КАНЦЕРОГЕН		**R4)** КРЫСА ОРАЛЬНО КАНЦЕРОГЕН	
R5) ХОМЯК ПОДКОЖНО КАНЦЕРОГЕН		**R6)** КРЫСА ОРАЛЬНО **НЕКАНЦЕРО-ГЕННО**	
R7) КРЫСА ОРАЛЬНО **НЕКАНЦЕРО-ГЕННО**		**R8)** КРЫСА ОРАЛЬНО **НЕКАНЦЕРО-ГЕННО**	
R9) КРЫСА ПОДКОЖНО **НЕКАНЦЕРО-ГЕННО**		**R10)** КРЫСА ПОДКОЖНО КАНЦЕРОГЕН	

ДСМ-системе предъявлены для прогноза два соединения:



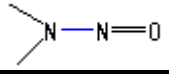
R11



R12 ,

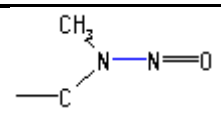
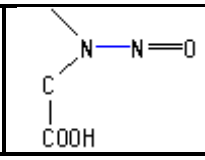
которые предполагается вводить ОРАЛЬНО КРЫСАМ.

Результатом информационного поиска в Базе Прецедентов системы мониторинга по индексам <КРЫСА & ОРАЛЬНО> будет выделение записей **R1-R4, R6-R8**. Предварительный анализ структур **R11** и **R12** с помощью стандартного “списка индексов” (т.е. путем сравнения с входящими в него “стандартными” структурными фрагментами) показал их принадлежность классу так называемых N-нитрозо-комплексов - т.е. классу, который релевантен стандартному индексу вида

R13)	

(фрагмент **R13** выделяется как подструктура для структур **R11** и **R12**). Таким образом, процесс структурного информационного поиска сужает множество подходящих прецедентов-аналогов до { **R2-R4, R6-R8** }.

Механизм адаптации прецедентов (анализа новых структур) реализуется в два этапа. Сперва средствами ДСМ-рассуждений структурная причина канцерогенности **R14** выделяются из объектов-прецедентов **R2-R4**.

R14)		**R15)**	

Затем на объектах-прецедентах **R7-R8**[9] порождается фрагмент **R15** - "структурная причина" антикацерогенности (соединений **R7-R8,** выделенная с точки зрения ДСМ-анализа предложенных системе мониторинга данных).

На втором этапе адаптации система проверяет возможности "переносимости" (средствами рассуждений "по аналогии", формализованных средствами ДСМ-метода) найденных ею структурных "причинных зависимостей" на предложенные ей для прогноза объекты. Здесь порождаются следующие результаты:

1) соединение **R12** прогнозируется как **НЕКАНЦЕРОГЕННОЕ** в предложенном для обучения системы мониторинга контексте;

2) в рамках ДСМ-рассуждений (на имеющемся для обучения контексте) не удается найти *достаточные основания*[10] для переноса порождаемых эмпирических зависимостей на соединение **R11**.

Наконец, выделенные в процессе работы системы мониторинга структурные фрагменты **R14** и **R15** (характеризующие "структурные причины" канцерогенности и неканцерогенности элементов Библиотеки Прецедентов), присоединяются к индексной структуре Библиотеки Прецедентов для последующего использования в анализе новых соединений.

П 2.3 Мониторинг потенциально опасных химических соединений.

Проблема мониторинга[11] потенциально опасных (т.е. канцерогенных, токсичных, мутагенных и т.п.) химических соединений - один из наиболее важных полигонов для применения современных компьютерных систем и информационных технологий. К концу XX века человечеству стали известны (например, описаны в соответствующих базах данных) около 20×10^6 химических соединений. Оборотной стороной технического прогресса человечества здесь оказывается постоянный прирост представленного множества структур ежегодно на 4×10^5 - 5×10^5 новых соединений. На сегодняшний день около 10 000 веществ выделены из общего количество и изучены (в ходе лабораторных экспериментов) как канцерогены. Примерно тот же порядок величины характеризует количество изученных на текущий момент токсинов. При этом полный цикл лабораторных экспериментов по изучению, например, (потенциальных -!) канцерогенных свойств нового соединения требует примерно трех лет исследований и финансирования в размере примерно 10-15 $\times 10^6$ US$. Легко видеть, что в таких условиях единственным практически осуществимым путем, позволяющим контролировать (т.е. выделять и изучать, прогнозируя проявления опасных физиологических свойств) потенциально опасные химические вещества в среде обитания человека,

---

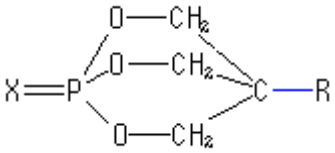
[9] Сходство всех трех структур **R6-R8** выражает уже иной - отличный от **R15** фрагмент.

[10] См. более подробно выше - **Главы 2-3**.

[11] Здесь имеется ввиду экологический мониторинг опасных химических соединений в окружающей человека природной и технологической среде.

оказывается построение интеллектуальных компьютерных систем экологического мониторинга [*N-YA 88, ИНТ 91, Забежайло 96* и др.]. Техника формализованных рассуждений, реализуемых в рамках интеллектуального анализа данных средствами ДСМ-метода может быть выбрана в качестве основания для, по-видимому, одной из наиболее адекватных промышленных технологий в обсуждаемой нами предметной области. Порождая (в процессе индуктивного обучения на примерах - известных опасных химических соединениях, в частности - канцерогенах - см. выше **Раздел П 2.2** - и контрпримерах - структурно похожих на известные канцерогены, но безопасных веществах) “причинные” зависимости, интеллектуальная система компьютерного мониторинга способна учесть их в анализе новых изучаемых соединений[12]. ДСМ-метод используется здесь как математическое основание при построении интеллектуальной компьютерной системы представленного вида [*Zabezhailo 95*].

В **Разделе П 8.1** демонстрируется пример экспериментов подобного типа. Дадим подробное описание послуживших для него основой результатов. Речь идет о изучении токсичности некоторого специального класса химических структур, проявляющих противосудорожную активность. Исходные данные - 36 соединений из ряда так называемых производных 1-фосфо-2,6,7-три-оксабицикло[2.2.2]октанов. Общий вид изучаемых структур



,

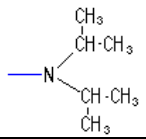
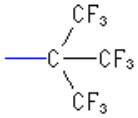
а точный вид радикалов X и R для каждого из 36 рассматриваемых соединений вместе с обсуждением их свойств можно найти ниже - в **Таблице 8.2.3.1** (см. также, например, работу [*Блинова 95*]). Предметом компьютерного анализа (средствами ДСМ-рассуждения) была токсичность исследуемых соединений: 13 из 36 изучаемых веществ были представлены как токсичные, 20 - как нетоксичные, а для оставшихся 3 было необходимо построить прогноз наличия или отсутствия искомого свойства - токсичности.

Таблица П.2.3.1

N	***Исходные данные***	(структуры)	***Свойства***[13] (токсичность)
	R	**X**	
1	$CH_3-CH_2-CH_2-$	O	+
2	$CH_3-CH(CH_3)-$	O	+
3	$CH_3-CH(CH_3)-$	S	+
4	$CH_3-CH(CH_3)-$	Se	+
5	$CH_3-CH(CH_3)-$	••	+
6	$CH_3-CH_2-CH_2-CH_2-$	O	+
7	$CH_3-CH_2-CH_2-CH_2-$	S	+
8	$CH_3-CH_2-CH_2-CH_2-$	Se	+
9	$CH_3-CH_2-CH_2-CH_2-$	••	+
10	CH_3-CH_2-	O	+

[12] Т.е. задействовать в процессе анализа новых прецедентов, когда для каждого нового соединения в Библиотеке прецедентов (например, базе данных об известных канцерогенах) порождаются необходимые “причинные” зависимости (т.е. порождаются соответствующие фармакофоры и антифармакофоры), которые и используются как инструмент прогноза потенциально опасных свойств у заранее не изученных соединений.

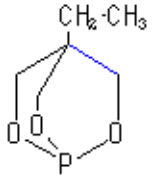
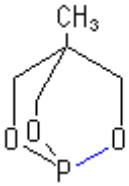
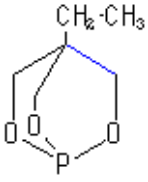
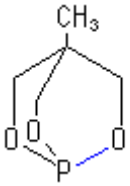
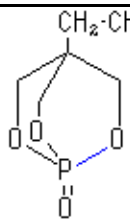
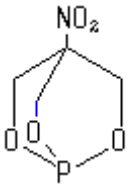
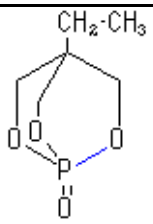
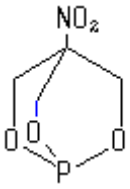
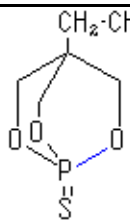
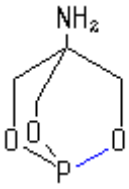
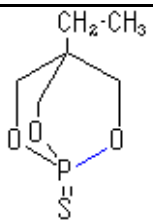
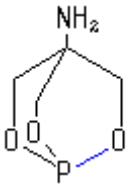
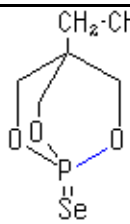
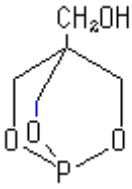
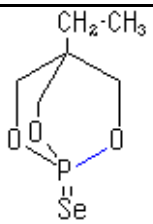
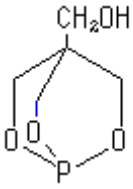
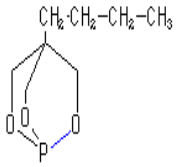
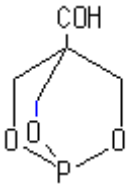
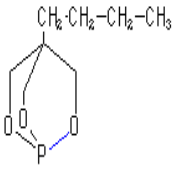
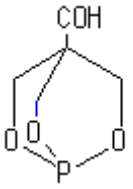
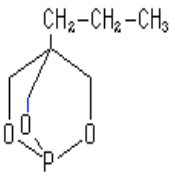
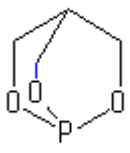
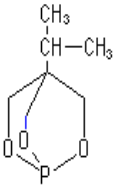
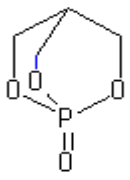
[13] Здесь (+) означает - токсично, (-) - нетоксично, а τ - требует прогноза.

11	CH_3-CH_2-	S	+
12	CH_3-CH_2-	Se	+
13	CH_3-CH_2-	••	+
14	CH_3-	O	–
15	CH_3-	S	–
16	CH_3-	Se	–
17	CH_3-	••	–
18	NO_2-	O	–
19	NO_2-	S	–
20	NO_2-	Se	–
21	NO_2-	••	–
22	NH_2-	O	–
23	NH_2-	S	–
24	NH_2-	Se	–
25	NH_2-	••	–
26	$-CH_2OH$	O	–
27	$-CH_2OH$	S	–
28	$-CH_2OH$	Se	–
29	$-CH_2OH$	••	–
30	$-COH$	O	–
31	$-COH$	S	–
32	$-COH$	Se	–
33	$-COH$	••	–
34	$-CF_3$	O	τ
35		O	τ
36		O	τ

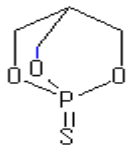
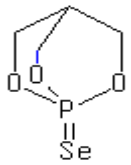
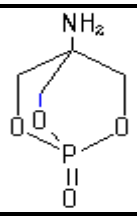
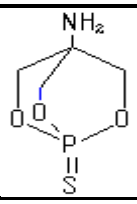
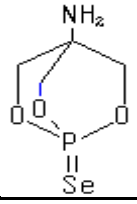
Обработка указанных данных средствами ДСМ-рассуждения осуществлялась в двух вариантах: для случая кодирования структур в дескрипторном языке ФКСП (см. выше в настоящем **Разделе**), и для случая представления структур как 2-D молекулярных графов.

Примеры порождаемых причинных зависимостей - *фармакофоров*[14] и *антифармакофоров*[15] - представлены ниже в **Таблице П 2.3.2**. При этом важно заметить, что для обоих видов представления данных общий вид причинных зависимостей (т.е. их структура в целом) совпадает.

Таблица П 2.3.2

	Ф К С П		***Молекулярные графы***	
N	позитивные	негативные	позитивные	негативные
1				
2				
3				
4				
5				
6				
7				

[14] Т.е. позитивных причинных зависимостей.
[15] Т.е. негативных причинных зависимостей.

8				
9				
10				
11				
12				

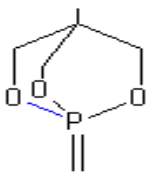
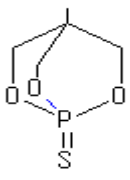
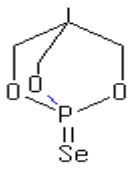
Кроме того, в случае обработки графов порождается больше гипотез (что естественно, т.к. используется более детальное описание данных), и это влияет на результаты прогноза: в ФКСП-эксперименте нет возможности осуществить ДСМ-прогноз свойств трех последних соединений в **Таблице П 2.3.1** на *достаточном основании*. В то же время в случае работы с молекулярными графами удается в рамках ДСМ-рассуждения построить на достаточном основании прогноз свойств всех трех тестируемых соединений:

- 34 соединение - нетоксично,
- 35 соединение - токсично,
- 36 соединение - нетоксично.

(Результаты этого компьютерного прогноза полностью подтверждаются биологическими испытаниями).

Ниже в **Таблице П 2.3.3** представлены примеры эмпирических зависимостей, которые порождаются на второй стадии экспериментов (т.е. в графовом представлении), но не порождаются на их первой стадии (т.е. при использовании кода ФКСП[16]).

Таблица П 2.3.3

Hpt₁	***Hpt₂***	***Hpt₃***
		

[16] Код ФКСП оказывается слишком “грубым”, чтобы представить именно эти фрагменты (без каких-либо структурных добавок.

Еще одной важной и поучительной сферой приложения ДСМ-подходов представленного класса стала разработанная совместными усилиями специалистов ВИНИТИ РАН, ГосНИИОХТ технология[17] прогноза так называемой отставленной нейротоксичности химических соединений (см., например, [*Johnson 69, Bondy 86, Hollingshaus 82, Aldridge 90, Хохлов 90*] и др.). Опасность развития отдаленных нейротоксических поражений человека в связи с широким применением всевозможных химических соединений в промышленности, сельском хозяйстве, быту - вот наиболее актуальные основания для осуществления обсуждаемого проекта.

Целями данного проекта были провозглашены:

- разработка информативных тест-систем ранней диагностики развития отдаленных нейротоксических эффектов;
- создание проблемно-ориентированной базы данных химических соединений, проявляющих отставленное нейротоксическое действие;
- разработка и создание программных средств для молекулярного моделирования и прогноза нейротоксических свойств соединений, проявляющих отдаленное нейротоксическое действие.

Особенностью проекта было использование при внеэкспериментальном прогнозировании отставленных нейротоксических эффектов оригинальной системы компьютерного дизайна физиологически активных веществ. Система имела гибридную модель представления знаний, накапливаемых в процессе индуктивного обучения на примерах.

Результатами исследования наряду с фундаментальными теоретическими разработками, касающимися создания базы данных, классификации отдаленных нейротоксических эффектов, установления корреляций между невропатологическими и биохимическими проявлениями интоксикации, должны были стать

- создание скрининговых тест-систем, отличающихся простотой, надежностью и дешевизной,
- внеэкспериментальное прогнозирование нейротоксических свойств соединений,
- проведение скрининга новых и широко применяемых химических веществ на отставленную нейротоксичность.

Перспективность избранного направления по осуществлению прогнозирования и экспресс-диагностики отдаленной нейротоксичности химических веществ отражена в ряде зарубежных и отечественных работ, а также в работах одного из соисполнителей проекта - ГосНИИОХТ.

Программа работ проекта состояла из 3 этапов, каждый из которых имел собственную научную и практическую ценность.

ПЕРВЫЙ ЭТАП был ориентирован на проведение аналитического исследования, направленного на разработку критериев оценки отставленной нейротоксичности и классификацию соединений по характеру специфического действия. Этап предполагал также формирование базы данных и осуществление синтеза модельных соединений.

ВТОРОЙ ЭТАП был связан с разработкой программной системы для внеэкспериментального прогнозирования нейротоксических свойств соединений, а также с разработкой информативных биохимических и нейрофизиологических критериев характеристики отдаленной нейротоксичности.

ТРЕТИЙ ЭТАП предполагал синтез соединений в соответствии с прогнозом и разработку тест-систем для экспресс-диагностики отдаленных нейротоксических эффектов по биохимическим параметрам.

Структура работ по этапам:

---

[17] Эти работы проводились, в частности, в рамках исследовательского проекта Московского Международного Научно-Технического Центра (ISTC) “Экспресс-диагностика развития отдаленных нейротоксических эффектов физиологически активных веществ” (регистрационный номер проекта в МНТЦ - 574).

А-1. Проведение аналитического исследования, направленного на разработку критериев оценки отставленной нейротоксичности и классификацию соединений по характеру специфического действия. Создание базы данных химических соединений, обладающих отставленным нейротоксическим действием и синтез модельных соединений.

А-1.1. Анализ литературы.
А-1.2. Разработка критериев оценки отставленной нейротоксичности.
А-1.3. Выбор соответствующего программного инструментария, разработка обоснования структуры базы данных.
А-1.4. Создание базы данных.
А-1.5. Классификация нейротоксикантов по механизму отставленного действия.
А-1.6. Синтез модельных соединений.
А-1.7. Выбор экспериментальной модели для исследования влияния химических веществ на отставленную нейротоксичность.
А-1.8. Создание специализированных программных средств для планирования и обобщения результатов токсикологических экспериментов.

Было запланировано, что результатом первого этапа исследования станет классификация веществ, обладающих нейротоксичностью в соответствии с их характером действия (по механизму - центральное либо периферическое действие, по времени проявления эффектов - острое действие, отставленные, отдаленные последствия).

Создание базы данных имеет собственное научное и практическое значение, поскольку такая БД может быть основой для целенаправленного синтеза соединений с заданными свойствами и разработки скрининговых тест-систем.

А-2. Создание программ для внеэкспериментального прогнозирования нейротоксических свойств соединений. Разработка информативных биохимических и нейрофизиологических критериев характеристики отставленной нейротоксичности.

А-2.1. Разработка методики каузального анализа и внеэкспериментального компьютерного прогноза нейротоксических свойств химических соединений.
А-2.2. Синтез соединений, позволяющих создать обучающие выборки для прогнозных исследований.
А-2.3. Выбор наиболее информативных физиологических методов, позволяющих оценить характер и тяжесть ОНТ. (Для изучения периферического действия было решено оценить электромио- и нейрографические методы а также методы, оценивающие состояние двигательной активности, нарушение координации, физической работоспособности животных. Для исследования центрального действия было решено использовать поведенческие, условно-рефлекторные методы, оценивать состояние памяти, обучаемости и т.п..)
А-2.4. Поиск биохимических показателей, изменения которых сопряжены с развитием ОНТ. Исследование ряда ферментов, в частности, β-галактозидазы, β-глюкуронидазы, и сдвигов в метаболических реакциях фосфолипидного обмена.
А-2.5. Исследование зависимости степени биохимических нарушений от тяжести интоксикации.
А-2.6. Выбор соответствующих программных инструментов и создание программной системы молекулярного дизайна химических соединений, ориентированной на проблему каузального анализа и прогноза отставленной нейротоксичности.

В результате второго этапа планировалось выбрать наиболее информативные биохимические и патофизиологические методы для количественной оценки степени ОНТ, установить корреляции между невропатологическими и биохимическими проявлениями

интоксикации с использованием соединений, обладающих отставленным центральным или периферическим действием.

Практическая ценность работ этого этапа связана прежде всего с разработкой программной системы молекулярного моделирования, каузального анализа и внеэкспериментального прогноза нейротоксических свойств химических соединений.

А-3. Синтез соединений с прогнозированными свойствами, обладающих отставленной нейротоксичностью, и разработка тест-систем для ранней диагностики отдаленных нейротоксических эффектов по биохимическим и патофизиологическим параметрам.

А-3.1. Внеэкспериментальный анализ свойств новых соединений средствами созданной на втором этапе компьютерной системы молекулярного дизайна.
А-3.2. Выбор моделирующего вида животных для скрининговых тест-систем.
А-3.3. Установление особенностей ОНТ, зависящих от способа введения, кратности воздействия и дозы веществ.
А-3.4 Разработка скрининговых тест-систем для ранней диагностики ОНТ на основе биохимических показателей.
А-3.5. Синтез прогнозированных соединений для экспериментальной проверки полученных результатов.
А-3.6. Проведение скрининга вновь синтезированных и широко применяемых в хозяйстве препаратов на предмет апробации разработанных тест-систем.

Результатом выполнения третьего этапа должен был стать синтез новых соединений на основе внеэкспериментального прогнозирования свойств веществ, разработка тест-систем для ранней диагностики отдаленных нейротоксических эффектов, скрининг новых и ранее известных соединений с помощью разработанных тест-систем.

Предполагалось, что осуществление данного проекта позволит специалистам, связанным в прошлом с изучением соответствующего механизма действия для целенаправленного синтеза высокотоксичных веществ, в рамках международных усилий по конверсии оборонных производств[18] применить свои знания с максимальной пользой для предсказания опасных свойств соединений, ограничения их использования в народном хозяйстве, защиты здоровья людей и окружающей природы. Также было понятно, что успешное завершение этой работы позволит разработать нормативную документацию, регламентирующую производство и применение веществ, обладающих ОНТ, в медицине, сельском хозяйстве и других областях народного хозяйства.

Подготовка работ по проекту (включавшая в себя не только создание и утверждение соответствующих документов, но и необходимые проработки исследовательского характера) заняла примерно полтора года. Полностью оформленное, прошедшее все необходимые экспертизы и официально поддержанное Правительством Российской Федерации проектное предложение было представлено в МНТЦ. Исследовательская часть работ по проекту была представлена (и вызвала одобрение научного сообщества) на ряде семинаров и конференций (в частности, на семинаре МНТЦ в ГосНИИОХТ в 1996 году и на проводившемся в декабре 1996 года в Лаборатории физических наук Университета штата Нью-Мексико, Las Crusas, NM, USA, Семинаре, посвященном Квази-аксиоматическим теориям). Тем не менее, проект не получил по линии МНТЦ необходимого финансирования[19] и остался незавершенным.

[18] А для решения именно этой задачи - для содействия конверсии оборонных производств - создавался МНТЦ.
[19] По мнению (разумеется, субъективному) руководителей проекта, такой ход событий был обусловлен (вполне понятным, однако неприемлемым российскими участниками) стремлением западных партнеров получить полный доступ к используемым в проекте уникальным технологиям без урегулирования надлежащим образом авторских и имущественных прав разработчиков.

Приложение 3.

Text-mining: текст-процессинг и задачи компьютерного анализа содержания полнотекстовых документов.

Сегодня поиск и отбор (в множестве найденных «кандидатов») полезной информации по той или иной теме (например, поиск в расположенных в интернете материалах сведений по тому или иному конкретном вопросу) уже невозможно себе представить без использования соответствующих компьютерных систем-«ассистентов». Сообщество специалистов, заинтересованных в разработке подобного рода инструментария, постоянно расширяется. Однако арсенал моделей и подходов, которые нашли результативное использование в прикладных программных системах, был и остается в своей основе мало изменившимся, будь то системы автоматического перевода с языка на язык, системы автоматического отнесения текстов к темам или же, например, системы автоматического реферирования (аннотирования) полнотекстовых документов. В самом общем виде используемая здесь «конструктивная схема» выглядит следующим образом:

- в качестве «атомарных» индексов – «носителей смысла» обрабатываемых текстов (представляющих элементы содержания текстов) выступают определенные слова и словосочетания;
- определенные комбинации таких «атомарных индексов» (смысловые комплексы индексов) формируются (реконструируются в ходе специализированного текст-процессинга) по наборам текстов: каждому набору сходных по содержанию текстов ставится в соответствие характерный (для их содержания) набор комплексов смысловых индексов;
- целью такого сопоставления (*тексты <=> смысловые индексы*) является использование сформированного на примерах[1] отношения индексирования текстов «носителями смысла» для решения прикладных задач;
- описания применения этого подхода несложно найти, например, в задачах автоматической классификации текстов по темам[2], автоматического аннотирования полнотекстовых документов (text summarization)[3] или автоматической оценки эмоциональной окрашенности текстовых сообщений [4].

С точки зрения обсуждаемой нами проблематики ИАД в связи с оценкой результатов, получаемых в подобном текст-процессинге, можно сформулировать ряд принципиальных вопросов, в том числе:

(1) Как быть с формированием утверждений о связности «смысловых индексов», порождаемых на заданном массиве текстов[5], если на одних признаках (порожденных одними средствами экспертизы) такая связность имеется, а на других – нет?

(2) Как быть с несовпадающими автоматически порожденными аннотациями: они «говорят» об одном и том же, или же о разном[6]?

(3) Как быть с множественностью средств экспертизы, которая естественным образом связана с устойчивостью ее результатов[7]?

---

[1] Частично-определенного - !

[2] См., например, модель связности текстов в работе [*Дейк 88*] др..

[3] См., например, [*Nenkova 05*] а также регулярно публикуемые материалы ACM SIGIR, SIGMOD и др..

[4] См., например, данные о проектах мониторинга «боевого духа армии» в работах [*Rosso 14*] и др. .

[5] Например, в представлении «смыслового объекта» как многословного выражения при переходе от отдельных слов к языковым выражениям.

[6] При дальнейшем углублении в тему здесь приходится также иметь дело с трудным понятием *детальности* представления «смыслов» текста, например – *совпадение* аннотаций следует понимать как совпадение в «главном»? Что такое *близость* (в том числе – в большей или меньшей степени) *текстов* по смыслу? Наконец, как быть, если лексика двух текстов близка, однако, тексты – о разном (например, оба – о банковском клиринге, однако, один – о его правовой базе, а другой – об особенностях алгоритмики)?

[7] В частности – при расширении используемого набора текстов.

Конечно, выделенные нами «шероховатости» обычно пребывают за пределами области внимания большей части исследовательского сообщества: обсуждаемые компьютерные средства автоматизации текст-процессинга в первую очередь призваны сократить объемы физически просматриваемого специалистом массива текстовых документов (что они, в общем-то, и позволяют сделать). Однако, если взглянуть на проблему, так сказать, с «академической» точки зрения, то трудно представить себе, как подчистка «вручную» полноты и точности результатов автоматического текст-процессинга могла бы стать приемлемой (т.е. реализуемой на практике надежной) процедурой в ситуациях, когда приходится иметь дело с (выполняемым в разумное время) исчерпывающим анализом больших коллекций полнотекстовых документов.

Переходя к представлению возможностей ДСМ-ИАД в рассматриваемых здесь задачах компьютерного текст-процессинга, представляется уместным выделить следующие принципиальные моменты:

- развиваемая в рамках ДСМ-метода техника представления данных (возможности согласованного перехода от теоретико-множественных – традиционных дескрипторных – языков представления данных к более выразительным конструкциям вида *множества индексов + отношения на них*[8]) позволяет варьировать доступные дескриптивные возможности интеллектуального анализ данных;
- используя выразительные возможности различных ДСМ-правил правдоподобного вывода а также возможности их комбинирования в те или иные стратегии, можно варьировать используемые инструменты ИАД (обеспечивая поиск наиболее приемлемых результатов при имеющихся исходных данных а также средствах экспертизы);
- в процессе описываемого варьирования средств экспертизы сравнивать между собой получаемые результаты, чтобы выбрать из них наилучшие,

а также

- вести анализ устойчивости порождаемых результатов текст-процессинга (экспертизы) при расширении исходной коллекции текстов, в том числе – и в поиске приемлемых ответов на сформулированные чуть выше вопросы (1)-(3).

При таком подходе повышается уровень доверия к порождаемым в ходе реализуемого текст-процессинга результатам, появляются дополнительные возможности для управления его полнотой и точностью.

***

Переходя к подробному обсуждению приложений техники ДСМ-рассуждений в задачах текст-процессинга (т.е. в задачах интеллектуальной обработки текстов на естественном языке), прежде всего мы обратимся к проблематике информационного поиска. Сегодня около 80% используемой обществом информации находится в неструктурированной форме (в том числе, значительная часть - в виде текстов на бумажных носителях), и лишь примерно 20% - структурированные (так называемые операционные) данные, основная часть которых - числовая информация в тех или иных БД. По мнению многих аналитиков в обозримом будущем эта пропорция не претерпит заметных изменений. Более того предполагается (см., например [*Gotthard 97*] и др.), что неструктурированные текстовые данные будут доминирующим типом информации, загружаемой в разного рода хранилища в режиме ONLINE (электронная почта и Internet, электронная коммерция, электронный документооборот, в том числе - документация по новым технологиям и оборудованию - вот лишь некоторые области, где уже сегодня ощущается особая роль компьютерного анализа полнотекстовых документов).

---

[8] Причем как *статическом* (характеризующем *со-встречаемость индексов*, которая и характеризует определенную *тему*), так и в *динамическом* (характеризующем *последовательности со-встречаемости* статических комплексов индексов в тексте, представляемые цепочками индексов или же их групп) вариантах.

Отсутствие достаточно глубоких представлений о внутренней структуре, отражающей содержание документов, вместе с громадными объемами (технически весьма эффективно поддерживаемых) компьютерных хранилищ информации ставят задачу автоматического “извлечения” неявно содержащейся в текстах важной (для той или иной задачи или ПО) информации в ряд первоочередных задач сегодняшнего и перспективного развития computer science.

Информационно-поисковые системы (т.е. вопросно-ответные системы, ориентированные на поиск информации в массивах текстовых документов) по праву считаются одним из типовых классов систем искусственного интеллекта. Впервые заявив о себе в начале 50-х годов (см., например, ставшую классической работу [*Ланкастер 72*]), интеллектуальные системы этого типа прошли длительный и поучительный путь эволюции как в совершенствовании реализуемых в них конкретных технологий, так и в осознании фундаментальных характеристических особенностей области их применения. В самом общем виде фундаментальные свойства обсуждаемой предметной области (ПО) могут быть представлены следующей цепочкой.

Во-первых, извлечение данных и знаний из неструктурированных документов - это, прежде всего, работа с текстами на естественном языке (ЕЯ). В свою очередь, ЕЯ - *открытая система*, изменяющаяся во времени (в том числе - пополняемая новыми данными и знаниями). Таким образом, формальное описание ПО при использовании компьютерных систем рассматриваемого типа требует применения специфических средств формализации - открытых теорий.

Во-вторых, качество поиска в текстах информации, соответствующей конкретному запросу, определяется качеством формализации *отношения релевантности* запросов и содержимого текстовой базы данных (БД), к которой они адресуются. В подавляющем большинстве практически важных случаев рассматриваемое отношение оказывается плохо формализованным (например, частично-определенным: известны конкретные примеры документов как релевантных, так и не являющихся релевантными некоторым запросам. В остальных случаях эксплицитно заданная информация о релевантности запросов и документов в БД попросту отсутствует).

В-третьих, недостаточный *уровень формализованности* делает процесс информационного поиска комбинацией “искусства” и “методов”, где успех применения того или иного “инструментального средства” существенным образом зависит от искусства выбора конкретной области его использования. Именно этим обстоятельством объясняется здесь роль аналогий в ведении информационного поиска компьютерной системой и человеком-специалистом.

Таким образом, сравнивая качество поиска нужных данных современными интеллектуальными информационными системами с качеством “ручного” поиска необходимых сведений специалистом, представляется естественным сфокусировать внимание на проблеме *анализа содержания* текстов на ЕЯ и возможностях ее решения (в первую очередь - приближенного) компьютерными средствами.

Следуя традиционному толкованию метафоры “искусственный интеллект”, мы воспользуемся (на сколько это возможно) параллелями

действия человека-эксперта	<==>	***их модель, реализованная в интеллектуальной компьютерной системе***

для прояснения исходных принципов нашего подхода. Поиск нужной информации в массивах текстов, осуществляемый экспертом, основан в первую очередь на понимании содержания просматриваемых документов. При этом мы будем считать, что понимать содержание текста означает, помимо прочего, уметь

- определить его тематическую направленность (т.е. отнести его к тем или иным конкретным темам, тематическим категориям), а также
- выделить в его содержании наиболее важные понятия[9] и отношения (сопоставив их обсуждаемым в тексте тематическим категориям).

Именно поэтому нас прежде всего будут интересовать возможности автоматической *реконструкции понятий* по текстам обсуждающих их документов, возможности *экспликации взаимосвязей* соответствующих понятий (и, возможно, отношений между ними) с темами обсуждающих их документов, и, наконец, возможности автоматического отнесения вновь рассматриваемых ЕЯ-текстов к конкретным темам (опираясь на автоматически извлекаемую из текстов понятийную структуру).

Итак, предмет нашего рассмотрения - интеллектуальные системы обработки текстов на ЕЯ, ориентированные на задачи информационного поиска и так называемого data mining'а[10] в больших текстовых БД. В стремлении продвинуться к автоматическому "пониманию" ЕЯ-текстов компьютерными системами мы сфокусируем внимание на проблеме автоматической реконструкции понятий по текстам обсуждающих их ЕЯ-документов. Предлагаемый нами подход объединяет машинное обучение (более точно: основанное на дометрическом[11] анализе сходства примеров и контрпримеров машинное обучение) и традиционный лингвистический анализ (морфологический, синтаксический и семантический анализ, использование специальных словарей и т.п.). Потенциальные возможности подхода мы проиллюстрируем обсуждением некоторых экспериментальных результатов. В заключение представлены некоторые практически важные содержательные задачи, которые могут быть эффективно формализованы обсуждаемыми средствами.

П 3.1. Проблема аппроксимации понятий в полнотекстовом информационном поиске.

Традиционна схема формализации отношения релевантности запросов и содержимого документов в текстовой БД, основанная на концепции поисковых образов документов, была принята еще в 50-е годы (см., например, уже упоминавшуюся работу [*Ланкастер 72*] и др.). В самом общем виде она представима следующей последовательностью шагов.

- Каждому помещаемому в БД документу сопоставляется соответствующее множество индексов некоторого дескрипторного языка (например, ключевых слов, терминов рубрикатора или какой-либо другой классификационной системы и т.п.).
- Вместе с хранимым (в той или иной форме) документом БД хранит сопоставленные ему индексы (так называемый поисковый образ документа).
- При поступлении запроса к БД таковой индексируется индексами используемого дескрипторного языка (т.е. строится так называемый поисковый образ запроса).
- Поиск в БД документов, релевантных запросу, осуществляется сравнением (например, анализом соответствующей вложимости или непустоты пересечения) поискового образа запроса и поисковых образов документов в БД.

---

[9] Понимаемых здесь в содержательном смысле. О возможностях их формального уточнения мы поговорим позднее.

[10] Data mining - термин, используемый в последнее время в специальной литературе для представления широкого круга задач поиска зависимостей и интеллектуального анализа данных (см. выше – **Главу 1** - и др.).

[11] Т.е. реализуемом более слабыми, чем метрические, средствами.

Приведенная схема и сегодня лежит в основе ряда активно используемых механизмов организации поиска в текстовых БД. (Достаточно, например, упомянуть семейство конкретных систем, использующих для поиска в полных текстах документов так называемый пословный полнотекстовый индекс[12] и порождаемые на его основе в современных промышленных СУБД конструкции).

Основным ограничением этой простой и прозрачной поисковой схемы оказались трудно преодолимые проблемы с полнотой и точностью[13] информационного поиска. Практика показала, что дескрипторный механизм индексирования не дает достаточно эффективных средств управления одновременно и полнотой и точностью поиска. Сегодня понятно, что использованные в этой схеме *выразительные средства* дескрипторных языков индексирования оказались слишком бедными для адекватного представления содержащихся в ЕЯ-текстах *понятий*. Показано (см., например, [*Хвостова 97, Финн 99б*] и др.), что в ряде предметных областей формализованная структура понятия может быть представлена наличием некоторого ”ядра” и его “окружения“ (характеризуемого смысловыми связями “ядра” с релевантной ему информацией). В такой ситуации дескрипторные языки индексирования способны представлять в лучшем случае лишь исходный материал (т.е. те или иные “составные части”) для возможного описания понятий, полностью игнорируя необходимые взаимосвязи между такими “составными частями”, а также не давая гарантий “полноты” используемого ими в каждом конкретном случае исходного множества “составных частей”.[14]. Учитывая приведенные соображения, легко видеть, что идеальной модификацией приведенной выше схемы формализации отношения релевантности и организации информационного поиска был бы переход к индексированию ЕЯ-текстов обсуждаемыми в них понятиями. Любые продвижения в этом направлении потребуют ответа на следующие вопросы:

- Какой лексический материал (и каким образом) будет выбираться для индексирования ЕЯ-текстов по предлагаемой новой схеме?
- Каким образом из выбранных “лексических примитивов” будут порождаться “комплексы” - те или иные приближенные представления обсуждаемых в анализируемых текстах понятий?

Обзору некоторых практически важных вариантов ответа на эти два вопроса мы и посвятим дальнейшее обсуждение.

Можно предложить целый ряд вариантов модернизации классической схемы информационного поиска, связанных с тем или иным конкретным вариантом учета структуры характерных для соответствующей предметной области (ПО) понятий. Сама по себе идея аппроксимации понятий, основанной на выделении в заданной (ограниченной ПО) “лексических примитивов” и специфических правил объединения их в “комплексы”, не нова. Как мы уже показали ранее, в той или иной форме параллельные ей идей развивались в различных направлениях лингвистических исследований - от генеративной лингвистики (см., например, [*Chomsky 82, Степанов 75*] и др.) и компонентного анализа (см., например, [*Кузнецов 86, Nida 75*] и др.), до работ в области синтагматики и порождающей семантики (см., например, [*Степанов 75, Апресян 74*] и др.). Интерес к этой проблематике достаточно высок

---

[12] Пословный полнотекстовый индекс - морфологически нормализованный список встречающихся в анализируемом тексте слов, исключающий так называемые стоп-слова.

