

ЗАКЛЮЧЕНИЕ ДИССЕРТАЦИОННОГО СОВЕТА Д 002.073.04 на базе федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН) по диссертации на соискание ученой степени кандидата технических наук.

аттестационное дело № _____

решение диссертационного совета от 05 июня 2017 г. № 9

О присуждении Ильвовскому Дмитрию Алексеевичу, гражданину Российской Федерации, ученой степени кандидата технических наук.

Диссертация «Методы и алгоритмы обработки текстовых данных на основе графовых дискурсивных моделей» по специальности 05.13.18 «Математическое моделирование, численные методы и комплексы программ» принята к защите 3 апреля 2017 г., протокол №7, диссертационным советом Д 002.073.04 на базе федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН), 119333, Москва, ул. Вавилова, д. 44, корп. 2, приказ Министерства образования и науки Российской Федерации № 747/нк от 22 июня 2016 г.

Соискатель – Ильвовский Дмитрий Алексеевич, 1986 г. рождения, в 2010 г. окончил государственное образовательное учреждение высшего профессионального образования «Московский авиационный институт (государственный технический университет) «МАИ»» по специальности «Прикладная математика и информатика»; в 2014 году окончил заочную аспирантуру федерального государственного автономного образовательного учреждения высшего профессионального образования «Национальный исследовательский университет «Высшая школа экономики» по специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ»; в 2016 г. был прикреплен для подготовки диссертации по научной специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ» к федеральному государственному автономному образовательному учреждению высшего образования «Национальный исследовательский университет «Высшая школа экономики». В настоящее время работает в федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский университет «Высшая школа экономики» в должности младшего научного

сотрудника международной научно-исследовательской лаборатории интеллектуальных систем и структурного анализа.

Диссертация выполнена в Департаменте анализа данных и искусственного интеллекта факультета компьютерных наук федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики».

Научный руководитель – доктор физико-математических наук, профессор Кузнецов Сергей Олегович, федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет «Высшая школа экономики», руководитель Департамента анализа данных и искусственного интеллекта.

Официальные оппоненты:

1. Богатырев Михаил Юрьевич, доктор технических наук, доцент, профессор кафедры информационной безопасности Института прикладной математики и компьютерных наук федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Тульский государственный университет».

Отзыв оппонента положительный, имеются замечания:

- 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.
- 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.
- Аналогичное замечание относится к разд. 3.2 «Пример расширения деревьев разбора».
- 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.

- 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.

2. Виноградов Дмитрий Вячеславович, кандидат физико-математических наук, старший научный сотрудник лаборатории интеллектуального анализа данных и автоматизированной поддержки научных исследований федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук»

Отзыв оппонента положительный, имеются замечания:

- В обзоре предыдущих подходов к проблеме не упомянуты работы отечественных исследователей.
- Описание математических деталей частных подходов (ядра и онтологии) перенести из первой главы в соответствующие главы (3 и 4, соответственно). Тогда первая глава станет компактнее, а читателю будет удобно читать о применении ядер для коротких текстов и поиске тождественных сущностей в онтологиях, ознакомившись с относящимися к делу определениями в первых параграфах соответствующих глав.
- Рецензент находит странным порядок перечисления использованных источников в списке литературы.
- В списке литературы некоторые отечественные работы процитированы на английском языке (В.К. Финн, С.А. Евтушенко, одна и та же работа которого цитируется 2 раза в [11] по-русски, а в [103] – по-английски).
- Исправления опечаток:
 - ✓ Страница 22: в определении 1.9 следует опустить индексы у знаков частичного порядка.
 - ✓ Страница 24: операция может быть алгоритмически (не асимптотически!) сложной.
 - ✓ Страница 27: в формуле для $\text{idf}(t,D)$ опечатка.

