На правах рукописи

Ильвовский Дмитрий Алексеевич

МЕТОДЫ И АЛГОРИТМЫ ОБРАБОТКИ ТЕКСТОВЫХ ДАННЫХ НА ОСНОВЕ ГРАФОВЫХ ДИСКУРСИВНЫХ МОДЕЛЕЙ

Специальность 05.13.18

Математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Москва - 2017

Работа выполнена в Департаменте анализа данных и искусственного интеллекта федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики»

Научный руководитель	доктор физико-математических наук, **Кузнецов Сергей Олегович**,
Официальные оппоненты:	**Богатырев Михаил Юрьевич,** доктор технических наук, профессор, Кафедра информационной безопасности Института прикладной математики и компьютерных наук Тульского государственного университета, профессор
	Виноградов Дмитрий Вячеславович, кандидат физико-математических наук, Федеральный исследовательский центр «Информатика и управление» Российской Академии Наук, старший научный сотрудник
Ведущая организация:	Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН)

Защита состоится «___» _____ 2017 г. в 11 часов на заседании диссертационного совета Д 002.073.04 при Федеральном государственном учреждении “Федеральный исследовательский центр “Информатика и управление” Российской академии наук” (ФИЦ ИУ РАН) по адресу: 117312, Москва, проспект 60-летия Октября, 9 (конференц-зал, 1-й этаж).

С диссертацией можно ознакомиться в библиотеке ФИЦ ИУ РАН, Москва, ул. Вавилова, д. 40.
Электронные версии диссертации и автореферата размещены на официальном сайте ФИЦ ИУ РАН http://www.frccsc.ru.
Электронная версия автореферата размещена на официальном сайте ВАК Министерства образования и науки РФ по адресу: http://vak.ed.gov.ru

Отзывы и замечания по автореферату в двух экземплярах, заверенные печатью, просьба высылать по адресу 117312, Москва, проспект 60-летия Октября, 9, ФИЦ ИУ РАН, диссертационный совет Д 002.073.04.

Автореферат разослан ____.____._____
Телефон для справок: +7(499) 135-51-64

Ученый секретарь диссертационного совета Д 002.073.04
Д.т.н., профессор Крутько В.Н.

Общая характеристика работы

Актуальность работы. Моделирование языковых процессов порождает значительное количество открытых проблем, связанных с развитием соответствующего математического аппарата, созданием и реализацией эффективных алгоритмов и комплексов программ. К настоящему моменту разработано значительное количество хорошо развитых моделей текста, позволяющих (помимо представления текста) вычислять сходство между текстами: «мешок слов», n-граммы, синтаксические деревья разбора и т.д. Среди исследователей, внесших значительный вклад в разработку и применение этих моделей в прикладных задачах (для английского языка), можно отметить C.Manning, H.Schutze, D.Jurafsky, S.Abney, M.Collins, A.Moschitti и многих других. Подавляющее большинство реализованных на практике моделей не полностью учитывает структурные особенности текста, ограничиваясь либо частотными характеристиками слов и n-грамм, либо синтаксическими связями внутри отдельных предложений. Эти модели не позволяют работать с текстом на уровне фрагментов, состоящих из нескольких связанных предложений – абзацев. К другому классу моделей относятся многочисленные лингвистические теории, в той или иной степени учитывающих семантические связи между предложениями. Здесь можно отметить работы таких исследователей как W.Mann, D.Marcu, J.Searle, I.Mel’cuk, H.Kamp, M.Recaesens, D.Jurafsky и многих других. Однако эти модели обладают уже другим недостатком: они носят по большей части теоретический характер, не имеют полного математического или алгоритмического описания и не могут напрямую быть использованы для решения прикладных задач. В то же время учет дискурсивных связей внутри абзаца является критическим фактором в таких важных задачах, как поиск по сложным и редким запросам, кластеризация поисковой выдачи по сложным запросам, классификация текстовых описаний. Всё это делает применение

существующих моделей текста затруднительным и требует разработки новой модели, которая была бы предназначена для решения перечисленных задач, одновременно обладала достаточной теоретической базой и была реализуема на практике.

Необходимость интеграции в модель сложных структурных описаний и применения модели для задач кластеризации, делает актуальным применение методов, позволяющих работать со структурным сходством и использовать эффективные приближения описаний. Методы теории решеток замкнутых описаний предоставляют удобный и эффективный математический аппарат для построения моделей в решении целого ряда важных научных и прикладных задач, в число которых входит и работа с текстами. Эта теория позволяет осуществлять концептуальную кластеризацию и находить сходство произвольного множества объектов (в частности, текстов). Включенный в теорию аппарат проекций позволяет эффективно работать с приближенными описаниями, в той или иной мере учитывающими основные свойства структуры и понижающими вычислительную и временную сложность обработки этих описаний.

Объект исследований – математические модели текстов на естественном языке. **Предмет исследований** – модели текстов на естественном языке, предназначенные для поиска, классификации и кластеризации текстовых данных.

Целью диссертационного исследования является разработка моделей и методов представления и обработки текстов на естественном языке, учитывающих синтаксическую и дискурсивную структуру текстового абзаца и ориентированных на применение в задачах поиска, классификации и кластеризации текстовых данных.

К **задачам исследования** относятся:

- Разработка структурной модели текстов на естественном языке, ориентированной на поиск, классификацию и кластеризацию текстов и использующей синтаксические и дискурсивные связи внутри текста;
- Применение построенной модели в задаче поиска сходства текстов с целью улучшения релевантности поиска по сложным запросам;
- Применение построенной модели в задаче классификации текстов с целью повышения качества существующих методов за счет использования дискурсивной информации;
- Построение на основе разработанной модели таксономического представления текстовых документов с использованием решеток замкнутых структурных описаний и применение представления в задаче кластеризации текстов;
- Разработка математической модели и метода для определения связи «та же сущность» в построенных на основе текстовых данных формальных описаниях и эффективная алгоритмическая реализация данной модели.
- Реализация разработанных моделей, методов и алгоритмов в виде программного комплекса.

К **методам**, использовавшимся в исследовании, относятся:

- Методы построения и анализа решёток замкнутых описаний;
- Методы фильтрации решеток понятий на основе индексов качества моделей;
- Методы построения проекций моделей на узорных структурах;
- Методы построения структурных моделей для текстовых данных;
- Методы построения синтаксических и дискурсивных моделей текста;
- Методы порождения моделей, основанных на графовом представлении.