[13] *Полнота* поиска характеризуется долей всего множества релевантных запросу документов, попавших в ответ на предложенный запрос. *Точность* поиска характеризуется долей релевантных запросу документов в ответе на него.

[14] Чтобы проиллюстрировать это обстоятельство, сошлемся на известный парадокс Лахути-Финна [*Финн 81, Финн 84*]: хорошая документальная поисковая система дескрипторного типа должна быть основана на достаточно “полной” системе индексов-дескрипторов, которая подразумевает наличие достаточно полного фактографического описания исследуемой ПО (т.е. существование в данной ПО достаточно эффективной фактографической поисковой системы). Но такое описание, в свою очередь, требует поиска и отбора подходящих документов в текстах о ПО, т.е. наличия подходящей документальной поисковой системы.

и в современных разработках. Переход от дескрипторных языков к поиску по понятиям декларируется и в ряде популярных коммерческих систем текст-процессинга (см., например, [*OCT, EXCALIBUR*] и др.). Например, предлагаемый компанией **EXCALIBUR Technologies Corp**. [*EXCALIBUR PS*] инструментарий для поиска текстовой информации ***RetrievalWare*** провозглашает одним из своих основных преимуществ по отношению к другим системам *поиск по понятиям*, а не просто по словам. Для работы с понятиями здесь поддерживается специальная семантическая сеть, объединяющая

- словарь (учитывающий морфологическую нормализацию) с перечнем смысловых значений его элементов; а также
- тезаурус, описывающий семантические связи между различными значениями слов из словаря (с учетом типов и силы таких связей).

Однако следует подчеркнуть, что представление понятий в такой сети носит лишь *виртуальный* характер: *границы* понятий явным образом не выделены, и пользователь при уточнении своего поискового запроса волен сам ограничить "глубину обзора" соответствующего фрагмента семантической сети, выбирая устраивающее его приближение в описании конкретного понятия.

Эксперименты по автоматическому порождению так называемых *составных ключевых слов* (в нашей терминологии - осмысленных комплексов лексических примитивов) представлены, в частности, в работе [*Mucoz 97*]. Здесь средствами машинного обучения (реализованного инструментарием нейронных сетей) по "гладкому[15]" тексту анализируемого документа автоматически порождаются *значимые отношения* (комбинации) между отдельными словами.

Обсуждаемая, в частности, в работе [*Кузнецов 97*] системе ***ДИАНА*** – пример неявного использование взаимосвязей между "лексическими примитивами" при переформулировании запроса пользователя. (Здесь, фактически, осуществляется автоматическое порождение фрагмента тезауруса, описывающего изучаемую ПО в той его части, которая может быть сопоставлена множеству "вариантов" запроса, полученных таким переформулированием последнего).

Наконец, автоматический поиск по ЕЯ-текстам так называемых "*лексических близостей*" (lexical affinities - взаимосвязанных групп слов, достаточно часто встречающихся в тексте на небольшом расстоянии друг от друга) - это ключевой момент развиваемого в рамках направления Data Mining & Knowledge Discovery[16]) так называемого *интеллектуального анализа* полнотекстовых документов. Здесь "лексические близости" служат исходным материалом для изучения понятийного и тематического сходства полнотекстовых документов.

Следуя принципу

- *сходство тем*, которым посвящены рассматриваемые документы, может быть прослежено и на *сходстве понятийных структур*, представленных в этих текстах,

можно уточнить термин "*понимание*" текста компьютерной системой как возможность в ходе автоматической обработки документа выделить в нем релевантные целям анализа (т.е. тематике документа) *понятия* и *отношения* между ними. При этом следует учесть, что полученные в ходе такой переработки конструкции будут приближенными описаниями понятий и отношений между ними. Качество приближения будет определяться по крайней мере двумя обстоятельствами:

- выразительными возможностями языка представления данных, в котором эффективным образом определимо сходство структурных описаний текстов;

---

[15] Т.е. не подвергавшемуся какой-либо предварительной обработке.

[16] Data Mining & Knowledge Discovery - направление исследований в области computer science, ориентированное на интеллектуальный анализ и поиск зависимостей в данных (см. более подробно - редактируемую G.Piatetsky-Shapiro web-страницу [*DM KDR*] и др.).

- выразительными возможностями формальных средств выделения такого сходства (инструментов разбиения множеств документов на категории - классы сходства - в соответствии с их внутрикатегорными подобиями и межкатегорными различиями).

Механизм компьютерной реконструкции понятий по текстам, основанный на технике ДСМ-рассуждений (см., например, [*Zabezhailo 97*, *Забежайло 97, Забежайло 98, Забежайло 98-а, Забежайло 98-б*] и др.), можно в самых общих чертах охарактеризовать следующей принципиальной схемой:

контекст + лингвистические трансформации + машинное обучение.

Задача предлагаемого механизма - построение разумных[17] аппроксимаций для понятий, обсуждаемых в конкретных (т.е. задаваемых примерами) наборах ЕЯ-текстов, а также использование порождаемых аппроксимаций для понятий в автоматическом отнесении рассматриваемых текстов к соответствующим тематическим категориям.

В общих чертах предлагаемый механизм может быть описан следующей схемой:

- порождение по тексту документа специального набора характеристических признаков (индексов в виде тех или иных лексических единиц - слов, словосочетаний, а также, возможно, некоторых отношений между ними);
- приведение порожденных признаковых структур (множеств индексов, графов индексов, связанных определенными отношениями и т.п.) к "каноническому" виду с использованием специализированных словарей и систем трансформационных преобразований;
- выделение в ходе машинного обучения на примерах текстов, принадлежащих заданной тематике (т.е. в ходе анализа сходств и различий представляющих рассматриваемые тексты признаковых структур), тех подструктур (подмножеств индексов, подграфов, ...), которые и являются соответствующими[18] аппроксимациями понятий - носителей содержания и тематических характеристик рассматриваемых текстов;
- тематическая классификация новых текстов (на базе выделяемых в них понятийных структур) с учетом ранее найденных аппроксимаций для понятий - носителей тематических характеристик предложенного для обучения множества текстов.

Принципиальной особенностью предлагаемой схемы, отличающей ее от обычно используемых в лингвистическом анализе вариантов обработки текстов (например, от стандартного пофразного семантического анализа, преобразующего каждое предложение ЕЯ-текста в соответствующий фрагмент семантической сети), является роль прагматической компоненты: из всего множества семантических структур, релевантных каждому конкретному тексту, выделяются лишь те характеристические признаки, которые имеют отношение к *цели* осуществляемого анализа. При этом упомянутое отношение релевантности текста и семантических структур само уточняется (как частично-определенная в исходном состоянии конструкция) в ходе машинного обучения на прецедентах. Речь, таким образом, идет о порождении *контекстно-синтагматического каркаса* (кс-каркаса) для каждого понятия, релевантного цели анализа.

Контекстно-зависимый характер кс-каркаса определяется, прежде всего, совокупностью текстов, отобранных (экспертом) как имеющие отношение к некоторой теме. Нетрудно убедиться, что один и тот же текст может приводить в обсуждаемом нами анализе к разным результатам как при изменении изучаемой темы, так и при замене совместно рассматриваемых с ним текстов (прецедентов, используемых в машинном обучении).

Синтагматические характеристики каркаса - его состав (т.е. конкретная группа "атомов" - лексических единиц, из которых строится исследуемый объект) определяется применением специальных вычислительных процедур. (Ниже мы более подробно поговорим о

[17] Т.е., в данном случае, семантически оправданных и эффективно конструируемых по заданному множеству текстов.

[18] Имеется ввиду соответствие выразительным возможностям выбранного языка представления знаний при реконструировании понятий.

позитивной и негативной компонентах кс-каркаса, т.е., соответственно, составляющих, которые входят в конкретный комплекс “атомов”, и о составляющих, которые релевантны исследуемой проблематике, но в конкретный комплекс “атомов” не входят).

Использованием термина *каркас* мы стремимся подчеркнуть возможности существования различных форм обсуждаемой аппроксимации. Множество “форм” каркаса может порождаться использованием различных языков представления знаний (ЯПЗ). Грубый ЯПЗ дает грубый каркас (например, каркас в относительно бедном выразительными возможностями языке дескрипторного типа), более детальный каркас может быть порожден более детальным ЯПЗ (например, дескрипторы-индексы вместе с отношениями на них - основание для порождения графового представления и т.п.). Наконец, возможны преобразования подобия для кс-каркасов (например, преобразования к “канонической” форме с использованием специальных словарей, либо проблемно-ориентированного анализа отношений сходства\различия на лексическом материале).

Начнем с уточнения способов представления основных элементов содержания текста - понятий и тем. Прежде всего примем во внимание следующее обстоятельство: в отличие от достаточно четких представлений о структуре понятий в точных науках (например, в математике), где практически каждое понятие имеет предикативную природу[19], подавляющее большинство понятий в гуманитарной сфере носит непредикативный характер. В.К.Финн показал (см., например, [*Хвостова 97, Финн 99б*] и др.), что в этом случае понятие характеризуется наличием некоторого “ядра” (т.е. своего рода “общей части” в описаниях экземплификаций понятия) и его “окружения” (характеризуемого смысловыми связями “ядра” с релевантной ему информацией). Однако, в обоих случаях (т.е. при анализе понятий как предикативного, так и непредикативного характера) анализируемые понятия имеют определенное (и характерное только для них) лексическое представление. Таким образом, мы имеем здесь достаточно полную аналогию с уже обсуждавшейся нами выше[20] гипотезой А.М.Бутлерова:

содержание соответствующих понятий
полностью характеризуется
их лексическим представлением в тексте документов.

Итак, говоря об обсуждающихся в тексте понятиях и темах (понимаемых как связанные отношениями семейства понятий), мы имеем возможность судить о них (т.е. понятиях и темах) по представлениям соответствующими совокупностям лексических единиц (слов, словосочетаний, ...).

С инструментальной точки зрения это означает, что достаточно выделить тот или иной набор лексических примитивов, чтобы далее строить (по обсуждавшейся выше схеме) иерархии формальных языков для представления содержания соответствующего массива полнотекстовых документов. Попробуем пояснить сказанное на примере.

Первый (простейший) уровень представления содержания документов может быть задан (формируемым экспертом) списком ключевых слов - смысловых примитивов представления знаний в соответствующей предметной области. Выделенный список может быть задействован как каталог индексов при построении формального описания каждого из рассматриваемых документов.

Следующий уровень - уровень простейших отношений между смысловыми примитивами - может быть организован по следующей схеме:

---

[19] Фактически, здесь возможны всего два вида определений конкретного понятия:

- понятие как имя специфического класса объектов (заданных, например, перечислением);
- понятие как экземплификация объекта, принадлежащего специфическому классу (заданному тем или иным условием).

[20] См. представленный в **Разделе 8.2** вариант изучения структурных описаний и свойств химических соединений как анализа специфических (структурных) химических “текстов” и их “содержания”.

- экспертом формируются “значимые” пары смысловых примитивов (т.е. такие пары, для которых в анализируемой предметной области (ПО) найдется объект, описываемый ими хотя бы частично. Например, для трех дескрипторов *телефонная сеть*, *диагностика* и *обучение* пара, составленная из первых двух дескрипторов может быть сопоставлена диагностике телефонной сети - реально существующему явлению, в то время как пара, образованная первым и третьим дескрипторами, не имеет осмысленной интерпретации в реальной ПО и не может быть отнесена экспертом к множеству “значимых”);
- выделение (в процессе анализа текста конкретного документа) вхождений как отдельных индексов из каталога смысловых примитивов, так и их “значимых” пар позволяет сформировать из множества всех “значимых” пар подмножества как имеющих, так и не имеющих отношения к содержанию анализируемого документа. Таким образом, появляется возможность сопоставить каждому просматриваемому документу характеристики понятий (из общего описания исследуемой ПО) как заведомо имеющих отношение к данному тексту (т.е. понятий, которые затрагиваются в данном тексте), так и список понятий (из исследуемой ПО !), не имеющих отношения к данному документу. Именно эти два списка (характеризуемые множествами “со-встречаемости” элементов каталога смысловых примитивов) и порождают новый - “структурный[21]” - индекс каждого анализируемого документа.

Нетрудно увидеть в таком способе порождения индекса, учитывающего связи между смысловыми примитивами (синтагмами - дескрипторами), аналогии как с механизмом построения кода ФКСП[22] (учет парных связей между смысловыми примитивами - активными центрами), так и с идеей построения покрытий графа химической структуры подграфами, порождаемыми “значимыми” парами смысловых примитивов, - “стандартными” подграфами-дескрипторами кода ФКСП. Последнее обстоятельство – механизм построения простейшего расширения обычного (т.е. основанного на индексировании по стандартному каталогу индексов) дескрипторного ЯПЗ о содержании полнотекстовых документов новыми дескрипторами “комплексного” характера - демонстрирует заложенные в рассматриваемом нами подходе возможности порождать более детальные (хотя и не очень сложные - т.е. сохраняющие полиномиальный характер сложности вычислений) средства представления знаний о содержании анализируемых документов. Таким образом может быть реализован следующий шаг в построении обсуждавшейся нами выше иерархии языков представления знаний, создаваемой для все более полного анализа содержания полнотекстовых документов.

Переход далее в рамках этой иерархии может основываться, в частности, на учете более глубоко формализованных связей между смысловыми примитивами, например, с учетом явным образом представленных в тексте отношений между элементами каталога смысловых примитивов. Очевидным примером такого варианта средств представления знаний о семантике текста может быть использование предложенного Д.А.Поспеловым механизма синтагм[23] в языках ситуационного управления (см., например, [*Поспелов 86*] и др.).

Особого внимания заслуживают возможные (т.е. наблюдаемые в анализируемых текстах) отношения на обсуждавшиеся нами подмножества “значимых” пар характеризующих документы смысловых примитивов. Нетрудно сообразить, что выделяя на множествах таких для пар документов общие части (подмножества), можно конструктивным образом

---

[21] Те. представимый конкретными подграфами графа, вершины которого - элементы каталога смысловых примитивов, а ребра - характеристики со-встречаемости смысловых примитивов в данной ПО, задаваемые “значимыми” парами смысловых примитивов.

[22] См. выше **Раздел 8.2**.

[23] Т.е. троек вида ***aRb***, где ***a*** и ***b*** - элементы каталога смысловых примитивов, а ***R*** - отношение из некоторого заранее заданного списка отношений ***R***.

определить отношение “смыслового” сходства между документами. (Это отношение - действительно будет отношением толерантности: свойства рефлексивности и симметричности являются прямым следствием использования для его формализации булевской операции пересечения множеств). Далее, естественно выделять классы сходства (толерантности) документов по их содержанию, опираясь на классы со-встречаемости (в формализованных описаниях документов) смысловых примитивов (рассматривая для этого не только попарные, но и потроечные, почетверочные и т.д. со-встречаемости элементов каталога дескрипторов в просматриваемых текстах). Порождаемые таким образом таксономии множеств индексов (в том числе - со сложными взаимными вложениями элементов таких таксономий) естественным образом характеризуют множество понятий и тем (понимаемых в оговоренном нами выше смысле), о которых говорится в исследуемом множестве текстов. Возможности учитывать при порождении таксономий не только “позитивные” (т.е. имеющие отношение к содержанию данного текста), но и “негативные” (т.е. не имеющие отношения к содержанию данного текста, но участвующие в описании всей анализируемой ПО) индексы предоставляют дополнительные формальные средства для более детального разделения характеризующих ПО понятий и тем. (Именно этот направление мы и рассмотрим более подробно ниже при обсуждении экспериментов по автоматической тематической классификации полнотекстовых документов).

Опишем в общих чертах основные механизмы выделения по тексту характеристических признаков, используемых затем при порождении кс-каркасов. При обработке полиязычных БД это естественно начинать с идентификации языка документа. Обычно такая идентификация осуществляется на основе машинного обучения на прецедентах (текстах, принадлежность которых к конкретному языку установлена заранее). Следующим шагом может быть выделение специальных лексических единиц - имен, составных терминов, сокращений и т.п., - которые обычно связаны с характерными словосочетаниями, представляющими понятия в конкретной ПО. Так анализ имен, например, подразумевает

- выделение вхождений имен в тексте;
- географическое и временное позиционирование;
- выделение названий организаций, названий работ (публикаций, ...) и т.п.;
- приведение выделенных характеристик к “каноническому” виду (в том числе - снятие возможных неоднозначностей и т.п.);
- разделение пар имен и парных имен (в том числе - с использованием обучения на контекстах);

и др..

Выделение терминов подразумевает

- выделение многословных, т.е. составных терминов (например, в ходе простейшего лингвистического анализа синтаксической связности слов);
- учет повторяемости терминов (говорящий о том, что появление этого термина в тексте - не “случайность”);
- учет вариантов формулировки термина (в том числе - принимая во внимание морфологические преобразования лексических компонент термина);

и др..

Особого внимания заслуживает анализ встречаемости в тексте сокращений, снятие неоднозначностей в их расшифровке[24] (т.е. согласование с уже выделенным ранее списком имен, названий и т.п.). Отдельных усилий требует обработка числовых характеристик - чисел (представляемых словами, долями, процентами и т.п.), датировок (которые необходимо уметь привести к каноническому виду, в том числе - с учетом точки референции), денежных сумм (которые также должны быть приведены к определенной “канонической” форме

[24] Часто возможное лишь при учете контекста документа (например, РФФИ как Российский Фонд Федерального Имущества или как Российский Фонд Фундаментальных Исследований и т.п.).

записи). В большинстве случаев перечисленные выше варианты автоматического выделения характеристик текстов могут быть реализованы с помощью тех или иных конкретных эвристик (см., например, примеры средств feature extraction в работе [*Gotthard 97*]) и, вообще говоря, не требуют тяжеловесных средств стандартного лингвистического анализа.

Возможно также порождение характеристических признаков в процессе использования тех или иных систем автоматического индексирования документов. (Достаточно известный пример - порождение так называемых векторов тем по гладкому тексту документа, осуществляемое уже упоминавшейся системой ***ORACLE ConText*** [*OCT*]).

Говоря о возможностях трансформаций подобия для различных представлений кс-каркаса понятий к "каноническому" виду, следует выделить два вида инструментов:

- проблемно-ориентированные системы отношений сходства\различия на совокупностях "атомов" используемого ЯПЗ и
- специализированные словари (семантические, комбинаторные, инверсно-тематические и т.п.).

Первый из них предназначен для расширения анализируемого при обработке кс-каркасов множества лексических едини за счет использования заранее описанных систем отношений сходства\различия. (Наиболее интересные примеры таких систем, используемые как стандартные расширения типового инструментария систем компьютерного текст-процессинга, - это словари синонимов, системы учета взаимосвязей гипонимы-гиперонимы и др.).

Задача специализированных словарей - поддерживать процедуры перевода текста с естественного языка на язык семантического представления (т.е. язык представления знаний) и обратно, а также обеспечивать преобразования подобия внутри уже порожденного семантического представления. Семантические словари (см., например, [*Лахути 97*]) призваны обеспечить смысловое развертывание лексических единиц в иерархии представлений от уровней семантических категорий и семантических типов к уровню семантических примитивов. Поддержка "навигации" в таких семантических структурах может стать существенным дополнением к традиционно используемым средствам лингвистической обработки - системам морфологического и синтаксического анализа.

Комбинаторные словари (см., например, [*Benson 86*] и др.) могут стать эффективным средством контроля результатов, получаемых программными системами формального порождения "комплексов" из лексических единиц. Именно они могут быть практическим инструментом отсечения артефактов, порождаемых соответствующими системами кластеризации и\или категоризации текстов по используемым в них лексическим комплексам.

Отдельного обсуждения заслуживает, на наш взгляд, использование инверсно-тематических словарей. Блестящий пример такого инструмента - Longman's Production Dictionary [*Longman PD*], где "ядерная" (т.е. используемая, как правило, независимо от конкретной области приложений) лексика английского языка сгруппирована вокруг примерно 1000 семантических классов - "родовых" терминов, объединяющих множества близких по значению лексических единиц. Словарь позволяет для каждого из таких "родовых" терминов выбрать (в соответствии с конкретным контекстом словоупотребления) тот или иной вариант лексического выражения. (Поддерживая преобразования в обе стороны - от "родовых" терминов к конкретной лексике и наоборот, словарь такого типа позволяет обращаться также к конкретным значениям лексических единиц, В частности, он может быть использован при снятии неоднозначностей в толковании выделяемых из текста характеристических признаков - имен, названий и т.п.).

Существенной особенностью структуры кс-каркаса понятий является наличие в нем как "позитивной" так и "негативной" составляющей. Позитивная часть представляет лексические компоненты, которые *входят* в каждый из рассматриваемых текстов из обучающего множества (для данного понятия) и являются (вместе с, возможно, имеющимися между ними отношениями) своего рода *носителем сходства* этих текстов. Негативная компо-

нента порождается лексикой (плюс, возможно, связывающими ее отношениями), релевантной рассматриваемой ПО, но не используемой в выбранной для обучения понятию группе документов.

Негативная компонента кс-каркаса играет в его представлении роль, аналогичную роли “негативной” компоненты булевского запроса в современных поисковых системах дескрипторного типа (т.е. той части запроса, которая входит в соответствующую дизъюнктивную нормальную форму в виде индексов с отрицаниями). Способы порождения негативной компоненты кс-каркаса (как и в случае его позитивной части) определяются выбором конкретного механизма машинного обучения, используемого для анализа тематического сходства текстов исходных документов. (Ниже мы приведем примеры таких процедур). Подчеркнем контекстную зависимость представленных компонент обоих знаков от параметров конкретного обучающего множества текстов-примеров.

Обратим внимание на существование (при заданном обучающем множестве текстов и фиксированном языке представления знаний о их содержании) кс-каркасов двух типов: составных и неразложимых. Диаграмма вложимости кс-каркасов (порождаемая, например, как отражение вложимости одних множеств индексов в другие, или как вложимость соответствующих подграфов в графы и т.п.) позволяет для заданного контекста порождать своего рода *тематические карты* - т.е. структуры, эксплицирующие (быть может неявно выраженные в исходных гладких текстах) взаимосвязи между понятиями, имеющими отношение к тематике обучающей выборки документов. Явный вид таких тематических карт может быть подспорьем при описании внутренней структуры соответствующей ПО (например, при построении тезаурусов, семантических сетей и т.п.).

Наконец, ключевую роль в предлагаемом нами механизме компьютерной реконструкции понятий играют конкретные механизмы машинного обучения на примерах, используемые для тематической категоризации документов. Все основные варианты технической формализации соотнесения анализируемых текстов и тематических категорий (иерархическая или неиерархическая кластеризация, индуктивное обучение правилам, поиск ближайших аналогов и т.п.) основаны на идее ДСМ-формализации и автоматического анализа *сходства* документов, а также тематической классификации новых документов с учетом выделенных на этапе обучения “формальных носителей” этого сходства.

П 3.2 Автоматическая классификация полнотекстовых документов.

Мы представим два примера использования предложенного подхода в задачах интеллектуального текст-процессинга [*Забежайло 97, Виноградов 97, Забежайло 98а, Забежайло 98б*]. В первом из них решалась задача автоматической классификации по темам специальным образом заиндексированных (при существенном участии эксперта) русскоязычных текстов. Во втором - задача автоматического индексирования и классификации по темам полнотекстовых англоязычных документов. Ключевым элементом обоих экспериментов было применение уже представленной техники ДСМ-рассуждения,

В первом примере была еще раз использована известная идея Р.С.Гиляревского (см., например, [*Гиляревский 98а*] о систематическом изучении (т.е. разработке соответствующих методов и их экспериментальном исследовании) “внутренней структуры” предметной области ***информатика***. Конкретными (см. также [*Виноградов 97*] и [*Забежайло 97*]) исходными данными эксперимента был массив аннотаций защищенных диссертаций в одном из направлений computer science. Для тематической компьютерной классификации были избраны два раздела: “информатика” (понимаемая в соответствии с предметной областью, охватываемой специальностью 05.13.17 - Теоретические основы информатики) и “не информатика”. При этом исходными текстами был массив документов, полученных из БД ВНТИЦентра по запросу “информ*” (т.е. все те, в которых встречается словоформа ”информ*”). Экспертом (профессором Р.С.Гиляревским) из всего массива диссертаций для эксперимента были выбраны 178 документов. Отобранные документы были разложены (им

же) на два множества: 100 имеющих отношение к теме “информатика” (они выполняли в машинном ДСМ-обучении роль “примеров”), и 78 - не имеющих. (Последние выполняли роль “контрпримеров”).

Для индексирования отобранных документов средствами дескрипторного ЯПЗ экспертами были построены:

- каталог индексов (ключевых слов и словосочетаний), характеризующих рассматриваемую предметную область и содержащий элементы как позитивного, так и негативного[25] характера;
- классы бинарной совстречаемости индексов из каталога в рассматриваемой предметной области. (Эти классы сыграли роль исходного материала для выделения негативных индексов для каждого рассматривавшегося документа: наличие некоторого индекса $\mathbf{i_1}$ в множестве положительных индексов вместе с наличием пары $<\mathbf{i_1},\mathbf{i_2}>$ в соответствующем[26] классе совстречаемости и отсутствием индекса $\mathbf{i_2}$ в множестве положительных индексов означает, что этот индекс может быть отнесен к множеству исходных элементов при порождении негативной части соответствующих кс-каркасов[27].)

Затем классифицируемые тексты были заиндексированы: позитивные примеры - индексами обоих типов - и негативные примеры - позитивными индексами - в ходе “ручного” индексирования экспертом, а негативными индексами - в ходе автоматического индексирования (по классам бинарной совстречаемости индексов из каталога). После этого была осуществлена автоматическая ДСМ-классификация документов (как “примеров”, так и “контрпримеров”). Следует подчеркнуть, что осуществляемая классификация учитывала индексы обоих типов (как “*позитивные*”, так и “*негативные*”). Технически были осуществлены следующие действия: программная система в процессе обучения на примерах и контрпримерах порождала совокупности кс-каркасов для затрагиваемых в тексте понятий (и системы классов сходства на анализируемых текстах), характеризующие релевантность документов теме, а затем использовала их для анализа новых документов.

Порождаемые в процессе ДСМ-обучения аппроксимации понятий позволяют иметь более объективное (т.е. основанное на объективном формальном анализе) представление о тематической природе исследуемой предметной области. Для оценки качества классификации в этом эксперименте были отобраны 97 “примеров” и 75 “контрпримеров”. Оставшиеся 3 текста из каждого из классов были исключены из процесса обучения и предложены программной системе для классификации. Все предъявленные для классификации объекты были правильным образом отнесены к соответствующим классам.

Во втором (см. также работы [*Забежайло 97, Забежайло 98а*]) - решалась задача автоматической классификации полнотекстовых документов на английском языке, относящихся к двум темам: научная фантастика и эротика. (Последняя была выбрана в связи с актуальностью проблемы автоматического выявления документов подобного рода в сети INTERNET). В базе примеров содержались 33 эротических и 23 фантастических текста, перекачанных из компьютерной библиотеки полнотекстовых документов. Для классификации произвольным образом были выбраны по два документа из каждого класса.

Формальное представление каждого из анализируемых текстов в виде так называемого вектора тем[28] было получено с помощью программной системы ***ORACLE ConText.***

---

[25] Т.е. индексы, характеризующие свойство “*не принадлежать к рассматриваемой теме*”.

[26] Т.е. построенном, в данном случае, экспертом, исходя из его представлений о существовании в анализируемой ПО объектов, которые могут быть характеризованы одновременным использованием этих двух индексов (плюс, возможно, некоторых дополнительных индексов в комплексе с ними).

[27] Вместо консультаций с экспертом в такой ситуации могут быть, в частности, задействованы специализированные словари: так например, используемый в системе ***ДИАНА*** (см. [*Кузнецов 97*]) словарь со-встречаемости лексических единиц, построенный мониторингом научно-технических текстов Реферативного журнала ВИНИТИ. В такой словарь попадают (с учетом тематики соответствующей ПО) только те пары слов, которые встречаются в каком-либо из конкретных текстов, прошедших через РЖ.

[28] Здесь мы описываем лишь один из вариантов эксперимента (такой, где рассматривался лишь “стандартный” вид вектора тем – 16 элементов без учета частот их встречаемости). В дополнение к этому в экспериментах

(Для осуществления этой работы были использованы специально разработанные программные интерфейсы, обеспечивающие сопряжение представленной нами системы текст-процессинга с системой ***ORACLE ConText***).

В результате работы программной системы по одному из каждых двух текстов по каждой из тем были правильно классифицированы. Два других текста попали в область противоречивой классификации. (При визуальном анализе этих текстов стало ясно, что каждый из них эксперту трудно однозначно отнести к одному из классов, принимая во внимание их полное содержание.)

Программная система[29], реализующая представленную схему ДСМ-анализа данных и принятия решений была разработана на базе интегральной инструментальной среды ***SAS System***, созданной компанией **SAS Institute** (NC, USA). Обсуждаемый программный комплекс представляет собой масштабируемый прототип, обладающий свойством переносимости как по отношению к вычислительной платформе (от конфигурации, выполнимой на вычислительных машинах класса *PC*, до конфигураций, выполнимых на *mainframe*'ах[30]), так и к операционному окружению. Программный комплекс совместим по экспорту\импорту данных практически со всеми широко используемыми на сегодняшний день системами управления базами данных.

В обоих экспериментах вычисления производились на ЭВМ ***SUN ULTRA***, там же хранилась и база данных, содержавшая индексные описания всех используемых в обучении и классифицируемых документов. Для классификации использовался достаточно выразительный фрагмент ДСМ-метода - метод простого сходства с запретом на контрпримеры.

Формальное представление анализируемых текстов в виде так называемого вектора тем порождалось с помощью программной системы ***ORACLE ConText***. (Для осуществления этой работы были использованы специально разработанные интерфейсы, обеспечивающие сопряжение обсуждаемого программного комплекса с системой ***ORACLE ConText***). Процедура классификации проходила в режиме КЛИЕНТ-СЕРВЕР. В процессе вычислений тексты с локальной машины копировались на сервер (***SUN ULTRA***). Далее для генерации векторов тем запускался ***ORACLE ConText***, который сохранял результаты своей работы в базе данных ***ORACLE***. После этого полученные векторы закачивались в таблицы ***SAS***, хранящиеся там же на сервере, где, собственно и проходил процесс классификации.

Использование автоматически извлекаемых из анализируемых текстов понятий (более точно - их кс-каркасов) вместе с представлением текстов в виде тематических карт позволяет предложить альтернативный подход в порождении сжатых вариантов изложения ЕЯ-текстов (так называемые abstracting и summarization). Наиболее часто на сегодняшний день здесь используется техника статистического анализа встречаемости в текстах тех или иных лексических единиц и выделения тех фрагментов текста, которые содержат статистически наиболее значимую лексику. Предлагаемая нами альтернатива - использовать для выделения значимых фрагментов текста порождаемые в ходе ДСМ-анализа кс-каркасы понятий и объединяющие их в диаграммы взаимосвязей тематические карты документов (см. ниже – **Приложение 3.3**).

использовался также и вариант "расширенного" вектора тем (где текст характеризовался не только 16 наиболее "важными", но и всеми остальными приписываемыми ему системой ***ORACLE ConText*** темами. При этом оказалось, что на конкретном множестве рассматривавшихся здесь текстов порождаются векторы тем, содержащие не более 40-45 элементов). Наконец, представленная в **Главах 2-3** ДСМ-техника работы с нумерованными множествами и числовыми векторами позволяет оперировать порождаемыми системой ***ORACLE ConText*** векторами тем с учетом вычисляемых ею попутно числовых характеристик "важности" соответствующих элементов – "весов" тем.

[29] Ее авторы - М.И.Забежайло, М.Н.Емеленко, Е.А.Липчинский и М.В.Максин.

[30] Использование в обсуждаемых экспериментах конкретной комбинации ***SAS System & SUN/Solaris & ORACLE*** обеспечивает простую масштабируемость (в плане вычислительных ресурсов) полученного программно-технического решения вдоль линейки технических решений компании **SUN Microsystems** в интервале от *рабочей станции* до ***SUN(CRAY)Ultra Enterprise 10000***.

Средствами предлагаемой техники может быть также формализована задача контроля тематической полноты документальных баз данных. Представление конкретных текстовых массивов в виде тематической карты дает возможность объективизированного сравнения содержания таких массивов с некоторым эталоном (для построения которого можно использовать, например, те или иные тезаурусы, либо достаточно полные коллекции документов, такие, как реферативные БД **ВИНИТИ РАН**, массивы нормативных документов **ВНИИКИ Госстандарта РФ** и т.п.).

Предлагаемая нами техника формального анализа содержания полнотекстовых документов дает возможности контроля “многоаспектности” текстов (имеющей важное значение, например, в технологии больших реферативных служб - **ВИНИТИ**, **ИНИОН** и т.п.). Здесь каждый из (заранее выделенных как требующий учета) тематических аспектов содержания документов может быть сопоставлен соответствующему набору лексических единиц (кс-каркасов) и взаимосвязей между ними (представляемых соответствующей тематической картой).

Наконец, обсуждаемая техника компьютерного текст-процессинга открывает новые возможности для автоматического контроля содержания динамически пополняемых текстовых БД. Наблюдая возможные изменения в структуре тематических карт соответствующего массива документов при поступлении новых документов, можно отслеживать изменения системы классов тематического сходства для рассматриваемого массива текстов. Выявление существенных структурных перемен будет означать, что получена новая информация, радикальным образом изменяющая содержание анализируемого массива документов.

Разумеется данный список может быть продолжен. Задачи управления поиском информации в сети INTERNET, задачи автоматического пополнения баз данных извлеченной из текстов фактографической информацией и др. - вот примеры направлений, заслуживающих отдельного обсуждения (выходящего, однако, за рамки данной работы).

Успех представленных экспериментов дает определенные аргументы в оценке перспектив использования предлагаемого подхода в задачах автоматического анализа и “понимания” документов компьютерной системой. Применение инструментов обсужденного типа открывает практические подходы к решению ключевых задач интеллектуализации информационных систем, повышению качества поиска информации в промышленных текстовых БД.

П 3.3 Тематические карты содержания полнотекстовых документов.

Задача конструктивного выделения объема понятий давно известна и достаточно хорошо изучена (см., например, [*Апресян 74, Финн 99, Финн 996*] и др.). Интересной, на наш взгляд, прикладной областью использования конкретных инструментов решения этой задачи является проблематика наукометрических исследований - выделение тенденций развития областей знания, оценки перспективности новых технологий и т.п.. В качестве примера можно привести известный ([*Гиляревский 88*, *Гиляревский 88а*] и др.) механизм поиска и оценки перспектив развития информационных технологий, основанный на явном выделении понятийной (т.е. тематической) структуры предметной области “*информатика*”, а также на отслеживании динамики ее изменений во времени. Легко видеть, что объем конкретных уточнений понятия “*информатика*”, очерчиваемый в этих экспериментах путем кластеризации специальных текстов[31] (представляющих те или иные аспекты изучаемой предметной области) по темам, а также динамические изменения во времени этого объема

[31] В работе [*Гиляревским 88а*] это аннотации диссертаций из информационного массива ВНТИЦентра, в текстах которых используется словоформа *информ**, за исключением тех, которые отброшены из этого множества экспертом (проф. Р.С.Гиляревским) как не имеющие отношения к предметной области “информатика” (понимаемой как объединение *information* и *computer sciences*).

- вот объективные показатели, характеризующие тенденции развития обсуждаемой предметной области.

Мы руководствовались аналогичными соображениями, ставя своей задачей уточнение понимания достаточно популярного сегодня термина *прикладная семиотика*. Обращение к этой тематике обусловлено тем интересом, который проявляется к ней в промышленно-развитых странах, и прежде всего - в США (см., например, монографию [*Meystel 95*], материалы конференций *[AS'95, SA'96*] или материалы ежегодной конференции [*ISAS'97*] в американском Национальном Институте Стандартов и Технологии - NIST, Gathersburg, MD, USA). Наш дополнительный интерес к этой работе связан также с тем, что среди ведущих отечественных исследователей в области информационных систем и искусственного интеллекта[32] давно сложилось мнение о вторичном характере[33] термина "*прикладная семиотика*" и о чисто организационной природе интереса к этому кругу вопросов в США (связанной, например, с перераспределением и переадресацией финансирования исследований в области искусственного интеллекта).

Выбранный нами механизм выделения тематической структуры предметной области *прикладная семиотика* опирается на известную схему осуществляемого средствами ДСМ-метода анализа текстов на естественном языке (см., наприме*р,* [*Виноградов 97, Забежайло 97, Забежайло 98, Забежайло 98а*]). Главная особенность используемого подхода - в порождении классов сходства (системы классов толерантности) текстов, релевантных данной теме (или множеству тем), осуществляемом с учетом анализа близких по лексике, но не являющихся релевантными теме (или множеству тем) текстов. Порождаемые в процессе ДСМ-обучения на примерах и контрпримерах текстов, релевантных теме, классы тематического сходства (т.е. множества ДСМ-сходных документов) вместе с являющимися своего рода "носителями" этого сходства понятиями (представленными выделенным в ходе ДСМ-анализа лексическим материалом) - вот искомые параметры, характеризующие тематическую структуру анализируемого корпуса текстов (и описываемой ими предметной области).

Исходным материалом нашего эксперимента были 10 подготовленных международной группой экспертов текстов (на английском языке), характеризующих различные аспекты относимых к области *прикладная семиотика* проблем. Все анализируемые тексты (см. **Таблицу П 3.3.1**) взяты в ДСМ-обработку после того, как были опубликованы [*GBG HT*] на web-странице ***Glass Bead Game*** компании **Highland Tehnologies** (MD, USA).

Перед порождением тематической "карты" исследуемого массива текстов каждый из документов был просмотрен системой лингвистического анализа ***ORACLE ConText*** [*OCT*]. В результате такой обработки каждому из документов был сопоставлен вектор из 16 наиболее важных (с точки зрения лингвистического процессора системы ***ORACLE ConText***) тем, затрагиваемых в данном тексте. (В **Таблицах 8.3.3.1-2** представлены соответствующие векторы тем, а также сводный каталог тем, характеризующих - с точки зрения системы лингвистического анализа ***ORACLE ConText*** - тематическое содержание рассматриваемых текстов и, соответственно, описываемой ими предметной области).