Ведущая организация – федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН) дала положительное заключение (заключение составил и подписал Цинман Леонид Львович, кандидат физико-

математических наук, ведущий научный сотрудник лаборатории компьютерной лингвистики ИППИ РАН им. А.А. Харкевича). Замечания:

1. В связи с тем, что в работе используется широкий спектр методов и технологий из разных областей, хотелось бы видеть их более подробное описание. Из текста диссертации часто понятно, как именно используется та или иная технология, но не до конца ясно, для чего она используется. Описание редко иллюстрируется содержательными примерами.
2. В определении понятия «чаши разбора» существенно используется парсер лингвистического процессора ЭТАП-3, разработанный в ИППИ РАН им. А.А. Харкевича. В частности, рис. 1.3 на стр. 40 прямо воспроизводит структуры, полученные с помощью этого парсера. Между тем никакой ссылки на указанный процессор в диссертации нет.
3. Глава 4 о поиске тождественных денотатов весьма отдалённо соотносится с основной тематикой диссертации. Несмотря на схожесть используемых методов (решётки формальных понятий), в данной главе речь идёт не об обработке текста, а об обработке формальных объектов, уже заранее сконструированных из текста некоторыми алгоритмами, описание которых не приводится. Более релевантным было бы рассмотреть задачу в комплексе: взяв текст, автоматически сформировать на его основе онтологию, а затем приведённым алгоритмом выявить в ней тождественные объекты, решив таким образом задачу кореферентности текстовых выражений.

Есть целый ряд замечаний к таблице 2.1 на странице 68 (Оценка релевантности поиска) и описанию, приведенному под этой таблицей:

4. В описании сравнивается метод проекций с другими методами, а в заголовках таблицы слово «проекция» не встречается, поэтому сложно понять о каких колонках таблицы идёт речь.
5. Автор заявляет «Нетрудно видеть, что с ростом сложности запроса увеличивался эффект от применения технологии обобщения». Не совсем понятно, на основе каких цифр сделано данное утверждение. Беглое сравнение колонок показывает, что это справедливо только для одного (второго) типа запроса из трёх. Было бы уместно провести сравнение или даже выразить «эффект» явными цифрами в таблице, но этого не сделано.
6. Автор утверждает также, что «другим важным выводом является незначительная потеря в точности при существенном выигрыше в скорости за

счет использования проекций». Однако данные по скорости в таблице не представлены.

Соискатель имеет 20 опубликованных научных работ, из них по теме диссертации опубликовано 15 научных работ общим объемом 10.15 п.л., в том числе, 3 статьи – в научных журналах и изданиях, которые включены в перечень российских журналов и изданий для опубликования основных результатов диссертаций, 12 статей - в рецензируемых трудах международных конференций. Личный вклад соискателя составляет 5 п.л.

Наиболее значимые работы по теме диссертации:

1. 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.
2. 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.)
3. 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.)
4. Ilvovsky D. Going beyond sentences when applying tree kernels // Proceedings of the Student Research Workshop.– ACL 2014.– P. 56-63. 0.75 п.л.
5. 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.)
6. 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.)

На диссертацию и автореферат поступило 4 отзыва, которые подписали:

1. Друза Алексей Валерьевич, кандидат физико-математических наук, разработчик-исследователь программного обеспечения ООО «Яндекс». Отзыв положительный, содержит 1 замечание:

- В диссертации показано, что возможность автоматически учитывать синтактико-дискурсивную структуру текста позволяет существенно повысить качество поиска. При этом сложность синтаксического и дискурсивного анализа текста не позволяет эффективно реализовать его на очень больших поисковых массивах, но позволяет сделать это при обработке, например, первых 50 или 100 выданных при поиске документов. В качестве недостатка работы стоит отметить, что в автореферате (в отличие от диссертации) это соображение не высказано в явном виде.
2. Лукашевич Наталья Валентиновна, доктор технических наук, ведущий научный сотрудник Научно-исследовательского вычислительного центра федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В. Ломоносова». Отзыв положительный, содержит 2 замечания:
- В автореферате неоднократно цитируется работа Alessandro Moschitti, однако нигде не дается ссылка на его конкретную работу
 - В автореферате десятичные числа записываются то через запятую, то через точку.
3. Петровский Михаил Игоревич, кандидат физико-математических наук, доцент факультета вычислительной математики и кибернетики федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В. Ломоносова». Отзыв положительный, содержит 4 замечания:
- Хотя работа посвящена в первую очередь работе с текстами на английском языке, автору стоило пояснить, почему рассматриваемые в работе модели не могут быть (полностью или частично) применены для других языков, насколько они являются универсальными.
 - Из реферата не вполне ясно, все ли перечисленные типы связей внутри абзаца использовались автором в практических приложениях. Этот момент стоило отметить явно.
 - При сравнении предложенных автором моделей текстовых данных с существующими не рассмотрены одни из наиболее популярных современных моделей типа word2vec и doc2vec.