Научная новизна. В диссертации получен ряд новых научных результатов, которые **выносятся на защиту**:

1. Разработана графовая модель текстов, использующая и обобщающая структурное синтактико-дискурсивное представление текстового абзаца (чащу разбора). Новизна модели заключается в совместном использовании синтаксических деревьев разбора и дискурсивных связей для представления текстовых абзацев на английском языке. Модель ориентирована на применение в задачах поиска, классификации и кластеризации текстов и позволяет описывать сходство текстов в терминах обобщения их структурных графовых и древесных описаний.
2. Предложенная модель применена в задаче поиска ответов по сложным запросам. Разработан численный метод, использующий разработанную модель. Применение метода позволяет улучшить качество поиска и устранить недостатки существующих моделей благодаря применению впервые введенной в работе операции структурного синтактико-дискурсивного сходства для запроса и ответов.
3. Разработанная модель применена в задаче классификации текстовых данных. На основе предложенной модели реализован численный метод, использующий ядерные функции. Применение модели позволяет устранить недостатки существующих моделей благодаря ранее не применявшемуся в задачах классификации абзацев использованию дискурсивной информации.
4. Разработано на базе предложенной модели таксономическое представление коллекции текстовых данных в виде решетки замкнутых структурных синтактико-дискурсивных описаний. Полученное представление применено в задаче кластеризации текстовых данных и позволяет улучшить результаты, достигаемые альтернативными моделями.
5. Разработана на основе модели текстов и теории решеток замкнутых описаний оригинальная модель тождественных денотатов для формальных

описаний. Предложены численный метод и алгоритм построения связей типа «та же сущность», использующие разработанную модель. Новизна метода заключается в использовании оригинального индекса ранжирования замкнутых формальных описаний для нахождения денотатов.

Теоретическая значимость работы заключается в разработке принципиально новых моделей и методов: синтактико-дискурсивной модели текстов, позволяющей представлять текстовые абзацы в виде графов (полное описание) и лесов (приближенное описание) и вычислять сходство между текстами, таксономического представления текстовых данных, модели и метода выявления тождественных денотатов для формальных описаний.

Практическая ценность подтверждена экспериментами по оценке релевантности поиска по сложным запросам, обучению на текстовых абзацах, выявлению тождественных денотатов. Эксперименты продемонстрировали улучшение по сравнению с существующими аналогами. Разработанные алгоритмы и методы были успешно **внедрены** в реальных проектах, а также использованы в преподавательской деятельности Департамента анализа данных и искусственного интеллекта Факультета компьютерных наук НИУ ВШЭ. Компания ООО «ФОРС-Центр разработки» применила метод классификации текстовых абзацев в проекте оценки пользовательских предпочтений. Компания Авикомп внедрила метод выявления тождественных денотатов для оптимизации прикладной онтологии. Все разработанные методы были реализованы в виде программного комплекса, предназначенного для решения исследовательских и прикладных задач.

Достоверность полученных результатов подтверждена строгостью построенных математических моделей, экспериментальной проверкой результатов численных расчетов и практической эффективностью программных реализаций.

Апробация результатов работы. Основные результаты работы обсуждались и докладывались на следующих научных конференциях и семинарах:

1. 9-й международной конференции «Интеллектуализация обработки информации» (ИОИ-2012), Будва, Черногория.
2. 1-м семинаре по анализу формальных понятий и информационному поиску (FCAIR-2013) в рамках 35-й европейской конференции по информационному поиску (ECIR-2013), Москва, Россия.
3. 11-й международной конференции по анализу формальных понятий (ICFCA-2013), Дрезден, Германия.
4. 8-й международной конференции по компьютерной лингвистике ДИАЛОГ-2013, Москва, Россия.
5. 3-м семинаре по представлению знаний в виде графов (GKR-2013) в рамках 23-й объединенной международной конференции по искусственному интеллекту (IJCAI-2013), Пекин, Китай.
6. 7-й международной конференции по компьютерной лингвистике RANLP-2013, Хисаря, Болгария.
7. 8-й международной конференции по компьютерной лингвистике RANLP-2015, Хисаря, Болгария.
8. 16-й международной конференции по интеллектуальному анализу данных AIMSA-2014, Варна, Болгария.
9. 14-й международной конференции по интеллектуальной обработке текста и компьютерной лингвистике CICLING-2014, Катманду, Непал.
10. 15-й международной конференции по интеллектуальной обработке текста и компьютерной лингвистике CICLING-2015, Каир, Египет.

11. 52-й международной конференции Ассоциации компьютерной лингвистики ACL-2014, Балтимор, США.
12. 53-й международной конференции Ассоциации компьютерной лингвистики ACL-2015, Пекин, Китай.

Публикация результатов. Основные результаты работы изложены в 15 научных статьях. 12 статей опубликованы в рецензируемых трудах международных конференций, 3 статьи опубликованы в журналах из списка ВАК.

Структура диссертации. Диссертация состоит из введения, пяти глав, заключения, списка использованной литературы и приложений. Общий объем диссертации – 250 с. машинописного текста (с приложениями). Основная часть работы изложена на 164 с. и содержит 16 рисунков и 11 таблиц. Библиография включает в себя 139 наименований.

Содержание работы

Во **введении** раскрывается актуальность темы диссертации, формулируются проблемы исследования, предмет исследования, определяется цель работы, описываются методы исследования, излагаются основные научные результаты, обосновывается теоретическая и практическая значимость работы, даётся общая характеристика исследования.

В **первой главе** рассматриваются теоретические основы используемых в дальнейшем моделей и методов и описываются особенности моделирования текстовых данных. Приводятся основные определения, связанные с частично упорядоченными множествами и решетками, анализом формальных понятий (АФП), решетками замкнутых описаний, синтаксическими и дискурсивными моделями представления текста. Также рассматриваются некоторые подходы к структурному обучению на текстовых данных. Вводится модель структурного представления текстовых абзацев – чаща разбора.