Далее была порождена система классов ДСМ-сходства[34], порождаемых на представляющих тексты векторах тем. Для вычислений использовалась та же программная система, которая уже была охарактеризована нами выше - в Разделе **8.3.2**.. Полный список и структура полученных классов тематического сходства представлены в **Таблицах П 3.3.1-3**.

Анализ полученных результатов позволяет сделать по крайней мере следующие два заключения:

---

[32] Более точно эту предметную область характеризует название серии конференций - "Семиотические аспекты формализации интеллектуальной деятельности" - см., например, [*САФИД 88*] и др.).

[33] Т.е. об объединении под этим названием ряда уже известных разделов computer & information sciences.

[34] Сходство объектов здесь формализуется операцией взятия пересечения множеств тем, входящих в соответствующие каждому тексту векторы тем.

- “понятийное ядро” - см. **Таблицы П 3.3.1-3** - изучаемой предметной области (т.е. наиболее крупные по числу входящих в них текстов классы сходства) не носит[35] оригинального характера и представляется набором синтагм[36], специфичных для ряда уже известных разделов computer science;
- полученная тематическая карта демонстрирует реальную[37] тематическую структуру анализируемой предметной области и позволяет более объективно судить как о ее структурных особенностях, так и о взаимосвязях со смежными областями (представлением знаний, базами знаний, формальными моделями и др.).

С точки зрения используемых в ДСМ-анализе правил правдоподобного вывода первое из приведенных заключений может быть представлено в еще более жесткой форме: если бы при построении тематической карты изучаемого множества текстов использовались не только позитивные примеры (т.е. тексты, релевантные теме), но и близкие по лексике, но не являющиеся релевантными теме документы (контрпримеры), то (учитывая, что используемая программная система реализует достаточно сильный – см. **Главу 3** - фрагмент ДСМ-метода - так называемый метод простого сходства с запретом на контрпримеры) мы вполне могли бы иметь ситуацию, когда в изучаемой предметной области не порождалось бы ни одного непротиворечивого (в смысле ДСМ-анализа) класса тематического сходства документов. (Так например, имеющийся в **Таблице П 3.3.3** класс, характеризуемый темами

<applications, control, engineering, methods, models, problems, technology>,

вполне может порождаться также и текстами, скажем, посвященными технологиям и приложениям систем автоматического управления, и т.п.).

Проведенные эксперименты позволяют с оптимизмом смотреть на перспективы использования представленного подхода в области текст-процессинга и информационных систем. Очевидными направлениями применения представленной технологии могут быть исследования, связанные, например,

- с выделением тематической структуры конкретных отраслей знания,
- более адекватным порождением систем *базовых элементов* описания предметной области (например, систем проблемно-ориентированных дескрипторов и т.п.) для фактографических информационных систем и, конечно же,
- с повышением качества информационного поиска.

Таблица П 3.3.1. Тематическое индексирование системой ***ORACLE ConText*** рассматриваемого множества документов. (*Векторы тем* документов).

Quasi Axiomatic Theory and Iconic Computation (*Paul S. Prueitt*)

architecture	bags	composure	computation
content	embedment	falsity	finiteness
heights	linguistics	observations	opposition
physicality	semantics	stratification	width

Consistency Support in Active Knowedge Base (Sergej K. Dulin)

acquisition	analysises	computers	consistency
direction	interactive systems	investigation	knowledge
knowledge bases	knowledge representation	methods	reduction
requirements	smarts	structure	support

[35] С точностью до подбора характеризующих предметную область в данном конкретном случае текстов.

[36] Это прежде всего: *control, content, knowledge, languages, methods, models, problems, representation, structure, technology* и др., при отсутствии синтагм *semiotic* и *applied semiotic*.

[37] И при этом несколько отличающуюся от подразумеваемой пропагандистами этого направления - см., например, [*Meystel 95*] и др. - в плане особенностей, отличающих предметную область “*прикладная семиотика*” от смежных дисциплин.

Developing models of a world with due regard for its dynamics General principles (Gennady S. Osipov)

Meta	argumentation	dynamics	formality
fundamentals	goals	humans	inference
knowledge	languages	models	prediction
problems	requirements	software	subjects

Hierarchical Integration of Indeterminancy (Galia V. Tzvetkova)

acquisition	auxiliaries	characteristics	complex systems
connection	consideration	content	falsity
hierarchy	indexes	models	ranks
structure	studies	symbolism	topology

Knowledge-Based Inference in Semiotic Systems (Vadim N. Vagin)

algorithms	deductive	inference	knowledge
knowledge bases	knowledge representation	layers	methods
models	problems	procedures	representation
semantics	signs	structure	topology

Making Applied Semiotics Work in the Development Software (Vladimir F. Khoroshevsky)

applications	computer software	connection	control
designs	engineering	implementation	languages
methods	models	problems	representation
semiotics	software engineering	structure	technology

The Autonomous Organization of Data through Semiotic Methods (Paul S. Prueitt)

Russians	analysises	causals	class
content	contexts	decomposition	highlands
human factors	Information technology	knowledge bases	metadatas
physical topics	support	technology	theories

Semiotic Models of Behavior (Dmitri A. Pospelov)

autonomy	behavior	charsets	content
deeds	estimation	falsity	goals
humans	interpretation	languages	models
motives	referents	representation	requirements

Applied Semiotic: New Approach To Modeling And Simulating In Control Systems (Alexander I. Ehrlich)

applied semiotic	applications	approaches	change
content	control	engineering	functionality
knowledge	methods	models	possibility
problems	solutions	subjects	technology

On Systems, Complexity and Simplicity (Shelia Guberman, Wita Wojtkowski)

Systems	apriority	cars	complex systems
consideration	construction	control	data processing
dependence	falsity	investigation	languages
partitions	problems	recognition	theories

Таблица П 3.3.2. Полный каталог тем, выделенных системой ***ORACLE ConText*** в рассматриваемом множестве документов.

Applied Semiotic	Meta	Russians	Systems
acquisition	algorithms	analysises	applications
approaches	apriority	architecture	argumentation
autonomy	auxiliaries	bags	behavior
cars	causals	change	characteristics

charsets	class	complex systems	composure
computation	computer software	computers	connection
consideration	consistency	construction	content
contexts	control	data processing	decomposition
deductive	deeds	dependence	designs
direction	dynamics	embedment	engineering
estimation	falsity	finiteness	formality
functionality	fundamentals	goals	heights
hierarchy	highlands	human factors	humans
implementation	indexes	inference	information technology
interactive systems	interpretation	investigation	knowledge
knowledge bases	knowledge representation	languages	layers
linguistics	metadatas	methods	models
motives	observations	opposition	partitions
physical topics	physicality	possibility	prediction
problems	procedures	ranks	recognition
reduction	referents	representation	requirements
semantics	semiotics	signs	smarts
software	software engineering	solutions	stratification
structure	studies	subjects	support
symbolism	technology	theories	topology
width			

Таблица П 3.3.3 Полная система классов тематического сходства рассматриваемых документов.

Сходство:	• ***investigation***
Орбита сходства:	• On Systems, Complexity and Simplicity • Consistency Support in Active Knowledge Base

Сходство:	• ***semantics***
Орбита сходства:	• Quasi Axiomatic Theory and Iconic Computation • Knowledge-Based Inference in Semiotic Systems

Сходство:	• ***theories***
Орбита сходства:	• The Autonomous Organization of Data through Semiotic Methods • On Systems, Complexity and Simplicity

Сходство:	• ***knowledge*** • ***knowledge bases*** • ***knowledge representation*** • ***methods*** • ***structure***
Орбита сходства:	• Knowledge-Based Inference in Semiotic Systems • Consistency Support in Active Knowledge Base

Сходство:	• ***analysises*** • ***knowledge bases*** • ***support***
Орбита сходства:	• The Autonomous Organization of Data through Semiotic Methods • Consistency Support in Active Knowledge Base

Сходство:	• ***knowledge bases***
Орбита сходства:	• Consistency Support in Active Knowledge Base • The Autonomous Organization of Data through Semiotic Methods • Knowledge-Based Inference in Semiotic Systems

Сходство:	• ***knowledge*** • ***requirements***

Орбита сходства:	• Developing models of a world with due regard for its dynamics General principles • Consistency Support in Active Knowledge Base

Сходство:	• ***goals*** • ***humans*** • ***languages*** • ***models*** • ***requirements***
Орбита сходства:	• Semiotic Models of Behavior • Developing models of a world with due regard for its dynamics General principles

Сходство:	• ***requirements***
Орбита сходства:	• Consistency Support in Active Knowledge Base • Developing models of a world with due regard for its dynamics General principles • Semiotic Models of Behavior

Сходство: (см. **текст выше**)	• ***applications*** • ***control*** • ***engineering*** • ***methods*** • ***models*** • ***problems*** • ***technology***
Орбита сходства:	• Applied Semiotic: New Approach To Modeling And Simulating In Control Systems • Making Applied Semiotics Work in the Development Software

Сходство:	• ***content*** • ***technology***
Орбита сходства:	• The Autonomous Organization of Data through Semiotic Methods • Applied Semiotic: New Approach To Modeling And Simulating In Control Systems

Сходство:	• ***technology***
Орбита сходства:	• The Autonomous Organization of Data through Semiotic Methods • Making Applied Semiotics Work in the Development Software • Applied Semiotic: New Approach To Modeling And Simulating In Control Systems

Сходство:	• ***content*** • ***falsity*** • ***models***
Орбита сходства:	• Semiotic Models of Behavior • Hierarchical Integration of Indeterminancy

Сходство:	• ***complex systems*** • ***consideration*** • ***falsity***
Орбита сходства:	• Hierarchical Integration of Indeterminancy • On Systems, Complexity and Simplicity

Сходство:	• ***falsity*** • ***languages***
Орбита сходства:	• Semiotic Models of Behavior • On Systems, Complexity and Simplicity

Сходство:	• ***content*** • ***falsity***

Орбита сходства:	• Semiotic Models of Behavior • Hierarchical Integration of Indeterminancy • Quasi Axiomatic Theory and Iconic Computation

Сходство:	• ***falsity***
Орбита сходства:	• Quasi Axiomatic Theory and Iconic Computation • Hierarchical Integration of Indeterminancy • On Systems, Complexity and Simplicity • Semiotic Models of Behavior

Сходство:	• ***inference*** • ***knowledge*** • ***models*** • ***problems***
Орбита сходства:	• Knowledge-Based Inference in Semiotic Systems • Developing models of a world with due regard for its dynamics General principles

Сходство:	• ***knowledge*** • ***models*** • ***problems*** • ***subjects***
Орбита сходства:	• Developing models of a world with due regard for its dynamics General principles • Applied Semiotic: New Approach To Modeling And Simulating In Control Systems

Сходство:	• ***knowledge*** • ***methods*** • ***models*** • ***problems***
Орбита сходства:	• Applied Semiotic: New Approach To Modeling And Simulating In Control Systems • Knowledge-Based Inference in Semiotic Systems

Сходство:	• ***knowledge*** • ***methods***
Орбита сходства:	• Consistency Support in Active Knowledge Base • Applied Semiotic: New Approach To Modeling And Simulating In Control Systems • Knowledge-Based Inference in Semiotic Systems

Сходство:	• ***knowledge*** • ***models*** • ***problems***
Орбита сходства:	• Applied Semiotic: New Approach To Modeling And Simulating In Control Systems • Knowledge-Based Inference in Semiotic Systems • Developing models of a world with due regard for its dynamics General principles

Сходство:	• ***knowledge***
Орбита сходства:	• Consistency Support in Active Knowledge Base • Knowledge-Based Inference in Semiotic Systems • Developing models of a world with due regard for its dynamics General principles • Applied Semiotic: New Approach To Modeling And Simulating In Control Systems

Сходство:	• ***languages*** • ***models***

	• ***problems***
Орбита сходства:	• Making Applied Semiotics Work in the Development Software • Developing models of a world with due regard for its dynamics General principles

Сходство:	• ***languages*** • ***models*** • ***representation***
Орбита сходства:	• Semiotic Models of Behavior • Making Applied Semiotics Work in the Development Software

Сходство:	• ***control*** • ***languages*** • ***problems***
Орбита сходства:	• On Systems, Complexity and Simplicity • Making Applied Semiotics Work in the Development Software

Сходство:	• ***languages*** • ***models***
Орбита сходства:	• Making Applied Semiotics Work in the Development Software • Developing models of a world with due regard for its dynamics General principles • Semiotic Models of Behavior

Сходство:	• ***languages*** • ***problems***
Орбита сходства:	• Making Applied Semiotics Work in the Development Software • On Systems, Complexity and Simplicity • Developing models of a world with due regard for its dynamics General principles

Сходство:	• ***languages***
Орбита сходства:	• Semiotic Models of Behavior • Making Applied Semiotics Work in the Development Software • Developing models of a world with due regard for its dynamics General principles • On Systems, Complexity and Simplicity

Сходство:	• ***methods*** • ***models*** • ***problems*** • ***representation*** • ***structure***
Орбита сходства:	• Making Applied Semiotics Work in the Development Software • Knowledge-Based Inference in Semiotic Systems

Сходство:	• ***methods*** • ***structure***
Орбита сходства:	• Consistency Support in Active Knowledge Base • Making Applied Semiotics Work in the Development Software • Knowledge-Based Inference in Semiotic Systems

Сходство:	• ***methods*** • ***models*** • ***problems***
Орбита сходства:	• Making Applied Semiotics Work in the Development Software • Applied Semiotic: New Approach To Modeling And Simulating In Control Systems • Knowledge-Based Inference in Semiotic Systems

Сходство:	• ***methods***
Орбита сходства:	• Consistency Support in Active Knowledge Base • Applied Semiotic: New Approach To Modeling And Simulating In Control Systems • Knowledge-Based Inference in Semiotic Systems • Making Applied Semiotics Work in the Development Software

Сходство:	• ***acquisition*** • ***structure***
Орбита сходства:	• Hierarchical Integration of Indeterminancy • Consistency Support in Active Knowledge Base

Сходство:	• ***models*** • ***structure*** • ***topology***
Орбита сходства:	• Hierarchical Integration of Indeterminancy • Knowledge-Based Inference in Semiotic Systems

Сходство:	• ***connection*** • ***models*** • ***structure***
Орбита сходства:	• Hierarchical Integration of Indeterminancy • Making Applied Semiotics Work in the Development Software

Сходство:	• ***models*** • ***structure***
Орбита сходства:	• Making Applied Semiotics Work in the Development Software • Hierarchical Integration of Indeterminancy • Knowledge-Based Inference in Semiotic Systems

Сходство:	• ***structure***
Орбита сходства:	• Hierarchical Integration of Indeterminancy • Consistency Support in Active Knowledge Base • Making Applied Semiotics Work in the Development Software • Knowledge-Based Inference in Semiotic Systems

Сходство:	• ***content*** • ***models***
Орбита сходства:	• Applied Semiotic: New Approach To Modeling And Simulating In Control Systems • Semiotic Models of Behavior • Hierarchical Integration of Indeterminancy

Сходство:	• ***content***
Орбита сходства:	• Quasi Axiomatic Theory and Iconic Computation • Applied Semiotic: New Approach To Modeling And Simulating In Control Systems • Semiotic Models of Behavior • Hierarchical Integration of Indeterminancy • The Autonomous Organization of Data through Semiotic Methods

Сходство:	• ***control*** • ***problems***
Орбита сходства:	• On Systems, Complexity and Simplicity • Applied Semiotic: New Approach To Modeling And Simulating In Control Systems • Making Applied Semiotics Work in the Development Software

Сходство:	• ***models*** • ***problems***

Орбита сходства:	• Applied Semiotic: New Approach To Modeling And Simulating In Control Systems • Developing models of a world with due regard for its dynamics General principles • Making Applied Semiotics Work in the Development Software • Knowledge-Based Inference in Semiotic Systems

Сходство:	• ***problems***
Орбита сходства:	• Making Applied Semiotics Work in the Development Software • On Systems, Complexity and Simplicity • Knowledge-Based Inference in Semiotic Systems • Developing models of a world with due regard for its dynamics General principles • Applied Semiotic: New Approach To Modeling And Simulating In Control Systems

Сходство:	• ***models*** • ***representation***
Орбита сходства:	• Making Applied Semiotics Work in the Development Software • Knowledge-Based Inference in Semiotic Systems • Semiotic Models of Behavior

Сходство:	• ***models***
Орбита сходства:	• Applied Semiotic: New Approach To Modeling And Simulating In Control Systems • Semiotic Models of Behavior • Making Applied Semiotics Work in the Development Software • Knowledge-Based Inference in Semiotic Systems • Hierarchical Integration of Indeterminancy • Developing models of a world with due regard for its dynamics General principles

П 3.4 Управление знаниями в системах электронного документооборота.

Системы электронного управления документооборотом (ЭУД) занимают сегодня одно из наиболее видных мест на рынке информационных технологий [*IDM*]. Объединив в унифицированную технологическую цепочку все необходимые средства поддержки жизненного цикла документов – инструменты сканирования и распознавания (оптический ввод документов), инструменты доступа к документам, представленным в различных электронных форматах (средства просмотра всех используемых форматов данных), инструменты маршрутизации документов в сети документооборота, инструменты управления доступом конкретных пользователей к конкретным категориям документов (средства защиты информации и поддержки санкционированного доступа), инструменты поддержки полнотекстового (пословного) информационного поиска в массивах вовлеченных в оборот документов, инструменты поддержки многоофисной работы на базе web-коммуникации и др., - системы такого типа создают и достаточно эффективно поддерживают для своих пользователей электронную версию своего рода “единого информационного пространства”. В литературе для их характеристики все чаще стал использоваться особый термин – системы поддержки “корпоративной памяти”, подразумевающий “процедурные” возможности создания, хранения и навигации в больших массивах электронных документов.

В последние пару лет в обсуждаемой области информационных технологий оформилась новая область интереса: все чаще употребляются, становясь все более популярными, термины “интеллектуальный” и “интеллектуализация”, “знания” и “управление знаниями”, и т.п.. Растущее внимание к подобной тематике легко объяснимо. Выступая в качестве “информационного основания” корпоративных систем управления, системы ЭУД эволюцио-

нируют в сторону задач поддержки принятия решений. Именно здесь необходимость оперировать с большими объемами плохо формализованной информации (ведь в подавляющем большинстве – это тексты на естественном языке, т.е. неструктурированные документы) заставляет специалистов обратиться к тому опыту, который наработан, в частности, в исследованиях в области искусственного интеллекта. Проблемы представления знаний, проблемы порождения новых знаний в рамках тех или иных формализованных рассуждений, и т.п. – вот наиболее актуальные аспекты функциональных возможностей следующего поколении систем ЭУД – систем управления знаниями (понимаемым как системы автоматического предоставления *всей* необходимой для *принятия решения* информации *в точности* тем, кто реально вовлечен в процесс принятия конкретного решения) [*KT 98*].

Если оставить в стороне "разговорные аспекты" (в частности, достаточно модные заявления о "роли знаний в информационной экономике постиндустриального общества" и т.п.), то среди факторов, стимулирующих рост интереса к технологиям управления знаниями следует (помимо проблематики ЭУД) отметить еще одну предметную область – построение больших информационно-управляющих систем (и, в первую очередь, одну из наиболее важных компонент этой деятельности - реинжиниринг бизнес-процессов). Здесь накоплен достаточно поучительный опыт "трудных случаев", когда, с одной стороны, не представляется возможным построить достаточно полное формальное описание объекта управления – его модель, а с другой стороны, значительная часть информации о задействованных механизмах принятия решений и оказывающих на них существенное влияние факторах представлена в виде текстовых документов (текстуальных описаний ситуаций-прецедентов). В таких случаях приходится оперировать схемами вида

СИТУАЦИЯ ➔ РЕШЕНИЕ (*),

причем описания конкретных ситуаций задаются совокупностями текстов на естественном языке. Именно здесь разработчики сталкиваются с необходимостью использовать механизмы автоматического анализа текстовых документов (автоматической классификации документов по заранее заданным тематическим рубрикам, автоматической реконструкции отношения релевантности текста конкретной тематике и т.п.). "Вызовом времени" оказываются умение порождать формализованные описания объекта управления по исходным плохо формализованным данным: порождение схем управления вида (*) в ходе обучения на прецедентах как успешных, так и ошибочных решений, а также автоматическое извлечение из текстов важной для описания СИТУАЦИЙ информации (так называемое feature extraction [*Gotthard 97*]). Важнейшей характерной особенностью этого "вызова" оказывается необходимость оперировать (в дополнение к хорошо известным процедурной и декларативной) еще одной формой представления знаний – понятийной. С "алгоритмической" точки зрения решающую роль в рассматриваемом варианте компьютерной обработки текстовой информации играет возможность автоматически идентифицировать обсуждаемые в текстах *понятия*, а также *отношения* между ними. От качества результатов, порождаемых используемыми здесь компьютерными инструментами, пожалуй, в наибольшей степени зависит окончательный результат управления.

Говоря об интеллектуальном управлении документами, обычно выделяют три основные группы процессов [*IDM, KT 98*]: системы "интеллектуальных" ссылок (intelligent linking), авто-маршрутизируемые документы (self-managing documents), и, наконец, интеллектуальные агенты (intelligent agents). В первом случае речь идет возможностях устанавливать и поддерживать динамически изменяемые связи между компонентами взаимозависимого (interlinked) множества документов. (При этом изменение одной компоненты может автоматически вызывать изменение другой, например, результаты "сводного отчета" автоматически изменяются при замене значений "обозреваемых" параметров в документах, охватываемых таковым "сводным отчетом"). Во втором случае речь идет о восприятии документа в компьютерной системе (управления знаниями) как объекта (в смысле объектно-

ориентированного подхода), т.е. как пары *документ* и *правила*, определяющие (в соответствии с его содержанием) жизненный цикл этого конкретного документа (маршрутизацию, дисциплины обработки и т.п.). В третьем случае имеется ввиду технология использования специальных программ (information agents, intelligent agents, knowledge agents и т.п.), задача которых – составить представление о содержании документа (путем опознания в нем тех или иных фрагментов, релевантных конкретным “шаблонам” - описаниям персоналий, упоминаниям географических объектов, упоминаниям конкретных технических устройств, упоминаниям конкретных финансовых инструментов, датировке событий и т.п.). По результатам работы интеллектуальных агентов с документом обычно порождается специальное множество индексов (атрибутов преимущественно фактографического характера), которые сопоставляются анализировавшемуся документу и используются в процессе проверки соответствия последнего профилю поисковых запросов того или иного пользователя. Именно таким путем выясняется, кому и какие документы должны быть адресованы системой ЭУД. В процессе описываемой адресации предлагается использовать также средства автоматического “понимания” документов (text abstracting and summarization tools), представления о смежных и альтернативных предметных областях. Предполагается, что таким путем на базе больших коллекций полнотекстовых документов могут строиться виртуальные хранилища информации (Virtual Data WareHouse), в которых названные выше три механизма “интеллектуализации” информации позволяют предоставить пользователю адекватные средства поддержки процесса принятия решений. (Фактически предполагается, что описание цели принятия решений в такой системе должно преобразовываться в структуры трех предложенные выше типов и далее давать возможность извлекать из хранилища информации все релевантные цели принятия решений документы, доставляя их нужному пользователю).

С точки зрения общепризнанных представлений об “интеллектуальности” информационных систем (см., например, развернутое обсуждение этой проблемы в работе [*Ивашко 86*]), предлагаемый в рамках систем интеллектуального управления документами комплекс технологий имеет реальные основания считаться действительно интеллектуальным. Тем не менее, анализ представленных в литературе реальных приложений (см., например, [*IDM, KT 98*] и др.) оставляет открытыми целый ряд конкретных вопросов. Наиболее трудные из них, в частности, связаны с *качеством* порождаемых здесь сегодня решений. Так, например, хорошо известная исследователям в области искусственного интеллекта проблема извлечения знаний (knowledge extraction) при работе с текстовыми документами выступает как проблема *автоматической* идентификации и *выделения* из неструктурированных текстов на естественном языке всех существенных элементов *содержания*. Целый ряд технологий здесь (в частности, автоматическая оценка релевантности текста и темы, автоматическая классификация полнотекстовых документов по их содержанию и затрагиваемой в них тематике, автоматическое реферирование полнотекстовых документов, и др.) находятся пока на стадии исследовательских проектов, где только намечены постановки интересных классов задач и сделаны всего лишь первые шаги в экспериментальной оценке эффективности предлагаемых для их решения подходов и алгоритмов (см., например, [*Забежайло 98а*] и др.). Здесь сегодня скорее уместно говорить о достаточно ясном понимании целей и перспектив развития обсуждаемого класса информационных технологий, чем о реальном “интеллектуальном поведении” используемых в промышленных приложениях систем. Цели ясны, первые шаги в их сторону дают разумные результаты, но для достаточно серьезного продвижения в выбранном направлении предстоит еще очень многое сделать.

Тем не менее, не сложно убедиться, что обсуждаемая нами техника текст-процессинга, осуществляемого средствами ДСМ-рассуждения, позволяет уже сегодня иметь целый ряд необходимых оснований для разработки систем интеллектуального управления документами (в названном выше понимании этого термина). Так, в частности, рассмотренные выше в **Приложениях 3.1-3** возможности ДСМ-технологии тематической идентификации и классификации текстов на естественном языке предоставляют практическую основу для

эффективного учета содержания обрабатываемых текстов в системах интеллектуальной маршрутизации документов или автоматической поддержки так называемых виртуальных групп экспертов.

Проиллюстрируем примером это соображение. Известно, что в больших корпорациях в последние годы все более интенсивно используются (для повышения обоснованности решений, принимаемых руководством фирмы) информационные структуры, обеспечивающие быстрый анализ экспертами в соответствующих предметных областях тех или иных конкретных проблем (см., например, [*FT 99*] и др.). Трудность здесь в том, что, с одной стороны, экспертов в корпорации бывает достаточно много, и каждый из них специализируется во вполне конкретном круге проблем, а с другой стороны, адекватный анализ каждой текущей группы вопросов должен быть выполнен достаточно быстро: руководству необходимы конкретные рекомендации, вообще говоря, в режиме реального времени. Итак, желательно, чтобы маршрутизация документов во внутри-корпоративной сети была устроена таким образом, чтобы каждый конкретный документ попадал *всем тем* и *только тем* экспертам, которые имеют заранее заявленную компетенцию по соответствующему кругу вопросов. (Понятно, что рассылка *всех* документов *всем* возможным адресатам не только бессмысленна по соображениям информационной перегрузки, но и недопустима в плане обеспечения информационной безопасности).

Легко представить себе, как в этой ситуации может быть построено решение, основанное на уже рассмотренных ДСМ-технологиях:

- каждый из экспертов заявляет список тем, определяющих область его профессиональной компетенции (и ответственности);
- соответствующие списки тем объединяются в сводный каталог тем (характеризующий область компетенции в принятии корпоративных решений);
- для каждой требующей экспертизы группы документов выполняется их тематическая классификация по сводному каталогу тем;
- по результатам построенной классификации система интеллектуального управления документооборотом реализует маршрутизацию необходимых материалов тем и только тем экспертам, которые компетентны в соответствующей проблематике.

Использование подобного инструментария позволяет руководителю практически мгновенно переадресовывать “горячие” материалы именно тем, кто должен быть привлечен к их анализу, а также контролировать процесс подготовки необходимых рекомендаций (например, альтернатив для принятия решения) практически в интерактивном режиме. Кроме того, сами эксперты в этой ситуации имеют возможность сразу же контактировать и консультироваться (если это необходимо) именно с теми коллегами, которые могут дать существенную информацию по анализируемому кругу вопросов.

Приложение 4.

Формализация метода геологических аналогий.

Наблюдаемое в последние три десятилетия все более и более широкое использование информационных технологий и средств компьютерного анализа данных во всевозможных индустриальных приложениях делает все более актуальной проблему *практического*[1] оценивания (т.е. возможностей оценить степень доверия к полученным результатам, предложить механизмы принятия или непринятия этих результатов и т.п.) найденных в ходе применения таких средств конечных результатов. Здесь в фокусе внимания оказывается хорошо известная, например - в экспериментальной физике, ситуация, когда для изучения явления используется сразу несколько различных инструментов (методик, экспериментальных установок, и т.п.), каждый из которых в достаточной степени корректен[2], но полученные общие результаты, частично "пересекаясь", все же не "совпадают" полностью. Таким образом, приходится конструировать "механизмы" сравнения частных результатов, учитывающие, с одной стороны, все "аргументы" ЗА и ПРОТИВ, и позволяющие, с другой стороны, получать "заслуживающие доверия" интегральные решения. Одним из подходов к решению этой проблемы является использование формальных моделей аргументации (см. также [*Финн 96, Финн 99, Хвостова 97*] и др.), о которых мы уже говорили выше. Здесь мы приведем пример использования конкретного фрагмента *процедурной компоненты* из представленной общей схемы аргументации средствами ДСМ-рассуждения.

Итак, предмет нашего интереса - модели и методы аргументации в системах компьютерного анализа данных. Аргументация средствами логического вывода (т.е. дедуктивного доказательства), к сожалению, применима лишь в специальных полностью формализованных областях, таких, например, как ряд разделов математики или теоретической физики. В естествознании (не говоря уже о гуманитарных дисциплинах) дела обстоят значительно хуже. Здесь во многих интересных практических приложениях приходится иметь дело с открытыми (незамкнутыми, пополняемыми) описаниями предметной области и объекта изучения. Кроме того исследователю, как правило, доступен набор разнородных (т.е. ориентированных на различные аспекты и "модели" объекта изучения и в то же время "равноправных" инструментов - приборов, программных систем анализа данных и т.п.. Свести воедино результаты использования таких инструментов при анализе конкретного объекта, аргументировать учет одних и отбрасывание других промежуточных результатов при порождении конечного решения - вот один из наиболее ответственных моментов работы исследователя, использующего компьютерные методы анализа данных.

При создании прикладных систем для применения в обсуждаемой области следует не упускать из вида еще ряд принципиально важных обстоятельств. Так для процедуры принятия решений в естественных науках достаточно характерна ситуация, когда конечное заключение представляет собой результат *качественного* анализа *количественных* (числовых) данных. В дополнение к этому, часто обсуждаемая процедура анализа данных оказывается контекстно зависимой: в одних контекстах механизм порождения конечного качественного заключения корректен, а в других - нет. Таким образом, формальная процедура аргументации получаемых результатов, вообще говоря, требует объединения в единой среде (в инструментальной среде логического рассуждения - !) символьных и числовых методов обработки данных. Кроме того, инструментарий поддержки процесса аргументации в обсуждаемых нами ситуациях должен иметь возможности для реализации "контекстного" анализ данных.

Мы представим потенциал ДСМ-средств конструктивной абдукции (правдоподобного вывода на достаточном основании, осуществляемого в рамках ДСМ-рассуждения) как

---

[1] Т.е. такой, которая могла бы иметь широкое использование на практике.

[2] Другими словами, имеются все необходимые основания утверждать, что результаты применения каждого такого инструмента "*заслуживают доверия*".

средства процедурной аргументации в ситуациях обсуждаемого типа. Возможности (см. [*Забежайло 96*, *Zabezhailo 96-b*, *Забежайло 97*]) подхода мы обсудим на примере формализации рассуждений в рамках так называемого метода геологических аналогий.

П 4.1 Метод геологических аналогий и ДСМ-рассуждения.

Метод аналогии (*методом геологических аналогий*) известен как один из наиболее распространенных и в то же время один из наименее формализованных практических приемов, используемых в реальном геологическом прогнозировании (см., например, [*Салле 78*] и др.). Идея метода, грубо говоря, состоит в попытках провести аналогию между изучаемым районом и уже известными нефтегазоносными районами. При этом в рассмотрение вовлекаются как *данные объективного характера* (данные геофизических - электромагнитных, гравитационных и т.п. измерений, данные геохимических анализов, результаты, полученные на пробных скважинах, и др.), так и *плохо формализованные данные* (например, представления о структурной схеме геологического строения района; сведения о последовательности геологических явлений, обусловивших возникновение залежей; гипотезы о характере отложения осадков, образовавших материнские породы, о характере их погружения и ходе процесса генерации углеводородов, об отложении осадков, образовавших коллекторы, о характере осадконакопления и литогенеза, и т.п.). Дополнительные сложности в анализе таких данных порождены “двухуровневой” схемой принятия решения: окончательное заключение о наличии (или отсутствии) углеводородов на изучаемом участке строится в процессе анализа “глобальных” признаков (как правило) качественного характера (таких, например, как история геологического развития района, наличие структур, их геологические особенности и др.). При этом, значительная часть качественных признаков “верхнего” уровня формируется в процессе оценки достаточно большого набора числовых параметров - результатов конкретных (геофизических, геохимических, гравитационных и др.) измерений, выполняемых на “нижнем” уровне анализа данных - при обработке, например, результатов так называемых “полевых” исследований. Именно поэтому центральной проблемой использования метода геологических аналогий является объективизация данных, включаемых в процесс принятия решений: объективизация (т.е. аргументация формальными средствами) перехода с “нижнего” на “верхний“ уровень принятия решений вместе с формализацией процедуры порождения собственно конечного решения. Таким образом, степень объективной (формальной-!) аргументированности рассуждений по аналогии есть главный компонент успеха в поиске новых залежей и месторождений обсуждаемым методом.

В **Главах 2-3** - уже было представлен логико-математический инструментарий, позволяющий осуществить формализацию метода геологических аналогий средствами машинного обучения (machine learning) и правдоподобного вывода (конструктивной абдукции) на достаточном основании. Технология предлагаемого нами подхода (как и обычно в приложениях ДСМ-рассуждений) может быть сведена к следующей совокупности шагов:

располагая информацией о примерах и контрпримерах анализируемых объектов[3]*, мы строим “причинные зависимости”, обуславливающие как наличие так и отсутствие искомых свойств анализируемых объектов*[4]*, после чего проверяем переносимость (т.е. возможность “распространить” порожденные “причинные зависимости” с исходно заданных примеров и контпримеров) на описание нового (изучаемого) района*[5]**.**

---

[3] Например, сведениями о точках на карте конкретного региона, в которых есть углеводороды в интересующем нас количестве (мы будем называть эти данные позитивными примерами), и сведения о точках, где таковых углеводородов нет (мы будем называть их негативными примерами или контрпримерами).

[4] Т.е. выделяем совокупности параметров (а также области их значений), характеризующие соответственно наличие и отсутствие искомых свойств анализируемых объектов.

[5] При этом новый район может быть охарактеризован как отдельными точками (по каждой из которых мы будем стремиться породить прогноз наличия или отсутствия искомых свойств), так и указанием на карте мест-

Центральным элементом подхода, основанного на методе геологических аналогий является формализация (формальное уточнение) используемого экспертами-геологами содержательного представления о сходстве данных - анализируемых объектов геологоразведки. Проблема объективизации сходства в геологоразведке хорошо известна (см., например, [*Воронин 83, Воронин 91*] и др.), имеет обширную литературу (см., например, библиографию к работе [*Воронин 91*] и др.) и целый ряд конкретных решений, использованных в серьезных практических приложениях (см., например, [*Лаудон 81*] и др.). Один из наиболее часто применяемых здесь подходов - переход от качественного содержательного сходства, которым оперируют эксперты-геологи, к математической формализации сходства как метрической конструкции[6]. Однако очевидные вычислительные достоинства такого подхода не снимают ряда присущих ему трудностей, наиболее серьезные из которых можно суммировать к следующим двум обстоятельствам:

- необходимо оперировать в том числе с качественными (неметрическими-!, а более точно - дометрическими[7]) описаниями анализируемых объектов;
- необходимо иметь возможность согласовать качественное сходство в “дометрическом” случае с числовым сходством в “метрическом” случае так, чтобы при формализации и последующих вычислениях не нарушалось содержательное представление о сходстве, удовлетворяющее эксперта-геолога.

Нетривиальность первого требования достаточно явно иллюстрируется необходимостью, например, формализовать в рамках метода геологической аналогии традиционные для геологоразведки месторождений углеводородов рассуждения о “структурных” характеристиках залежи - “единстве” так называемых “ловушки”, “покрышки”, “коллектора” и “нефтегазоматеринской породы”.

Нетривиальность второго требования естественным образом обусловлена пожеланиями экспертов-геологов иметь в своем распоряжении такую систему поддержки принятия решений, в которой “метрический” инструментарий позволял бы “уточнять и детализировать” представления об объекте исследования, полученные “дометрическими” средствами, а “дометрический” инструментарий позволял бы, в свою очередь, порождать по “точной” числовой модели объекта исследования ее обобщенное качественное описание (которое можно было бы использовать как содержательный «образ» соответствующего количественного описания анализируемых данных и порождаемых результатов).

Оба приведенных нами обстоятельства тесным образом связаны с более общей проблемой - проблемой выбора языка представления данных и знаний для принятия решений. В задаче геологического прогнозирования проблема выбора языка представления - это, прежде всего, проблема выбора

- совокупности содержательных свойств[8] анализируемых объектов, принимаемых во внимание геологом-экспертом в процессе принятия решения;
- совокупности наблюдаемых (измеряемых, вычисляемых, ...) параметров, представляющих выделенные для анализа содержательные свойства изучаемых объектов;
- совокупности соответствующих шкал измерения значений наблюдаемых параметров; и, кроме того,

---

ности односвязных областей, где с точки зрения предложенного нами способа формализации метода геологических аналогий следует (положительный прогноз) или не следует (отрицательный прогноз) ожидать наличия искомых углеводородов.

[6] При таком подходе исходные анализируемые данные некоторым путем “погружаются” в подходящее метрическое пространство, где сходство анализируемых объектов определяется их близостью в терминах соответствующей метрики (т.е., попросту, расстоянием между ними).

[7] Т.е. такими описаниями, когда данных для определения соответствующей метрики еще просто недостаточно.

[8] При этом необходимо учитывать как *простые* (“атомарные”), так и *комплексные* (т.е. наделенные конкретной - возможно, нетривиальной - внутренней структурой) свойства. Не следует забывать, что такие свойства могут оцениваться как прямыми наблюдениями, так и анализом косвенных данных.