- В автореферате не приведен или другое формальное описание расчета предложенной автором ядерной (kernel) функции для оценки близости документов, хотя это один из важных результатов.

4. Чеповский Андрей Михайлович, доктор технических наук, профессор кафедры информационной безопасности факультета бизнес-информатики федерального государственного автономного образовательного учреждения высшего профессионального образования «Национальный исследовательский университет «Высшая школа экономики». Отзыв отрицательный, ниже приводятся замечания.

В диссертации заявляется о разработке модели текста (страница 10, п.1). Математические модели в компьютерной лингвистике могут быть построены на базе теории формальных грамматик, например. Но этого в диссертации нет. Модели на основе графов используются для описания связей объектов, выделяемых из текстов с целью анализа взаимодействия таких объектов. Такого в диссертации опять не наблюдается. Синтаксические деревья разбора не являются математическими моделями, а лишь визуализируют установленные в результате разбора связи между словами. Тем более нельзя утверждать, что создается некая новая модель, если деревья разбора в рамках одного абзаца объединили дополнительными связями (глава 3). Представления такого вида известны в прикладной лингвистике уже полвека и не рассматриваются обычно как математические модели, но могут служить дополнительным аппаратом в трансформационной грамматике, что далеко от текста диссертации Ильвовского Д.А. По приведенным причинам я утверждаю, что в диссертации Ильвовского Д.А. нет математических моделей, а тем более новых оригинальных предложений в области математического моделирования. Диссертационная работа Ильвовского Дмитрия Алексеевича на тему «Методы и алгоритмы обработки текстовых данных на основе графовых дискурсивных моделей» не соответствует Паспорту научной специальности 05.13.18 «Математическое моделирование, численные методы и комплексы программ» Номенклатуры специальностей научных работников, а именно формуле специальности в части необходимости присутствия в работе оригинальных результатов в области математического моделирования.

На странице 50 диссертации Ильвовского Д.А. при попытках описать «графовую модель текстовых абзацев» утверждается, что «для запросов,

состоящих из нескольких предложений, довольно трудно построить ранжирование, основанное на данных о пользовательских «кликах» по результатам...». Автор полагает, что ранжирование работает только на «кликах». Это неверное утверждение. Далее (второй абзац страницы 50) Автор говорит по сути по метапоисковой системе. Но выше по тексту он утверждает, что проблема в плохом качестве ранжирования. Тогда о каком переупорядочивании документов идет речь, если уже нашлись не те документы. Автор диссертации не понимает базовых механизмов функционирования поисковых систем (предварительная фильтрация результатов, многоуровневое ранжирование и т.д.). Таким образом, бессмысленны исходные предпосылки для построения методики (или модели) исследуемой предметной области. Весь последующий текст диссертации посвящен ошибочным задачам, что подтверждает следующий пример. На страницах 68-70 описывается некий эксперимент: «Попробуем оценить, как обобщение чаш разбора может улучшить поиск». На самом деле берется десять результатов поисковой системы Bing, которые просто меняются местами в выдаче. Никаких оснований считать, что верный документ оказался на 1-ой позиции в тексте не приводятся. Нельзя верить такому эксперименту, в основу которого положено наличие искомого документа в десяти полученных в результате запроса. Нельзя говорить об улучшении поиска, если результаты самого поиска не меняются. При этом понятие «точности» в обсуждаемом разделе 2.6 диссертации не определено, а понятие «полнота» не используется. Все это указывает на недобросовестность проведения эксперимента и представления результатов. Диссертация не соответствует требованиям п.9 Положения о присуждении ученых степеней, предъявляемым к кандидатским диссертациям в части содержания решения задачи, имеющей значение для развития соответствующей отрасли знаний в силу того, что нельзя решать задачу, которая не определена корректно. Кроме вышеизложенного в диссертационной работе Ильвовского Дмитрия Алексеевича присутствует следующий плагиат:

1. Со страницы 19 (начиная с «Определение 1.1 ...») и до страницы 22 текст полностью взят со страниц 8-11 (с точностью до удаления из текста названия «диаграммы Хассе») из диссертации Самохина Михаила Валерьевича «Машинное обучение на узорных структурах» (Диссертация на соискание

ученой степени кандидата технических наук по специальности 05.13.17 – «Теоретические основы информатики». Москва. ВИНТИ РАН, 2006 г. – 124с.). Ссылки на диссертацию Самохина М.В. или другую работу с данным текстом в диссертации Ильвовского Д.А. нет.

2. Начиная с абзаца «Теория риторических структур...» страницы 32 до страницы 37 текст является переводом на русский язык работы: William C. Mann and Sandra A. Thompson «Rhetorical Structure Theory: Toward a functional theory of text organization», 1988, Text 8 (3), 243-281. (<http://www.cis.upenn.edu/~nenkova/Courses/cis700-2/rst.pdf>). Ссылок на указанную статью в тексте диссертации Ильвовского Д.А. нет. В начале пункта есть упоминание других работ, включающих тех же авторов ([74, 75]): Mann, W., Matthissen, C., Thompson, S. Rhetorical Structure Theory and Text Analysis. Discourse Description: Diverse Linguistic analyses of a fund-raising text / ed by W.C.Mann and S.A.Thompson.-Amsterdam.-1992.-P.39-78. Но указаний на использование текста из какой-либо работы нет. Из этих примеров следует, что диссертация не соответствует требованиям п.14 Положения о присуждении ученых степеней, предъявляемым к кандидатским диссертациям.

По причинам того, что диссертационная работа не соответствует Паспорту специальности и не соответствует требованиям ВАК РФ, предъявляемым к кандидатским диссертациям и не соответствует критериям, изложенным в пунктах 9 и 14 «Положения о присуждении ученых степеней» (Постановление правительства РФ от 24.09.2013 г. №842), предъявляемым к кандидатским диссертациям, Ильвовскому Дмитрию Алексеевичу не может быть присуждена ученая степень кандидата технических наук по специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ».

На заседании диссертационного совета 05 июня 2017 г. отрицательный отзыв был **тщательно проанализирован**, в результате чего были сделаны следующие выводы:

1. Диссертация Ильвовского Дмитрия Алексеевича **полностью соответствует** Паспорту научной специальности 05.13.18 «Математическое моделирование, численные методы и комплексы программ» в части необходимости присутствия в работе оригинальных результатов в области математического моделирования.

2. Диссертация Ильвовского Дмитрия Алексеевича **полностью соответствует** требованиям п.9 Положения о присуждении ученых степеней, предъявляемым к кандидатским диссертациям в части содержания решения задачи, имеющей значение для развития соответствующей отрасли знаний. Все задачи, описываемые в работе, поставлены корректно и являются актуальными для рассматриваемой области.
3. Диссертация Ильвовского Дмитрия Алексеевича **полностью соответствует** п.14 Положения о присуждении ученых степеней. Диссертационная работа не содержит плагиата и недобросовестных заимствований. Результаты, полученные в работе, являются оригинальными, общий процент заимствований по тексту находится в пределах допустимой нормы. Все приведенные в отзыве фрагменты не являются плагиатом или недобросовестным заимствованием, поскольку в тексте диссертации во всех случаях явно указаны источники заимствования и присутствуют все необходимые ссылки. Все приведенные фрагменты в отзыве содержатся в обзорной части диссертации и не содержат описания оригинальных результатов, полученных соискателем.