Решетка – частичный порядок (антисимметричное транзитивное рефлексивное бинарное отношение), для любых двух элементов которого существуют инфимум и супремум. Решетки замкнутых описаний, называемые также *узорными структурами* (pattern structures) предназначены для работы со сложными данными. *Узорная структура* – это тройка $\left(G,\left(D,\sqcap\right),\delta\right)$, где G – множество объектов, $\left(D,\sqcap\right)$ – полная полурешетка всевозможных описаний, а $\delta: G \to D$ – функция, которая сопоставляет каждому объекту из множества G его описание из D. Операция $\sqcap$ позволяет вычислить сходство между двумя описаниями. *Проекция узорной структуры* – это функция $\psi: D \to D$, которая является монотонной $x \le y \Rightarrow \psi\left(x\right) \le \psi\left(y\right)$, сжимающей $\psi\left(x\right) \le x$ и идемпотентной $\psi\left(\psi\left(x\right)\right) = \psi\left(x\right)$. Для получения проекции узорной структуры мы должны спроецировать функцию – описание объектов, а также полурешетку описаний:

$\psi\left(\left(G,\left(D,\sqcap\right),\delta\right)\right) = \left(G,\left(D_\psi,\sqcap_\psi\right),\psi \circ \delta\right)$, где

$D_\psi = \psi\left(D\right) = \left\{d \in D \mid \exists d^* \in D : \psi\left(d^*\right) = d\right\}$ и $\forall x, y \in D,\ x \sqcap_\psi y = \psi\left(x \sqcap y\right)$.

Теория решеток замкнутых описаний находит своё применение в нескольких областях, в частности, она может быть использована для обработки текста на естественном языке. Автор приводит несколько основных способов представления текстовых данных, применяемых для этой обработки.

Модель «мешка слов» («bag-of-words») дает упрощенное представление текста, применяемое, в частности, в задаче информационного поиска. В этой модели текст представляется как неупорядоченный набор слов (или словосочетаний) без учета грамматики и порядка слов.

Дерево синтаксического разбора (syntactic parse tree) – это упорядоченное дерево, которое отражает синтаксическую структуру предложения или строки согласно некоторой формальной грамматике. Выделяют два основных класса:

деревья составляющих (constituency tree) и деревья зависимостей (dependency tree). Деревья синтаксического разбора используются и для компьютерных языков, и для обработки текстов на естественных языках.

Если рассматривать более объемные тексты, например, абзацы, состоящие из нескольких предложений, то использования синтаксической информации недостаточно. В этом случае источником структурных связей могут служить дискурсивные теории, учитывающие смысловые отношения между фрагментами текста. В работе используется несколько типов таких связей, описание которых приводит автор: *кореферентные связи* (coreference), *таксономические отношения* («та же сущность», гипоним, гипероним и т.д.), риторические отношения (*теория риторических структур*), связи между коммуникативными действиями (*теория речевых актов*). Также приводится краткое описание нескольких теорий, позволяющих устанавливать связи между предложениями, но не включенных в модель: теории семантической организации данных, теории представления дискурса и т.д.

Используя дискурсивные теории, позволяющие установить связи внутри текста, состоящего из нескольких предложений, можно обобщить понятие дерева синтаксического разбора на случай текстового абзаца.

Определение 1.1. *Чащей разбора* текстового абзаца называется множество деревьев разбора предложений абзаца и связи нескольких типов, устанавливаемых между вершинами этих деревьев. Каждая связь – это упорядоченная пара вершин деревьев разбора.

Со структурной точки зрения, чаща представляет собой ориентированный граф, который включает в себя деревья разбора, а также дуги, соответствующие несинтаксическим связям.

В исследовании также используются так называемые ядерные функции, применяемые в задаче классификации коротких текстов в сочетании с широким классом линейных классификаторов, использующих скалярное произведение в

векторных пространствах. Одним из таких методов является Метод Опорных Векторов (Support Vector Machine). Применение ядер позволяет использовать данный метод для объектов, имеющих сложную структуру и очень большое число свойств, не прибегая к явному выделению этих признаков. В частности, он применим к деревьям синтаксического разбора, для которых также вводятся функции ядра.

Во **второй главе** описывается графовая модель текстовых абзацев и её применение в задаче информационного поиска (для английского языка). Рассматриваются методы вычисления полного и приближенного структурного сходства текстовых абзацев, определяется проекция структурного представления текстового абзаца в виде расширенных синтаксических групп. Проводится анализ полученных результатов, демонстрируется преимущество, достигаемое за счет вычисления сходства на абзацах, производится сравнение методов, основанных на полном и приближенном сходстве. Также в главе описывается применение построенной модели для иерархической кластеризации текстовых абзацев, источником которых может служить, например, поисковая выдача.

В рамках расширения модели «чащи разбора» автором вводится ассоциативная и коммутативная операция обобщения (или сходства) текстов. Если представить текстовые абзацы P_1 и P_2 в виде ориентированных графов («чащ разбора») G_1 и G_2, то операция обобщения этих абзацев $P_1 \sqcap P_2$ наиболее естественным образом определяется как $\{H_i\}$ - множество всех максимальных по вложению (с учетом меток на вершинах и ребрах) общих подграфов графов из G_1 и G_2. Если рассматривать абзацы как *объекты*, а чащи разбора как их *описания*, то операция обобщения или сходства – это полурешеточная операция пересечения.

Используя несинтаксические связи, автор расширяет понятие синтаксической группы на случай нескольких предложений. Дискурсивные

связи между вершинами деревьев разбора позволяют объединять несколько групп из разных предложений или из одного предложения между собой. Такие связи при обходе группы условно позволяют «перескакивать» с одного дерева разбора на другое. В работе рассматриваются следующие типы групп:

- Синтаксические, или регулярные группы;
- Группы, включающие кореферентные и таксономические связи. Они также называются *чащевыми* группами.
- Риторические группы (RST). Две группы (каждая из них может быть и чащевой, и синтаксической), соединенные риторическим отношением.
- Коммуникативные группы (CA).

Для удобства все объединенные несинтаксическими связями синтаксические группы (чащевые, RST, CA) называются *расширенными группами*.

Выполнение операции обобщения на полных описаниях является NP-трудной задачей, поэтому для эффективного вычисления с сохранением свойств операции можно воспользоваться механизмом *проекций*. Определение проекции допускает существование большого числа способов её задания. Автор определяет проекцию чащи как *множество всех максимальных по вложению синтаксических и расширенных групп*, вычисленных для данного абзаца. Со структурной точки зрения, такая проекция – это максимальные по вложению поддеревья графа с дополнительными свойствами. В работе приводится алгоритм формирования всех расширенных групп для текстового абзаца.