- “информативных”[9] (т.е. “осмысленных”, “разумных”, “существенных” и т.п.) подмножеств свойств в совокупности ранее выделенных содержательных свойств анализируемых объектов[10],

и др..

Таким образом задача формализации метода геологических аналогий требует математического инструментария, пригодного во-первых для объединения в единой вычислительной среде качественных и числовых (метрических) средств анализа данных и принятия решений. Во-вторых, обсуждаемый нами математический инструментарий должен при работе с числовыми данными предоставить возможность оперировать с данными в рамках объектов, которые мы вслед за [*Воронин 91*] и [*Сокал 80*] будем называть *разношкальными квазипространствами*[11]. Последнее обстоятельство является очевидным следствием сформулированных нами выше требований к возможностям языка представления данных: действительно, когнитивная природа метода геологических аналогий требует возможностей обработки *разнородных* (символьных и числовых, качественных и количественных, ...) *данных* при помощи *однородных* (однотипных, согласованных) *вычислительных средств*.

Отличительной особенностью применения ДСМ-метода в задаче анализа и прогноза нефтегазоносности является его ориентированность на анализ *"причинности"* наличия искомых углеводородов в соответствующих регионах и геологических структурах. Используя в качестве входной информации различные экспериментальные (или экспертные) данные о соответствующей местности, выделяются (там, где это возможно:

а) то подмножество (исходно заданного множества) параметров в описании входных данных, которое (с точки зрения ДСМ-метода) может характеризовать *"причины"* наличия углеводородного сырья в исследуемой местности (и геологических структурах)[12];

б) *интервалы изменения* релевантных (см. пункт **а)** выше) *параметров* (т.е. интервалы числовых значений отобранных на предыдущем шаге параметров, характеризующие “*допустимые*" - с точки зрения предполагаемой *"причинности"* наличия углеводородов - *области изменения* соответствующих числовых величин).

П 4.2 Пример ДСМ-прогноза нефтегазоносности региона.

В качестве исходных данных для ДСМ-обучения и прогноза нами выбраны сведения по 34 точкам достаточно хорошо изученной области одного из нефтегазоносных регионов Венгрии[13], при этом в 21 из названных точек имелись искомые углеводороды в достаточном количестве (и мы будем рассматривать данные по каждой из этих точек как *позитивные примеры*), а в оставшихся 13 точках необходимые нам свойства не наблюдалось (что позволило нам рассматривать описания этих точек как *негативные примеры* для проведения геологических аналогий в рамках ДСМ-метода).

В качестве языка представления геологических данных и знаний, ориентированного на решаемую нами задачу прогноза, был выбрана система показателей, разработанная группой экспертов под руководством Б.В.Ермакова (ИППИ РАН). Затем исходные данные, полученные от наших венгерских коллег, были преобразованы в возглавляемой В.Г.Гитисом

---

[9] Информативных в смысле конкретной решаемой задачи геологического прогнозирования.

[10] Причем такое выделение должно, разумеется, учитывать и выбор “информативных” интервалов значений наблюдаемых параметров, с помощью которых характеризуются содержательные свойства анализируемых объектов.

[11] Пример объектов такого типа представлен, в частности, в работе [*Zabezhailo 93*], где обсуждаются проблема анализа связи ***СТРУКТУРА==>СВОЙСТВА*** в ситуации, когда пространство возможных числовых значений параметров, характеризующих ***СТРУКТУРУ*** анализируемых объектов, разбивается на области “постоянства механизма осуществления” анализируемых ***СВОЙСТВ***. Внутри каждой из таких областей на основе (в процессе) анализа соотвествующих исходных данных порождается своя специфическая метрика (характеризующая анализируемые объекты по сходству ***СТРУКТУР*** и механизма осуществления изучаемых ***СВОЙСТВ***).

[12] Мы будем называть эти параметры *релевантными* цели прогнозирования.

[13] Конкретные данные предоставлены ***ALL*** - *Applied Logic Laboratory*, Budapest, Hungary.

группе геологического прогнозирования ИППИ РАН в названный проблемно-ориентированный язык представления данных и знаний. (Ниже - в **Таблицах П 4.2.1,2** и **3** - представлены: список включенных в рассмотрение параметров исходных данных, а также конкретные значения соответствующих параметров для позитивных и негативных примеров)[14].

Расчеты велись лишь на параметрах, принимающих числовые значения, т.е. была задействована лишь версия ДСМ-метода для числовых данных (см., выше – **Главы 2-3**).

Для порождения конкретных результатов нами была использована версия для числовых данных так называемого метода простого сходства с запретом на контрпримеры (см., также **Главу 3**). Такой выбор ДСМ-стратегии вычислений был обусловлен тем, что характер задействованных в нем правил правдоподобного вывода гарантирует непротиворечивость порождаемых результатов: процесс вычислений устроен так, что ни один из отрицательных примеров не попадает ни в одну из областей позитивного прогноза (ни одна из точек, характеризующих отрицательные примеры, не попадает ни в одну из областей на карте исследуемого региона, соответствующую какой-либо позитивной гипотезе - т.е. выделенной в процессе прогнозирования так, что на этой области в соответствии с методом геологических аналогий следует ожидать наличия искомых углеводородов в необходимом количестве), и, аналогично, ни один из позитивных примеров не попадает в область негативного прогноза.

Результат ДСМ-ИАД - семейство гипотез (позитивных и негативных), представленного в **Таблице П 4.2.4** вида. Каждая из гипотез характеризуется данными четырех типов:

- информацией о “знаке” (а именно: позитивная гипотеза характеризует область, где в соответствии с методом геологических аналогий[15] следует ожидать наличия углеводородов; отрицательная - область, где следует ожидать отсутствия таковых углеводородов);
- списком исходно заданных объектов, ДСМ-сходство которых и выражает рассматриваемая гипотеза;
- списком параметров, значения которых (изменяясь “сходным” образом) и породили рассматриваемую гипотезу;
- интервалом допустимых (т.е. сходных в смысле выбранной стратегии ДСМ-вычислений) значений каждого из порождающих рассматриваемую гипотезу параметров.

Легко видеть, что каждая из порождаемых в процессе ДСМ-вычислений гипотез выделяет соответствующую область на карте анализируемого региона. Действительно, допустимые значения каждого из параметров, входящих в гипотезу задают соответствующую полосу (интервал между подходящими линиями уровня на карте значений рассматриваемого параметра). Пересечение[16] полос такого вида и определяет ту область на карте региона, где в соответствии с методом геологических аналогий можно ожидать наличия (случай позитивного прогноза) или отсутствия (случай негативного прогноза) искомых углеводородов. Объединение всех (т.е. соответствующих всем порожденным ДСМ-гипотезам) выделенных таким образом областей и составляет конечный результат ДСМ-прогноза нефтегазоносности исследуемого региона по методу геологических аналогий[17]. Интересно отметить, что

---

[14] Следует отметить нетривиальный характер этого преобразования: так, например, при картографировании ряда параметров некоторые конкретные их значения пришлось строить с использованием сложной интерполяционной техники, реализованной в программной системе **GEO**, созданной под руководством В.Г.Гитиса в ИППИ РАН (см., например, работы [*Gitis 95, Gitis 95a, Гитис 04*] и др.).

[15] “Буквальное” понимание аналогии в данном случае будет означать сходство изначально заданных для обучения системы данных с описанием любых новых точек (или областей) на карте региона, удовлетворяющих ограничениям, определяемым конкретным списком параметров (вместе с областями изменения их значений) в соответствующей ДСМ-гипотезе.

[16] Т.е. общая часть карты региона, входящая в каждую из порождаемых рассматриваемой гипотезой полос значений параметров.

[17] Визуализация (т.е. представление в виде цветных карт соответствующего масштаба) рассчитанных таким образом областей прекрасным образом может быть реализована средствами уже упоминавшейся нами программной системы **GEO**, созданной в ИППИ РАН (см., например, работы [*Gitis 95, Gitis 95a, Гитис 04*] и др.).

выбор стратегии ДСМ-вычислений[18] в соответствии с методом простого сходства с запретом на контрпримеры гарантирует, что на карте исследуемого региона области позитивного и негативного прогноза *не пересекаются*, т.е. наш прогноз обладает свойством содержательной *непротиворечивости*.

Особого рассмотрения заслуживает ряд обстоятельств, связанных с качеством и точностью измерения предлагаемых для обучения ДСМ-системы начальных данных. Представляемый нами эксперимент показал, в частности, сколь важную роль в реализации метода геологических аналогий могут играть аккуратно учитываемые ошибки измерений и другие традиционные для обработки экспериментальных числовых данных характеристики.[19]. Отмечая важность и глубину возникающих здесь проблем, мы посчитали уместным лишь обозначить их место в представленной работе, полагая, что они заслуживают отдельного анализа и обсуждения.

***

Таким образом, мы представили подход, позволяющий предложить полномасштабную формализацию метода геологических аналогий.

Геолог пытается провести аналогию (см., например, [*Салле 78*]) между изучаемым им районом и уже известными нефтеносными районами, полагая, что чем больше общих черт выявляет такое сравнение, тем больше надежд, что изучаемый район тоже окажется нефтеносным. Использование метода геологических аналогий в значительной степени оправдывается стремлением задействовать десятилетиями накапливаемый в этой отрасли промышленности практический опыт. Однако центральная проблема применения обсуждаемого метода состоит в том, что в большинстве случаев реализация рассуждений в рамках геологических аналогий осуществлялась лишь субъективным полуэмпирическим путем. Использование ДСМ-метода позволяет предложить математическую формализацию[20] представленного нами способа обработки геологических данных.

ДСМ-инструментарий позволяет объединить в унифицированной среде

- обработку качественных данных (информации качественного характера о “структурных” свойствах изучаемого геологического района - о характере стратиграфических разрезов, структур, ловушек, исторических особенностях геологического развития изучаемого региона и т.п.)[21], учитываемых геологами в принятии решений по методу аналогий;
- обработку числовых данных средствами “дометрического” анализа[22] (т.е. качественного анализа числовых данных), результатами применения которых являются
  - а) отбор из всего доступного множества числовых данных лишь тех параметров, которые релевантны цели исследования - конкретной задаче геологического прогнозирования;
  - б) выделение областей значений отобранных параметров, релевантных цели прогнозирования;
- обработку числовых данных стандартными метрическими средствами (т.е. обычными средствами анализа извлекаемых из данных числовых зависимостей).

Именно таким образом ДСМ-инструментарий позволяет в рамках формализации метода геологических аналогий получить решение одной из наиболее интересных проблем современной computer science - осуществить объединение (т.е. взаимно *согласованную реализацию* в *унифицированной вычислительной среде*) символьных и числовых расчетов.

---

[18] Т.е. использование в качестве базового элемента правдоподобного вывода ДСМ-правила, реализующего метод простого сходства с запретом на конртпримеры (см. выше).

[19] Так, например, доступность вместе с конкретными исходными данными аккуратно промерянных и учтенных ошибок измерений, позволяет перейти от “точечной” к “интервальной” версии ДСМ-метода обработки числовых данных, что в свою очередь позволяет повысить качество порождаемых в процессе ДСМ-прогнозирования результатов.

[20] Реализуемую как в виде логико-математической модели рассуждений и принятия решений в соответствии с геологическими аналогиями, так и в виде компьютерных систем геологического анализа и прогнозирования.

[21] Такая обработка реализуется средствами структурного ДСМ-метода (см., выше **Главы 2-3**).

[22] Такая обработка реализуется средствами представленного выше числового ДСМ-метода.

Таблица П 4.2.1.

Список параметров, принимавшихся во внимание при анализе данных.

Предложенные специалистами ***ALL*** для обработки и прогноза исходные данные по конкретному региону Венгерской Республики были описаны экспертами (руководитель группы - Б.В.Ермаков) в ИППИ РАН (коллектив исследователей под руководством В.Г.Гитиса) в терминах специального языка представления геологических данных и знаний, в основе которого - известная методика так называемого ***WEGA-D*** анализа (см., например, [*Dzwinel 83, Dzwinel 81]* и др.).

Параметры[23], которые были использованы в обсуждаемом нами примере.

1. Девиация фазы вектора напряженности электрического поля в приемнике, ориентированном по оси X, для двух эталонных значений частот - 0.3 Гц и 0.6 Гц - излучателя типа 1.
2. Усредненная амплитуда векторов напряженности электрического поля в области нижних частот для излучателя типа 3.
3. Усредненная амплитуда векторов напряженности электрического поля в области верхних частот для излучателя типа 3.
4. Девиация фазы вектора напряженности электрического поля в приемнике, ориентированном по оси Y, для двух эталонных значений частот - 0.3 Гц и 0.6 Гц - излучателя типа 3.
5. Усредненная амплитуда векторов напряженности электрического поля в области верхних частот для излучателя типа 1.
6. Усредненная амплитуда векторов напряженности электрического поля в области нижних частот для излучателя типа 1.
7. Пространственная производная от величины параметра 2.
8. Количественная характеристика гравитационного поля.
9. Результат сглаживания значений гравитационного поля (с вычислением лаплассиана).
10. Девиация фазы вектора напряженности электрического поля в приемнике, ориентированном по оси X, для 2 эталонных значений частот - 0.3 Гц и 0.6 Гц - излучателя типа 3.
11. Девиация фазы вектора напряженности электрического поля в приемнике, ориентированном по оси Y, для 2 эталонных значений частот - 0.3 Гц и 0.6 Гц - излучателя типа 1.

Таблица П 4.2.2.

Таблица значений анализируемых параметров для позитивных примеров.

			1	2	3	4	5	6	7	8	9	10	11
+1	1	100	84	821	825	99	769	757	11	1144	66	73	2
+2	2	100	79	820	826	109	768	760	4	1135	37	115	1
+3	5	100	74	814	822	48	766	756	-13	1133	-13	45	2
+4	6	100	7	813	820	17	766	757	-15	1149	34	19	-6
+5	7	100	48	812	818	52	765	755	-21	1137	-48	48	-1
+6	9	100	73	816	823	57	768	756	8	1140	40	45	-2
+7	11	100	92	814	818	78	766	755	3	1145	46	71	10
+8	13	100	73	817	822	48	768	760	-2	1155	-23	42	3
+9	14	100	42	818	822	1	775	765	5	1178	15	-23	-2
+10	15	100	43	814	815	-70	764	753	-11	1151	-64	-72	0
+11	16	100	44	813	813	-38	762	752	-8	1191	57	-52	-5
+12	17	100	5	816	818	-70	761	752	-8	1156	38	-87	-11
+13	19	100	-17	813	816	1	764	755	1	1193	59	-21	-7
+14	21	100	8	804	805	-178	748	745	11	1115	56	-189	-3
+15	22	100	-77	816	809	-127	759	752	13	1131	-10	-117	-10
+16	23	100	-27	812	810	-52	771	763	-15	1165	-85	-18	-36
+17	24	100	36	809	806	-30	772	760	-15	1140	37	49	-13
+18	28	100	79	806	813	37	763	750	-7	1143	26	65	18
+19	30	100	115	806	815	13	763	749	-12	1141	67	48	24
+20	31	100	75	811	816	57	764	754	-13	1137	48	64	12
+21	32	100	89	805	811	24	763	750	-25	1143	34	77	24

При этом:

[23] Т.е. параметры, числовые значения которых для соответствующих точек на карте изучаемого региона были использованы в анализе нефтегазоносности.

- *первый ненумерованный столбец* представляет порядковую нумерацию соответствующих положительных примеров;
- *второй ненумерованный столбец* представляет их нумерацию в исходном списке примеров;
- *третий ненумерованный столбец* - данные о наличии углеводородов;
- *столбцы с номерами 1-11* содержат числовые значения соответствующих параметров (см. **Таблицу П 4.2.1**) для соответствующих примеров.

Таблица П 4.2.3.

Таблица значений анализируемых параметров для негативных примеров.

			1	2	3	4	5	6	7	8	9	10	11
-1	3	-200	31	827	822	174	767	759	40	1148	23	164	-9
-2	4	-200	152	817	824	93	770	760	-14	1113	-43	99	12
-3	8	-200	-3	814	817	79	763	755	-4	1148	56	71	-7
-4	10	-200	79	815	819	106	763	755	-6	*	*	81	3
-5	12	-200	134	809	812	89	763	751	-9	1146	-18	93	17
-6	18	-200	-308	810	819	-31	762	755	-7	1183	22	-20	-6
-7	20	-200	80	823	821	-18	769	763	9	1117	21	-28	15
-8	25	-200	56	814	823	-49	754	752	-2	*	*	-34	-1
-9	26	-200	-102	827	834	-217	768	761	12	1088	-74	-193	-4
-10	27	-200	203	878	823	153	774	767	2	*	*	158	17
-11	29	-200	112	815	817	97	765	758	-17	1038	-6	88	14
-12	33	-200	130	809	812	97	759	749	-4	1140	11	100	19
-13	34	-200	150	811	816	73	763	753	2	1136	66	91	21

При этом:

- *первый ненумерованный столбец* представляет порядковую нумерацию соответствующих отрицательных примеров;
- *второй ненумерованный столбец* представляет их нумерацию в исходном списке примеров;
- *третий ненумерованный столбец* - данные об отсутствии углеводородов;
- *столбцы с номерами 1-11* содержат числовые значения соответствующих параметров (см. **Таблицу П 4.2.1**) для соответствующих примеров[24].

Таблица П 4.2.4.

Примеры порожденных по методу геологических аналогий ДСМ-гипотез.

4.1. Позитивная гипотеза: <***21.18.3.8*** || ***1.2.3.4.5.6.9.10***> .

Порождена ДСМ-сходством **3, 8, 18** и **21** позитивных примеров по значениям параметров **1, 2, 3, 4, 5, 6, 9** и **10**, попадающим, соответственно в следующие числовые интервалы:

1: [**89_73**] 2: [**805_817**] 3: [**811_822**] 4: [**24_48**]
5: [**763_768**] 6: [**750_760**] 9: [**34_-23**] 10: [**77_42**].

«Визуализация" м-башни, соответствующей рассматриваемой нами гипотезе:

			1	2	3	4	5	6	7	8	9	10	11
+14	21	100	8	804	805	-178	748	745	11	1115	56	-189	-3
+15	22	100	-77	816	809	-127	759	752	13	1131	-10	-117	-10
+10	15	100	43	814	815	-70	764	753	-11	1151	-64	-72	0
+12	17	100	5	816	818	-70	761	752	-8	1156	38	-87	-11
+16	23	100	-27	812	810	-52	771	763	-15	1165	-85	-18	-36
+11	16	100	44	813	813	-38	762	752	-8	1191	57	-52	-5
+17	24	100	36	809	806	-30	772	760	-15	1140	37	49	-13
+9	14	100	42	818	822	1	775	765	5	1178	15	-23	-2
+13	19	100	-17	813	816	1	764	755	1	1193	59	-21	-7
+19	30	100	115	806	815	13	763	749	-12	1141	67	48	24
+4	6	100	7	813	820	17	766	757	-15	1149	34	19	-6
+21	32	100	89	805	811	24	763	750	-25	1143	34	77	24
+18	28	100	79	806	813	37	763	750	-7	1143	26	65	18
+3	5	100	74	814	822	48	766	756	-13	1133	-13	45	2
+8	13	100	73	817	822	48	768	760	-2	1155	-23	42	3
+5	7	100	48	812	818	52	765	755	-21	1137	-48	48	-1

[24] При этом * означает, что соответствующее значение параметра попросту отсутствует.

+6	9	100	73	816	823	57	768	756	8	1140	40	45	-2
+20	31	100	75	811	816	57	764	754	-13	1137	48	64	12
+7	11	100	92	814	818	78	766	755	3	1145	46	71	10
+1	1	100	84	821	825	99	769	757	11	1144	66	73	2
+2	2	100	79	820	826	109	768	760	4	1135	37	115	1
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**

Ось монотонности - столбец **4**.

4.2. Позитивные гипотезы: **<*20.5.13.7.3.6.1* || *2.6*>** и **<*13.7.3.6.1* || *2.3.5.6*>.**

Порождены ДСМ-сходством соответствующих подмножеств множества позитивных примеров {***20,5,13,7,3,6,1***} по значениям соответствующих подмножеств множества числовых параметров {***2,3,5,6***}, попадающим в следующие числовые интервалы:

2: [**811_821**] 3: [**816_825**] 5: [**764_769**] 6: [**754_757**]

“Визуализация” м-башен, соответствующих рассматриваемым нами гипотезам:

			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
+14	21	100	8	804	805	-178	748	745	11	1115	56	-189	-3
+19	30	100	115	806	815	13	763	749	-12	1141	67	48	24
+18	28	100	79	806	813	37	763	750	-7	1143	26	65	18
+21	32	100	89	805	811	24	763	750	-25	1143	34	77	24
+11	16	100	44	813	813	-38	762	752	-8	1191	57	-52	-5
+12	17	100	5	816	818	-70	761	752	-8	1156	38	-87	-11
+15	22	100	-77	816	809	-127	759	752	13	1131	-10	-117	-10
+10	15	100	43	814	815	-70	764	753	-11	1151	-64	-72	0
+20	31	100	75	811	816	57	764	754	-13	1137	48	64	12
+5	7	100	48	812	818	52	765	755	-21	1137	-48	48	-1
+13	19	100	-17	813	816	1	764	755	1	1193	59	-21	-7
+7	11	100	92	814	818	78	766	755	3	1145	46	71	10
+3	5	100	74	814	822	48	766	756	-13	1133	-13	45	2
+6	9	100	73	816	823	57	768	756	8	1140	40	45	-2
+1	1	100	84	821	825	99	769	757	11	1144	66	73	2
+4	6	100	7	813	820	17	766	757	-15	1149	34	19	-6
+2	2	100	79	820	826	109	768	760	4	1135	37	115	1
+8	13	100	73	817	822	48	768	760	-2	1155	-23	42	3
+17	24	100	36	809	806	-30	772	760	-15	1140	37	49	-13
+16	23	100	-27	812	810	-52	771	763	-15	1165	-85	-18	-36
+9	14	100	42	818	822	1	775	765	5	1178	15	-23	-2

Ось монотонности - столбец 6.

4.3. Негативная гипотеза: <***12.5.13.3.11*** || ***2.3.5.6***> .

Порождена ДСМ-сходством **12, 5, 13, 3** и **11** негативных примеров по значениям параметров **1, 2, 3, 4, 5, 6, 9** и **10**, попадающим, соответственно в следующие числовые интервалы:

2: [**809_815**] 3: [**812_817**] 5: [**759_765**] 6: [**749_758**]

“Визуализация” м-башни, соответствующей рассматриваемой нами гипотезе:

			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
-6	18	-200	-308	810	819	-31	762	755	-7	1183	22	-20	-6
-9	26	-200	-102	827	834	-217	768	761	12	1088	-74	-193	-4
-3	8	-200	-3	814	817	79	763	755	-4	1148	56	71	-7
-1	3	-200	31	827	822	174	767	759	40	1148	23	164	-9
-8	25	-200	56	814	823	-49	754	752	-2	*	*	-34	-1
-4	10	-200	79	815	819	106	763	755	-6	*	*	81	3
-7	20	-200	80	823	821	-18	769	763	9	1117	21	-28	15
-11	29	-200	112	815	817	97	765	758	-17	1038	-6	88	14
-12	33	-200	130	809	812	97	759	749	-4	1140	11	100	19
-5	12	-200	134	809	812	89	763	751	-9	1146	-18	93	17
-13	34	-200	150	811	816	73	763	753	2	1136	66	91	21
-2	4	-200	152	817	824	93	770	760	-14	1113	-43	99	12
-10	27	-200	203	878	823	153	774	767	2	*	*	158	17

Ось монотонности - столбец **1**.

Приложение 5.

Банковский клиринг как задача управления и оптимизации в больших системах.

Известно, что ряд задач оптимизации больших систем не имеет осуществимых на практике (по соображениям вычислительной сложности поиска ответа) точных решений - время поиска соответствующего экстремума средствами современной вычислительной техники оказывается недопустимо большим. Ситуация достаточно часто усугубляется также тем, что не представляется возможным предложить адекватную модель оптимизируемой большой системы (формализованную в виде, например, системы балансовых соотношений и т.п.). Своего рода "классическими" примерами таких оптимизационных проблем стали, например, задачи управления ядерными станциями, сложными военными комплексами[1], глобальными компьютерными сетями и др.. Внимание работающих в этой области специалистов сегодня сфокусировано на подходе, который часто называют семиотическим анализом[2] (см., например, [*AS'95, SA'96, ISAS'97,*]. и др.). В основе используемой здесь техники лежит предположение о взаимосвязанности структурных описаний изучаемых объектов и проявляемых ими свойств[3].

П 5.1. Проблема управления в больших системах комплексного характера.

В крупномасштабных приложениях обсуждаемого типа (т.е. в управлении системами большой размерности - комплексами подсистем, оперирующих разнотипными данными и

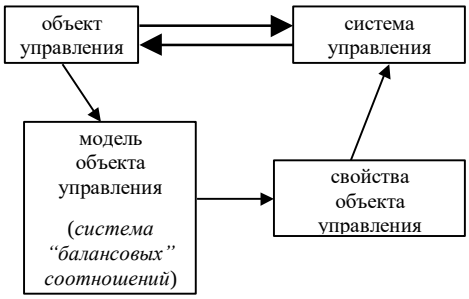


Рис. П 5.1.1. Схема управления, где по *объекту управления* строится в *модель* (т.е. система балансовых соотношений, например, уравнений с ограничениями и т.п.), по которой анализируются *свойства объекта управления* (выделенные на модели), что позволяет сформировать *систему управления объектом* (предназначенную для контроля его свойств)

[1] Содержательный иллюстративный пример комплекса такого типа - крупная группировка военно-морских сил, решения по оперативному управлению которой определяются не только (достаточно детально формализованными в показаниях соответствующих приборов и локальных систем управления) данными о текущем состоянии оборудования и технических средств - кораблей, систем вооружения, средств поддержки, инженерной инфраструктуры и т.п., но и неформализованными данными докладов и сообщений персонала, касающихся как состояния собственных сил, так и оперативного взаимодействия с противником.

[2] В нашей стране это направление сложилось вокруг работ Д.А.Поспелова по ситуационному управлению - см., например, [*Поспелов 86*] и др..

[3] Читатель легко обнаружит здесь взаимосвязи как с известными работами в области формальной семантики (логико-математическими, структурно-лингвистическими, философскими и т.п.), так и, например, с обсуждавшимися нами выше проблемами анализа структурной химической информации, где гипотеза Бутлерова о полной определимости свойств конкретного вещества его химической структурой есть фундаментальная предпосылка всех обсуждавшихся построений.

использующих разнородные механизмы обработки этих данных), фактически, приходится балансировать между двумя крайностями: с одной стороны, классический инструментарий моделирования (грубо говоря - техника разного рода балансовых соотношений) допускает разрешимость в режиме реального времени лишь для моделей ограниченной[4] размерности; с другой - при более детальном[5] описании объекта управления оказывается актуальной уже не только проблема вычислений в реальное время, но и более общий вопрос о выборе адекватных средств моделирования[6]. Отсутствие возможностей использовать (удовлетворительным образом) обычную технику моделирования - т.е. схему Рис. П.5.1.1 - стимулировало обращение к технике так называемого *семиотического моделирования* (см. [*AS'95, Поспелов 86*] и др.). Здесь делается попытка судить о свойствах объекта управления, минуя процесс порождения (в явном виде) его модели: предметом исследования становится отношение ***ОПИСАНИЕ ОБЪЕКТА УПРАВЛЕНИЯ ==> СВОЙСТВА ОБЪЕКТА УПРАВЛЕНИЯ***. В рамках такого подхода считается, что описание объекта управления содержит (хотя, быть может, и косвенным образом[7]) достаточный объем данных, определяющих как наличие, так и отсутствие анализируемых свойств у объекта управления. При этом допускаются достаточно широкие рамки для выбора языков описания как "структуры" объекта управления, так и его свойств[8]: важно лишь, чтобы выразительные возможности конкретного языка позволяли представить (а затем выделить из общего описания и анализировать) все факторы, определяющие "причинность" наличия и отсутствия изучаемых свойств.

В семиотическом моделировании, фактически, реализуется схема (Рис П 5.1.2):

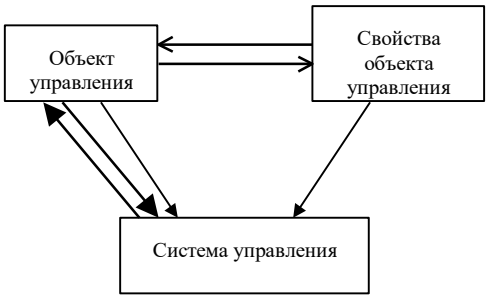


Рис. П 5.1.2. Схема управления, где *система управления* отражает (например, в виде соответствующих таблиц решений) частично определенное отношение ***ОПИСАНИЕ ОБЪЕКТА УПРАВЛЕНИЯ*** **==>** ***ОПИСАНИЕ*** **ЕГО** ***СВОЙСТВ***.

Более точно: множество всех возможных описаний объекта управления - т.е. множество "*ситуаций*", в которых может оказаться объект управления - разбивается на подмножества - *состояния*, каждому из которых сопоставляется подходящий набор управляющих воздействий - *решение*, "гарантирующее" наличие (или отсутствие) необходимых свойств. Проектирование (или настройка) системы управления при таком подходе состоит в порождении *решающих правил*, позволяющих для произвольной *ситуации* (конкретного описания

[4] И, как правило, не удовлетворяющей требованиям сегодняшних реальных приложений.

[5] Т.е. пригодном для моделирования более "тонких" эффектов в поведении объекта управления.

[6] Поучительным примером ситуаций такого типа может служить практика управления современными ядерными станциями, где оператору необходимо отслеживать для принятия решений такое число конкретных процессов и систем, что только на центральный пульт управления станции приходится выводить показания нескольких десятков сотен датчиков (см., например, [*AS'95*], стр. 410-413).

[7] Другими словами, в описании объектов заведомо содержатся данные, определяющие "причинность" наличия и отсутствия анализируемых свойств объекта управления. В то же время эти данные смешаны с информацией побочного характера, и их выделение (из общего описания объекта управления) - специальная задача.

[8] Сборник [*AS'95*], например, содержит описания конкретных приложений, где описания анализируемых объектов были представлены, например, на естественном языке (в частности - в системах ситуационного управления) или языке структурных формул (т.е. 2-D или 3-D молекулярных графов) химических соединений (в частности - в задачах экологического мониторинга потенциально опасных химических соединений в окружающей человека среде) и др.).

объекта управления) найти адекватное текущей *ситуации решение*[9] (используя, когда необходимо, опыт обучения на примерах - заранее описанных[10] парах *ситуация==>решение)*.

Возвращаясь к затронутой нами проблеме оптимизации больших систем, сфокусируем внимание на проблеме сложности (т.е. объемах) вычислений при поиске оптимальных решений. Принадлежность многих возникающих здесь задач к классу трудных вычислительных проблем (***NP***-трудных, ***NP***-полных, перечислительно полных и т.п. - см., например [*Гэри 82*] и др.), делает практически неприемлемым[11] использование таких стандартных средств "точной" оптимизации, как, например, процедуры типа метода ветвей и границ. Обращение к средствам недетерминистского поиска (например, методам теории расписаний или теории массового обслуживания), зачастую способно обеспечить лишь асимптотически оптимальные решения (что оказывается неудовлетворительным при анализе одной конкретной *ситуации* - описания объекта управления). В то же время практически важным оказывается выделение существенных для конкретных приложений быстро разрешимых подклассов рассматриваемых трудных задач.

Обзор реальных приложений в обсуждаемой нами области показал, что одним из работоспособных здесь подходов могут оказаться так называемые "структурные" методы - поиск практически приемлемых (не только оптимальных, но и квази-оптимальных) решений в рамках так называемых Case Based Reasoning и ряда других формализаций рассуждений в духе "здравого смысла" (см., например, [*Josephson 94, Gabbay 96*] и др.). Используемая тут техника может быть представлена следующей схемой:

- опыт решения конкретных задач оптимизации аккумулируется в Библиотеку Прецедентов (описаний *ситуаций* и принимавшихся в них *решений*);
- описание новой *ситуации* сопоставляется с описаниями "похожих на нее" *ситуаций* в Библиотеке Прецедентов, причем
- *решения*, принимавшиеся в "похожих" *ситуациях* используются для генерации *решения* в текущей анализируемой *ситуации*.

Фактически, сведения о подходящих решениях-прецедентах используются как подсказка (стартовая точка), с которой может быть начат процесс оптимизации текущей ситуации. В большинстве случаев такая информация оказывается достаточной для поиска в режиме реального времени удовлетворительного (т.е. если и не оптимального, то хотя бы достаточно "близкого" к нему) решения.

Мы проиллюстрируем возможности представленного подхода на примере одной известной задачи оптимизации больших сетевых структур. Неформальная постановка рассматриваемой задачи сводима к следующей схеме [*Воронов 96, Забежайло 96, Забежайло 96-a, Zabezhailo 96-a*]:

Дано: - имеется набор "персонажей (т.е. действующих лиц некоторого процесса взаимодействий)", каждый из которых располагает некоторым "ресурсом" (причем в начальном состоянии процесса взаимодействия заданы конкретные объемы "ресурса", имеющиеся в распоряжении каждого из "персонажей");

- между "персонажами" назначены попарные взаимодействия, причем так, что если некоторая пара "персонажей" вовлечена в такие взаимодействия, то задан набор конкретных "связей" между рассматриваемыми персонажами. Связи определены так, что каждая из них характеризуется значением соответствующего параметра - "интенсивностью связи", а также "направлением" (сведениями о "персонаже", передающем "ресурс", и "персонаже", получающем "ресурс");

---

[9] Другими словами: по примерам пар *ситуация ==> решение*, принадлежащих и не принадлежащих частично определенному отношению ***СОСТОЯНИЯ==>РЕШЕНИЯ***, порождается процедура, позволяющая по предъявленному описанию конкретной *ситуации* подобрать подходящее *решение*, а также проверить их (*ситуации* и *решения*) принадлежность представленному выше частично определенному отношению.

[10] Известных из предыдущего опыта работы системы, заданных экспертом-"учителем" и т.п..

[11] Из-за большой размерности требующих оптимизации в практических приложениях систем и необходимых для их анализа гигантских объемов вычислений.

- процесс взаимодействия “персонажей” состоит в их взаимном обмене ресурсом в соответствии с его объемом, имеющимся у каждого из “отправителей ресурса” на текущий момент, и соответствующей “интенсивностью связи”, реализующей конкретное парное взаимодействие;
- процесс взаимодействия может регулироваться ограничениями на текущие объемы “ресурсов” у выделенных “персонажей”[12], или на порядок реализации конкретных попарных связей между “персонажами”[13];
- процесс взаимодействия стартует с заданного набора “персонажей” (вместе с конкретными значениями доступного им “ресурса”) и попарных связей между персонажами. Процесс взаимодействия заканчивается в ситуации, когда ни одна из имеющихся попарных связей между “персонажами” не может быть реализована либо из-за имеющихся ограничений[14], либо ввиду отсутствия необходимых объемов перераспределяемого ресурса у “персонажей” взаимодействия.

Требуется: организовать процесс взаимодействия (т.е., например, указать последовательность выполнения попарных взаимодействий) так, чтобы “персонажи” максимизировали объемы передаваемого в процессе взаимодействия ресурса.

Легко видеть, что к обсуждаемому классу можно отнести и следующие три важные прикладные задачи:

- оптимизация трафика в больших компьютерных сетях;
- оптимизация трафика в больших транспортных структурах (например, в транспортной системе мегаполиса и т.п.);
- оптимизация порядка выполнения банковских платежей (так называемый технический клиринг).

При этом важно заметить, что часто решение обсуждаемых нами задач должно быть получено в режиме ***реального времени***[15]. Последнее обстоятельство налагает дополнительные ограничения параметры получаемых в процессе оптимизации решений.

Если обратиться к анализу вычислительной природы этих задач, то следовало бы в первую очередь обратить внимание на следующие три обстоятельства:

- на комплексный характер оптимизации (оптимизацию со статическими и динамическими ограничениями);
- на динамический (пошаговый) характер процесса;
- на специфику вычислений в режиме реального времени (т.е. на специфику многокритериальной оптимизации, поиск так называемых квазиоптимальных решений и т.п.).

Все множество подходов, используемых для решения обсуждаемого класса прикладных задач можно условно разделить на две большие группы:

- "стандартные" подходы (т.е. те, которые реализуют прямой поиск экстремума оптимизируемого функционала качества), и
- "нестандартные" подходы, ориентированные на поиск приближенных решений (в том числе - с использованием тех или иных эвристических методов).

Говоря о первой группе подходов (“стандартные” методы), следует учитывать известные трудности, которые приходится преодолевать при поиске точного оптимума соответствующих функционалов качества. Наиболее существенную роль здесь играют

а) размерность задачи,

б) конкретный способ организации численной оптимизации в условиях невозможности получить точное решение в аналитическом виде,

---

[12] Например, часто предполагается, что в результате процесса взаимодействия текущий объем ресурса, которым располагает каждый из “персонажей”, не может оказаться ниже наперед заданного “порога”, и т.п..

[13] Например, часть связей может быть объявлена приоритетными, что означает их выполнение (а значит - и перераспределение соответствующих ресурсов) в первую очередь.

[14] Например, из-за требований, обсуждавшихся выше.

[15] Это вытекает, прежде всего, из необходимости управлять соответствующими системами - транспортными потоками, компьютерными сетями и др. - в режиме реального времени.

в) рост объемов вычислений, порождаемый проблемой "хорошей" сходимости соответствующих численных методов, и др..

Рассмотрим чуть подробнее каждое из только что сформулированных обстоятельств.

Размерность задачи. Переборный характер поиска решений для рассматриваемого нами класса прикладных проблем отражается в высокой комбинаторной сложности алгоритмов, порождающих такие решения[16]. Это, в свою очередь, требует объемов вычислений, растущих недопустимо[17] быстро с ростом характерных параметров входных данных (например, числа элементов анализируемой структуры). Таким образом, в значительном количестве задач реальной практики их точное решение в рамках названного подхода (рассчитываемое в режиме реального времени) требует вычислительных ресурсов, сопоставимых с ресурсами суперЭВМ.

