ОТЗЫВ

официального оппонента на диссертацию
Ильвовского Дмитрия Алексеевича на тему:
«Методы и алгоритмы обработки текстовых данных на основе графовых дискурсивных моделей»,
представленной на соискание учёной степени кандидата технических наук
по специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ».

1. **Актуальность темы диссертации.**

Современное состояние информационного общества характеризуется лавинообразным нарастанием объёмов данных, среди которых текстовые данные занимают всё большую часть, являясь основным видом данных в сети Интернет. Это требует создания новых методов обработки текстовых данных, позволяющих выполнять более глубокий, семантический анализ текстов с целью автоматического извлечения из них знаний, воспринимаемых и обрабатываемых компьютерными системами.

Данная диссертация относится к направлению исследований, обозначаемому термином Text Mining, что в русскоязычной терминологии соответствует терминам «Интеллектуальный анализ текстовых данных», «Разработка текстовых данных» или даже «Понимание текста». Последнее обусловлено тем, что решение фундаментальной проблемы понимания текста компьютером является стратегической целью развития данного направления.

Все методы анализа текстов, применяемые в Text Mining, классифицируются по двум направлениям: методы, использующие статистики встречаемости слов в текстах, и методы, основанные на применении семантических моделей текста. Данная работа относится ко второму, семантическому направлению. В ней используются методы Анализа формальных понятий (АФП) и дискурсивных теорий. Преимуществом методов Анализа формальных понятий является строгая математическая основа предлагаемых решений и их универсальность. Формальный контекст, - основная модель АФП, определён на произвольных множествах, поэтому может применяться к данным любой природы, не обязательно к текстам. Применение методов АФП к текстовым данным является новым направлением, в котором не так много результатов. Особенно это касается построения семантических моделей текста, выходящих за рамки отдельного предложения. В данной работе как раз создана такая модель. Это определяет актуальность темы диссертации. Разрабатываемые в ней модели текстовых данных на основе графовых

дискурсивных моделей позволяют на новом уровне решать задачи поиска, классификации и кластеризации текстов.

2. **Содержание диссертации**.

Диссертация состоит из введения, пяти разделов, заключения, включает список литературы и семь приложений.

Во введении обосновывается актуальность темы исследования, формулируются задачи исследования, описываются применяемые для решения методы, излагаются основные результаты работы, даётся оценка их новизны, научной и практической ценности, приводятся сведения об апробации и внедрении результатов работы, а также обзор содержания работы по разделам.

Первый раздел носит вводный характер и содержит необходимые сведения из теории решёток, АФП, дискурсивных теорий. Приводятся модели представления текста, среди которых "мешок слов», деревья синтаксического разбора, дискурсивные модели, чащи разбора. Рассматривается задача машинного обучения на текстовых данных и применение ядерных функций в её решениях. Материал раздела имеет объем, необходимый для представления содержания работы в последующих разделах, изложен математически корректно, с необходимыми ссылками на литературу.

Второй раздел работы посвящён описанию разработанных в диссертации моделей и методов поиска ответов на сложные запросы. Обосновывается необходимость разработки таких методов для случаев, когда поисковые запросы представляют собой несколько предложений. Далее вводится модель текстового абзаца, основанная на понятии чащи разбора. Для сравнения текстовых абзацев применяется операция их обобщения, в которой используются несинтаксические связи. Правильно отмечается, что выполнение операции обобщения на полных описаниях является *NP*-трудной задачей. Для эффективного вычисления обобщения с сохранением свойств данной операции предлагается воспользоваться механизмом проекций чащ разбора. Изложенные положения и утверждения далее используются для получения конкретных результатов: алгоритма вычисления сходства для двух абзацев с применением проекций, алгоритма кластеризации текстов, использующего узорные решётки. В разделе имеются примеры оценки релевантности поиска по сложным запросам и кластеризации текстов по предложенному алгоритму.

Третий раздел диссертации содержит результаты решения задачи обучения с учителем на текстовых абзацах, в котором используются ядерные функции в методе

опорных векторов. Раздел начинается с достаточно информативного обзора существующих методов и результатов машинного обучения на текстовых данных. Далее описываются подходы к построению ядер в методе опорных векторов и приводятся решения задачи поиска ответов на сложные запросы. Также в разделе приведены результаты решения задачи классификации технических документов. В разделе подробно описаны условия проведения вычислительных экспериментов с данными, организация тестовых данных, применяемые программные средства, даны корректные ссылки на внешние сетевые ресурсы. Выводы об эффективности предложенных решений, помещённые в конце раздела, подтверждаются приведёнными в разделе результатами экспериментов.

Четвёртый раздел работы посвящён поиску тождественных денотатов в онтологиях и формальных контекстах. Даётся определение денотата и приводится алгоритм поиска тождественных денотатов. Рассматриваются также альтернативные методы решения данной задачи. Алгоритм поиска тождественных денотатов основан на преобразовании онтологии в формальный контекст, в котором идентифицируются понятия. Онтология представляет собой сложно организованный объект, поэтому в разделе предлагается ряд решений, позволяющих применить аппарат АФП к решению задачи поиска тождественных денотатов в онтологиях. Раздел содержит подробное описание предлагаемых решений. Далее в разделе приведены описания и результаты экспериментов по проверке эффективности решения задачи поиска тождественных денотатов.