Работа с проекциями позволяет добиться экономии по сложности (переход к работе с деревьями) без значимого ущерба для качества результата (группы учитывают все необходимые лингвистические связи внутри абзаца).

В работе формулируется алгоритм вычисления сходства для двух абзацев с использованием проекций:

1. Выполнить их фрагментацию и извлечь все синтаксические группы из каждого предложения.
2. Найти дискурсивные связи внутри абзаца.
3. Используя семантические связи, построить на основе синтаксических групп расширенные группы.
4. Провести обобщение для каждого из четырех типов групп, заключающееся в поиске множества наибольших общих подгрупп для каждой пары групп одного и того же типа.

Построенная модель применяется для решения задачи информационного поиска. Использование абзацев текста в качестве запросов применяется, например, в основанных на поиске рекомендательных системах. Рекомендательные агенты отслеживают действия пользователей чатов, блогов и форумов, комментарии пользователей на торговых сайтах и предлагают наиболее релевантные веб-документы и их фрагменты, относящиеся к решениям о покупке товара.

В экспериментах сначала вычисляется сходство между вопросом и потенциальными ответами, затем ответы ранжируются по вычисляемому на базе сходства числовому значению. В случае использования полного описания значение вычисляется как размер максимального общего подграфа. Для проекций сначала вычисляется максимальный размер (количество вершин) среди наибольших общих подгрупп для каждого типа групп, а затем эти значения суммируются. На различных наборах данных новый подход сравнивается с несколькими альтернативными методами:

- Применение ключевых слов: базовый подход, в котором тексты представляются в виде «мешка слов», а затем вычисляется набор общих ключевых слов / N-грамм и их частот.

- Попарное сравнение предложений: применяются синтаксические обобщения для каждой пары предложений, полученные результаты суммируются.

Таблица 2.1. Оценка релевантности поиска

Тип запроса	Сложность запроса	Релевантность исходного поиска в Bing, %,	Релевантность поиска с использованием обобщений для отдельных предложений, %,	Релевантность поиска с помощью чащ, построенных на **фрагментах**, %,	Релевантность поиска с помощью чащ, построенных на **оригинальных абзацах**, %,	Релевантность поиска с использованием обобщения чащ на **графах**, %
Поиск рекомендаций по товарам	1 составное предложение	62.3	69.1	72.4	72.9	73.3
	2 предложения	61.5	70.5	71.9	72.8	71.6
	3 предложения	59.9	66.2	72.0	73.4	71.4
	4 предложения	60.4	66	68.5	69.2	66.7
Поиск рекомендаций по путешествиям	1 составное предложение	64.8	68	72.6	74.7	74.2
	2 предложения	60.6	65.8	73.1	76.9	73.5
	3 предложения	62.3	66.1	70.9	70.8	72.9
	4 предложения	58.7	65.9	72.5	73.9	71.7
Поиск рекомендаций контента на Facebook	1 составное предложение	54.5	63.2	65.3	68.1	67.2
	2 предложения	52.3	60.9	62.1	63.7	63.9
	3 предложения	49.7	57	61.7	63.0	61.9
	4 предложения	50.9	58.3	62.0	64.6	62.7
Средние показатели		**58.15**	**64.75**	**68.75**	**70.33**	**69.25**

Таблица демонстрирует, что с ростом сложности запроса увеличивался эффект от применения технологии обобщения. Метод с использованием абзацев превосходит ключевые слова и предложения. Другим важным результатом является незначительная потеря качества при существенном выигрыше в скорости за счет использования проекций.

Помимо собственно улучшения релевантности результатов поиска, существенным аспектом является их интерпретация – одно из важнейших

направлений в промышленном информационном поиске. В работе приводится описание применения модели для задачи иерархической концептуальной кластеризации текстов, одним из частных случаев которой является представление результатов поиска в виде решетки замкнутых множеств (кластеров), а не в виде линейного списка. Структурным описанием каждого текста является чаща разбора или её проекция. Решеточная операция пересечения – это операция сходства чащ разбора.

Кластеризация в случае использования полного описания выглядит следующим образом:

1. Взять множество текстов (поисковую выдачу) *T*.
2. Для каждого результата $t_i \in T$ построить чащу разбора $p_i \in P$.
3. Используя операцию обобщения чащ разбора в качестве решеточной операции пересечения $\sqcap$, построить узорную решетку $\left(T,\left(P,\sqcap\right),\delta\right)$ для всех текстов с помощью любого стандартного алгоритма (например, AddIntent или Замыкай-По-Одному).
4. Получить иерархические кластеры – узорные понятия решетки.

При использовании приближенного представления алгоритм немного модифицируется:

1. Взять множество текстов (поисковую выдачу) *T*.
2. Для каждого результата $t_i \in T$ построить проекцию чащи разбора $\psi\left(p_i\right) \in \psi\left(P\right)$.
3. Используя операцию обобщения проекций в качестве решеточной операции пересечения, построить проекцию узорной решетки $\left(T,\left(P_\psi,\sqcap_\psi\right),\psi\circ\delta\right)$
4. Для всех текстов с помощью любого стандартного алгоритма (например, AddIntent или Замыкай-По-Одному).
5. Получить иерархические кластеры – проекции узорных понятий решетки.

В **третьей** главе описывается применение построенной модели для задачи обучения с учителем на текстовых абзацах (для английского языка), основанное на использовании ядерных функций (kernels) в методе опорных векторов (SVM). Производится сравнение с существующими моделями (Moschitti, «мешок слов»), не использующими дискурсивную информацию о связях между предложениями абзаца. Демонстрируется преимущество применения новой модели в задаче классификации поисковых результатов и в задаче классификации технических документов.

Функция ядра (convolution kernel) на деревьях задает пространство признаков, состоящее из возможных типов поддеревьев деревьев разбора, и подсчитывает количество общих подструктур в качестве синтаксической близости между деревьями. В исследовании применяется подход к построению ядра, базирующегося более чем на одном дереве разбора: ядра для леса деревьев. Сравниваются два подхода:

1. Существующий подход. Обучение на лесе, сформированном из деревьев разбора для всех предложений абзаца (Moschitti);
2. Модифицированный подход. Обучение на лесе, сформированном из обычных деревьев разбора, дополненных *расширенными* деревьями. Каждое расширенное дерево включает в себя одну дискурсивную связь («перескок» между деревьями). Такой лес представляет собой альтернативный вариант задания проекции чащи разбора.