Численная оптимизация и проблема сходимость алгоритмов поиска решений. В условиях многокритериальной оптимизация при наличии сложных неоднородных ограничений часто оказывается невозможным получить точное аналитическое выражение для значений варьируемых параметров, доставляющих оптимум функционалу качества. В таких случаях обычно используются методы численной оптимизации. Последние же , в свою очередь имеют ряд особенностей. Наиболее значительные опасности здесь вязаны с итеративным характером большинства методов численной оптимизации: нередко итеративный алгоритм, начиная с некоторого шага[18], переходит в стадию “биений” в окрестности экстремума (что объясняется попаданием процесса вычислений в область медленной сходимости алгоритма оптимизации к искомому экстремуму). В такой ситуации даже небольшие улучшения оценок оптимизируемого функционала качества требуют, как правило, значительных дополнительных объемов (а значит и времени - !) вычислений.

П 5.2. ДСМ-рассуждения в оптимизации банковского клиринга.

Обратимся к более детальному анализу представленного нами класса прикладных задач. В качестве конкретного объекта исследования выберем задачу оптимизации банковских платежей (часто называемую также задачей оптимизации технического клиринга).

В начале и середине 90-х годов одним из наиболее тревожных процессов в экономике России стал лавинообразно нарастающий кризис взаимных неплатежей. Попытки проанализировать природу цепочек невыполняемых (по разным причинам) платежей привели специалистов по управлению экономикой к пониманию важности и актуальности задачи выявления возможных цепочек взаимозачета платежных обязательств.

Решение задачи оптимизации взаимозачета платежных обязательств позволяет:

- минимизировать объемы “реальных” обязательств, сводя их (в идеальном случае) к так называемым чистым нетто-позициям;
- получить детальное представления о структуре взаимных платежей (например, дать представления о минимальных кредитных ресурсах, необходимых для “развязки” цепочек взаимных задолженностей, а также об оптимальном адресном распределении таких кредитных ресурсов и др.);
- оптимизировать использование различного рода платежных обязательств (векселей и др.) в процессе осуществления взаимозачета неплатежей;
- понизить уровни системных рисков в платежной системе (за счет оптимизации порядка выполнения платежей и снижения суммарного объема средств, находящихся в текущем обороте, а также задействования ряда других факторов управления рисками).

---

[16] Известно (см., например, уже цитированную выше монографию [*Гэри 82*] и др.) что в задачах обсуждаемого типа ряд соответствующих комбинаторных проблем попадает в классы так называемых **NP**-полных или **#P**-полных (перечислительно полных) проблем, т.е. требуют для своего решения в общем случае экспоненциально сложных алгоритмов [*Гэри 82*].

[17] Например, экспоненциально быстро.

[18] Как правило, эффект возникает с момента, когда размер “продвижения к оптимуму” на конкретном шаге становится сопоставимым с параметрами чувствительности алгоритма.

Особенностью задачи управления порядком прохождения платежей (и, в том числе, оптимизации взаимозачетов платежных обязательств) является ее доказуемо "трудный" комбинаторный характер (Так, например, ряд конкретных оптимизационных подзадач здесь не могут быть решены, грубо говоря, быстрее, чем полным перебором возможных вариантов. Таким образом, в практических ситуациях порождение точных решений для них в режиме реального времени *н е в о з м о ж н о*). Именно этим обусловлен ряд известных проблем, с которыми сталкиваются в своей практической деятельности, например, клиринговые палаты и другие учреждения платежной системы страны, призванные выполнять известные функции по управлению потоками взаимных платежей.

Предметом нашего детального рассмотрения будет нижеследующая технология (см., например, [*Воронов 96*] и др.):

Банки-участники клиринговых расчетов в течение некоторого заранее оговоренного промежутка времени передают свои платежные документы в клиринговый центр. По истечении этого времени проводится предварительный клиринговый сеанс, в процессе которого рассчитываются чистые позиции[19] *каждого банка-участника по результатам клиринга. Далее отводится некоторое время для урегулирования чистых дебетовых*[20] *позиций. Урегулирование своих чистых дебетовых позиций в общем случае банки-участники могут производить как за счет средств, полученных на рынке межбанковских краткосрочных кредитов, так и за счет средств, полученных в качестве дневных ссуд в Центральном Банке. В случае если хотя бы одному банку-участнику не удается урегулировать свою чистую дебетовую позицию, предварительный клиринговый сеанс повторяется. При этом все платежные документы банков-участников, которые не смогли урегулировать свои чистые дебетовые позиции, а также все платежные документы в пользу этих банков-участников изымаются и в дальнейших клиринговых расчетах не принимаются во внимание.*

Прежде чем перейти к обсуждению комбинаторной природы рассматриваемой нами задачи оптимизации клиринговых расчетов приведем пример точной (формализованной) постановки анализируемой нами проблемы[21]

Исходные данные:

- Список $\{\mathbf{P_1}, \ldots, \mathbf{P_n}\}$ участников клиринга, где каждый из участников $\mathbf{P_i}$ характеризуется списком

$$\{\mathbf{T_{i,1}}^{+}, \ldots, \mathbf{T_{i,m}}^{+}, \mathbf{T_{i,1}}^{-}, \ldots, \mathbf{T_{i,m}}^{-}\}$$

  необходимых ему платежей. При этом каждый элемент $\mathbf{T_{i,j}}^{\alpha}$ имеет вид $\mathbf{t}^{\alpha}_{\mathbf{i,r,s}}$, где $\alpha \in \{+,-\}$, r есть номер платежа участника $\mathbf{P_i}$ участнику $\mathbf{P_S}$ в случае, когда α = +, или номер платежа участника $\mathbf{P_S}$ участнику $\mathbf{P_i}$, когда α = -.
- Список $\underline{\boldsymbol{Pr}} = \{\boldsymbol{Pr_1}, \ldots, \boldsymbol{Pr_l}\}$ списков $\boldsymbol{Pr_i}$, характеризующих приоритеты на множестве платежей. При этом:

$$(1)\ \boldsymbol{Pr_i} \subseteq \bigcup_{i=1}^{n} \{\mathbf{T_{i,1}}^{+}, \ldots, \mathbf{T_{i,m}}^{+}, \mathbf{T_{i,1}}^{-}, \ldots, \mathbf{T_{i,m}}^{-}\}$$

$$(2)\ \boldsymbol{Pr_i} \cap \boldsymbol{Pr_j} = \varnothing \quad \text{при } \mathbf{i} \neq \mathbf{j}\ .$$

  (Значение индекса **i** в имени $\boldsymbol{Pr_i}$ есть номер приоритета, присвоенного соответствующему множеству платежей).

---

[19] Т.е. сумма остатка на текущем счете после исполнения всех платежных распоряжений.

[20] Т.е. сумм "перерасходованных" средств.

[21] В специальной литературе эту процедуру часто называют построением комбинаторной интерпретации соответствующей содержательной (неформализованной) задачи.

- Список $\{\mathbf{S_1}, \ldots, \mathbf{S_n}\}$ сумм $\mathbf{S_i}$, которыми в исходный момент времени каждый участник $\mathbf{P_i}$ располагает на своем корреспондентском счете. При этом каждая из таких сумм по крайней мере не меньше, чем некоторый фиксированный остаток $\mathbf{S_0}$:

$$\forall \mathbf{i}\ [(\mathbf{i} \in \{1,\ldots,n\}) \supset (\mathbf{S_i} > \mathbf{S_0})].$$

- Исходные ограничения (***Constr***):

  а) в процессе клиринга недопустима ситуация, когда на заключительном шаге взаимных платежей на счете хотя бы одного из участников клиринга $\mathbf{P_i}$ текущий остаток $\mathbf{S_{i0}}$ окажется меньше заранее заданной величины $\mathbf{S^*_{i0}}$.

  б) в процессе клиринга недопустима ситуация, когда в результате вычислений хотя бы один из приоритетных платежей окажется нереализованным, в то время как хотя бы один из неприоритетных платежей одного из участников будет реализован.

Формальная постановка задачи.

1. Сопоставим предъявленным исходным данным процесса клиринга граф $\underline{\boldsymbol{Cl}} = \langle \underline{\boldsymbol{P}}, \underline{\boldsymbol{T}} \rangle$, множество вершин которого $\underline{\boldsymbol{P}} = \{\mathbf{P_1}, \ldots, \mathbf{P_n}\}$ порождается множеством участников клиринга, а каждый элемент множества ребер $\underline{\boldsymbol{T}}$ порожден соответствующим платежом из множества всех необходимых платежей:

$$\bigcup_{i=1}^{n} \{\mathbf{T_{i,1}}^{+}, \ldots, \mathbf{T_{i,m}}^{+}, \mathbf{T_{i,1}}^{-}, \ldots, \mathbf{T_{i,m}}^{-}\}$$

(при этом направленность ребра определяется направленностью соответствующего платежа: от плательщика к получателю).

2. Каждой вершине $\mathbf{P_i}$ из множества $\underline{\boldsymbol{P}}$ сопоставим неотрицательное число $\mathbf{s_i}$ - ее вес, который в начальный момент времени определяется соотношениями:

$$\mathbf{s_i} = \mathbf{S_i}\ , \text{ где } \mathbf{i} \in \{1, \ldots, n\}.$$

3. Каждому ребру $\mathbf{T_j}$ из множества $\underline{\boldsymbol{T}}$ сопоставим неотрицательное число $\mathbf{t_j}$ - его пропускную способность, - и сделаем это так , что величина $\mathbf{t_j}$ в точности соответствует размеру (сумме) конкретного платежа, представленного соответствующим ребром.

4. Таким образом, мы имеем раскрашенный граф

$$\underline{\boldsymbol{Cl^*}} = \langle \underline{\boldsymbol{P^*}}, \underline{\boldsymbol{T^*}} \rangle,$$

где $\underline{\boldsymbol{P^*}} = \langle \underline{\boldsymbol{P}}, \mathbf{s} \rangle$ и $\underline{\boldsymbol{T^*}} = \langle \underline{\boldsymbol{T}}, \boldsymbol{t} \rangle$, а

$$\mathbf{s} = \{\mathbf{s_i}\} \text{ и } \boldsymbol{t} = \{\mathbf{t_j}\}.$$

5. Процедуру ***Trans***, перерабатывающую графы вида $\underline{\boldsymbol{Cl^*}}$ в графы аналогичного вида, определим следующим образом:

формула $\underline{\boldsymbol{Trans}}\ (\underline{\boldsymbol{Cl^*}}\mathbf{a}) = \underline{\boldsymbol{Cl^*}}_{\mathbf{B}}$ означает, что

5.1. В графе выбрано некоторое ребро $\mathbf{T_{ij}}$, соединяющее вершину $\mathbf{P_i}$, имеющую вес $\mathbf{s_i}$, с вершиной $\mathbf{P_j}$, имеющей вес $\mathbf{s_j}$, при этом пропускная способность $\mathbf{t_{ij}}$ удовлетворяет ограничению $\mathbf{s_i} > \mathbf{t_{ij}}$.

5.2. Граф $\underline{\boldsymbol{Cl^*}}_{\mathbf{B}}$ получен из графа $\underline{\boldsymbol{Cl^*}}\mathbf{a}$ следующими заменами:

а) $\mathbf{s_{iB}} = \mathbf{s_{ia}} - \mathbf{t_{ij}}$,

б) $\mathbf{s_{jB}} = \mathbf{s_{ja}} + \mathbf{t_{ij}}$,

в) ребро $\mathbf{T_{ij}}$ в графе $\underline{\boldsymbol{Cl^*}}_{\mathbf{B}}$ отсутствует.

6. Будем говорить, что граф $\underline{\boldsymbol{Cl^*}}_{\mathbf{B}}$ локально подобен графу $\underline{\boldsymbol{Cl^*}}\mathbf{a}$, если $\underline{\boldsymbol{Cl^*}}_{\mathbf{B}} = \underline{\boldsymbol{Trans}}(\underline{\boldsymbol{Cl^*}}\mathbf{a})$. Будем говорить, что граф $\underline{\boldsymbol{Cl^*}}\mathbf{A}$ подобен графу $\underline{\boldsymbol{Cl^*}}\mathbf{B}$, если существует цепочка графов $\boldsymbol{G}$

$$\boldsymbol{G} = \langle \underline{\boldsymbol{Cl^*}}\mathbf{qA}\ ,\ \underline{\boldsymbol{Cl^*}}\mathbf{q1}\ ,\ \ \ldots\ ,\ \underline{\boldsymbol{Cl^*}}\mathbf{q(C\text{-}1)}\ ,\ \underline{\boldsymbol{Cl^*}}\mathbf{qC} \rangle$$

таких, что

$$\underline{\boldsymbol{Cl^*}}\mathbf{qA} = \underline{\boldsymbol{Cl^*}}\mathbf{A}\ ,\ \underline{\boldsymbol{Cl^*}}\mathbf{qC} = \underline{\boldsymbol{Cl^*}}\mathbf{B}\ ,$$

и для всякого $\mathbf{i} \in \{1,...,C\}$ граф $\underline{\boldsymbol{Cl^*_{qi}}}$ локально подобен графу $\underline{\boldsymbol{Cl^*_{q(i+1)}}}$.

7. Будем считать, что для каждого подобного заданному графу $\underline{\boldsymbol{Cl^*}}$ графа $\underline{\boldsymbol{Cl^*_i}}$ множество $\boldsymbol{Trf}(\underline{\boldsymbol{Cl^*_i}})$ содержит все те ребра $\mathbf{t_{ij}} \in \underline{\boldsymbol{T^*}}$ исходного графа $\underline{\boldsymbol{Cl^*}}$, которые были отброшены в процессе порождения подобного графа $\underline{\boldsymbol{Cl^*_i}}$ применением процедуры $\underline{\boldsymbol{Trans}}$.

8. Пусть $\underline{\boldsymbol{G}} = \underline{\boldsymbol{G}}(\underline{\boldsymbol{Cl^*}})$ есть все множество графов, подобных заданному графу $\underline{\boldsymbol{Cl^*}}$.

9. Для каждого графа

$$\underline{\boldsymbol{Cl^*_i}} \in \underline{\boldsymbol{G}}(\underline{\boldsymbol{Cl^*}})$$

вычислим величины $\boldsymbol{Sum_V}(\underline{\boldsymbol{Cl^*_i}})$ и $\boldsymbol{Sum_N}(\underline{\boldsymbol{Cl^*_i}})$, удовлетворяющие соотношениям:

$$\boldsymbol{Sum_V}(\underline{\boldsymbol{Cl^*_i}}) = \sum_{Trf} (\mathbf{t_{ij}}), \text{ и}$$

$$\boldsymbol{Sum_N}(\underline{\boldsymbol{Cl^*_i}}) = |\, \boldsymbol{Trf}\,(\underline{\boldsymbol{Cl^*_i}})\,|,$$

т.е. представляющие собой, соответственно, сумму весов дуг, входящих в множество $\boldsymbol{Trf}(\underline{\boldsymbol{Cl^*_i}})$, и число элементов в множестве $\boldsymbol{Trf}(\underline{\boldsymbol{Cl^*_i}})$.

Теперь мы имеем возможность дать формальную постановку задачи об оптимизации клиринга.

Задача ***Clir***:

Дано:

раскрашенный граф

$$\underline{\boldsymbol{Cl^*}} = \langle \underline{\boldsymbol{P^*}}, \underline{\boldsymbol{T^*}} \rangle,$$

где $\underline{\boldsymbol{P^*}} = \langle \underline{\boldsymbol{P}}, \mathbf{s} \rangle$ и $\underline{\boldsymbol{T^*}} = \langle \underline{\boldsymbol{T}}, \boldsymbol{t} \rangle$, а

$$\mathbf{s} = \{ \mathbf{s_i} \} \text{ и } \boldsymbol{t} = \{ \mathbf{t_j} \}.$$

Найти:

в множестве $\underline{\boldsymbol{G}}$ (графов, подобных исходно заданному) элемент

$$\underline{\boldsymbol{Cl^*_0}} \in \underline{\boldsymbol{G}}(\underline{\boldsymbol{Cl^*}}),$$

удовлетворяющий условиям

$$(1)\ \boldsymbol{Sum_V}(\underline{\boldsymbol{Cl^*_0}}) = \max_{\underline{\boldsymbol{Cl^*_0}} \in \underline{\boldsymbol{G}}(\underline{\boldsymbol{Cl^*}})} (\boldsymbol{Sum_V}(\underline{\boldsymbol{Cl^*_i}}))$$

или

$$(2)\ \boldsymbol{Sum_N}(\underline{\boldsymbol{Cl^*_0}}) = \max_{\underline{\boldsymbol{Cl^*_0}} \in \underline{\boldsymbol{G}}(\underline{\boldsymbol{Cl^*}})} (\boldsymbol{Sum_N}(\underline{\boldsymbol{Cl^*_i}})),$$

не нарушив в процессе вычислений ограничения ***Constr.***

***

Перейдем к анализу комбинаторной природы задачи оптимизации клиринга. Прежде всего уточним понимание используемых далее терминов.

П 5.2.1. Определение

Пусть граф $\underline{\boldsymbol{Cl^*_{fin}}} \in \underline{\boldsymbol{G}}(\underline{\boldsymbol{Cl^*}})$ получен из исходного графа $\underline{\boldsymbol{Cl^*}}$ применением процедуры $\underline{\boldsymbol{Trans}}$. Будем называть граф $\underline{\boldsymbol{Cl^*_{fin}}}$ тупиком при оптимизации $\underline{\boldsymbol{Cl^*}}$, если имеет место

$$\underline{\boldsymbol{Trans}}\,(\underline{\boldsymbol{Cl^*_{fin}}}) = \underline{\boldsymbol{Cl^*_{fin}}},$$

т.е. является неподвижной точкой оператора $\underline{\boldsymbol{Trans}}$.

o

П 5.2.2. Определение

Будем называть стратегией ***BT*** (*все тупики*) следующий алгоритм оптимизации:

Шаг 1. Располагая исходным графом $\underline{\boldsymbol{Cl^*}}$, построим множество $\underline{\boldsymbol{G_{fin}}}(\underline{\boldsymbol{Cl^*}})$ всех его тупиков.

Шаг 2. Для каждого элемента $\underline{\boldsymbol{Cl^*_{fin}}}$ множества $\underline{\boldsymbol{G_{fin}}}(\underline{\boldsymbol{Cl^*}})$ вычислим значение соответствующего ($\boldsymbol{Sum_V}(\underline{\boldsymbol{Cl^*_{fin}}})$ или $\boldsymbol{Sum_N}(\underline{\boldsymbol{Cl^*_{fin}}})$) функционала качества.

Шаг 3. На множестве $\underline{\boldsymbol{G_{fin}}}(\underline{\boldsymbol{Cl^*}})$ найдем элемент $\underline{\boldsymbol{Cl^*_0}}$, удовлетворяющий соответствующему критерию оптимизации

$$(1)\ \boldsymbol{Sum_V}(\underline{\boldsymbol{Cl^*_0}}) = \max_{\underline{\boldsymbol{Cl^*_0}} \in \underline{\boldsymbol{G}}(\underline{\boldsymbol{Cl^*}})} (\boldsymbol{Sum_V}(\underline{\boldsymbol{Cl^*_i}}))$$

или

(2) $\mathbf{Sum_N(\underline{Cl^*_0}) = \max_{\underline{Cl^*_0} \in \underline{G}(\underline{Cl^*})} (Sum_N(\underline{Cl^*_i}))}$.

Шаг 4. СТОП.

o

П 5.2.3. Определение

1. Будем называть задачей $\mathbf{Clir_N}$ задачу **Clir** об оптимизации графа (произвольного исходно заданного графа - !) клиринга $\underline{Cl^*}$ с критерием

(2) $\mathbf{Sum_N(\underline{Cl^*_0}) = \max_{\underline{Cl^*_0} \in \underline{G}(\underline{Cl^*})} (Sum_N(\underline{Cl^*_i}))}$.

2. Будем называть задачей $\mathbf{Clir_V}$ задачу **Clir** об оптимизации графа клиринга $\underline{Cl^*}$ с критерием

(1) $\mathbf{Sum_V(\underline{Cl^*_0}) = \max_{\underline{Cl^*_0} \in \underline{G}(\underline{Cl^*})} (Sum_V(\underline{Cl^*_i}))}$

o

Представление о взаимной зависимости задач $\mathbf{Clir_N}$ и $\mathbf{Clir_V}$ дает

П 5.2.4. Лемма

Вычислительная сложность задачи $\mathbf{Clir_V}$ мажорируется снизу вычислительной сложностью задачи $\mathbf{Clir_N}$.

Доказательство.

Положив в задаче $\mathbf{Clir_V}$ все $\mathbf{t_j}$ (для соответствующих ребер $\mathbf{T_j}$ из $\underline{T^*} = <\underline{T},t>$) равными 1, мы получим задачу $\mathbf{Clir_N}$. Таким образом, задача $\mathbf{Clir_N}$ (как переборная проблема) есть частный случай задачи $\mathbf{Clir_V}$.

o

Таким образом, мы располагаем возможностью оценить сложность рассматриваемой нами вычислительной проблемы оптимизации клиринга, рассмотрев в качестве базовых

- задачу $\mathbf{Clir_N}$

и

- стратегию поиска оптимума ***BT*** (*все тупики)*.

Определение П 5.2.5.

Будем называть задачей ***ПТ*** (*поиск тупика*) следующую комбинаторную задачу:

Дано:

раскрашенный граф $\underline{Cl^*} = <\underline{P^*},\underline{T^*}>$, где $\underline{P^*} = <\underline{P},S>$ и $\underline{T^*} = <\underline{T},t>$, а

$S = \{ s_i \}$ и $t = \{ t_j \}$.

Найти:

в множестве $\underline{G}$ (графов, подобных исходно заданному) элемент

$\underline{Cl^*}fin \in \underline{G}(\underline{Cl^*})$,

удовлетворяющий условию

$\underline{Trans}\ (\underline{Cl^*}fin) = \underline{Cl^*}fin$,

т.е. найти в множестве $\underline{G}$ тупик в смысле **Определения 7.3.2.1**.

o

П 5.2.6. Лемма

Задача ***ПТ*** (*поиск тупика*) принадлежит классу **P** полиномиально разрешимых (см. например, [*Гэри 82*] и др) задач.

Доказательство.

Легко видеть, что каждый тупик может быть построен применением процедуры $\underline{Trans}$, стартующей с произвольной вершины графа $\underline{Cl^*} = <\underline{P^*},\underline{T^*}>$, просмотром ровно один раз множества $\underline{T^*}$ всех ребер графа $\underline{Cl^*}$, т.е. с помощью алгоритма, число шагов которого линейно зависит от характерных размеров графа $\underline{Cl^*}$.

o

Окончательное представление о свойствах анализируемой задачи ***ПТ*** дает

П 5.2.7. Теорема

Задача ***ПТ*** (*поиск тупика*) принадлежит к классу **#P** перечислительно полных задач.

Доказательство.

Известно (см., например, [*Valiant 79*] и др.), что задача ***BF-MON*** о выполнимости монотонной булевской функции принадлежит классу **#P** (перечислительно полна). Мы покажем, что задача ***BF-MON*** полиномиально сводима к частному случаю задачи ***ПТ*** (*поиск тупика*), которая[22] (по **Лемме 7.3.2.6** – см. ниже) также лежит в классе **P**.

Прежде чем продемонстрировать механизм сводимости, договоримся об удобной терминологии, дав

П 5.2.7.1. Определение

Пусть предъявлена монотонная булевская функция

$$\Phi(x_1, \ldots, x_n)$$

вида

$$\Phi(x_1, \ldots, x_n) = (x_1 \vee x_{i1}) \,\&\, \ldots \,\&\, (x_{im} \vee x_n).$$

Входящую в нее дизъюнкцию $Q_j(x_{j1},x_{j2})=(x_{j1} \vee x_{j2})$ мы будем называть вырожденной, если в **Ф** найдется такая отличная от Q_j дизъюнкция $Q_{j'}=Q_{j'}(x_{j1},x_{j2})$, что среди переменных x_{j1},x_{j2} найдется совпадающая либо с x_{j1}, либо с x_{j2}. Повторяющуюся в дизъюнкции $Q_{j'}$ переменную мы будем называть *порождающей вырождение* соответствующей дизъюнкции Q_j.

o

Перейдем к описанию процедуры сводимости и анализу ее свойств.

1. Преобразование ***REDUCT***, позволяющее по предъявленной монотонной функции **Ф**

(1) $\Phi(x_1, \ldots, x_n) = (x_1 \vee x_{i1}) \,\&\, \ldots \,\&\, (x_{im} \vee x_n)$

построить граф ***Cl*Ф*** (формализующий соответствующий частный случай задачи клиринга ***Clir***), мы определим следующим образом:

- преобразование ***REDUCT*** строит по функции **Ф** граф ***Cl*Ф***, перебирая дизъюнкции Q_i в представлении **Ф**, стартовав с первой - Q_1 и закончив свою работу анализом последней - Q_m;
- перед обработкой первой дизъюнкции Q_1 граф ***Cl*Ф*** представляет собой пустое множество вершин и пустое множество ребер;
- получив на вход текущую дизъюнкцию $Q_i = (x_{i1} \vee x_{i2})$, процедура ***REDUCT*** делает следующие шаги:

- в граф ***Cl*Ф*** добавляются четыре новые вершины
  - q_i (сопоставлена текущей дизъюнкции Q_i),
  - q_{i1} (сопоставлена переменной x_{i1}),
  - q_{i2} (сопоставлена переменной x_{i2}),
  - q'_i (сопоставлена текущей дизъюнкции Q_i, представляя одновременную выполнимость переменных x_{i1} и x_{i2}),

[22] В случае поиска какого-либо одного решения (т.е. какого-либо одного тупика).

при этом вершины $\mathbf{q_{i1}}$ и $\mathbf{q_{i2}}$ добавляются лишь в том случае, когда мы рассматриваем первое вхождение соответствующих переменных в конъюнкцию (1); а в случае повторных вхождений мы будем использовать для последующих построений уже внесенные в граф вершины, сопоставленные соответствующим переменным;

- в граф $\underline{\boldsymbol{Cl^*\Phi}}$ добавляются шесть новых ребер
  - $\mathbf{T_{i1}(q_i,q_{i1})}$, которое направлено от вершины $\mathbf{q_i}$ к вершине $\mathbf{q_{i1}}$),
  - $\mathbf{T_{i2}(q_{i1},q_i)}$, которое направлено от вершины $\mathbf{q_{i1}}$ назад к вершине $\mathbf{q_i}$),
  - $\mathbf{T_{i3}(q_i,q_{i2})}$, которое направлено от вершины $\mathbf{q_i}$ к вершине $\mathbf{q_{i2}}$),
  - $\mathbf{T_{i4}(q_{i2},q_i)}$, которое направлено от вершины $\mathbf{q_{i2}}$ назад к вершине $\mathbf{q_i}$),
  - $\mathbf{T_{i5}(q'_i,q_{i1})}$, которое направлено от вершины $\mathbf{q'_i}$ к вершине $\mathbf{q_{i1}}$),
  - $\mathbf{T_{i6}(q'_i,q_{i2})}$, которое направлено от вершины $\mathbf{q'_i}$ к вершине $\mathbf{q_{i2}}$);
- если дизъюнкция $\mathbf{Q_i}$ невырождена, то добавляем ребро
  - $\mathbf{T_{i7}(q_i,q'_i)}$, которое направлено от вершины $\mathbf{q_i}$ к вершине $\mathbf{q'_i}$), после чего переходим к следующей дизъюнкции $\mathbf{Q_{i+1}}$ как к текущей;
- если дизъюнкция $\mathbf{Q_i}$ вырождена, то (последовательным просмотром вершин, порождающих вырождение) для каждой переменной $\mathbf{x_{j1}}$[23], порождающей вырождение дизъюнкции $\mathbf{Q_i}$, осуществляется нижеследующее пополнение графа $\underline{\boldsymbol{Cl^*\Phi}}$:

  - в него вносятся еще три вершины $\mathbf{q_j}$, $\mathbf{q_{j2}}$ и $\mathbf{q'_j}$ (по правилам порождения вершин вида $\mathbf{q_i}$, $\mathbf{q_{i1}}$, $\mathbf{q_{i2}}$ и $\mathbf{q'_i}$, представленным выше)[24] и еще шесть ребер $\mathbf{T_{j8}}$, ..., $\mathbf{T_{j13}}$ (также по правилам, представленным выше для ребер $\mathbf{T_{i1}}$, ... , $\mathbf{T_{i6}}$);

  - рядом с вершиной $\mathbf{q_{j2}}$ достраиваются еще две вершины $\mathbf{q^*_{j2}}$ и $\mathbf{q^{**}_{j2}}$ вместе с ребрами
    - $\mathbf{T_{i14}(q_{j2},q^*_{j2})}$, направленным от вершины $\mathbf{q_{j2}}$ к вершине $\mathbf{q^*_{j2}}$,
    - $\mathbf{T_{i15}(q^*_{j2},q^{**}_{j2})}$, направленным от вершины $\mathbf{q^*_{j2}}$ к вершине $\mathbf{q^{**}_{j2}}$,
    - $\mathbf{T_{i16}(q^{**}_{j2},q'_i)}$, направленным от вершины $\mathbf{q^{**}_{j2}}$ к вершине $\mathbf{q'_i}$,

    (при этом, если вершина $\mathbf{q_{j2}}$ уже была построена ранее в ходе обработки какого-либо другого вырождения, то новую вершину $\mathbf{q^*_{j2}}$ вместе с новым ребром $\mathbf{T_{i14}(q_{j2},q^*_{j2})}$ порождать не будем, а воспользуемся ранее построенными при обработке другого вырождения);

  - рядом с вершиной $\mathbf{q_{i2}}$ (в полном соответствии с построением только что описанной "окрестности" вершины $\mathbf{q_{j2}}$) добавляем в граф $\underline{\boldsymbol{Cl^*\Phi}}$ вершины $\mathbf{q^*_{i2}}$ и $\mathbf{q^{**}_{i2}}$ вместе с ребрами
    - $\mathbf{T_{i17}(q_{i2},q^*_{i2})}$, направленным от вершины $\mathbf{q_{i2}}$ к вершине $\mathbf{q^*_{i2}}$,
    - $\mathbf{T_{i18}(q^*_{i2},q^{**}_{i2})}$, направленным от вершины $\mathbf{q^*_{i2}}$ к вершине $\mathbf{q^{**}_{i2}}$,
    - $\mathbf{T_{i19}(q^{**}_{i2},q'_j)}$, направленным от вершины $\mathbf{q^{**}_{i2}}$ к вершине $\mathbf{q'_j}$,
  - рядом с вершиной $\mathbf{q_i}$ добавляем в граф $\underline{\boldsymbol{Cl^*\Phi}}$ две вершины $\mathbf{q^*_i}$ и $\mathbf{q^{**}_i}$ в месте с ребрами
    - $\mathbf{T_{i20}(q_i,q^*_i)}$, направленным от вершины $\mathbf{q_i}$ к вершине $\mathbf{q^*_i}$,
    - $\mathbf{T_{i21}(q^*_i,q^{**}_i)}$, направленным от вершины $\mathbf{q^*_i}$ к вершине $\mathbf{q^{**}_i}$,
    - $\mathbf{T_{i22}(q^{**}_{i2},q'_i)}$, направленным от вершины $\mathbf{q^{**}_{i2}}$ к вершине $\mathbf{q'_j}$,
    - $\mathbf{T_{i23}(q^{**}_{i2},q^{**}_{j2})}$, направленным от вершины $\mathbf{q^{**}_{i2}}$ к вершине $\mathbf{q^{**}_{j2}}$;

---

[23] Здесь ради упрощения обозначений мы будем считать, что переменная $\mathbf{x_{j1}}$ в дизъюнкции $\mathbf{Q_j} = (\mathbf{x_{j1}} \vee \mathbf{x_{j2}})$ совпадает, например, с переменной $\mathbf{x_{i1}}$ в дизъюнкции $\mathbf{Q_i} = (\mathbf{x_{i1}} \vee \mathbf{x_{i2}})$, т.е. $\mathbf{x_{j1}} = \mathbf{x_{i1}}$.

[24] При этом вершина $\mathbf{q_{j1}}$ уже порождена при обработке дизъюнкции $\mathbf{Q_i}$.

при этом если вершина $\mathbf{q_i}$ уже участвовала ранее в обработке какого-либо вырождения, то вершину $\mathbf{q^*_i}$ и ребро $\mathbf{T_{i20}}$ мы заново не порождаем, воспользовавшись их аналогами, порожденными ранее при обработке соответствующего предыдущего вырождения;

- рядом с вершиной $\mathbf{q_j}$ (в полном соответствии с построением только что описанной "окрестности" вершины $\mathbf{q_i}$) добавляем в граф ***Cl*Φ*** две вершины $\mathbf{q^*_j}$ и $\mathbf{q^{**}_j}$ в месте с ребрами
  - $\mathbf{T_{i24}(q_j,q^*_j)}$, направленным от вершины $\mathbf{q_j}$ к вершине $\mathbf{q^*_j}$,
  - $\mathbf{T_{i25}(q^*_j,q^{**}_j)}$, направленным от вершины $\mathbf{q^*_j}$ к вершине $\mathbf{q^{**}_j}$,
  - $\mathbf{T_{i26}(q^{**}_j,q'_j)}$, направленным от вершины $\mathbf{q^{**}_j}$ к вершине $\mathbf{q'_j}$,
  - $\mathbf{T_{i27}(q^{**}_j,q^{**}_{i2})}$, направленным от вершины $\mathbf{q^{**}_j}$ к вершине $\mathbf{q^{**}_{i2}}$;

- рядом с вершиной $\mathbf{q'_i}$ добавляем в граф ***Cl*Φ*** две вершины $\mathbf{q'^*_i}$ и $\mathbf{q'^{**}_i}$ в месте с ребрами
  - $\mathbf{T_{i28}(q'_i,q'^*_i)}$, направленным от вершины $\mathbf{q'_i}$ к вершине $\mathbf{q'^*_i}$,
  - $\mathbf{T_{i29}(q'^*_i,q'^{**}_i)}$, направленным от вершины $\mathbf{q'^*_i}$ к вершине $\mathbf{q'^{**}_i}$,
  - $\mathbf{T_{i30}(q'^{**}_i,q_i)}$, направленным от вершины $\mathbf{q'^{**}_i}$ к вершине $\mathbf{q_i}$,
  - $\mathbf{T_{i31}(q'^{**}_i,q_j)}$, направленным от вершины $\mathbf{q'^{**}_i}$ к вершине $\mathbf{q_j}$;

  при этом если вершина $\mathbf{q'_i}$ уже участвовала ранее в обработке какого-либо вырождения, то вершину $\mathbf{q'^*_i}$ и ребро $\mathbf{T_{i28}}$ мы заново не порождаем, воспользовавшись их аналогами, порожденными ранее при обработке соответствующего предыдущего вырождения;

- рядом с вершиной $\mathbf{q'_j}$ (в полном соответствии с построением только что описанной "окрестности" вершины $\mathbf{q'_i}$) добавляем в граф ***Cl*Φ*** две вершины $\mathbf{q'^*_j}$ и $\mathbf{q'^{**}_j}$ в месте с ребрами
  - $\mathbf{T_{i32}(q'_j,q'^*_j)}$, направленным от вершины $\mathbf{q'_j}$ к вершине $\mathbf{q'^*_j}$,
  - $\mathbf{T_{i33}(q'^*_j,q'^{**}_j)}$, направленным от вершины $\mathbf{q'^*_j}$ к вершине $\mathbf{q'^{**}_j}$,
  - $\mathbf{T_{i34}(q'^{**}_j,q_j)}$, направленным от вершины $\mathbf{q'^{**}_j}$ к вершине $\mathbf{q_j}$,
  - $\mathbf{T_{i35}(q'^{**}_j,q_i)}$, направленным от вершины $\mathbf{q'^{**}_j}$ к вершине $\mathbf{q_i}$;

- после чего переходим к обработке новой текущей дизъюнкции $\mathbf{Q_{i+1}}$, пока не пройдем все $\mathbf{m}$ компонент $\mathbf{\Phi}$;

- "пропускные способности" ребер задаются следующими соотношениями:

(2) $\mathbf{t_{i1} = t_{i3} = t_{i8} = t_{i10} = t_{i30} = t_{i34} = 6}$

(3) $\mathbf{t_{i2} = t_{i4} = t_{i5} = t_{i6} = t_{i9} = t_{i11} = t_{i12} = t_{i13} = t_{i14} = t_{i17} = t_{i23} = t_{i27} = t_{i28} = t_{i32} = 1}$

(4) $\mathbf{t_{i16} = t_{i19} = t_{i31} = t_{i35} = 11}$

(5) $\mathbf{t_{i7} = t_{i20} = t_{i24} = 5}$

(6) $\mathbf{t_{i22} = t_{i26} = 4}$

(7) $\mathbf{t_{i15} = 11 \times V_1 - 1}$

(8) $\mathbf{t_{i18} = 11 \times V_2 - 1}$

(9) $\mathbf{t_{i21} = V_3 + 4}$

(10) $\mathbf{t_{i25} = V_4 + 4}$

(11) $\mathbf{t_{i29} = 11 \times V_3 + 6 - 1 = 11 \times V_3 + 5}$

(12) $\mathbf{t_{i33} = 11 \times V_4 + 6 - 1 = 11 \times V_4 + 5}$,

где (для каждой текущей дизъюнкции $\mathbf{Q_i}$):

1. $\mathbf{V_1}$ есть число дизъюнкций $\mathbf{Q_k}$ с вырождением (в представлении функции **Ф**), где выбранная нами (выбором текущей дизъюнкции $\mathbf{Q_i}$) переменная $\mathbf{x_{j2}}$ соседствует с порождающей соответствующее вырождение переменной $\mathbf{x_{k1}}$ в дизъюнкции $\mathbf{Q_k} = (\mathbf{x_{k1}} \vee \mathbf{x_{j2}})$,
2. $\mathbf{V_2}$ есть число дизъюнкций $\mathbf{Q_k}$ с вырождением (в представлении функции **Ф**), где выбранная нами (выбором текущей дизъюнкции $\mathbf{Q_i}$) переменная $\mathbf{x_{i1}}$ соседствует с порождающей соответствующее вырождение переменной $\mathbf{x_{k1}}$ в дизъюнкции $\mathbf{Q_k} = (\mathbf{x_{k1}} \vee \mathbf{x_{j2}})$,
3. $\mathbf{V_3}$ есть число (отличных от текущей) дизъюнкций $\mathbf{Q_k}$ в представлении функции **Ф**, в которые входит порождающие вырождение (дизъюнкции $\mathbf{Q_i}$) переменные $\mathbf{x_{i1}}$ и $\mathbf{x_{i2}}$;
4. $\mathbf{V_4}$ есть число (отличных от $\mathbf{Q_j}$, выбранной вместе с текущей $\mathbf{Q_i}$) дизъюнкций $\mathbf{Q_k}$ в представлении функции **Ф**, в которые входит порождающая вырождение (дизъюнкции $\mathbf{Q_j}$) переменная $\mathbf{x_{i1}}$;

- веса вершин графа ***Cl*Ф*** в исходном состоянии распределяются следующим образом:

- для каждой вершины $\mathbf{q_i}$ ($\mathbf{i} \in \{1,...,m\}$) имеем

$$\mathbf{s(q_i)} = \mathbf{9};$$

- для каждой вершины $\mathbf{q^*_{j2}}$ ($\mathbf{j} \in \{1,...,m\}$) имеем

$$\mathbf{s(q^*_{j2})} = \mathbf{11} \times \mathbf{V_1(q_{j2})}^{25} - \mathbf{2};$$

- для каждой вершины $\mathbf{q^*_i}$ ($\mathbf{i} \in \{1,...,m\}$) имеем

$$\mathbf{s(q^*_i)} = \mathbf{V_3(q^*_i)}^{26} - \mathbf{1};$$

- для каждой вершины $\mathbf{q'^*_i}$ ($\mathbf{i} \in \{1,...,m\}$) имеем

$$\mathbf{s(q'^*_i)} = \mathbf{11} \times \mathbf{V_3} + \mathbf{4}.$$

Описание преобразования ***REDUCT*** завершено.