Выбор официальных оппонентов и ведущей организации обуславливается их соответствием требованиям Положения о присуждении ученых степеней, широкой известностью своими достижениями в области моделирования и анализа текстовых данных, искусственного интеллекта, наличием публикаций в соответствующей сфере исследования и способностью определить научную и практическую ценность диссертации.

Диссертационный совет отмечает, что на основании выполненных соискателем исследований:

1. **Разработана** графовая модель текстов, использующая и обобщающая структурное синтактико-дискурсивное представление текстового абзаца (чащу разбора).
2. **Предложен** численный метод, использующий разработанную модель в задаче поиска ответов по сложным запросам.
3. **Реализован** численный метод, использующий ядерные функции в задаче классификации текстовых данных.

4. **Разработано** на базе предложенной модели таксономическое представление коллекции текстовых данных в виде решетки замкнутых структурных синтактико-дискурсивных описаний.
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.

Теоретическая значимость исследования обоснована тем, что были разработаны принципиально новые модели и методы: синтактико-дискурсивная модель текстов, позволяющая представлять текстовые абзацы в виде графов (полное описание) и лесов (приближенное описание) и вычислять сходство между текстами, таксономическое представление текстовых данных, модель и метод выявления тождественных денотатов для формальных описаний.

Значение полученных соискателем результатов исследования для практики подтверждается тем, что:

1. Разработан и внедрен в компании ООО «ФОРС-Центр разработки» метод классификации текстовых абзацев в проекте оценки пользовательских предпочтений.
2. Разработан и внедрен в компании ООО «Авикомп сервисез» метод выявления тождественных денотатов для оптимизации прикладной онтологии.
3. Проведены эксперименты по оценке релевантности поиска по сложным запросам, обучению на текстовых абзацах, выявлению тождественных денотатов, демонстрирующие улучшение по сравнению с существующими аналогами.
4. Разработанные методы и алгоритмы реализованы в виде программного комплекса, предназначенного для решения исследовательских и прикладных задач, связанных с анализом текстовых данных.

5. Разработанные алгоритмы и методы могут использоваться в преподавании и изучении ряда математических дисциплин, таких как «Автоматическая обработка текстов» и «Анализ неструктурированных данных».

Оценка достоверности результатов исследования выявила, что идея диссертации базируется на использовании соискателем корректного математического аппарата дискретной математики (в частности, теории решеток замкнутых описаний), на анализе международной литературы в области анализа и обработки текстовых данных и на обобщении передового опыта по разработке численных методов в этой области; что результаты диссертации являются воспроизводимыми, а проведенные эксперименты подтверждают превосходство предложенного способа моделирования.

Личный вклад соискателя состоит в:

1. разработке графовой модели представления текстового абзаца;
2. разработке численных методов поиска ответов по сложным запросам, классификации текстовых данных, проведении апробации предложенных методов.
3. разработке таксономического представления коллекции текстовых данных в виде решетки замкнутых структурных описаний;
4. разработке модели тождественных денотатов для формальных описаний;
5. разработке численного метода и алгоритма выявления тождественных денотатов и проведении экспериментальных исследований на прикладной онтологии;
6. реализации разработанных моделей, методов и алгоритмов в виде программных комплексов.

Диссертация охватывает основные вопросы поставленной научной задачи и соответствует критерию внутреннего единства, что подтверждается наличием последовательного плана исследования и взаимосвязанностью выводов.

На заседании 05 июня 2017 г. диссертационный совет пришел к выводу о том, что диссертация представляет собой научно-квалификационную работу, которая соответствует критериям, установленным Положением о порядке

совета Д 002.073.04, д.ф.-м.н., профессор

совета Д 002.073.04, д.т.н., профессор

В.Н. Крутько

«05» июня 2017 г.