Автор формулирует алгоритм построения расширенных деревьев для абзаца. Итоговые деревья не являются корректными деревьями синтаксического разбора, однако формируют адекватное пространство признаков для ядер на деревьях. В исследовании приводятся результаты экспериментов, демонстрирующие выигрыш при использовании множества расширенных деревьев в задаче поиска с помощью классификации и в задаче классификации технических документов.

Задача поиска с помощью классификации представляет собой разбиение множества поисковых результатов по двум классам: релевантные и нерелевантные. Соответствующая обучающая выборка формируется как множество ответов с высоким рейтингом (положительные примеры) и множество ответов с низким рейтингом (отрицательные примеры). Тестовая выборка формируется из оставшегося множества путем случайного выбора. Для каждого результата используется его «сниппет» (выдаваемый поисковой системой фрагмент), а также соответствующий ему фрагмент текста, извлеченный со страницы (два независимых эксперимента). Этот эксперимент базируется на предположении, что верхние (нижние) результаты, выдаваемые Bing, так или иначе релевантны (нерелевантны) исходному запросу, несмотря на то что они могут быть неверно упорядочены.

Таблица 3.1. Результаты для запросов, связанных с мнением о продуктах. Обучение на текстах со страниц

Продукты	*Метод Moschitti*	*Дисс. метод*
Точность	56,8	**58,7**
Полнота	75,2	**84,6**
F-мера	64,9	**67,5**

Таблица 3.2. Результаты для запросов, связанных с мнением о продуктах. Обучение на поисковых сниппетах

Продукты	*Метод Moschitti*	*Дисс. метод*
Точность	56,3	**63,2**
Полнота	78,4	**83,1**
F-мера	61,7	**67**

Таблица 3.3. Результаты для запросов, сформированных на базе вопросов из *Yahoo Answers*. Обучение на текстах со страниц

Yahoo Answers	*Метод Moschitti*	*Дисс. метод (только кореферентные связи)*	*Дисс. метод*
Точность	51,7	50,8	**54,4**
Полнота	73,6	79,2	**83,3**
F-мера	60,1	54,6	**62,8**

Таблица 3.4. Результаты для запросов, сформированных на базе вопросов из *Yahoo Answers*. Обучение на поисковых сниппетах

Yahoo Answers	*Метод Moschitti*	*Дисс. метод (только кореферентные связи)*	*Дисс. метод*
Точность	59,5	62,6	**67,9**
Полнота	73,3	74,9	**79**
F-мера	62,5	64,3	**70,7**

Эксперименты демонстрируют, что добавление новых признаков без изменения схемы эксперимента улучшает качество классификации существующего подхода. Это улучшение колеблется в диапазоне от 2 до 8 % для текстов из нескольких областей, имеющих различную структуру. При этом улучшение и внедрение дополнительных признаков не требуют доработки самого алгоритма обучения на деревьях.

В задаче классификации технических документов рассматриваются документы, относящиеся к двум классам:

1. Action-plan (описание оригинальной разработки) - документ, который содержит четкое и хорошо структурированное описание того, как построить конкретную систему в какой-либо области.
2. Meta-document (мета-описание) – документ, объясняющий, как писать документы, относящиеся к первому классу, например, инструкция, учебник, технический стандарт и т.д.

Данная задача важна с практической точки зрения. «Мета-документы», как правило, содержат общедоступную информацию и могут распространяться свободно. Описание же оригинальных разработок является собственностью компаний и не может передаваться и копироваться без их разрешения. Эксперименты проводились на наборе из нескольких тысяч документов, имеющих сходную тематику и практически совпадающие ключевые слова. Эти данные были разбиты на 3 группы для проведения обучения и тестирования по методу кросс-валидации

Таблица 3.5. Результаты классификации технических документов.

Метод	**Точность, %**	**Полнота,%**	**F-мера,%**
*«Ближайшие соседи» (на основе TF*IDF)*	53.9	62	57.67+-0.62
Наивный Байесовский	55.3	59.7	57.42+-0.84
Ядра на синтаксических деревьях	71.4	76.9	74.05+-0.55
Ядра на расширенных деревьях (только анафора), ***Дисс.***	77.8	81.4	79.56+-0.70
Ядра на расширенных деревьях (только RST), ***Дисс.***	80.1	80.5	80+-1.03
Ядра на расширенных деревьях (анафора +RST), ***Дисс.***	**83.3**	**83.6**	**83.45+-0.78**

Эксперименты демонстрируют, что обучение на расширенных деревьях существенно превосходит по качеству классификации альтернативные подходов, основанные на использовании синтаксических деревьев и «мешка слов» (наивный байесовский классификатор и метод ближайших соседей). Преимущество по F-мере по сравнению с обучением на синтаксических деревьях составляет 9%, по сравнению с классификаторами на основе «мешка слов» - 30%.

В **четвертой** главе рассматривается задача выявления тождественных денотатов для случая формальных описаний, построенных на основе предварительно обработанных текстовых данных. Предлагается модель тождественных денотатов для формальных описаний и метод, позволяющий устанавливать семантические связи типа «та же сущность» (кореферентные связи) между формальными описаниями, выделяемыми из текста. Метод основан на применении фильтрации решеток формальных понятий. Производится сравнение данного метода с альтернативными методами на нескольких наборах данных: сгенерированных и полученных из реального приложения. Демонстрируется улучшение, достигаемое за счет применения нового метода.

Одним из типов дискурсивных связей, используемых в исследовании для соединения фрагментов текста, является отношение «та же сущность». Обнаружение такого рода связей является отдельной задачей, известной также под названием *выявления тождественных денотатов*. В работе

рассматривается частный случай проблемы, когда имеются формальные описания денотатов, построенные с помощью предварительной обработки текстовых данных.

Одной из наиболее универсальных и популярных моделей представления структурированных данных являются прикладные онтологии. При автоматической или полуавтоматической генерация онтологии из текстовых данных на основе заранее подготовленного набора правил возникает проблема появления нескольких описаний одних и тех же объектов реального мира (денотатов). В работе приводится и поэтапно описывается алгоритм поиска тождественных денотатов в прикладной онтологии. На вход алгоритм принимает прикладную онтологию. На выходе алгоритм выдает списки объектов, которые были идентифицированы им как тождественные.