2. Нетрудно видеть, что преобразование ***REDUCT*** по предъявленной функции **Ф** полиномиально быстро порождает граф ***Cl*Ф***: действительно

- число его вершин не превосходит 2m+n+4n+2n=2m+7n,
- число его ребер не превосходит 35 x m,

при этом каждая вершина и каждое ребро порождаются в результате не более чем однократного просмотра всех дизъюнкций в представлении функции **Ф**.

Таким образом можно утверждать, что справедлива

П .2.7.2. Лемма

Преобразование ***REDUCT*** принадлежит классу **P** (т.е представляет собою полиномиально сложное преобразование).

o

3. Будем называть вершины вида $\mathbf{q_i}$, $\mathbf{q^*_i}$ ($\mathbf{i} \in \{1,...,m\}$) и $\mathbf{q^*_{j2}}$ ($\mathbf{j} \in \{1,...,m\}$) графа ***Cl*Ф*** основными. Остальные вершины графа ***Cl*Ф*** мы будем называть вспомогательными. Нетрудно видеть, что старт преобразования ***Trans*** на графе ***Cl*Ф*** означает выбор одного из трех ребер, исходящих из каждой вершины вида $\mathbf{q_i}$. Также легко видеть, что при этом

---

[25] Т.е. значение величины $\mathbf{V_1}$, рассчитанной в (находящейся в соответствующей "окрестности" вершины $\mathbf{q_{j2}}$) вершине $\mathbf{q^*_{j2}}$

[26] Т.е. значением величины $\mathbf{V_3}$, рассчитанной в (находящейся в соответствующей "окрестности" вершины $\mathbf{q_i}$) вершине $\mathbf{q^*_i}$.

концы ребер вида $\mathbf{T_{i1}}$, $\mathbf{T_{i3}}$ или $\mathbf{T_{i5}}$,$\mathbf{T_{i6}}$ указывают набор вершин вида $\mathbf{q_{ij}}$, соответствующий набору соответствующих переменных $\mathbf{x_{ij}}$, выполняющим функцию **Ф**.

С другой стороны, каждый из наборов, выполняющих функцию **Ф**, порождает выбор одного из трех ребер, исходящих из каждой вершины $\mathbf{q_i}$. Таким образом, каждому из выполняющих функцию **Ф** наборов значений переменных соответствует тупик преобразования ***Trans*** на графе ***Cl*Ф***. Не составляет труда убедиться, что конкретный способ определения преобразования **Ф** гарантирует графическое совпадение всех возможных тупиков, порождаемых одним выполняющим **Ф** набором значений переменных: каждый из тупиков представляет собой суперпозицию "полных подграфов", порождаемых каждой невырожденной дизъюнкцией из **Ф** или каждой парой вида

<*вырожденная дизъюнкция* $\mathbf{Q_i}$;

дизъюнкция $\mathbf{Q_j}$, *содержащая*

переменную $\mathbf{x_{j1}}$, *которая порождает вырождение* $\mathbf{Q_i}$>.

Таким образом имеет место

П 5.2.7.3. Лемма

1. Преобразование ***REDUCT*** позволяет сопоставить каждому набору значений переменных, выполняющих функцию **Ф**, единственный тупик преобразования ***Trans*** на графе ***Cl*Ф*** = ***REDUCT***(**Ф**).
2. Преобразование ***REDUCT*** позволяет сопоставить каждому тупику преобразования ***Trans*** на графе ***Cl*Ф*** = ***REDUCT***(**Ф**) единственный набор значений переменных, выполняющий функцию **Ф**.

o

Мы завершили демонстрацию полиномиальной сводимости задачи ***BF-MON*** о выполнимости произвольной монотонной булевской функции **Ф** к задаче ***ПТ*** (*поиск тупика*) о перечислении всех тупиков (т.е. неподвижных точек) преобразования ***Trans*** на графе клиринга ***Cl****.

oo

Таким образом в организации платежей процесса клиринга возможны ситуации, когда число тупиков, требующих просмотра при поиске оптимальной последовательности осуществляемых платежей, может оказаться экспоненциально большим[27]. Содержательное представление об этой ситуации дает

П 5.2.8. Пример

Известно, что в середине 90-х годов в России действовало порядка 2000 банков. Рассмотрим модельную ситуацию, в которой между ними должны быть осуществлены платежи по представленной нами схеме клиринга. Предположим (для простоты вычислений), что просмотр каждого возможного тупика в соответствующем графе клиринга может быть осуществлен за одну “макрооперацию” (т.е. за один шаг некоторого специального алгоритма). Следствием **Теоремы П 5.2.7** является существование таких исходных ситуаций клиринга, которые порождают

$$|\mathbf{N_{туп}}| = \mathbf{O(2^{2000})}$$

тупиков. Таким образом, для просмотра множества $\mathbf{N_{туп}}$ с помощью компьютера с быстродействием **spd = 200-300** млн. операций в секунду потребуется примерно

$$\mathbf{O(2^{2000})} : (3600 \text{ x } 24 \text{ x } 365 \text{ x } \mathbf{spd}) =$$

$$= \mathbf{O(2^{10 \text{ x } 200})} : (3600 \text{ x } 24 \text{ x } 365 \text{ x } \mathbf{spd}) =$$

[27] Таким же большим, как и множество наборов, выполняющих монотонную булевскую функцию - см. **Теорему П 5.2.7**.

$$= \mathbf{O(10^{3} \text{ x } 200)} : (3600 \text{ x } 24 \text{ x } 365 \text{ x } \mathbf{spd}\ ) =$$

$$= \mathbf{O(10^{600})} : \mathbf{O(10^{18})} = \mathbf{O(10^{582})}$$

т.е. примерно $\mathbf{10^{582}}$ лет.

o

Следует также заметить, что любой другой (т.е. отличный от представленного выше и основанного на стратегии ***BT*** <*все тупики*>) алгоритм оптимизации клиринга должен порождать сопоставимые с названными выше объемы вычислений. Основанием для такого вывода служит следующее простое соображение: каждый алгоритм, использующий то или иное условие локальной оптимизации (т.е. алгоритм, анализирующий в каждый момент времени какой-либо локальный фрагмент описаний исходного графа клиринга или подобного ему графа), способен породить лишь локальный экстремум (в нашем описании - тупик). Поиск глобального экстремума требует просмотра всех локальных экстремумов, т.е. приводит к перебору представленного выше (возможно - экспоненциально большого) множества тупиков.

Рассмотренное нами обстоятельство порождает также

П 5.2.9. Следствие

Задача ***Clir*** не принадлежит ***NPC*** - классу ***NP***-полных задач.

Доказательство.

Принадлежность задачи ***Clir*** классу ***NPC*** означала бы, что по предъявленному графу $\underline{\boldsymbol{Cl^{*}_{0}}} \in \underline{\boldsymbol{G}}(\underline{\boldsymbol{Cl^{*}}})$ можно было бы полиномиально быстро проверить выполнимость условий

$$(1)\ \boldsymbol{Sum_{V}}(\underline{\boldsymbol{Cl^{*}_{0}}}) = \max_{\underline{\boldsymbol{Cl^{*}_{0}}} \in \underline{\boldsymbol{G}}(\underline{\boldsymbol{Cl^{*}}})} (\boldsymbol{Sum_{V}}(\underline{\boldsymbol{Cl^{*}_{i}}}))$$

или

$$(2)\ \boldsymbol{Sum_{N}}(\underline{\boldsymbol{Cl^{*}_{0}}}) = \max_{\underline{\boldsymbol{Cl^{*}_{0}}} \in \underline{\boldsymbol{G}}(\underline{\boldsymbol{Cl^{*}}})} (\boldsymbol{Sum_{N}}(\underline{\boldsymbol{Cl^{*}_{i}}})),$$

Проверка того, является ли граф $\underline{\boldsymbol{Cl^{*}_{0}}}$ тупиком преобразований подобия на графе $\underline{\boldsymbol{Cl^{*}}}$, действительно осуществима алгоритмом полиномиальной вычислительной сложности (достаточно по одному разу просмотреть каждое из ребер графа $\underline{\boldsymbol{Cl^{*}}}$). Проверка же “глобальности” экстремума в тупике $\underline{\boldsymbol{Cl^{*}_{0}}}$ требует, вообще говоря, сравнений его со всеми другими локальными экстремумами, т.е. - просмотра всех тупиков преобразования подобия на графе $\underline{\boldsymbol{Cl^{*}}}$. А их (см. **Теорему П 5.2.7**) может оказаться экспоненциально много.

o

Подробно рассмотренная выше группа “негативных” свойств задачи ***Clir*** заставляет задуматься об изменении тактики оптимизации при создании прикладных систем (например, систем оптимизации реальных клиринговых расчетов). Разумным представляется отказаться (по крайней мере при работе в режиме *реального времени*) от поиска точного решения задачи ***Clir***, перейдя к использованию ряда подходов, которые мы условно назовем “нестандартными”. К ним, на наш взгляд, в первую очередь следует отнести:

- выделение паретовых множеств при многокритериальной оптимизации;
- оптимизацию с "оракулом", т.е. использование опыта ранее построенных решений
  - (а) для выбора "начального приближения"("стартовой точки") процесса перевода анализируемой ситуации в соответствующее паретово множество;
  - (б) для декомпозиции исходной ситуации к задаче меньшей размерности;
- "структурную" оптимизацию в рамках так называемых рассуждений, основанных на анализе прецедентов и другие варианты реализации формализованного рассуждения.

Рассмотрим несколько подробнее каждое из представленных направлений.

Идея использования *паретовых множеств*[28] при многокритериальной оптимизации состоит, грубо говоря, в реализации двух шагов:

[28] Мы будем называть паретовым множество, содержащие все точки, доставляющие оптимальное значение соответствующему функционалу качества.

(1º) незначительном “ослаблении” требований критерия оптимизации (когда не нужно искать исключительно точный оптимум функционала качества: достаточно найти соответствующую “ε-окрестность” оптимума, при “попадании” в которую гарантировано выполнение “ослабленных” критериев оптимизации);
(2º) использовании алгоритмов меньшей (чем для построения точного оптимума) вычислительной сложности для поиска “пути” в заданную “ε-окрестность” оптимума.

Оптимизация с “оракулом”, т.е. использование “подсказок (например, советов эксперта или опыта ранее решавшихся задач), может дать интересные результаты как при “быстром” поиске нового решения, базирующегося на варианте ранее использовавшегося решения в качестве “стартовой точки” для применения подходящей итеративной процедуры оптимизации, так и при декомпозиции решаемой задачи к задаче меньшей размерности путем удаления из ее описания фрагмента, в точности соответствующего ранее “обсчитанной” конфигурации.

Последняя процедура может быть реализована с помощью развиваемой в последние годы техники оперирования с объектами сложной структуры (в том числе - графами) использующей рассуждения, основанные на ДСМ-анализе прецедентов. Особенностью этого подхода является возможность опираться в процессе принятия решений на опыт как положительных, так и отрицательных решений-прецедентов, обработанных на предыдущих этапах и собранных в специальную Базу Прецедентов. Последняя соединена с соответствующей информационно поисковой системой (призванной при попадании на вход системы принятия решений описания текущей ситуации автоматически подбирать в Базе Прецедентов описания элементов, “схожих” с анализируемой ситуацией, вместе с описанием принятых в них - как позитивных, так и негативных - решений), и интеллектуальной системой автоматического порождения возможных альтернатив для принятия решения. Таким образом, решения, порожденные ранее на ситуациях-прецедентах, удается использовать эффективно в технике *оптимизации с “оракулом*”, что, в свою очередь, позволяет получить ощутимый выигрыш в необходимых для порождений квазиоптимального (близкого[29] к оптимальному) решения объемах вычислений.

В качестве конкретного механизма "структурной" оптимизации используется уже представленный ДСМ-метод автоматического порождения гипотез в базах данных с неполной информацией.

Мотивация подхода.

В основе используемого подхода лежит идея соединения техники так называемого ситуационного анализа[30], рассуждений, основанных на анализе прецедентов, и методов машинного обучения на примерах (*machine learning*), базирующегося на анализе сходства (*similarity based learning*) в описаниях изучаемых объектов (примеров).

Общая структура ДСМ-рассуждения в данном случае.

Структура ДСМ-рассуждения в данном случае повторяет все основные элементы ранее представленной в **Главах 2** и **3** конструкции и определяется следующей схемой:

изучаются зависимости вида ***ОБЪЕКТ*** ***==>СВОЙСТВА***, при этом

- заданы множество $\boldsymbol{Ex}^{+}$ примеров (описаний исходно заданных объектов), обладающих анализируемым свойством, и множество $\boldsymbol{Ex}^{-}$ объектов, не обладающих анализируемым свойством;
- сходство анализируемых объектов формализовано как алгебраическая операция (результат применения которой к паре объектов есть фрагмент описания каждого из двух исходных объектов, представляющий их “общую часть” - “сходство”);

[29] В терминах строго заданного уточнения понятия “близости” (например, в терминах соответствующей метрики и т.п.).
[30] Основные идеи ситуационного анализа, предполагающего использование логико-лингвистических моделей в теории и практике управления, можно найти, например в монографии [*Поспелов 86*] и др..

- задана система правил правдоподобного вывода

$$M = M_1 \cup M_2 = (M^+_1 \cup M^+_2) \cup (M^-_1 \cup M^-_2),$$

содержащая правила $M^+=(M^+_1 \cup M^+_2)$ анализа множества Ex^+ и правила $M^-=(M^-_1 \cup M^-_2)$, анализа множества Ex^-; при этом

- правила из множества M^{α}_1 используются для порождения зависимостей вида

<фрагмент (описания соответствующих объектов) есть причина наличия у них искомого свойства>

в случае, когда α=+ , и зависимостей вида

<фрагмент (описания соответствующих объектов) есть причина отсутствия у них искомого свойства>

в случае, когда α= **-;**

- правила из множества M^{α}_2 используются для переноса (там, где это возможно) на вновь анализируемые случаи[31];
- применение правил правдоподобного вывода обоих знаков регулируется специальным условием “контроля достаточности оснований для правдоподобного вывода”[32];
- процесс правдоподобного вывода стартует с исходно заданных множеств Ex^+ и Ex^-, порождая “причинные зависимости” и используя их для прогнозирования свойств исходно заданного множества объектов Ex^0, про которые не известно, обладает ли каждый из них искомым свойствам. При этом “доопределенные” на соответствующем шаге вычислений элементы множества Ex^0 добавляются (в соответствии со “знаком” доопределения) к множествам Ex^+ и Ex^-, после чего
- процесс итерируется до состояния стабилизации (в котором либо множество Ex^0 пусто, либо нет возможности породить ни одной новой причинной зависимости).

<u>База "прецедентов".</u>

Как уже отмечалось выше, База Прецедентов (исходно заданные, либо порожденные на соответствующем шаге процесса оптимизации множества Ex^+ и Ex^-) для задачи оптимизации клиринга содержит конфигурации структур вида

<исходный граф , результирующий граф[33], сведения об эффективности[34] решения >

<u>Формализация сходства "прецедентов"</u>

Сходство прецедентов формализуется как операция выделения общей части соответствующих графов, представляющих совокупность взаимных расчетов участников в процессе сеанса клиринга. При этом возможны варианты, варьирующие степень “строгости” сходства, и определяемые необходимостью

- учитывать приоритеты на вершинах графа клиринга (приоритеты на множестве “персонажей”, участвующих в процессе клиринга);
- учитывать приоритеты на ребрах графа клиринга (т.е. на типах платежей)

и т.п..

---

[31] Результатом переноса зависимости о <u>причинах наличия</u> искомого свойства является вывод о наличии этого свойства у вновь анализируемого объекта (результат применения +-правила). Результатом переноса зависимости о <u>причинах отсутствия</u> искомого свойства есть вывод об отсутствии этого же свойства у анализируемого объекта (результат применения --правила).

[32] Специальным критерием верификации результатов, порождаемых алгоритмом вычисления.

[33] При этом возможно хранение как финальных, так и важных промежуточных конфигураций графов, подобных заданному исходному.

[34] Т.е. пометку, что случай “позитивный”, если он разрешен “быстрыми” вычислениями, или пометку, что случай “негативный”, если разрешить его “быстрыми” вычислениями не удалось.

Выделение решений-"аналогов" в базе "прецедентов".

Процедура выделения аналогов для текущей (анализируемой в данный момент) ситуации представляет собой просмотр Базы Прецедентов и выделение в ней элементов, “сходных” - в смысле представленной выше операции сходства - с текущей анализируемой ситуацией.

Анализ переносимости решения с "прецедентов" на оптимизируемую ситуацию.

Анализ переносимости вариантов решения с ситуации-аналога на текущую оптимизируемую ситуацию регулируется математической техникой ДСМ-метода и в общем виде сводим к следующей схеме:

- версия решения может быть перенесена с ситуации аналога на текущую ситуацию тогда и лишь тогда, когда возможна вложимость в граф текущей оптимизируемой ситуации соответствующей подструктуры хотя бы одного “позитивного” прецедента, и не возможна вложимость соответствующих подструктур ни одного из “негативных” прецедентов,
- в случае одновременной вложимости “позитивных” и “негативных” подструктур имеется альтернатива:
  а) исследуемый вариант может в грубом анализа не рассматриваться как “позитивно доопределимый”;
  б) исследуемый вариант может быть подвергнут более тонкому анализу средствами ДСМ-метода с целью устранения (в тех случаях, когда это возможно) возникнувшего так называемого “эмпирического противоречия (см. **Главы 2-3**)”.

Следует подчеркнуть, что вычислительные характеристики используемого фрагмента ДСМ-метода не препятствуют разработке “быстрого” варианта предлагаемой технологии.

Предлагаемый при анализе сходства “структурных” описаний механизм выделения общих частей соответствующих графов клиринга, в отличие от общего случая, может быть реализован посредством быстрых полиномиальных алгоритмов: использование имен вершин графа ***Cl**** в ситуации, когда каждое из имен имеет уникальный характер, позволяет при анализе сходства графов такого вида точно определить способ наложения графа на граф по схеме “вершина на вершину”, причем реализовать каждое такое наложение всего один раз при помощи однократного просмотра списков вершин соответствующих графов. Также отметим независимость предлагаемой технологии оптимизации клиринговых расчетов от конкретного – централизованного (а) или распределенного (б) - вида организационной структуры, их поддерживающей. Можно показать, что предлагаемая процедурная

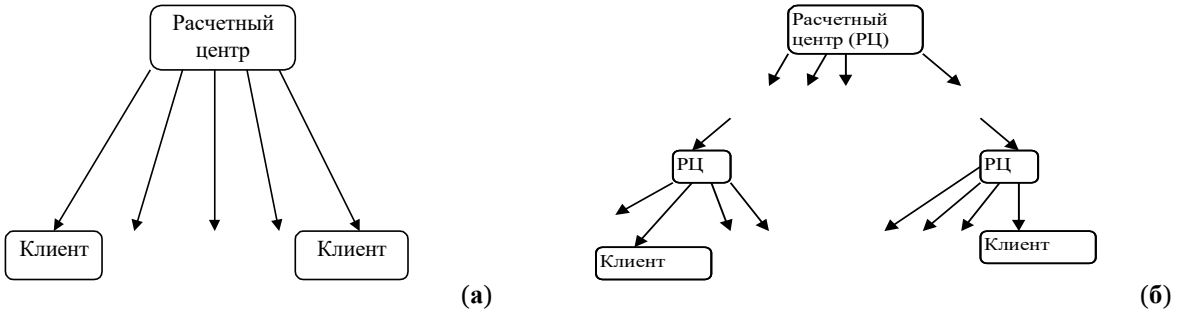


схема может быть реализована в обеих архитектурах обработки платежных документов.

Предложенный механизм оптимизации, как показали эксперименты на реальных данных[35], обеспечивает значительное сокращение объемов (в среднем на обрабатывавшихся данных - это десятки процентов) связанных в системе управления взаимозачетами финансовых средств при вполне приемлемых затратах вычислительных ресурсов (особенно на стадии, когда База Прецедентов системы уже содержит достаточно большой объем “обсчитанного” материала).

---

[35] Исследования были осуществлены в рамках Проекта создания системы **БАРС** по заданию Центрального Банка Российской Федерации.

Приложение 6.

Экспертиза сложных технических проектов.

Интересной новой областью приложений моделей и технологий ДСМ-ИАД оказалась проблематика экспертизы сложных технологических проектов, в частности - при проектировании и разработке (или же строительстве) высокотехнологичных объектов комплексного характера. Содержательная задача здесь в самом общем виде может быть сформулирована следующим образом:

- принято решение о необходимости создать тот или иной сложный технический комплекс (нефте- или газопровод, технологический объект, телекоммуникационную систему и т.п.). Определен заказчик разворачивающихся работ и объявляется конкурс на исполнение соответствующего подряда;
- предполагаемые работы еще только предстоит «формализовать»: сформировать не только «стартовые»[1] требования к создаваемому объекту а также к генеральному подрядчику «запускаемого» проекта, но и соответствующие действующим типовым требованиям и стандартам комплекты проектной и технической документации;

тем не менее

- участие в конкурсе (на исполнение подрядных работ) требует оценки соответствующих «интересов» сторонами[2] в оперативном режиме: длительность типового конкурса, как правило, - 1.5-2 месяца, в то время как процесс штатного проектирования (разработки комплекта «гостированной» проектной документации) по такого рода проектам длится, как правило, не менее 6 месяцев;

итак:

- ожидаемый успех «запускаемого» проекта требует исключения ошибок в оценке предложений и выборе действительно оптимального варианта действий.

Таким образом, одной стороне – исполнителю - (еще до выполнения необходимого цикла проектных работ и формирования аргументированных смет, «карт» компетенций и т.п.) требуется оценить ресурсы и прибыльность предлагаемого проекта, а другой стороне – заказчику – иметь основания для объективной экспертизы предложений участников конкурса (ведь простой выбор предложения с минимальной ценой может оказаться неверным, если представивший его претендент сам ошибся в соответствующих расчетах).

Оказалось, что в ряде ситуаций обсуждаемого типа имеется возможность воспользоваться опытом ранее уже выполнявшихся проектов аналогичного характера (опытом проектирования и строительства, технического конструирования и т.п.). Именно такая возможность «обучаться на прецедентах» и была использована при решении нескольких задач для конкретных прикладных областей, два примера которых и представлены ниже.

А.М.Гаджиев [*Гаджиев 12*][3] продемонстрировал возможности использования функционала ИАД при решении обсуждаемого класса задач «экспресс»-оценки проектов ДСМ-средствами восстановления зависимостей в *числовых* данных. На примере конкретного предприятия, работающего в сфере строительства сложных технических объектов для нефтегазовой отрасли (газопроводов, компрессорных станций и т.п.), показано как при участии в тендерах, которые используют так называемую предварительную проектную документацию для реализации того или иного конкретного объекта, каким образом на базе ана-

---

[1] Которые, как правило, характеризуются неполнотой описания соответствующих целей и задач. (Обычно это отражается в процедурах последующих пополнений исходных данных для проекта новыми, иногда – весьма существенными для *уже сформированной* структуры соответствующего проекта – элементами).

[2] Заметим: не только потенциальными исполнителями, но и заказчиком (ведь возможные ошибки в оценке необходимых компетенций и ресурсов одинаково критичны для обеих сторон разворачивающегося процесса).

[3] Эта работа была отмечена в июне 2012 года Государственной Экзаменационной Комиссией Факультета ВМК МГУ как одна из лучших магистерских диссертаций года.

лиза прецедентов (значений соответствующих параметров в нескольких ранее уже выполненных компанией подобных проектах) можно принимать *обоснованные* решения об участии (или же, наоборот, не участии) в конкретном новом проекте[4].

При этом процедура принятия основного решения по участию в проекте состоит фактически в оценке разницы между суммой, обозначенной заказчиком, о прогнозной оценкой[5] себестоимости работ предлагаемого проекта. Оказалось, что на практике оценки с точностью порядка 10% позволяют исполнителю в отсутствии полной информации о параметрах анализируемого проекта не только избежать участия в убыточных проектах, но и обеспечивают определенные конкурентные преимущества по отношению к тем участникам конкурса, кто средств экспертизы такой точности в своем распоряжении не имеют.

На материале уже выполненных компанией 17 конкретных проектов, представленных разбитыми на 12 «однородных» групп 51 параметром, а также с учетом специальных коэффициентов «настройки» на соответствующие погодные условия, особенности географического ландшафта проекта и др., была сформирована (и введена в эксплуатацию в планово-финансовом департаменте этой компании) компьютерная система[6] – ассистент финансового аналитика, далее успешно используемая при решении текущих бизнес-задач.

При этом было экспериментально показано, что формальные результаты счета на реальных данных имеют вполне адекватную *содержательную интерпретацию*: по результатам экспериментов стало понятно, что подобные результаты при анализе имеющихся эмпирических данных мог бы получить и человек-эксперт, однако, компьютерная система, работающая по предложенной А.М.Гаджиевым технологии, порождает их несравненно быстрее (имея при этом возможности масштабирования для обработки недоступных человеку по своим объемам обучающих выборок исходных данных).

В работе Е.В.Синяковой [*Синякова 13, Забежайло 14*] рассматривалась задача быстрого прогнозирования оценочных характеристик объектов в проектах производства сложных высокотехнологичных изделий в условиях исходно неполной информации о разрабатываемых изделиях, то есть до выполнения (и завершения) их детального проектирования.

Обсуждаемая задача представляет собой один из востребованных типов задач поддержки принятия решения в области информационных систем управления предприятием. Актуальность рассматриваемой проблематики обусловлена тем, что ситуации, в которых возникает потребность в быстром оценивании соответствующих характеристик разрабатываемого изделия (а именно ситуации, срок принятия решения о запуске или же отказе от запуска соответствующего проекта в которых заведомо меньше штатных сроков выполнения полного проектирования целевого технического объекта) достаточно распространены в производственной деятельности современных высокотехнологичных предприятий. Примерами таких ситуаций могут быть, в частности, заключение договора с заказчиком, участие в различных конкурсных процедурах[7] и т.д.

В рассматриваемой прикладной задаче техническими изделиями, характеристики проектов производства которых необходимо прогнозировать, является особый тип инженерного оборудования – специальные климатические шкафы, используемые в телекомму-

---

[4] Информация в конкурсной документации обычно характеризует оценки как допустимой цены работ, так и основных параметров будущего объекта, формулируемые заказчиком (например: общую протяженность магистрали, количество переходов через водные объекты, прочие геодезические характеристики местности, требования к самой магистрали - диаметр, давление и т.п., - требования к инфраструктуре и прочее).

[5] Напомним, формируемой в ситуации, когда *детальный* анализ себестоимости каждого вида работ, детализируемого до уровня элементарных компонентов объекта (конкретных часов работы исполнителя соответствующей квалификации, конкретных объемов и типов используемых материалов, конкретных часов работы машин и механизмов, себестоимость которых предусмотрена соответствующими отраслевыми нормами и правилами), в сроки проведения конкурса попросту *невозможно* выполнить.

[6] На промышленной программной платформе для решения задач бюджетирования Oracle Hyperion Planning

[7] Разумеется, ограниченных в своей длительности по времени лишь требованиями заказчика.

никационной индустрии. Характеристикой, которую здесь необходимо прогнозировать, является класс длительности производства данного оборудования. Известной на момент принятия решения неполной информацией об объекте, на основе которой может быть осуществлён прогноз, является технический проект изделия[8], отражающий внутреннюю структуру («архитектуру») объекта без учета ряда особенностей изготовления *нетиповых* аппаратно-схемных или же программных компонентов. Дополнительными исходными данными является накапливаемое множество фактов, представляющее собой данные о уже имеющихся прецедентах производства подобных технических изделий за определённый срок.

Основными особенностями решаемой прикладной задачи, характерными в общем случае для задач поддержки принятия решения в рассматриваемой области, являются:

- необходимость оперировать техническими объектами сложной структуры, наличие неполной информации о структуре объекта на этапе принятия решения о запуске проекта,
- малая и, как правило, статистически незначимая выборка уже завершённых проектов, и
- наличие скрытых (неявно выраженных в описании соответствующих прецедентов) причинно-следственных зависимостей в накапливаемом множестве фактов.

Для ее решения использовалась технология ДСМ-ИАД: были выполнены настройки стандартных средств ДСМ-метода на особенности выбранной предметной области. Программная реализация разработанных алгоритмов была оформлена в виде законченного инструментального программного средства.

Так (следуя стандартной схеме применения ДСМ-метода):

1. Множеству объектов X поставлено в соответствие множество структурных схем (проектов) P изделий рассматриваемой прикладной задачи. Формализованное представление объектов (а именно структурных схем технических изделий) отражает множество элементов[9] номенклатора, входящих в состав рассматриваемого объекта, а также их количество. Так как возможно кратное вхождение элементов номеклатора в состав изделия, формализованное описание объекта представляет собой нумерованное множество. Для представления объекта (нумерованного множества) стандартными теоретико-множественными средствами было введено следующее формальное взаимно-однозначное соответствие:

O – множество всех объектов номенклатора (множество образующих);

$O^* = \{o_{i,n}\}$, (где $i = 1,...,|O|$, $n \in N$, N – множество натуральных чисел) – множество, соответствующее расширению используемого множества образующих O для кодировки кратных вхождений элементов в нумерованные множества.

Таким образом, $P_i \subseteq O^*$.

2. Множеству признаков V объекта поставлено в соответствие множество фрагментов структурных схем F ($F_i \subseteq O^*$) изделий рассматриваемой прикладной задачи.

3. Множеству свойств W объектов поставлено в соответствие множество классов длительности проектов C: $C = \{c_1, c_2, c_3, c_4, c_5\}$, где каждому свойству $c_i \in C$ соответствует упорядоченная пара $\{t_{\min}, t_{\max}\}$, соответствующая интервалу длительности $[t_{\min}..t_{\max}]$ производства изделий.

«Настройки» ДСМ-инструментария, подбираемые в ходе итераций (последовательно проводившихся при решении рассматриваемой прикладной задачи) ИАД-экспериментов, затронули в первую очередь используемое формальное определение сходства. Так в ситуации, когда:

а) объект рассматривается как множество типовых фрагментов его структуры (составных блоков), т.е. когда задействованный тип данных – множества, используемое сходство (рассматриваемых) объектов – теоретико-множественная операция пересечения $\cap$;

---

[8] Комплект проектной документации

[9] Подструктур стандартного вида.

а в ситуации, когда:

б) объект рассматривается как множество типовых фрагментов его структуры (составных блоков) с дополнительным учетом кратности вхождения соответствующих составных блоков в структуру описываемого объекта (т.е. когда задействованный тип данных – нумерованные множества), уместно провести расширение исходного алфавита $\boldsymbol{U}$ по следующей схеме: каждой встречающейся кратности k вхождения составного блока S_i в описание текущего объекта A_j сопоставляется (в расширенном алфавите $\boldsymbol{U^*}$) новая образующая S_{ik}. Легко видеть, что расширенный алфавит $\boldsymbol{U^*}$ для каждого конкретного множества примеров $\boldsymbol{\Omega}$ восстанавливается достаточно легко. Далее: сходство объектов, представленных в расширенном алфавите $\boldsymbol{U^*}(\boldsymbol{\Omega})$, вновь будет определяться как теоретико-множественная операция $\cap$ пересечения, что будет эквивалентно выделению для пары сходных объектов всех общих составных блоков S_{ik} совпадающей кратности.

Итак, в рассматриваемой ситуации отражение структурной «природы» обрабатываемых объектов в используемом языке представления данных об объектах (а именно – описанное представление их стандартными теоретико-множественными средствами) позволило определить сходство двух объектов как сходство нумерованных множеств, формализуемое операцией теоретико-множественного пересечения[10].

Правила правдоподобного вывода ДСМ-метода при работе с объектами в обсуждаемой прикладной задаче выбирались

- сперва в стандартной конфигурации;
- а в ситуациях, когда применение обычных ППВ не давало адекватно интерпретируемых результатов, расширялись дополнительными логическими условиям, отражающими некоторые специальные метрические соображения[11].

Следуя далее стандартной схеме применения ДСМ-метода, пришлось определиться с представлением в базе фактов позитивных примеров, негативных примеров, а также примеров неопределённости изучаемого свойства. В самом общем виде здесь множество позитивных примеров – множество объектов, принадлежащих определённому классу ((+)-примеры), множество негативных примеров – объекты, не принадлежащие определённому классу ((–)-примеры). Множество недоопределенных примеров – объекты, класс длительности создания которых не определён.

Для рассматриваемой прикладной задачи выделение объектов, принадлежащих определённому классу, и объектов, класс которых не определён, не вызывает трудностей. Наличие пяти классов длительности, на которые подразделяются проекты выполненных изделий, накапливаемых во множестве фактов, автоматически обеспечивает наличие позитивных примеров для каждого из классов длительности. Примерам неопределённости изучаемого явления соответствуют проекты ещё не выполненных изделий, для которых необходимо прогнозировать соответствующий класс длительности.

Однако, выделение объектов, не принадлежащих определённому классу ((–)-примеров класса или контрпримеров), вызвало определенные затруднения. Оказалось, что можно предложить различные варианты определения контрпримеров, которые в общем случае могут иметь различное влияние на интерпретируемость получаемых результатов, и прежде всего – на содержательную интерпретируемость в терминах решаемой прикладной задачи выявленных гипотез (т.е. «реконструируемых» ДСМ-методом причин принадлежности/непринадлежности объектов к определённому классу длительности). Так, в частности, (–)-примерами определённого класса длительности можно считать проекты-представители классов длительности, отличных от рассматриваемого, или же проекты-представители

---

[10] Более точно: нумерованные множества представлены обычными множествами (каждая возможная кратность вхождения конкретного элемента в нумерованное множество кодируется отдельным элементом обычного множества), сходство которых и задается теоретико-множественной операцией пересечения.

[11] Количества соответствующих случаев.

классов длительности, отличных от рассматриваемого и относящихся к наименьшим (наибольшим) классам длительности по сравнению с рассматриваемым классом.

Для выбора наилучшего из возможных вариантов определения понятия контрпример, наиболее подходящего с точки зрения возможностей содержательной интерпретации полученных гипотез, был произведён ряд вычислительных экспериментов.

Так при заданных пяти классах длительности анализируемых проектов T_1, T_2, … , T_5 и уточнении (для каждого класса T_i) множества примеров – как элементов этого класса T_i, а множества контрпримеров – как множества { T_1, T_2, … , T_5 }\ T_i, критически важным «дефектом» при порождении ДСМ-результатов оказались «межклассовые» сходства для классов слева и справа от T_i. (Разумной содержательной интерпретации таким сходствам предложить не удалось).

При отнесении к множеству примеров вместе с текущим T_i также всех классов меньшей длительности, а к множеству контрпримеров – оставшихся классов из множества { T_1, T_2, … , T_5 } объемы не имеющих адекватной содержательной интерпретации ДСМ-гипотез хоть и сократились, однако, как и ранее не удалось полностью избежать соответствующих «смысловых дефектов».

И только при явном запрете межклассовых сходств по схеме: внутри каждого класса сходство выражает операция пересечения $\cap$, а при применении операции сходства к представителям разных классов результат по определению полагается равным $\varnothing$, удалось (применением стандартных ДСМ-ППВ) сформировать полностью интерпретируемые результаты (обзор которых и представлен ниже).

В рамках проводимых экспериментов был реализован простейший вариант схемы ДСМ-ИАД (интеллектуального анализа данных средствами ДСМ-метода). Основными задачами, решёнными при реализации данной схемы, были:

- выбор варианта определения понятия контрпример определённого класса длительности. Так, в качестве первого варианта было принято решение рассматривать в качестве (–)-примеров фиксированного класса длительности все проекты-представителей классов длительности, отличных от рассматриваемого;
- уточнение синтаксиса формулировки основных понятий ДСМ-метода, а именно синтаксиса формулировки решающих предикатов и правил правдоподобного вывода первого и второго рода (возможность использования стандартных формулировок, представленных при формализации метода сходства, лежащего в основе используемого для решения поставленной прикладной задачи ДСМ-метода АПГ без внесения модификаций, оказалось ограниченной узким кругом рассматриваемых здесь задач);
- выполнение модификации логических и алгоритмических средств ДСМ-метода в соответствии с особенностями прикладной задачи. Так была осуществлена настройка алгоритмики ДСМ-метода (а именно алгоритма порождения гипотез и алгоритма классификации), учитывающая содержательные особенности предметной области, отражённые в частности при уточнении синтаксиса формулировки основных понятий ДСМ-метода. Кроме того, была осуществлена разработка модифицированной версии быстрого алгоритма поиска глобальных сходств «Замыкай по одному» [*Кузнецов 93*], нивелирующей недостатки данного алгоритма и обеспечивающей возможности его применения в рамках ДСМ-метода;
- выполнение программной реализации разработанных алгоритмов в виде законченного программного средства.