Пятый раздел диссертации содержит описание программных решений, применяемых в вычислительных экспериментах. В начале раздела помещён краткий обзор существующих программных средств АФП. Далее описывается программный комплекс FCART для анализа данных методами АФП, разработанный при участии автора. Приведены базовые понятия, отражённые в функциях комплекса: аналитические артефакты, решатели, визуализаторы, отчёты. Показана программная архитектура комплекса, его пользовательский интерфейс, примеры работы на конкретных данных. В разделе чётко разделены программные решения, разработанные автором, и применяемые им сторонние программные средства.

В целом содержание диссертации позволяет составить исчерпывающее представление о разработанных в ней методах и результатах их применения. Текст диссертации написан грамотно, математические определения и результаты изложены достаточно строго. В тексте диссертации не обнаружены опечатки.

Список литературы достаточно полно отражает современное состояние в выбранной области исследований, включает как классические работы, монографии, так и последние статьи. Выборочный контроль не обнаружил в списке работ, на которых нет ссылок в тексте.

Автореферат диссертации в полной мере отражает её структуру, содержание, положения и выводы.

2. **Обоснованность научных положений, выводов и рекомендаций, сформулированных в диссертации**.

Автором исчерпывающе описана исходная проблематика в выбранной области исследования и правильно сформулированы задачи данного исследования. Для решения поставленных задач выбраны соответствующие им подходы и методы. Полученные результаты изложены полно и детально проиллюстрированы в работе. Выводы об эффективности полученных результатов не подлежат сомнению.

3. **Достоверность и новизна исследования, полученных результатов, сформулированных в диссертации.**

Диссертация в целом представляет собой новое исследование в области Анализа формальных понятий, имеющее практическое значение. В работе получены два новых научных результата: графовая модель текстов, использующая и обобщающая структурное синтактико-дискурсивное представление текстового абзаца в виде чащ разбора и модель тождественных денотатов для формальных описаний. Достоверность полученных результатов подтверждена строгостью используемых для их получения моделей, экспериментальной проверкой результатов численных расчётов и продемонстрированной практической эффективностью программных реализаций. Также достоверность полученных результатов подтверждается их публикациями в научной печати уровня, соответствующего требованиям ВАК. Публикации автора отражают содержание и результаты диссертации достаточно полно.

4. Значимость для науки и практики полученных автором результатов.

Полученные автором результаты носят междисциплинарный характер. С одной стороны, разработано новое приложение методов АФП к сложным текстам в виде абзацев. С другой

стороны, полученные результаты важны в математической лингвистике, где существует проблема построения семантических моделей текста, не ограниченных одним предложением. Алгоритмы и их код, разработанные в диссертации, позволяют применять найденные решения в технологиях анализа и обработки текстовых данных, выводя их на новый уровень. В этом состоит практическое значение данной работы.

5. Замечания по диссертационной работе.

1. В работе на стр. 68 имеется ссылка на параметры проведения эксперимента, описанные в работе [73]. Однако среди авторов данной работы нет соискателя. В целом, авторство соискателя не вызывает сомнения, но следовало бы сослаться на другую работу с его участием.
2. Примеры с текстами в разд. 2.3.2 «Различные подходы к выявлению сходства между текстовыми абзацами» изложены недостаточно подробно. Из них неясно, как выполняется попарное обобщение абзацев.
3. Аналогичное замечание относится к разд. 3.2 «Пример расширения деревьев разбора».
4. В разделе 2.7 оценка вычислительной сложности даётся на примере и приведены ссылки на литературу. Следовало бы привести в данном разделе известные аналитические результаты, касающиеся оценки вычислительной сложности, в том числе и из литературы, на которую выполнены ссылки.
5. Нецелесообразно помещать в приложения (Приложения 1-6) код программ, тем более фрагменты кода без подробных комментариев. Лучше было бы раскрыть в приложениях детали, касающиеся экспериментальной части работы.

6. Заключительная оценка работы

Указанные замечания не влияют на общую положительную оценку диссертационной работы. Диссертация Ильвовского Дмитрия Алексеевича на соискание учёной степени кандидата наук является законченной научно-квалификационной работой, в которой создана новая графовая модель семантики текста, разработаны алгоритмы её реализации в системах поиска ответов на сложные текстовые запросы и в системах классификации текстов. В работе также получено новое решение задачи нахождения тождественных денотатов для формальных описаний. Результаты данной научной работы имеют практическое значение, поскольку создают основу для построения новых

информационных технологий более глубокого анализа текстовых данных. В целом, в работе сделан практический вклад в решение проблемы понимания текста.

Считаю, что данная диссертационная работа полностью отвечает требованиям, предъявляемым ВАК к кандидатским диссертациям согласно п. 7 «Положения о порядке присуждения учёных степеней», соответствует специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ», а её автор, Ильвовский Дмитрий Алексеевич заслуживает присуждения ему учёной степени кандидата технических наук.

Официальный оппонент Богатырев Михаил Юрьевич,
доктор технических наук, доцент, 22.05. 2017
профессор кафедры Информационной безопасности.
Адрес электронной почты: okkambo@mail.ru
Телефон: +79207552479
Федеральное государственное бюджетное образовательное
учреждение высшего образования «Тульский государственный университет»
Адрес: 300012, г. Тула, пр. Ленина, 92
http://tsu.tula.ru

Подпись профессора М.Ю. Богатырева подтверждаю:
Начальник
Административно-кадрового управления М. В. Метелищенкова
22.05.17