Алгоритм состоит из двух этапов (второй этап может рассматриваться как самостоятельный алгоритм поиска тождественных денотатов в формальном контексте):

1. Преобразование онтологии в формальный контекст.

1.1 Преобразование онтологии в многозначный контекст;

1.2 Преобразование многозначного контекста в формальный контекст;

2. Поиск тождественных денотатов в формальном контексте.

2.1 Построение множества формальных понятий с помощью алгоритма AddIntent.

2.2 Фильтрация множества формальных понятий.

2.3 Формирование списков тождественных объектов в автоматическом или полуавтоматическом (с участием эксперта) режиме.

В исследовании предлагается числовой критерий (индекс) для фильтрации формальных понятий. Чем выше значение индекса, тем выше

уверенность в том, что понятие содержит тождественные объекты. При подборе критериев были учтены основные свойства, которыми должны обладать эти понятия. Во-первых, критерий должен принимать б*о*льшее значение, если, при прочих равных, число признаков, которыми отличаются объекты понятия, будет меньше:

$$I_1(A,B)=\frac{|A||B|}{\sum_{g\in A}\left|\{g\}'\right|}$$

Второе свойство - увеличение значение индекса при увеличении числа общих признаков (при прочих равных). При этом распространенный признак делает меньший вклад в значение критерия, чем редкий признак, так как чем признак более распространен, тем больше шансов, что понятие с данным признаком возникло из-за случайного пересечения признаков:

$$I_2(A,B)=\sum_{m\in B}\frac{|A|}{\left|\{m\}'\right|}$$

Итоговый критерий *DII* представляет собой комбинацию (используются два варианта) данных индексов:

$$DII_{+}=I_1+kI_2$$

$$DII_{*}=I_1*I_2^{\,k}$$

Неизвестный коэффициент может определяться с помощью экспертной оценки, так как интерпретация обоих индексов достаточно легка для понимания. Алгоритм предусматривает два режима работы: автоматическое принятие решения и полуавтоматический режим с привлечением эксперта-аналитика.

В исследовании проводится сравнительный анализ разработанного метода с методами попарного сравнения объектов на основе *расстояния*

Хэмминга и *абсолютного сходства*, применявшимися для решения задачи до проведения исследования, а также с методом, основанным на *индексе экстенсиональной устойчивости* понятия.

В исследовании приводятся результаты экспериментов на реальных данных (онтология) и на искусственно сгенерированных данных (формальные контексты) с заранее известными тождественными объектами. При генерации данных учитывались особенности прикладной онтологии. Для оценки качества метода использовались: полнота, точность, среднее значение полноты алгоритма при 100% значении точности, Mean Average Precision (MAP):

$$Map(K) = \frac{\sum_{i=1}^{|K|} AveP(K_i)}{|K|}$$

$$AveP(k) = \frac{\sum_{c \in C_k} (P(c))}{|C_k|},$$

где K - множество контекстов, C_k - множество релевантных формальных понятий контекста k, $P(c)$ - доля релевантных понятий среди всех понятий, имеющих ранг не ниже, чем у понятия c.

Результаты экспериментов демонстрируют, что алгоритм на основе обоих вариантов нового индекса показал более высокие результаты, чем рассмотренные альтернативы. Основной отличительной особенностью метода является весьма небольшое падение точности алгоритма (до 90%) при росте полноты вплоть до 70%. Вариант DII_* был менее стабилен на больших порогах.

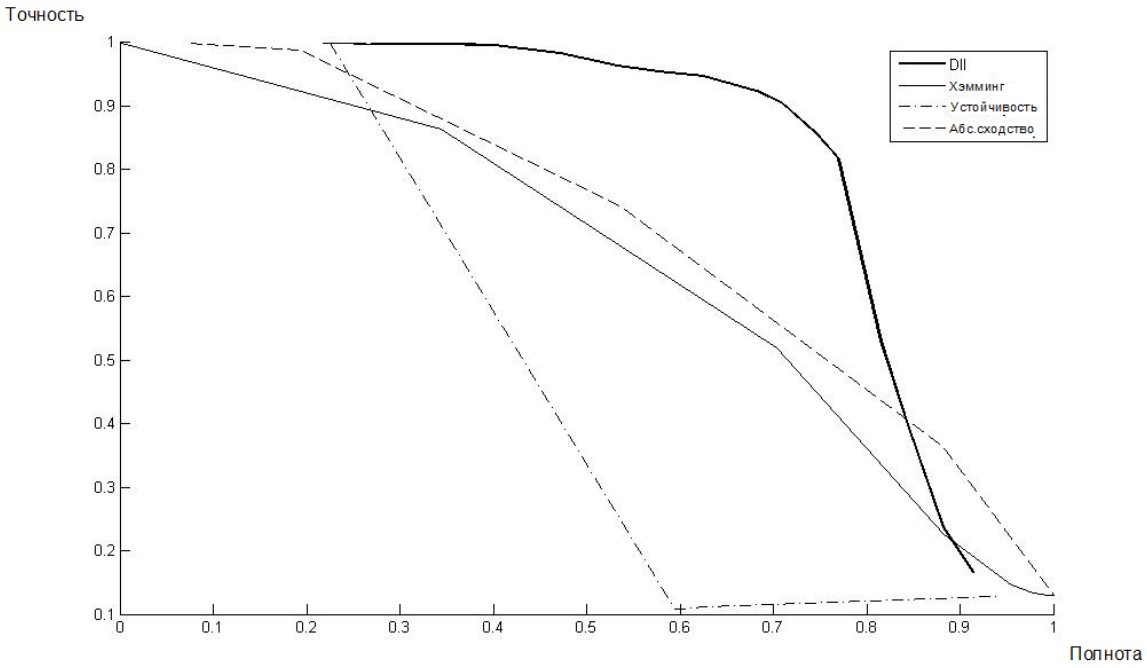

Рис. 4.1. Зависимость точности алгоритмов от полноты.