Большинство гипотез, полученных с помощью разработанного начального варианта схемы ДСМ-ИАД в выбранной прикладной предметной области, являлись необоснованными и плохо интерпретируемыми: наблюдалось появление ошибочных[12] (0)-гипотез (противоречивых гипотез, являющихся как причинами принадлежности, так и причинами

[12] С точки зрения содержательной интерпретации в терминах исследуемой прикладной предметной области.

непринадлежности объектов к определённому классу длительности) во всех классах длительности и ошибочных (–)-гипотез (гипотез, являющихся причинами непринадлежности объектов к определённому классу длительности) в старших классах длительности. Данные ошибочные (0)- и (–)-гипотезы должны были являться только (+)-гипотезами (гипотезами, отражающими причины принадлежности объектов к определённому классу длительности) для младших классов длительности. Выявленные трудности содержательной интерпретации полученных гипотез повлекли за собой также недостаточную обоснованность результатов классификации, полученных на основе применения данных гипотез («причинных» зависимостей).

Одной из особенностей порождения гипотез при поиске глобальных сходств объектов различных классов длительности оказалось выявление в старших классах сходств (т.е. подструктур изделий), характерных для младших классов. Данный факт, при использовании выбранного в данном варианте схемы ДСМ-ИАД способа определения понятий «пример» и «контрпример», оказался причиной «определения» большей части (+)-гипотез младших классов (0)-гипотезами, а также появления в старших классах необоснованных (0)- и (–)-гипотез.

Для устранения влияния выявленных псевдо-причин (в частности устранения возможности появления в старших классах необоснованных (0)- и (–)-гипотез) была предпринята попытка модификации используемых решающих средств, а именно модификации определения понятий «(+)-пример» и «(–)-пример» определённого класса длительности. При проведении второго эксперимента в качестве (+)-примеров фиксированного класса длительности были выбраны все проекты, длительность производства изделий по которым не превосходила длительность производства, характерную для рассматриваемого фиксированного класса. В качестве (–)-примеров фиксированного класса длительности были выбраны все проекты, длительность производства изделий по которым превосходила длительность производства изделий, характерную для рассматриваемого фиксированного класса. Введение данной модификации потребовало также внедрения некоторых изменений в правила правдоподобного вывода второго рода.

Анализ результатов применения данной модифицированной схемы ДСМ-ИАД в выбранной прикладной предметной области показал, что один из недостатков, которые были характерны для предыдущего варианта, появилась возможность устранить: в старших классах стало возможным избежать появления необоснованных (0)- и (–)-гипотез, характерных для младших классов. Причиной для устранения этого недостатка стало изменение способа выбора (–)-примеров, а именно принадлежность (–)-примеров старшим относительно рассматриваемого классам. Однако для данного варианта реализации метода также оказался характерен ряд существенных недостатков. Так, например, сохранилось появление ошибочных (0)-гипотез в младших классах, наблюдавшееся в предыдущем варианте, принадлежность (+)-примеров к различным классам повлекла за собой появление в качестве необоснованных (+)-гипотез большого количества межклассовых сходств (т.е. сходств, в образовании которых участвуют по одному проекту-представителю как минимум двух разных классов длительности).

Тем не менее, недостатки данного варианта оказались достаточно легко устранимыми. Так, для предотвращения появления ошибочных (0)-гипотез в младших интервалах было решено определить данные гипотезы как (+)-гипотезы слабой силы. Для устранения появления межклассовых сходств в качестве необоснованных гипотез был введён запрет (как пустота по определению соответствующей операции сходства) на рассмотрение сходств, в образовании которых участвуют по одному проекту-представителю как минимум двух разных классов длительности. Кроме того, данный вариант модификации за счёт устранения понятия (0)-гипотезы и введения ранжирования по силе (+)-гипотез позволил также существенно упростить правила правдоподобного вывода 2-ого рода (правила классификации).

Таким образом, в результате проведения ряда «настроечных» экспериментов удалось разработать проблемно-ориентированную ИС, результаты применения которой в выбранной прикладной предметной области (и прежде всего – полученные гипотезы и результаты классификации) являются обоснованными и ясно интерпретируемыми.

Достаточным основанием для принятия индуктивных выводов ДСМ-рассуждения является выполнение аксиомы каузальной полноты, которое требует корректной доопределимости всех объектов исходной базы фактов порожденными из нее[13] ДСМ-гипотезами применением правил правдоподобного вывода второго рода. Если условие каузальной полноты не выполняется, то исходную базу фактов (примеров и контрпримеров для ДСМ-обучения) следует пополнять новыми примерами.

Результаты классификации, полученные с помощью применения разработанной схемы ИАД-ДСМ, показали, что из 69 проектов, составляющих обучающую выборку, 64 проекта были классифицированы верно, 5 проектов классифицировано неверно, неклассифицированных проектов нет. Следовательно, несмотря на то, что была выявлена верно большая часть гипотез о принадлежности/непринадлежности проекта к определённому классу длительности, аксиома каузальной полноты, которая позволяет говорить о том, что индуктивный вывод был сделан на достаточном основании, осталась невыполненной. (Комментарий: зато степень каузальной полноты 64/69 очень высокая – больше 0.9 с учетом того, что принято считать «хорошим» значение больше 0.8).

Однако, анализ гипотез, которыми были неверно классифицированы объекты, позволяет направленным образом с помощью расширения базы фактов улучшать результаты прогноза. Так, для улучшения качества классификации следует формировать недостающие гипотезы с помощью направленного добавления новых проектов в те классы, где нет достаточного объема данных для корректного обучения ДСМ-системы, и понижать ранги некорректных гипотез в остальных классах также с помощью добавление новых фактов (также – вполне определенного вида). Предлагаемое в такой схеме направленное расширение обучающей выборки следует вести до ее «насыщения» – ситуации, когда все имеющиеся в ней (например, на текущий момент ДСМ-рассуждения) эмпирические зависимости о причинах изучаемых явлений полностью (пусть и в неявном виде – через описания соответствующих прецедентов) представлены имеющимися данными для ДСМ-обучения.

Таким образом, с помощью добавления новых фактов для формирования корректных гипотез и понижения ранга некорректных гипотез в нужных классах удалось обеспечить верную классификацию изначально неверно классифицируемых проектов обучающей выборки и, как следствие, добиться выполнимости аксиомы каузальной полноты.

Программная реализация разработанных алгоритмов в виде законченного программного средства обладает рядом интересных характеристик. Использование разработанного быстрого алгоритма поиска глобальных сходств позволяет обеспечить быстрый отклик ИС на запрос (о порождении каузальных зависимостей и прогнозировании свойств новых объектов), а также высокую производительность. Применение технологии Java повышает надёжность системы в целом, за счёт встроенных в Java средств контроля типов, автоматического сборщика мусора и т.п. Организация независимости от используемой СУБД за счёт использования технологии JPA, наличие хорошо-структурированной архитектуры ИС, а также использование объектно-ориентированного подхода предоставляют возможности для гибкого расширения или модификации функционала системы (например, при возникновении новых требований или же при появлении новых ДСМ-правил правдоподобного вывода). Также имеются возможности развития системы при необходимости до корпоративного уровня (т.е. до уровня информационной системы, используемой в качестве штатного инструментального средства при решении на предприятии производственных задач обсуждаемого типа), а также переноса части разработанного функционала системы на мобильные платформы.

---

[13] Из этой базы фактов.

Приложение 7.

О некоторых комбинаторных свойствах ассоциативных зависимостей.

Сегодня в ряде практических приложений так называемый *анализ ассоциативных правил* (см., например, http://www.basegroup.ru/library/analysis/association_rules/ и др.) представляет собою достаточно развитую отдельную область успешного применения моделей и методов computer science. Однако, начиная еще с самых ранних работ в этом направлении (см., например, [*Agrawal 93*] и др.) описания ряда комбинаторных свойств развиваемых здесь моделей и алгоритмов содержат определенные пробелы. Целью предпринятого ниже рассмотрения является стремление сузить область неопределенности, характеризующей некоторые аспекты сложности обсуждаемого *анализа ассоциативных правил*.

Пусть задано некоторое непустое множество символов (букв) $\mathbf{U} = \{a_1, a_2, \dots, a_n\}$. Будем называть его алфавитом. Каждое непустое подмножество $\varnothing \neq S \subseteq \mathbf{U}$ букв из алфавита **U** будем называть словом. Пусть $\mathbf{\Omega} = \{S_1, S_2, \dots, S_m\} \subseteq 2^{\mathbf{U}} \setminus \varnothing$ - некоторое заданное множество слов.

П 7.1. Определение

Пусть

1) $X \subseteq \mathbf{U}$ и $Y \subseteq \mathbf{U}$ – два (непустых) подмножества букв алфавита **U**;
2) в множестве слов **Ω** найдется подмножество $\mathbf{\Omega}^{\sim} = \{S_{i1}, S_{i2}, \dots, S_{ik}\}$ при $(k \geq 1)$ таких, что

$$(X \cap Y = \varnothing)\ \&\ \forall j\{[(X \subseteq S_j)\ \&\ (\ (S_j = S_{i1}) \vee (S_j = S_{i2}) \vee \ \dots \vee (S_j = S_{ik})\ )] \rightarrow (Y \subseteq S_j)\}$$

т.е. множества X и Y не имеют общих букв, а кроме того всякий раз, когда множество букв X содержится в слове S_j, в этом же слове S_j также содержится и множество букв Y, и при этом подобная совместная вложимость имеет место для всех тех и только тех слов S_j из множества **Ω**, которые перечислены в множестве $\mathbf{\Omega}^{\sim}$.

В таких случаях будем говорить, что имеет место ассоциативная зависимость

$$(X \sim Y)_{\mathbf{\Omega}}\ ,$$

т.е. множества X и Y *ассоциативно* связаны (*ассоциированы*) в *контексте* **Ω**.

o

П 7.2. Определение

Определим две числовые характеристики ассоциативных зависимостей вида $(X \sim Y)_{\mathbf{\Omega}}$:

(1) $\boldsymbol{supp}(X \sim Y)_{\mathbf{\Omega}} = \left| \mathbf{\Omega}_{X \cup Y} \right|$, где

$$\mathbf{\Omega}_{X \cup Y} = \{\ S_j \mid [(X \subseteq S_j)\ \&\ (Y \subseteq S_j)\ \&\ (S_j \in \mathbf{\Omega})]\ \}$$

т.е. $\boldsymbol{supp}(X \sim Y)_{\mathbf{\Omega}}$ отображает число таких элементов S_j в множестве слов **Ω**, которые содержат одновременно множества букв X и Y.

(2) $$\boldsymbol{conf}(X \sim Y)_{\mathbf{\Omega}} = \frac{\boldsymbol{supp}(X \sim Y)_{\mathbf{\Omega}}}{\boldsymbol{supp}(X)_{\mathbf{\Omega}}}$$

т.е. $\boldsymbol{conf}(X \sim Y)_{\mathbf{\Omega}}$ отображает долю содержащих одновременно множества букв X и Y элементов S_j в множестве слов **Ω** по отношению к доле всех слов из **Ω**, которые содержат множество букв X.

o

П 7.3. Определение

(1) Класс ассоциаций определяется каждой парой множеств $<\mathbf{U}^{\Diamond},\mathbf{\Omega}^{\Diamond}>$, где

$$\mathbf{U}^{\Diamond}= \{a^{\Diamond}_{\mathbf{1}}, a^{\Diamond}_{\mathbf{2}} , \ , a^{\Diamond}_{\mathbf{p}_{\Diamond}}\} \subseteq \mathbf{U} ,$$

$$\mathbf{\Omega}^{\Diamond} = \{S_{\mathbf{j1}}, S_{\mathbf{j2}}, \ , S_{\mathbf{jq}_{\Diamond}}, \} \subseteq \mathbf{\Omega} ,$$

и при этом:

$\mathbf{U}^{\Diamond}$ есть множество *всех* букв из исходного алфавита **U**, которые одновременно входят во все слова $S_{\mathbf{ji}}$ из множества $\mathbf{\Omega}^{\Diamond}$, а других (не входящих в $\mathbf{\Omega}^{\Diamond}$) слов $S_{\mathbf{0}}$ в множестве **Ω,** для которых все буквы из $\mathbf{U}^{\Diamond}$ также вкладываются в $S_{\mathbf{0}}$, в множестве **Ω** больше нет (т.е. $\mathbf{U}^{\Diamond}$ перечисляет те и только те буквы из исходного алфавита **U**, которые одновременно входят в те и только те слова из исходного множества слов **Ω**, которые собраны в сопоставленном множеству букв $\mathbf{U}^{\Diamond}$ множестве слов $\mathbf{\Omega}^{\Diamond}$).

(2) Будем называть каждую такую (см. (1) выше) пару $<\mathbf{U}^{\Diamond},\mathbf{\Omega}^{\Diamond}>$ *основанием* класса ассоциаций $\boldsymbol{A}(\mathbf{U}^{\Diamond},\mathbf{\Omega}^{\Diamond})$, все элементы которого – ассоциативные зависимости вида $(X^{\Diamond}\sim Y^{\Diamond})$ – определяются множеством всех пар вида $<X^{\Diamond},Y^{\Diamond}>$, где

$$(X^{\Diamond} \cup Y^{\Diamond}) \subseteq \mathbf{U}^{\Diamond} \text{ , и при этом } (X^{\Diamond} \cap Y^{\Diamond}) = \varnothing,\ X^{\Diamond} \neq \varnothing \text{ и } Y^{\Diamond} \neq \varnothing.$$

o

Следует учесть, что для заданной пары исходных данных $<\mathbf{U},\mathbf{\Omega}>$ могут существовать пересекающиеся классы ассоциаций. Пример: пусть

$$\mathbf{U} = \{a,b,c,d_1,d_2\},$$

$$\mathbf{\Omega} = \mathbf{\{}\ \{ a,b,c,d_1\},\ \{ a,b,c,d_2\},\ \{ a,b\}\ \mathbf{\}},$$

тогда в классе ассоциаций $\boldsymbol{A}(\mathbf{U}^{\Diamond 1},\mathbf{\Omega}^{\Diamond 1})$, где

$$\mathbf{U}^{\Diamond 1} = \mathbf{U}^{\Diamond} \text{ , a } \mathbf{\Omega}^{\Diamond 1} = \mathbf{\{}\ \{ a,b,c,d_1\},\ \{ a,b,c,d_2\}\ \mathbf{\}}$$

имеется ассоциативная зависимость $\boldsymbol{\varphi_1}$ вида

$$a\sim b$$

с характеристикой ***supp*** $(\boldsymbol{\varphi_1}) = 2$.

В свою очередь, в классе ассоциаций $\boldsymbol{A}(\mathbf{U}^{\Diamond 2},\mathbf{\Omega}^{\Diamond 2})$, где

$$\mathbf{U}^{\Diamond 1} = \mathbf{U}^{\Diamond} \text{ , a } \mathbf{\Omega}^{\Diamond 1} = \mathbf{\{}\ \{ a,b,c,d_1\},\ \{ a,b,c,d_2\},\ \{ a,b\}\ \mathbf{\},}$$

имеется ассоциативная зависимость $\boldsymbol{\varphi_2}$ вида

$$a\sim b$$

с характеристикой ***supp*** $(\boldsymbol{\varphi_2}) = 3$.

Можно показать, что структуру пересечений классов ассоциаций для заданной пары исходных множеств $<\mathbf{U},\mathbf{\Omega}>$ отражают:

- *Диаграмма* вложимости неподвижных точек *Замыканий Галуа*, сформированных на исходном алфавите **U** и множестве примеров **Ω (**см. **Разделы 5.2-4**);
- *Псевдо-Дерево* (восстановленное по буквам – образующим его корня) – см. также **Разделы 5.2-4** ;
- *Каркас* этого Псевдо-Дерева, а также последовательность дополнений этого каркаса до соответствующего Псевдо-Дерева (см. там же).

Обратимся к анализу характеризующих ассоциативные зависимости комбинаторных свойств параметров ***supp***$(X)_{\mathbf{\Omega}}$ и ***conf***$(X\sim Y)_{\mathbf{\Omega}}$.

П 7.4. Определение

Задачи 1, **2** и **3** определим следующим образом:

Дано: < **U** , **Ω** > - алфавит и некоторое множество слов в этом алфавите, а также натуральное число $N_0 \geq 2$ [1].

Найти: существует ли в множестве $\boldsymbol{A}_{<U,\Omega>}$ всех ассоциаций вида $(X\sim Y)_{\Omega}$ такая ассоциация

$$(X_0\sim Y_0)_{\Omega},$$

для которой:

(**Задача 1**) $\boldsymbol{supp}(X_0\sim Y_0)_{\Omega} \leq N_0$.
(**Задача 2**) $\boldsymbol{supp}(X_0\sim Y_0)_{\Omega} \geq N_0$.
(**Задача 3**) $\boldsymbol{supp}(X_0\sim Y_0)_{\Omega} = N_0$.

o

П 7.5. Утверждение

Задачи 1 и **2** полиномиально разрешимы.

Доказательство:

1. Построим по исходному множеству слов **Ω** модифицирующее его множество $\boldsymbol{\Omega}^{\Delta}$, добавив в нем к каждому входящему в множество **Ω** слову S_j еще один экземпляр S_j.: $\boldsymbol{\Omega}^{\Delta} = \boldsymbol{\Omega} \cup \boldsymbol{\Omega}$.
2. Найдем все попарные пересечения слов из $\boldsymbol{\Omega}^{\Delta}$, которых будет ровно

$$m(m-1)\backslash 2$$

штук. Каждый непустой элемент $Z_{ij} = S_i \cap S_j$ этого множества попарных пересечений при наличии в нем не менее двух букв из алфавита **U**

$$|Z_{ij}| \geq 2$$

порождает соответствующий класс ассоциаций – $\boldsymbol{A}^{(2)}_{<U,\Omega>}$, в котором каждая из принадлежащих этому классу ассоциативных зависимостей строится выделением из Z_{ij} двух непустых непересекающихся подмножества букв X_{ij} и Y_{ij}:

$$(X_{ij} \cup Y_{ij}) \subseteq Z_{ij},$$

однако $X_{ij} \cap Y_{ij} = \varnothing$, при $X_{ij} \neq \varnothing$ и $Y_{ij} \neq \varnothing$.

Если сформированное множество $\{Z_{ij}\}$ всех таких Z_{ij} пусто:

$$\{Z_{ij}\} = \varnothing,$$

или же содержит лишь однобуквенные подмножества, то **Задача 1** для заданных **U** и **Ω** решений не имеет.

3. Каждому слову S_j из **Ω** сопоставим букву b_j нового алфавита $\mathbf{U}^*$

$$\mathbf{U}^* = \{ b_1 , b_2 , \ , b_m \}$$

и сформируем новое множество слов $\boldsymbol{\Omega}^*$

$$\boldsymbol{\Omega}^* = \{S^*_1, S^*_2 , \ldots, S^*_n\},$$

положив каждое слово S^*_j множеством всех таких b_{ij} , что $a_i \in S_j$ в **Ω**.

4. Выполнив с новым множеством слов $\boldsymbol{\Omega}^*$ те же процедуры, что описаны выше в п.п. 1-2, получим:

- набор некоторых решений **Задачи 2** - некоторых классов ассоциаций в **Ω** , для которых

$$\boldsymbol{supp}(X_0\sim Y_0)_{\Omega} \geq N_0,$$

либо

[1] Нетрудно убедиться, что при $N_0 = 1$ решение обсуждаемых задач носит тривиальный характер.

- подтверждение того, что таких ассоциаций на множестве **Ω** попросту нет (при пустоте соответствующих попарных пересечений, либо отсутствии среди попарных пересечений слов из $\mathbf{\Omega}^*$ такого, которое содержало бы не менее чем N_0 букв из нового алфавита $\mathbf{U}^*$).

o

П 7.6. Утверждение.

Задача 3 лежит в классе *NP*-полных задач – ***NPC***.

Доказательство:
Демонстрируется сводимость к изучаемой задаче доказуемо принадлежащей к классу ***NPC*** задачи о ***ВЫПОЛНИМОСТИ БУЛЕВСКОЙ ФУНКЦИИ*** (см. подробности в **Разделе 4.2 Главы 4**)

o

П 7.7. Определение

Задача 4 определяется следующими условиями:

Дано: < **U** , **Ω** > - алфавит и некоторое множество слов в этом алфавите.

Найти: сколько элементов - ассоциаций вида $(X\sim Y)_{\mathbf{\Omega}}$ - существует в множестве $\boldsymbol{A}_{<\mathbf{U},\mathbf{\Omega}>}$ всех ассоциаций, порождаемых на заданных **U** и **Ω** .

o

П 7.8. Утверждение

Задача 4 лежит в классе ***#PC*** перечислительно полных задач.

Доказательство:
Демонстрируется сводимость к изучаемой задаче доказуемо принадлежащей к классу ***#PC*** задачи о ***ВЫПОЛНИМОСТИ МОНОТОННОЙ БУЛЕВСКОЙ ФУНКЦИИ*** в виде 2-КНФ. (см. подробности в **Разделе 4.2 Главы 4**).

o

П 7.9. Определение

Пусть $\mathbf{U} = \{a_1, a_2, \dots, a_n\}$ – алфавит, а $\mathbf{\Omega} = \{S_1, S_2, \dots, S_m\} \subseteq 2^{\mathbf{U}}\backslash\varnothing$ есть множество слов (т.е. непустых множеств букв из алфавита **U).** Рассмотрим алфавит $\mathbf{U^*} = \{a^*_1, a^*_2, \dots, a^*_m\}$ и множество объектов

$$\mathbf{\Omega^*} = \{S^*_1, S^*_2, \dots, S^*_n\} \subseteq 2^{\mathbf{U^*}}\backslash\varnothing ,$$

построенных над алфавитом **U*** по следующим правилам:

- пусть $S_i = \{a_{i1}, a_{i2}, \dots, a_{ip}\} \subseteq \mathbf{U}$ есть i-тый элемент множества слов **Ω:** $S_i \in \mathbf{\Omega}$. **;**
- каждому слову S_i из множества **Ω** сопоставляется буква a^*_i из алфавита **U***, а каждому слову S^*_i из множества **Ω*** - буква a_i из алфавита **U**;
- каждое слово S^*_i в множестве **Ω*** сформировано теми и только теми буквами a^*_i из множества **U***, для которых сопоставленные им слова S_i содержат сопоставленную слову S^*_i букву a_i.

Будем называть множества **U*** и **Ω*** (соответственно) *сопряженным* алфавитом и *сопряженным* множеством слов над ним.

o

Как и ранее - в **Разделе 4.2**, определим пару отображений $\boldsymbol{f}$ и $\boldsymbol{g}$, связывающих множества **U** и **Ω** следующим образом:

$$\boldsymbol{f}: 2^{\mathbf{\Omega}} \to 2^{\mathbf{U}}$$

и для

$$\forall \mathbf{X} = \{S_{i1}, S_{i2}, \dots, S_{ip}\} \in 2^{\mathbf{\Omega}}$$

имеем

$$\boldsymbol{f}(\mathbf{X}) = S_{i1} \cap S_{i2} \cap \dots \cap S_{ip};$$

и, соответственно,

$$\boldsymbol{g} : 2^{\mathbf{U}} \to 2^{\boldsymbol{\Omega}}$$

и для

$$\forall \mathbf{Y} = \{a_{i1}, a_{i2}, \dots , a_{iq}\} = \{S^*_{i1}, S^*_{i2}, \dots , S^*_{iq}\} \in 2^{\mathbf{U}}$$

имеем

$$\boldsymbol{g}(\mathbf{Y}) = S^*_{i1} \cap S^*_{i2} \cap \dots \cap S^*_{iq}$$

Несложно убедиться, что отображения $\boldsymbol{f}$ и $\boldsymbol{g}$ удовлетворяют соотношениям:

$$[\mathbf{X}_1 \subseteq \mathbf{X}_2] \supset [\boldsymbol{f}(\mathbf{X}_2) \subseteq \boldsymbol{f}(\mathbf{X}_1)]$$
$$[\mathbf{Y}_1 \subseteq \mathbf{Y}_2] \supset [\boldsymbol{g}(\mathbf{Y}_2) \subseteq \boldsymbol{g}(\mathbf{Y}_1)],$$

т.е. представляют собой соответствие Галуа [*Гусакова 87, Кон 68*]. При этом отображения

$$\boldsymbol{f}^*\boldsymbol{g}(\mathbf{X}) = \boldsymbol{g}(\boldsymbol{f}(\mathbf{X})) \quad \text{и} \quad \boldsymbol{g}^*\boldsymbol{f}(\mathbf{Y}) = \boldsymbol{f}(\boldsymbol{g}(\mathbf{Y}))$$

являются операторами замыкания, соответственно, на множествах $\boldsymbol{\Omega}$ и $\mathbf{U}$.

П 7.10 Определение[2].

1. Упорядоченное по взаимному вложению подмножеств букв исходного алфавита $\mathbf{U}$ множество $\boldsymbol{GC}(\mathbf{U},\boldsymbol{\Omega})$ всех неподвижных точек оператора $\boldsymbol{g}^*\boldsymbol{f}(Z) = \boldsymbol{f}(\boldsymbol{g}(Z))$[3] будем называть *Диаграммой вложимости замыканий Галуа* для пары $< \mathbf{U} , \boldsymbol{\Omega} >$.

2. *Базисом замыканий Галуа* для заданной пары $< \mathbf{U} , \boldsymbol{\Omega} >$ будем называть множество

$$\{ [a_1]_{<\mathbf{U},\boldsymbol{\Omega}>}, [a_2]_{<\mathbf{U},\boldsymbol{\Omega}>}, \quad , [a_n]_{<\mathbf{U},\boldsymbol{\Omega}>} \}$$

замыканий всех однобуквенных подмножеств текущего алфавита $\mathbf{U}$.

3. *Псевдо-деревом замыканий Галуа* (сформированным на слове S) будем называть поддиаграмму $\boldsymbol{GC}(\mathbf{U},\boldsymbol{\Omega})\big|_S$ для Диаграммы $\boldsymbol{GC}(\mathbf{U},\boldsymbol{\Omega})$, сформированную следующим образом: слово S есть максимальный элемент поддиаграммы $\boldsymbol{GC}(\mathbf{U},\boldsymbol{\Omega})\big|_S$, а ее минимальные элементы – множества (по крайней мере - одноэлементные) букв из $\mathbf{U}$, представляющие собою минимальные по вложению элементы Базиса замыканий Галуа, сформированные для всех входящих в S букв исходного алфавита $\mathbf{U}$.

4. Букву a_j из алфавита $\mathbf{U}$ будем называть *наклеенной* на букву a_i из этого же алфавита $\mathbf{U}$ в Диаграмме $\boldsymbol{GC}(\mathbf{U},\boldsymbol{\Omega})$ тогда и только тогда, когда

$$[a_i]_{<\mathbf{U},\boldsymbol{\Omega}>} \subseteq [a_j]_{<\mathbf{U},\boldsymbol{\Omega}>}, \quad \text{однако } a_j \notin [a_i]_{<\mathbf{U},\boldsymbol{\Omega}>}$$

5. Буквы a_i и a_j из алфавита $\mathbf{U}$ будем называть *склеенными* в Диаграмме $\boldsymbol{GC}(\mathbf{U},\boldsymbol{\Omega})$ тогда и только тогда, когда

$$[a_i]_{<\mathbf{U},\boldsymbol{\Omega}>} = [a_j]_{<\mathbf{U},\boldsymbol{\Omega}>}$$

o

П 7.11. Определение

Задачи 5, 6 и **7** определим следующим образом:

Дано: $< \mathbf{U} , \boldsymbol{\Omega} >$ - алфавит и некоторое множество слов в этом алфавите, а также такое положительное число α, что $1 \geq \alpha > 0$.

Найти: существует ли в множестве $\boldsymbol{A}_{<\mathbf{U},\boldsymbol{\Omega}>}$ всех ассоциаций вида $(X\sim Y)_{\boldsymbol{\Omega}}$ такая ассоциация

$$(X_0\sim Y_0)_{\boldsymbol{\Omega}},$$

для которой:

(**Задача 5**) $\quad \boldsymbol{conf}(X_0\sim Y_0)_{\boldsymbol{\Omega}} \leq \alpha.$

---

[2] См. также соответствующие определения в **Главах 4-5**.

[3] $\mathbf{Z}$ – непустое подмножество букв исходного алфавита $\mathbf{U}$: $\mathbf{Z} \subseteq \mathbf{U}$.

(**Задача 6**) $\quad$ ***conf*** $(X_0 \sim Y_0)_{\Omega} \geq \alpha$.
(**Задача 7**) $\quad$ ***conf*** $(X_0 \sim Y_0)_{\Omega} = \alpha$.

o

П 7.12. Утверждение

В множестве $A_{<U,\Omega>}$ всех ассоциаций, порождаемых на заданных **U** и **Ω** , ассоциация вида $(a_2 \sim a_1)_{\Omega}$, обладающая свойством

$$\textbf{\textit{conf}}\ (a_2 \sim a_1)_{\Omega} = 1,$$

существует тогда и только тогда, когда в исходном алфавите **U** в заданном контексте **Ω** найдется пара букв a_1 и a_2 , таких (см. **Определение П 7.10**), что a_2 оказывается *наклеенной* на a_1 буквой либо a_1 и a_2 - *склеенные* буквы в Диаграмме замыканий Галуа ***GC*(U,Ω)**.

Доказательство:
1. Для случая *склеенных* переменных результат утверждения очевиден.
2. Рассмотрим ситуацию, когда a_2 оказывается *наклеенной* на a_1 буквой.
(<==) Если a_2 наклеена на a_1 в смысле ***GC*(U,Ω)**, то

$$\textbf{\textit{conf}}(a_2 \sim a_1)_{\Omega} = \frac{\textbf{\textit{supp}}(a_2 \sim a_1)_{\Omega}}{\textbf{\textit{supp}}(a_2)_{\Omega}} = \frac{\textbf{\textit{supp}}(a_2 \cup a_1)_{\Omega}}{\textbf{\textit{supp}}(a_2)_{\Omega}} = 1$$

(==>) Если a_2 и a_1 таковы, что $\textbf{\textit{conf}}(a_2 \sim a_1)_{\Omega} = 1$, тогда (при условии, что a_1 и a_2 не являются склеенными буквами алфавита **U** в Диаграмме ***GC*(U,Ω)**, – иначе результат очевиден)

$$1 = \textbf{\textit{conf}}(a_2 \sim a_1)_{\Omega} = \frac{\textbf{\textit{supp}}(a_2 \sim a_1)_{\Omega}}{\textbf{\textit{supp}}(a_2)_{\Omega}} = \frac{\textbf{\textit{supp}}(a_2 \cup a_1)_{\Omega}}{\textbf{\textit{supp}}(a_2)_{\Omega}} = 1$$

таким образом, a_2 наклеена на a_1 в смысле ***GC*(U,Ω)**.

o

П 7.13. Замечание

Наличие *наклеенных* (а также *склеенных*) букв алфавита **U** в Диаграмме ***GC*(U,Ω)** легко (полиномиально быстро) диагностируется при формировании Базиса замыканий Галуа для исходных данных <**U,Ω**>.

o

П 7.14. Утверждение

Задача 5 из **Определения П 7.11** принадлежит к классу полиномиально разрешимых.

Доказательство:
1. Для каждой буквы a_i из **U** построим замыкание $[a_i]_{<U,\Omega>}$.
2. Будем последовательно перебирать слова $S_1, S_2, \ldots, S_m$ из множества **Ω** :
   2.1. Для каждого текущего $S_i = \{a_{i1}, a_{i2}, \ldots, a_{ip}\}$ рассмотрим множество

$$\{ [a_{i1}]_{<U,\Omega>}, [a_{i2}]_{<U,\Omega>}, \ldots, [a_{ip}]_{<U,\Omega>} \}$$

   замыканий однобуквенных подмножеств слова S_i.
   2.2. Выберем из только что построенных замыканий такое $[a^{i}_{0}]_{<U,\Omega>}$, которое порождается максимальным по числу элементов из **Ω** множеством слов $\{S^{i0}_{1}, S^{i0}_{2}, \ldots, S^{i0}_{r}\}$.
   2.3. Рассмотрим ассоциацию $(\{a^{i}_{0}\} \sim \{S_i \setminus \{a^{i}_{0}\}\})_{\Omega}$, которая отражает ассоциативную связь однобуквенного множества $\{a^{i}_{0}\}$ и слова S_i, из которого удалена именно эта буква a^{i}_{0}.
   2.4. Найдем

$$\alpha_i = \frac{supp\ (\ (\{a^i_0\}\sim\{S_i\setminus\{a^i_0\}\})_\Omega\ )}{supp(\{a_{i0}\})_\Omega}$$

2.5. Заметим, что для некоторых ***i*** в множество слов, обеспечивающих поддержку этой зависимости (т.е. формирование значения соответствующего показателя ***supp***), вместе с S_i могут входить также некоторые другие слова из Ω.

3. Найдем по всем $S_i \in \Omega$ минимальное значение α_i :

$$\min_{\Omega}(\alpha_i) = \alpha_0$$

4. Если $\alpha_0 \le \alpha$ то соответствующая этому α_0 ассоциация вида

$$(\{a^{i0}_0\}\sim\{S_{i0}\setminus\{a^{i0}_0\}\})_\Omega$$

и есть ответ[4] на вопрос **Задачи 5**. В противном случае (т.е. при $\alpha_0 > \alpha$) такой ассоциации на исходных данных <**U**,**Ω**> попросту нет.

o

П 7.15. Утверждение

Задача 6 из **Определения П 7.11** принадлежит к классу полиномиально разрешимых.

Доказательство:

1. Вновь воспользуемся инструментарием Базиса замыканий Галуа: для каждой буквы a_i из **U** построим замыкание $[a_i]_{<U,\Omega>}$.
2. Вновь организуем перебор всех S_i из множества Ω :
   2.1. Для каждого $S_i = \{a_{i1}, a_{i2}, \dots, a_{ip}\}$ рассмотрим множество замыканий всех пар букв из этого слова S_i - $\{\ [a^i_1, a^i_2]_{<U,\Omega>}, [a^i_1, a^i_3]_{<U,\Omega>}, \dots, [a^i_{p-1}, a^i_p]_{<U,\Omega>}\ \}$.
   2.2 Для каждой пары $\{a^i_{j1}, a^i_{j2}\}$ входящих в слово S_i букв a^i_{j1} и a^i_{j2} из исходного алфавита **U** таких, что $[a^i_{j1}]_{<U,\Omega>}$ есть минимальный элемент Псевдо_Дерева $GC(\mathbf{U},\Omega)\,|_{S_i}$, вычислим значение

$$\alpha^i_{j1,j2} = \frac{supp\ (\{a^i_{j1}, a^i_{j2}\})}{supp(\{a^i_{j1}\})_\Omega\ .}$$

   Здесь если a^i_{j1} и a^i_{j2} - склеенные буквы, либо a^i_{j2} - наклеенная на a^i_{j1} буква в Диаграмме $GC(\mathbf{U},\Omega)$, то

$$\alpha^i_{j2} = \frac{supp\ (\{a^i_{j1}, a^i_{j2}\})}{supp(\{a^i_{j2}\})_\Omega\ .} = 1,$$

   и требуемое **Утверждение** доказано.
   2.3. Для заданного S_i по всем соответствующим парам $<a^i_{j1}, a^i_{j2}>$ найдем

$$\alpha^i = \max_{\text{по всем парам } <j1,j2>}(\alpha^i_{j1,j2})$$

3. Найдем $\alpha^0 = \max_{\text{по всем } i}(\alpha^i)$.
4. Если $\alpha^0 \ge \alpha$, то соответствующая этому α^0 ассоциация и есть ответ[5] на вопрос **Задачи 6**. В противном случае такой (имеющей соответствующее значение параметра ***conf***) ассоциации в множестве всех ассоциаций, порождаемых на исходных данных < **U** , **Ω** >, попросту нет.

o

П 7.16. Утверждение

Задача 7 из **Определения П 7.11** принадлежит к классу ***NP***-полных задач ***NPC***.

---

[4] В общем случае – один из возможных ответов.

[5] В общем случае – один из таких ответов.