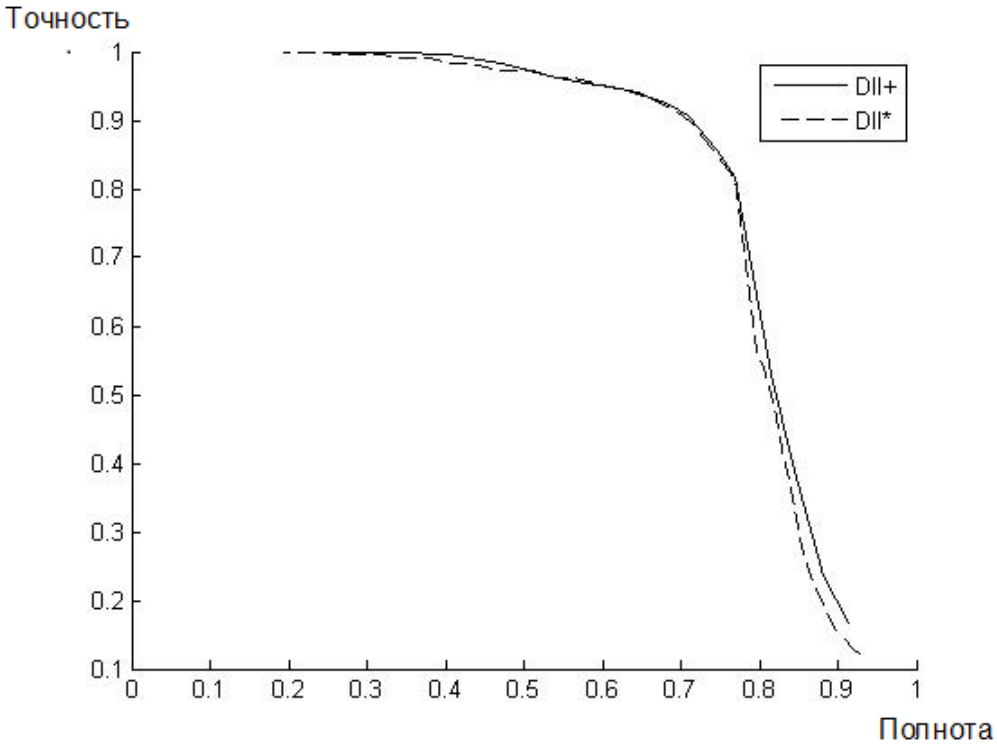

Рис. 4.2. Зависимость точности от полноты для двух вариантов нового индекса *DII*

Таблица 4.1. Максимальная полнота алгоритмов при максимальной точности

Алгоритм	**Максимальная полнота при точности 100% (на экспериментальных данных)**
Алгоритм на основе абсолютного расстояния	6.22%
Алгоритм на основе расстояния Хэмминга	0.56%
Алгоритм на основе индекса устойчивости	22.44%
Алгоритм на основе индекса DII_+	21.78%
Алгоритм на основе индекса DII_*	9.49%

При сравнении методов на основе индекса экстенсиональной устойчивости и вариантов нового индекса DII_+ и DII_* по мере *MAP* очевидное преимущество имеет новый индекс (таблица 4.2).

Таблица 4.2. Результаты по метрике Mean Average Precision.

Алгоритм	**MAP**
Алгоритм на основе индекса устойчивости	0.499
Алгоритм на основе нового индекса DII_+	0.935
Алгоритм на основе нового индекса DII_*	0.938

Для каждого метода был подобран оптимальный порог (максимальное расстояние от начала координат кривой «точность-полнота»), при котором алгоритм имеет оптимальную полноту при минимальных потерях точности (таблица 4.3). Новый метод продемонстрировал преимущество перед альтернативами.

Таблица 4.3. Оптимальные пороги для методов и качество поиска

Алгоритм	**Порог в алгоритме**	**Полнота**	**Точность**
На основе абсолютного расстояния	3.50	19.35%	98.82%
На основе расстояния Хэмминга	0.50	34.37%	86.32%
На основе индекса устойчивости	0.50	22.44%	100%
На основе нового индекса DII_+	1.15	40.09%	99.58%
На основе нового индекса DII_*	0.90	31.80%	99.55%

Для экспериментов на реальных данных использовалась онтология, построенная по новостным документам политической направленности (12006 объектов различных классов), в которой проводился поиск тождественных денотатов среди объектов классов «Персона» и «Компания» (9821 объект). Признаки формального контекста строились с использованием всех объектов и связей в онтологии. Алгоритм на основе индекса *DII* (использовался вариант DII_+) выделил 905 групп объектов, размеры которых варьируются от 2 до 41. 98% выделенных групп полностью состоят из тождественных объектов, что соответствует точности в 98%.

В **пятой главе** описывается построенный в рамках исследования программный комплекс, предназначенный для обработки текстовых данных с помощью чащ разбора и включающий в себя оригинальные модули, предназначенные для работы с чащами разбора, для построения узорных структур на чащах разбора и их проекциях, для поиска и обучения на текстах и т.д.

В проекте используются следующие технологии и программные средства:

- OpenNLP/Stanford NLP парсеры – для построения деревьев синтаксического разбора;
- Stanford NLP Coreference – для построения кореферентных связей;
- Bing API – для реализации базового поиска;
- Apache SOLR – для обеспечения интеграции с поисковыми системами;
- Риторический парсер Joty – для автоматического построения дискурсивных деревьев на основе машинного обучения.
- TK-Light – для обучения на деревьях с использованием ядер.

Архитектура комплекса предусматривает возможность интеграции с другими системами, в том числе с поисковыми приложениями.

Также в работе рассматривается *Formal Concept Analysis Research Toolbox* (*FCART*) – программный комплекс для анализа данных методами АФП. В рамках исследования данный комплекс был модифицирован автором. В комплекс было добавлено вычисление индексов устойчивости, отделимости, а также нового индекса *DII*, позволяющего выявлять тождественные денотаты.

В **заключении** приводятся основные выводы и ближайшие планы развития исследования.

Основные результаты работы

1. Разработана новая математическая модель текстовых данных, включающая в себя графовое синтактико-дискурсивное представление текста (чащу разбора) и операцию обобщения на чащах разбора.
2. Реализован численный метод повторного ранжирования результатов информационного поиска по сложным запросам, использующий предложенную модель. Применение метода повысило релевантность поиска по сравнению с альтернативными подходами. На нескольких наборах реальных интернет-данных было продемонстрировано, что вычисление сходства между текстами на уровне абзацев (обобщение чащ разбора) позволяет улучшить релевантность поиска по сравнению с использованием сходства на уровне отдельных предложений и отдельных слов (модель «мешка слов»).
3. Реализован численный метод классификации абзацев, использующий предложенную модель. Было показано, что добавление новых признаков без изменения схемы эксперимента позволяет повысить качество классификации, достигаемое при применении существующей модели, не использующей дискурсивную информацию.

4. Получено таксономическое представление текстовых данных на основе решетки замкнутых структурных описаний. Показана применимость полученного представления в задаче кластеризации текстовых абзацев.
5. Разработаны новая модель и численный метод поиска тождественных денотатов в прикладной онтологии и формальном контексте, предназначенные для построения дискурсивных отношений «та же сущность» в текстовых данных. Предложен индекс, позволяющий ранжировать формальные понятия по степени уверенности в том, что объекты данного понятия тождественны. Эксперименты на сгенерированных и реальных данных показали преимущество перед использовавшимися ранее альтернативами и высокую точность работы нового метода.
6. Разработан программный комплекс для работы с текстовыми данными, реализующий предложенные методы и алгоритмы. Комплекс применен в нескольких реальных задачах, связанных с поиском и классификацией текстовых данных.

Публикации по теме диссертации

Публикации в журналах, входящих в **перечень ВАК**:

1. Ильвовский Д., Климушкин М. Выявление дубликатов объектов в прикладных онтологиях с помощью методов анализа формальных понятий. // НТИ, Сер. 2. – 2013. - № 1. - С.10-17, 0.75 п.л. (вклад автора 0.5 п.л.)
2. Ильвовский Д., Применение семантически связанных деревьев синтаксического разбора в задаче поиска ответов на вопросы, состоящие из нескольких предложений. НТИ. Сер.2 - 2014. - № 2. - С.28-37, 0.9 п.л.
3. Ильвовский Д. А., Черняк Е. Л. Системы автоматической обработки текстов // Открытые системы. СУБД. 2014. № 01. С. 51-53. 0.45 п.л. (вклад автора 0.3 п.л.)

Прочие публикации:

1. Neznanov A., Ilvovsky D. A., Kuznetsov S. FCART: A New FCA-based System for Data Analysis and Knowledge Discovery. Contributions to the 11th International Conference on Formal Concept Analysis. Dresden: Qucoza, 2013. P.31-44. 0.6 п.л. (вклад автора 0.2 п.л.)
2. Kuznetsov S. O., Strok F. V., Ilvovsky D. A., Galitsky B. Improving Text Retrieval Efficiency with Pattern Structures on Parse Thickets // Proceedings of the FCAIR. Vol. 977. M.: CEUR Workshop Proceeding, 2013. P.6-21. 0.77 п.л. (вклад автора 0.25 п.л.)
3. Galitsky B., Ilvovsky D. A., Kuznetsov S. O., Strok F. V. Parse thicket representations of text paragraphs // По материалам ежегодной Международной конференции «Диалог». Т.1. Вып. 12 (19). М.: РГГУ, 2013. С. 134-145. 0.73 п.л. (вклад автора 0.2 п.л.)
4. Ильвовский Д. А., Климушкин М. А. Выявление дубликатов объектов в прикладных онтологиях на основе методов анализа формальных понятий // Сборник докладов 9-й международной конференции ИОИ-2012. Торус Пресс, 2012. С.625-628. 0.2 п.л. (вклад автора 0.1 п.л.)
5. Galitsky B., Ilvovsky D., Kuznetsov S. O., Strok F. Matching sets of parse trees for answering multi-sentence questions // Proceedings of the RANLP 2013. – INCOMA Ltd. – 2013. – P. 285–294. 0.8 п.л. (вклад автора 0.3 п.л.)
6. Galitsky, B. A., Ilvovsky, D., Kuznetsov, S. O., Strok, F. Finding Maximal Common Sub-parse Thickets for Multi-sentence Search. // Graph Structures for Knowledge Representation and Reasoning. Springer. – 2014. – P. 39-57. 1.1 п.л. (вклад автора 0.3 п.л.)
7. Ilvovsky D. Going beyond sentences when applying tree kernels // Proceedings of the Student Research Workshop.– ACL 2014.– P. 56-63. 0.75 п.л.

8. Galitsky B., Ilvovsky D., Kuznetsov S. Rhetoric map of an answer to compound queries.– ACL-IJCNLP 2015.– Vol. 2: Short papers. Beijing: 2015. P. 681-686. 0.6 п.л. (вклад автора 0.2 п.л.)

9. Galitsky B., Ilvovsky D., Kuznetsov S. O. Text Classification into Abstract Classes Based on Discourse Structure. Proceedings of the RANLP 2015. Hissar: 2015. P. 201-207. 0.8 п.л. (вклад автора 0.3 п.л.)

10. Galitsky B., Ilvovsky D., Kuznetsov S. Text integrity assessment: Sentiment profile vs rhetoric structure. 16th International Conference, CICLing 2015, Cairo, Proceedings, Part II. Vol. 9042. Springer International Publishing, 2015. P.126-139. 0.7 п.л. (вклад автора 0.3 п.л.)

11. Mahalova T. N., Ilvovsky D., Galitsky B. Pattern structures for news clustering. Proceedings of the International Workshop FCA4AI at IJCAI 2015. Buenos Aires: 2015. Ch. 5. P.35-42. 0.6 п.л. (вклад автора 0.2 п.л.)

12. Strok F. V., Galitsky B., Ilvovsky D. Pattern Structure Projections for Learning Discourse Structures. 16th International Conference, AIMSA 2014, Varna, Bulgaria, 2014. Proceedings. Vol. 8722. L., NY, Dordrecht, Heidelberg, Springer, 2014. P.254-260. 0.5 п.л. (вклад автора 0.2 п.л.)

Лицензия ЛР № 020832 от 15 ноября 1993 г.

Подписано в печать _____ 2017 г.

Формат 60 х 84/16

Бумага офсетная.

Печать офсетная.

Усл. печ. л. 1

Тираж 100 экз. Заказ №_____

Типография издательства

Национального исследовательского университета – Высшей школы экономики,

125319, г. Москва, Кочновский проезд, д.3