Доказательство:
Демонстрируется сводимость к изучаемой задаче доказуемо уже принадлежащей к классу ***NPC*** **Задачи 3** (см. **Утверждение П 7.6** выше).
1. Каждой входящей в исходный алфавит $\mathbf{U} = \{a_1, a_2, \dots, a_n\}$ букве a_i сопоставим соответствующую букву b_i (при этом считаем, что также как и все a_i, все b_i попарно различны). По исходному множеству слов $\mathbf{\Omega} = \{S_1, S_2, \dots, S_m\}$ сформируем еще одно - модифицирующее его - множество примеров $\mathbf{\Omega}^{\Delta} = \{S^{\Delta}_1, S^{\Delta}_2, \dots, S^{\Delta}_m\}$, где каждое из слов S^{Δ}_j формируется по соответствующему S_j добавлением к вхождению каждой буквы a^j_i также и дополнительной буквы b^j_i : для $S_i = \{a^i_1, a^i_2, \dots, a^i_p\}$ имеем $S^{\Delta}_i = \{a^i_1, a^i_2, \dots, a^i_p, b^i_1, b^i_2, \dots, b^i_p\}$.
2. Будем рассматривать ситуации, когда в множестве слов $\mathbf{\Omega}$ найдутся ровно $\boldsymbol{k}$ штук слов $\{S_{i1}, S_{i2}, \dots, S_{ik}\}$, на которых порождается некоторая ассоциативная зависимость $(X_0 \sim Y_0)_{\mathbf{\Omega}}$, которая имеет показатель

$$\boldsymbol{supp}(X_0 \sim Y_0)_{\mathbf{\Omega}} = \boldsymbol{k} .$$

3. Целевое множество слов $\mathbf{\Omega}^{\Sigma}$ сформируем по следующему правилу: соберем его из $\mathbf{p}$ экземпляров множества $\mathbf{\Omega}$ и $\mathbf{q}$ экземпляров множества $\mathbf{\Omega}^{\Delta}$. Таким образом:

$$\mathbf{\Omega}^{\Sigma} = \{(\mathbf{\Omega}, \ , \mathbf{\Omega})_{\mathbf{p}\text{-раз}}, (\mathbf{\Omega}^{\Delta}, \ , \mathbf{\Omega}^{\Delta})_{\mathbf{q}\text{-раз}}\},$$

при этом $\mathbf{p}$ и $\mathbf{q}$ – минимальные натуральные числа, связанные следующим соотношением:

$$\frac{\boldsymbol{k} \times \boldsymbol{p}}{\boldsymbol{k} \times (\boldsymbol{p}+\boldsymbol{q})} = \alpha. \qquad (!)$$

Покажем, что в множестве $\mathbf{\Omega}$ существует ассоциативная зависимость с показателем

$$\boldsymbol{supp}(X_0 \sim Y_0)_{\mathbf{\Omega}} = \boldsymbol{k} .$$

тогда и только тогда, когда в множестве слов $\mathbf{\Omega}^{\Sigma}$ существует ассоциативная зависимость с показателем

$$\boldsymbol{conf}\ (X^{\Sigma}_0 \sim Y^{\Sigma}_0)_{\mathbf{\Omega}} = \alpha.$$

3.1. (=>) Если в $\mathbf{U}$ и $\mathbf{\Omega}$ найдется соответственно пара таких множеств $\langle \mathbf{U}^{\Diamond}, \mathbf{\Omega}^{\Diamond} \rangle$, что для

$$\mathbf{\Omega}^{\Diamond} = \{S^{\Diamond}_1, S^{\Diamond}_2, \dots, S^{\Diamond}_k\}$$

имеем:

$$\mathbf{U}^{\Diamond} = S^{\Diamond}_1 \cap S^{\Diamond}_2 \cap \dots \cap S^{\Diamond}_k$$

при этом $\mathbf{U}^{\Diamond} \neq \varnothing$ и для всякого $S_0 \in (\mathbf{\Omega} \setminus \mathbf{\Omega}^{\Diamond})$ имеет место

$$S_0 \cap \mathbf{U}^{\Diamond} \subset \mathbf{U}^{\Diamond}$$

тогда в множестве слов $\mathbf{\Omega}^{\Sigma}$ (в силу процедуры его построения) также найдется ровно $(\boldsymbol{k} \times (\boldsymbol{p}+\boldsymbol{q}))$ слов, содержащих одновременно все буквы из множества $\mathbf{U}^{\Diamond}$, и ровно $(\boldsymbol{k} \times \boldsymbol{p})$ слов, содержащих вместе с каждой $\mathbf{a}^{\Diamond}_i$ из $\mathbf{U}^{\Diamond}$ также ее собственного «дублера» $\mathbf{b}^{\Diamond}_i$. Таким образом, если $\mathbf{Y}^{\Diamond}$ есть множество всех этих $\mathbf{b}^{\Diamond}_i$, для которых $\mathbf{a}^{\Diamond}_i$ входят в $\mathbf{U}^{\Diamond}$, то для ассоциативной зависимости $(\mathbf{U}^{\Diamond} \sim \mathbf{Y}^{\Diamond})_{\mathbf{\Omega}^{\Sigma}}$ имеем показатель

$$\boldsymbol{conf}\ (\mathbf{U}^{\Diamond} \sim \mathbf{Y}^{\Diamond})_{\mathbf{\Omega}^{\Sigma}} = \frac{(\boldsymbol{k} \times \boldsymbol{p})}{(\boldsymbol{k} \times (\boldsymbol{p}+\boldsymbol{q}))} = \alpha. \qquad (!!)$$

3.2. (<=). Если в множестве слов $\mathbf{\Omega}^{\Sigma}$ нашлась ассоциативная зависимость $(\mathbf{U}^{\Diamond} \sim \mathbf{Y}^{\Diamond})_{\mathbf{\Omega}^{\Sigma}}$ со свойствами, определяемым показателем (**!!**), то в силу использованного способа построения множества $\mathbf{\Omega}^{\Sigma}$ (путем соответствующим образом «взвешенного» смешивания требуемых «порций» множеств $\mathbf{\Omega}$ и $\mathbf{\Omega}^{\Delta}$) то в $\mathbf{\Omega}^{\Sigma}$ должно найтись ровно $(\boldsymbol{k} \times \boldsymbol{p})$ слов, содержащих вместе со всеми буквами $\mathbf{a}_i$ из

некоторого множества $\mathbf{U^a}$ одновременно и ровно такое же множество $\mathbf{U^b}$ букв-«дублеров» $\mathbf{b_i}$, а вместе с этим и ровно $\boldsymbol{k}$ x ($\boldsymbol{p}$+$\boldsymbol{q}$) слов, содержащих только буквы $\mathbf{a_i}$ из множества $\mathbf{U^a}$. Таким образом, выбрав множество $\mathbf{U^a}$ в качестве искомого множества букв $\mathbf{U^{\Diamond}}$, а из множества всех слов из $\mathbf{\Omega^{\Sigma}}$, содержащих только буквы из $\mathbf{U^a}$ (и не содержащих буквы из множества $\mathbf{U^b}$), а затем для каждых $\boldsymbol{p}$ копий соответствующего слова оставив лишь один «эталонный» экземпляр, мы получим множество $\mathbf{\Omega(U^a)}$, такое что пара <$\mathbf{U^a}$, $\mathbf{\Omega(U^a)}$> и образует *основание* соответствующего класса ассоциации с $\left|\mathbf{\Omega(U^a)}\right| = \boldsymbol{k}$, что и будет означать, что для любых непустых и не пересекающихся X и Y – подмножеств множества $\mathbf{U^a}$ получим

$$\boldsymbol{supp}(X\sim Y)_{\mathbf{\Omega(Ua)}} = \boldsymbol{k}\ .$$

4. Нетрудно убедиться (по построению множества $\mathbf{\Omega^{\Sigma}}$ из множеств $\mathbf{\Omega}$ и $\mathbf{\Omega^{\Delta}}$ путем «дублирования»[6] вхождений соответствующих букв в соответствующие слова из $\mathbf{\Omega^{\Delta}}$),что между соответствующими множествами

- $\{S_{i1}, S_{i2}, \ldots , S_{ik}\}$ и
- сопоставленным ему в $\mathbf{\Omega^{\Sigma}}$ «*композитом*», который содержит $\boldsymbol{p}$ экземпляров этого же множества $\{S_{i1}, S_{i2}, \ldots , S_{ik}\}$ а также $\boldsymbol{q}$ экземпляров сформированного по представленному выше способу множества $\{S^{\Delta}_{i1}, S^{\Delta}_{i2}, \ldots , S^{\Delta}_{ik}\}$,

в части наличия «связанных» пар классов ассоциации, имеющих *основания* соответствующих[7] размеров, имеет место взаимно однозначное соответствие.

o

Таким образом, также справедливо

П 7.17. Следствие

Задача 7 лежит в классе ***#PC*** перечислительно полных задач.

o

* * *

В заключение отметим, что для заданной пары <**U**,$\mathbf{\Omega}$> алгоритмика эффективного поиска *ассоциативных зависимостей* заданного вида (например, зависимостей,

- содержащих *конкретные комбинации букв* из заданного подмножества алфавита **U**,
- удовлетворяющих тем или иным *количественным ограничениям* на структуры множеств букв, связываемых порождаемыми ассоциациями,
- удовлетворяющих тем или иным *ограничениям на вхождение* в множество «слов поддержки» порождаемого класса ассоциаций тех или иных *конкретных элементов* исходного множества слов $\mathbf{\Omega}$, или же, например, групп таких слов, удовлетворяющих заданным ограничениям на *количественные размеры этих групп*, и др.)

может быть построена на основе уже представленной нами в выше в **Главе 5** техники формирования соответствующих Диаграмм взаимной вложимости замыканий Галуа, в основе которой – формирование и использование соответствующих Базисов замыканий Галуа а также методы и алгоритмы восстановления Псевдо-Деревьев, образующих рассматриваемые Диаграммы. Как уже было продемонстрировано ранее в **Главе 5**, такие Диаграммы строятся в процессе целенаправленного и контролируемого применения специально разработанных процедур последовательной инкрементальной (по-шаговой) реконструкции по их Каркасам[8] всех вкладывающихся в эти Диаграммы Псевдо_Деревьев.

---

[6] Т.е. замены вхождения каждого $\mathbf{a_i}$ на вхождение пары $\mathbf{a_i,b_i}$.

[7] Регулируемых соотношением (**!**).

[8] При этом каждый такой Каркас – это полиномиально быстро формируемая структура.

Приложение 8

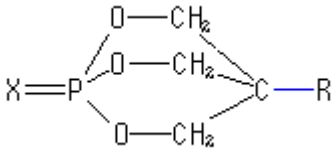
ДСМ-ИАД: как это работает

П 8.1. ДСМ-метод на примере.

Теперь мы дополним построения **Главы 3** подробным обсуждение конкретного примера. Предметом нашего рассмотрения будут структура ДСМ-рассуждения и средства представления знаний в задаче прогноза токсичности одного класса противосудорожных препаратов[1].

Итак, наша цель - проиллюстрировать основные аспекты ДСМ-технологии на примере причинного анализа и прогноза свойств физиологически активных веществ. Иллюстративный материал - результаты компьютерных экспериментов по изучению токсичности некоторого специального класса химических структур, проявляющих противосудорожную активность.

Исходные данные - 36 соединений из ряда так называемых производных 1-фосфо-2,6,7-три-оксабицикло[2.2.2]октанов. Общий вид изучаемых структур



,

а точный вид радикалов X и R для каждого из 36 рассматриваемых соединений вместе с обсуждением их свойств можно найти в **Таблице П.2.3.1** ниже, в **Приложении 8.2** (см. также работу [*Блинова 95*]). Предметом компьютерного анализа (средствами ДСМ-рассуждения) здесь была токсичность исследуемых соединений: 13 из 36 изучаемых веществ были представлены как токсичные (они составили множество $\boldsymbol{Exm^{+}}$), 20 - как нетоксичные (они составили множество $\boldsymbol{Exm^{-}}$), а для оставшихся 3 было необходимо построить прогноз наличия или отсутствия искомого свойства - токсичности (они составили множество $\boldsymbol{Exm^{\tau}}$).

В общих чертах (следуя схеме, представленной в **Разделе 3.2**) описание ДСМ-технологии может быть сведено к следующим основным шагам:

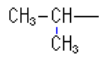
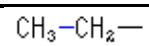
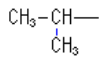
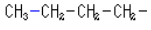
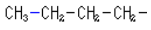
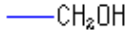
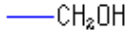
- применением ППВ типа **I** осуществляется анализ структурного сходства позитивных примеров (активных соединений);
- применением ППВ типа **I** осуществляется анализ структурного сходства негативных примеров (неактивных соединений);
- применением ППВ типа **I** осуществляется порождение позитивных и негативных гипотез (т.е. выделение сходств; проверка их “непротиворечивости”; наконец, проверка дополнительные условий, определяемых соответствующими решающими предикатами - см. условия ***cn8.1,...,cn8.6*** в **Разделе 3.3**);
- выполняется проверка условий критерия достаточности оснований для осуществления правдоподобного вывода в рамках ДСМ-рассуждения;
- средствами ППВ типа **II** осуществляется прогноз свойств новых (предложенных для прогноза) соединений (т.е. анализ переносимости на новые соединения эмпирических зависимостей, представленных порожденными гипотезами и характеризующих “причинность” наличия и отсутствия целевых свойств у изучаемых соединений).

---

[1] Приводимые построения будут здесь играть роль лишь иллюстративного (хотя и вполне реального) материала для представленной в **Разделах 3.2-3** формальной конструкции. Подробное описание исходного эксперимента в целом дано в **Приложении 8.2**.

Перейдем к анализу структуры ДСМ-рассуждения. Рассмотрим (в качестве стартовой) выборку, образованную 8 соединениями[2] из представленной ниже **Таблицы П 8.1.1**:

Таблица П 8.1.1

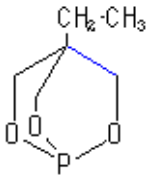
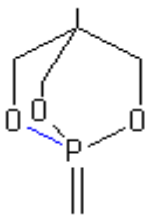
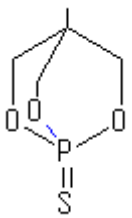
	R	***X***
C1) токсично		O
C2) токсично		S
C3) токсично		Se
C4) токсично		S
C6) *требует прогноза*		O
C7) нетоксично		O
C8) нетоксично		S
C9) нетоксично		Se
C10) нетоксично		S
C12) *требует прогноза*		••[3]

Нетрудно убедиться, что структурное сходство (понимаемое в смысле Миллевского метода сходства, формализованного средствами ДСМ-рассуждения - см. выше **Разделы 3.2 - 3**) позитивных примеров (токсичных химических соединений из исходной выборки) выражают (см. ниже **Таблицы П 8.1.2 - 3**) подструктуры ***Hp1***, ***Hp2***, ***Hp3*** и ***Hp4***. При этом, сходство

- примеров из множества {***C1-C3***} представляет[4] подструктура ***Hp1***,
- примеров из множества {***C1-C4***} представляет подструктура ***Hp2***,
- примеров из множества {***C2,C4***} представляет подструктура ***Hp3***

следующего вида:

Таблица П 8.1.2

Hp1) токсично		**Hp2) токсично**		**Hp3) токсично**	

[2] При этом каждое из них закодировано в так называемом Фрагментарном Коде Суперпозиции Подструктур - ФКСП, представляющем собой специальный проблемно-ориентированный язык дескрипторного типа для представления знаний о биохимических свойствах физиологически активных веществ (см., например, работы [*Авидон 74*, *Лейбов 91*] и др.).

[3] Таким образом обозначается ситуация, когда X есть ∅ (а смежный с ним фосфор имеет соответствующую валентность).

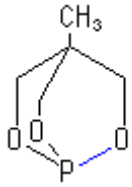
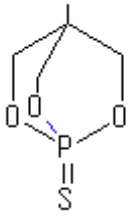
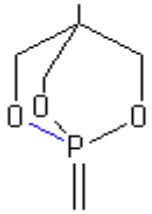
[4] С точностью до кодирования исходных данных в языке ФКСП - здесь и везде ниже до особо объявленного случая перехода к другому языку представления знаний о связи ХИМИЧЕСКАЯ СТРУКТУРА ==> ФИЗИОЛОГИЧЕСКИЕ АКТИВНОСТИ веществ.

Таким же образом структурное сходство негативных примеров (нетоксичных соединений) выражают подструктуры ***Hp4***, ***Hp5*** и ***Hp6***. При этом сходство

- примеров из множества {***C7-C9***} представляет подструктура ***Hp4***,
- примеров из множества {***C7-C10***} представляет подструктура ***Hp5***
- примеров из множества {***C8,C10***} представляет подструктура ***Hp6***

следующего вида:

Таблица П 8.1.3

Hp4) нетоксично		**Hp5) нетоксично**		**Hp6) нетоксично**	

Легко проверить, что подструктура ***Hp1*** не позволяет построить прогноз токсичности соединения ***C6*** (из-за различий в радикалах[5] $\mathbf{CH_2\text{-}CH_3}$ в ***Hp1*** и $\mathbf{C_4H_9}$ в ***C6*** нет вложимости ***Hp1*** в ***C6***).

Также не составляет труда проверить, что подструктура ***Hp4*** не позволяет сделать заключение об отсутствии токсичности у соединения ***C12*** (из-за различий в радикалах $\mathbf{CH_3}$ в ***Hp4*** и $\mathbf{CH_2OH}$ в ***C12*** нет вложимости ***Hp4*** в ***C12***).

Можно показать, что фрагмент ***Hp2*** не может быть использован в доопределении свойств структуры ***C6***, так как порождается сходством не только положительных, но и отрицательных примеров из множества {***C7-C10***}, представленным фрагментом ***Hp6*** (т.е. в терминах ДСМ-рассуждения выражает противоречивое сходство примеров и контрпримеров[6]). По тем же соображениям фрагмент ***Hp6*** не может быть использован в доопределении свойств структуры ***C12*** (т.к. также представляет противоречивое сходство).

По аналогичным соображениям фрагмент ***Hp3*** также не может быть использован в доопределении свойств структуры ***C6***, так как выражает противоречивое сходство примеров и контрпримеров (порождается сходством не только положительных примеров из множества {***C2, C4***}, но и совпадает с фрагментом ***Hp5***, выражающим сходство отрицательных примеров из множества {***C8, C10***}). По тем же соображениям и фрагмент ***Hp5*** не может быть использован в доопределении свойств структуры ***C12*** (т.к. также представляет противоречивое сходство).

Итак, процесс ДСМ-фальсификации[7] результатов выделения позитивных и негативных сходств приводит к следующим результатам:

- множество непротиворечивых позитивных гипотез, порождаемых на исходной обучающей выборке **C** = {***C1-C4, C7-C10***} имеет вид $\mathbf{H^+}$ = {***Hp1***};
- множество непротиворечивых негативных гипотез, порождаемых на исходной обучающей выборке **C** = {***C1-C4, C7-C10***} имеет вид $\mathbf{H^-}$ = {***Hp4***}.

Следующий шаг ДСМ-рассуждения - проверка условий критерия достаточности оснований для осуществления правдоподобного вывода в рамках ДСМ-метода. Как уже говорилось, этот критерий требует, чтобы порожденные в процессе обучения на примерах (и

---

[5] Воспринимаемых кодом ФКСП как единые комплексы (см. в [*Авидон 74, Лейбов 91*] характеристику учитываемых кодом ФКСП активных центров).

[6] Фактически, этот пример иллюстрирует реализованный в ДСМ-методе механизм фальсификации (в смысле К.Поппера) порождаемых анализом сходства гипотез.

[7] Т.е. выделения “эмпирических противоречий”.

контрпримерах) гипотезы “объясняли” (т.е. позволяли осуществить соответствующий ДСМ-прогноз) свойства всех примеров из исходной обучающей выборки.

Легко видеть, что условия этого критерия не выполняются ни для одного из построенных множеств $\mathbf{H}^+$ и $\mathbf{H}^-$ (позитивных и негативных гипотез):

- позитивная гипотеза ***Hp1*** не может быть использована для доопределения свойств входящего в исходную обучающую выборку позитивного примера (токсичного соединения) ***C4***: из-за различий в радикалах $\mathbf{CH_2\text{-}CH_3}$ в ***Hp1*** и $\mathbf{C_4H_9}$ в ***C4*** нет вложимости ***Hp1*** в ***C4***;
- негативная гипотеза ***Hp4*** не может быть использована для доопределения свойств входящего в исходную обучающую выборку негативного примера (нетоксичного соединения) ***C10***: из-за различий в радикалах $\mathbf{CH_3}$ в ***Hp4*** и $\mathbf{CH_2OH}$ в ***C10*** нет вложимости ***Hp4*** в ***C10.***

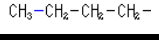
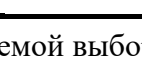
Наблюдаемая ситуация “необъясняемости” позитивного примера ***C4*** и негативного примера ***C10*** порождаемыми из выборки **C** = {***C1-C4,C7-C10***} гипотезами (элементами, соответственно, множеств $\mathbf{H}^+$ и $\mathbf{H}^-$) есть следствие их “уникальности”[8] в используемом ДСМ-системой для обучения материале.

Теперь, следуя стандартным требованиям организации ДСМ-рассуждения, исходная обучающая выборка должна быть пополнена новыми примерами, сходными, соответственно, с ***C4*** и ***C10***. Процесс пополнения не должен прекращаться по крайней мере до тех пор, пока не будет выполнено условие критерия достаточности оснований для осуществления правдоподобного вывода в рамках ДСМ-рассуждения.

В демонстрационных целях мы реализуем такое пополнение за один шаг.

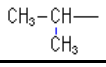
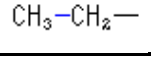
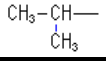
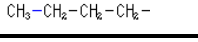
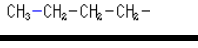
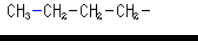
Пополним стартовую выборку соединений двумя новыми структурами - токсичным соединением ***C5*** и нетоксичным соединением ***C11*** :

Таблица П 8.1.4

	R	***X***
C5) токсично		Se
C11) нетоксично		O

Таким образом, расширенный вариант изучаемой выборки будет иметь следующий вид:

Таблица П 8.1.5

	R	***X***
C1) токсично		O
C2) токсично		S
C3) токсично		Se
C4) токсично		S
C5) токсично		Se
C6) *требует прогноза*		O

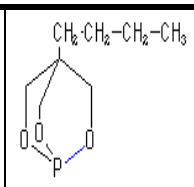
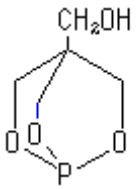
[8] Т.е. отсутствия (выражаемого какими-либо гипотезами) сходства с другими аналогичными по свойствам соединениями, входящими в исходную обучающую выборку.

C7) нетоксично	CH_3-	O
C8) нетоксично	CH_3-	S
C9) нетоксично	CH_3-	Se
C10) нетоксично	$-CH_2OH$	S
C11) нетоксично	$-CH_2OH$	O
C12) ***требует прогноза***	$-CH_2OH$	••[9]

В расширенном варианте изучаемой выборки (в дополнение к представленным выше структурным фрагментам)

- на позитивных примерах ***C4*** и ***C5*** порождается структурный фрагмент ***Hp7***, а
- на негативных примерах ***C10*** и ***C11*** порождается структурный фрагмент ***Hp8***:

Таблица П 8.1.6

Hp7) токсично		**Hp8) нетоксично**	

Нетрудно проверить, что и ***Hp7*** и ***Hp8*** представляют собой непротиворечивые сходства, соответственно, позитивных и негативных примеров из обучающей выборки, т.е. могут быть внесены в расширения $\mathbf{H^+_{ext}}$ и $\mathbf{H^-_{ext}}$ ранее порожденных множеств гипотез $\mathbf{H^+}$ и $\mathbf{H^-}$. Для расширенных множеств гипотез $\mathbf{H^+_{ext}}$ = {***Hp1,Hp7***} и $\mathbf{H^-_{ext}}$ = {***Hp4,Hp8***} выполняются условия критерия достаточности оснований для осуществления правдоподобного вывода в рамках ДСМ-метода: действительно, ***Hp7*** “доопределяет” в расширенной обучающей выборке соединения ***C4*** и ***C5***, а ***Hp8*** - соединения ***C10*** и ***C11***.

Теперь мы располагаем всем необходимым для того, чтобы осуществить прогноз свойств соединений ***C6*** и ***C12*** средствами ДСМ-рассуждения. Вложимость в ***C6*** гипотезы ***Hp1*** из $\mathbf{H^+_{ext}}$ (и невложимость[10] ***Hp4*** и ***Hp8*** из $\mathbf{H^-_{ext}}$) позволяют сделать в рамках ДСМ-рассуждения вывод о токсичности соединения ***C6***. Аналогичное рассуждение, учитывающее вложимость ***Hp8*** вместе с невложимостью ***Hp1*** и ***Hp7***, позволяет сделать вывод о нетоксичности соединения ***C12***. Итак, построен непротиворечивый прогноз свойств двух “неизученных” соединений, осуществленный в условиях достаточности оснований для правдоподобного вывода средствами ДСМ-рассуждения. Построенный нами прогноз полностью соответствует сведениям о свойствах (т.е. о токсичности) этих соединений, содержащимся в работе [*Блинова 95*].

[9] Таким образом обозначается ситуация, когда X есть ∅ (а смежный с ним фосфор имеет соответствующую валентность).

[10] Напомним, что в терминах ФКСП комплекс **-CH_3**, соединенный с соответствующей циклической частью (см. структуры обсуждаемых графов) не вкладывается в соединение комплекса **-CH_2-CH_2-CH_2-CH_3** с этой же циклической частью.

П 8.2. К архитектуре организации промышленных ДСМ-вычислений.

Задача разработки промышленных ДСМ-ИАД-вычислителей (т.е. проблемно-ориентированных программно-технических комплексов, способных восстанавливать эмпирические зависимости из накапливаемых экспериментальных данных и поддерживать процедуры принятия управленческих решений, например, процесс надежного[1] управления операционной деятельностью за счет оперативной идентификации текущих угроз и формирования рекомендаций Лицам, Принимающим Решения, по выбору эффективных средств противодействия таким угрозам и рискам) требует уточнения типовой архитектуры подобных программно-технических систем и, в частности, – создания перечня «типовых» программных модулей (компонентов прикладной ИАД-системы) а также описания основных процедур их взаимодействия (интеграции[2] в единую программно-техническую среду).

Один из наиболее ярких примеров подобной проблематики – задачи мониторинга ИТ-ресурсов и сервисов (проблематика так называемых *IT Resource* and *Service Management* – ***ITRM\ITSM***), ориентированные на оперативную идентификацию и устранение сбоев в нормальной работе сложных программно-технических систем[3]. С методической точки зрения критически важных проблем здесь – две:

- огромные *объемы* накапливаемых *данных* (см., например, специализированные технологические базы данных, в которых журналируется текущая операционная информация – многопараметрические логи, трэйсы, …) и
- необходимость выделять *каузальные факторы*, совместным воздействием которых и обусловлено возникновение той или иной нештатной ситуации. Не имея представления о полном наборе таких факторов в каждой конкретной ситуации, вообще говоря, нет возможности организовать *эффективный* процесс *противодействия* актуальным рискам и угрозам непрерывности нормального функционирования соответствующего программно-технического комплекса.

Для первой из только что названных проблем характерны три критичные особенности:

- в силу очевидной производственной необходимости процесс анализа текущих данных необходимо организовать в режиме *реального* времени;
- для большинства промышленных задач такого типа характерна ситуация класса *Big Data* (объемы релевантных анализируемой ситуации данных весьма велики и постоянно пополняются по результатам текущей операционной деятельности);
- относительное[4] количество сбоев в промышленном оборудовании ведущих индустриальных производителей настолько мало, что для анализа, например, наиболее популярными в инженерной среде – статистическими - средствами эти «коллекции» сбоев представляет собою так называемые *статистически незначимые* выборки (т.е. не позволяет порождать на них статистически значимые результаты).

В свою очередь для второй проблемы существенен следующий набор обстоятельств:

- огромные объемы имеющихся (накапливаемых в технологических ITRM\ITSM базах) данных не оставляет иных возможностей для их (по крайней мере - первичного) анализа кроме использования статистических средств. Наиболее распространенный здесь прием – восстановление тех или иных регрессионных зависимостей в накопленных данных. Однако, такая математическая техника порождает лишь *объяснительные* (а не *каузальные*) модели: результатом восстановления той или иной регрессии в данных оказывается построение *приближенного описания* накопленного

---

[1] В смысле сохранения непрерывности бизнеса (бесперебойной работы оборудования, соблюдения Соглашений об Уровне Сервиса (SLA), защиты оборудования от внешних вредоносных воздействий и т.п.).

[2] Содержащий как средства анализа данных и поддержки принятия решений, так и средства управления такими процессами (т.е. компоненты управления ИТ-ресурсами и сервисами).

[3] В специальной литературе эти задачи часто идентифицируется по ключевому термину – ***troubleshooting***.

[4] В соотнесении с общим массивом регистрируемых эксплуатационных данных.

эмпирического материала, сформированное путем подбора подходящих «*объяснительных»* переменных и связывающей их (с заданной точностью аппроксимации исходных данных) *аналитической зависимости*;

- при этом подобранные «*объяснительные»* переменные вовсе не обязаны отражать *реальные причинные факторы*, влиянию которых свои появлением и обязан исследуемый эффект (сбой в работе оборудования). Таким образом, для организации *противодействия* влиянию возможных угроз (т.е. для предотвращения возникновения нештатных ситуаций) влияния на факторы, соответствующие «объяснительным» переменным, вообще говоря, может оказаться недостаточно. Необходим детальный *казуальный* анализ: только располагаю пониманием полной картины (т.е. исчерпывающего перечня факторов а также описания взаимосвязей между входящими в него компонентами) «*причинных обусловленностей*» можно организовать эффективный процесс противодействия их влиянию; однако,
- детерминистские процедуры точного[5] анализа данных и выявления каузальной природы фиксируемых сбоев в работе оборудования, как правило, характеризуются экспоненциально быстро растущими оценками сложности вычислений. Таким образом, прямое применение подобных методов в значительной мере ограничено размерами коллекций данных, которые они способны обрабатывать в промышленном режиме.

Переходя к описанию собственно архитектуры программного комплекса для выполнения промышленных ДСМ-вычислений, прежде всего сфокусируем наше внимание на двух основополагающих аспектах:

- процедурах вычисления сходств и
- управлении (разумеется, там, где это возможно) перебором[6] при их формировании.

Итак, технология порождения сходств, варьируемая в зависимости от выбора того или иного конкретного способа представления данных, - главный компонент обсуждаемого нами архитектурного решения. При ее реализации важно не только иметь соответствующий набор программных модулей (вычисляющих сходство двух множеств признаков[7], двух плоских симметризованных графов, двух цепочек символов и т.п.), собственно исполняющих соответствующую операцию сходства, но также и эффективные средства контроля повторяемости порождаемых сходств[8], вместе с инструментами формирования «дополнения» только что порожденного сходства до состояния исчерпываемости (т.е. расширения множества его «родителей» до состояния неподвижной[9] точки). Не менее существенно иметь программные инструменты быстрого восстановления исчерпывающих сходств по их фрагментам (например, восстановление в булевском случае исчерпывающего сходства по заданному подмножеству его образующих).

К сожалению, в ряде случаев (см. подробнее **Главу 4**) процесс формирования исчерпывающих ДСМ-сходств характеризуется экспоненциально быстрым (по отношению к линейному росту объемов исходных данных) ростом объемов необходимых вычислений. Для минимизации влияния проблем вычислительной сложности на возможности получать за *приемлемое время* решения для конкретных приложений полезным оказывается использование еще одного набора вычислительных ДСМ-процедур – тех, которые позволяют строить границы (нижние и верхние по вложимости) множества всех исчерпывающих сходств, извлекаемых из заданной для ДСМ-обучения коллекции примеров и контрпримеров.

---

[5] В противовес расчетам *в среднем*, характерным для статистических подходов.

[6] Сложностью вычислений.

[7] Булевских векторов.

[8] Т.е. средства быстрой проверки, порождалось ли ранее в процессе анализа имеющихся данных именно такое сходство (а из этого следует, что только что порожденное сходство не требует повторной обработки, например, на соответствие дополнительным логическим условиям – запрету на контрпримеры и т.п., и что можно перейти к текущему анализу следующего сходства).

[9] В контексте имеющихся для ДСМ-обучения примеров и контрпримеров.

Наконец, в обсуждаемый архитектурный ландшафт следует включить также и комплекс процедур, которые обеспечивают целенаправленный перебор множества всех эмпирических зависимостей, восстанавливаемых из данных средствами приближенного ДСМ-метода (см. подробнее **Главу 5**).

Итак, типовой набор вычислительных модулей для порождения ДСМ-ЭЗ включает специализированные компоненты для формирования и анализа следующих конструкций:

- дерево всех [10]сходств (пересечений) примеров и контрпримеров, предложенных для обучения и восстановления зависимостей средствами ДСМ-метода;
- диаграмма вложимости замыканий Галуа;
- псевдо-деревья;
- каркасы псевдо-деревьев;
- проверка выполнимости дополнительных логических условий (характеризующих решающие предикаты и стратегии)

Важно учесть, что процедурная конструкция приближенного ДСМ-метода предоставляет возможности для распараллеливания процесса восстановления ДСМ-ЭЗ из данных за счет выделения из диаграммы вложимости всех исчерпывающих ДСМ-сходств независимо друг от друга обрабатываемых псевдо-деревьев. При этом обработка каждого из подобных псевдо-деревьев может быть организована средствами *типового набора* программных модулей (реализующих вычисление сходств и замыканий, восстановление псевдо-деревьев по их каркасам, проверку дополнительных логических условий и т.п.), допускающих возможности «упаковки» в соответствующий набор *виртуальных машин*. Подобные программные объекты легко тиражируемы, а в рамках одного «цикла» ДСМ-порождения зависимостей из данных имеются стандартные[11] возможности управлять их актуальным количеством (числом копий, задействованных в данный момент вычислений). Именно таким образом можно обеспечить в каждом конкретном случае эффективное управление требуемыми для реализации ДСМ-ИАД вычислительными ресурсами. Более того развитие облачных платформ вычислений и SDI-технологий позволяет организовать подобные вычисления в территориально-распределенном режиме, обеспечивая *балансировку* необходимых вычислительных нагрузок в зависимости как от географических, так и временных характеристик загрузки конкретного системно-технического ландшафта.

Для поддержки интеграции технологий ДСМ-ИАД и статистических методов восстановления зависимостей из данных можно воспользоваться схемой, предложенной в работе [*ВВГУ*]. Здесь на входе системы анализа данных имеется заданная таблицей функция

<значение *целевого параметра*> = ***f*** (<значения *наблюдаемых* и *управляемых параметров*[12]>),

для которой на первой стадии вычислений подбираются подходящие «объяснительные» переменные и строится аналитическая зависимость – регрессия специального вида, обеспечивающая (за счет обучения на части имеющихся эмпирических данных и экзамена на их оставшейся части) *приближение наперед заданной точности* имеющейся табличной функции $y=\boldsymbol{f}(x)$ вновь формируемой аналитической функцией $y=\boldsymbol{F}(x)$ от выбранных «объяснительных» переменных.

Одновременно ведется мониторинг значений тех факторов (причем некоторые из них, возможно, будут учтены при выборе «объяснительных» переменных), которые способны оказать влияние на целевой параметр. Т.е. строится своего рода ***Карта Рисков***, отражающая каузальные влияния на исследуемый эффект. Для каждого из таких факторов ведется

---

[10] По-парных, по-троечных и т.д. .

[11] См., например, практику реализации облачных вычислений на базе технологий Software Defined Infrastructure ([*SDI*] и др.).

[12] В работе [*ВВГУ*] это таблица остатков на текущих счетах клиентов коммерческого банка на конкретные даты.

(в тех случаях, когда это возможно) мониторинг текущих значений для соответствующих значений релевантных факторов риска – наблюдаемых и управляемых параметров (например, в цитируемых выше работах - на соответствующие календарные даты).

При прогнозе значения целевого параметра $y_0=\boldsymbol{f}(x_0)$ на значениях соответствующих параметров x_0 сперва вычисляется аналитическое приближение $y_0=\boldsymbol{F}(x_0)$, а затем - специальные *контекстные* поправки к нему, использующие следующую процедуру обучения на «контекстах» (примерах и контрпримерах):

(1) на всей обучающей выборке значений параметров x выделяются три типа областей (подвыборок – *контекстов* принятия решений):
   a. области[13] совпадения пиков табличной и аналитической функции, характеризующие поправки типа (0),
   b. области, где пики табличной функции расположены выше пиков аналитической (поправки типа (+)), и, наконец,
   c. области, где пики табличной функции расположены ниже пиков аналитической (поправки типа (-));

(2) на подвыборках каждого из трех выделенных типов средствами ДСМ-метода восстанавливаются те контекстные зависимости в наблюдаемом взаимодействии соответствующих факторов риска, которые релевантны наблюдаемым в новой «точке» x_0 значениям риск-параметров (т.е. позволяют сформировать ДСМ-прогноз для новой «точки» x_0). В случае непротиворечивости формируемого таким образом ДСМ-прогноза аналитическое приближение $y_0=\boldsymbol{F}(x_0)$ корректируется соответствующей «контекстной» поправкой. В противном случае решение о выборе необходимой поправки к значению $y_0=\boldsymbol{F}(x_0)$ должен выбрать эксперт[14] на основании дополнительных[15] соображений.

Использование при вычислении соответствующих ДСМ-поправок процедурной техники приближенного ДСМ-метода (а также уже представленных выше возможностей облачных вычислений) позволяет организовать этот процесс с необходимой степенью исчерпываемости перебора соответствующих ДСМ ЭЗ при расчете (прогнозе) финального значения целевого параметра.

Наличие на рынке информационных технологий надежных программных систем статистического анализа данных[16] предоставляет известную свободу выбора конкретных инструментальных средств регрессионного моделирования при формировании архитектуры промышленной системы ИАД, интегрирующей статистические и детерминистские средства анализа данных по осуждаемой нами схеме.

Дополнительно отметим, что ряд вычислительных процедур, выполняемых в рамках ДСМ-ИАД, может быть поддержан использованием специализированных аппаратно-схемных решений (например, базирующихся на технологиях TCAM[17] или же на использовании комплексов программируемых интегральных микросхем класса FPGA[18] и т.п.).

---

[13] Размер таких областей на практике, как правило, легко устанавливается экспертом, исходя из соображений однородности или же, наоборот, неоднородности (смены) локального контекста влияний риск-факторов на анализируемое явление. (Например, для цитированной выше ситуации анализа динамики остатков на счетах клиентов и управления риском разрыва мгновенной ликвидности подобные решения эксперта обычно базируются на учете как макроэкономических, так и текущих операционных – период выдачи зарплаты, период налоговых платежей, период платежей по кредитам и т.п. – факторов).

[14] Ведь именно его интеллектуальные возможности призвана усилить обсуждаемая интеллектуальная система анализа данных и поддержки принятия решений.

[15] По отношению к предложенным процедурам восстановления регрессии и ДСМ-обучения.

[16] Прекрасный образец такого инструментария – программный комплекс ***SAS-system*** компании ***SAS Institute*** (см., например, [*SASI*] и др.).

[17] См., например, [TCAM] и др. .

[18] См., например, [FPGA] и др. .