

УТВЕРЖДАЮ

Проректор Национального
исследовательского университета
«Высшая школа экономики»

к.э.н., доцент Рошин Сергей Юрьевич

«22» ноября 2016 г.

ЗАКЛЮЧЕНИЕ

федерального государственного автономного образовательного
учреждения высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

Диссертация Ильвовского Дмитрия Алексеевича на тему: «Методы и алгоритмы обработки текстовых данных на основе графовых дискурсивных моделей» выполнена в Департаменте анализа данных и искусственного интеллекта факультета компьютерных наук федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики».

В период подготовки диссертации соискатель ученой степени Ильвовский Дмитрий Алексеевич работал в федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский университет «Высшая школа экономики» на должности младшего научного сотрудника международной научно-исследовательской лаборатории интеллектуальных систем и структурного анализа.

В 2010 г. Ильвовский Дмитрий Алексеевич 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.

С 24 октября 2016 г. по настоящее время 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».

Удостоверение о сдаче кандидатских экзаменов выдано в 2014 г. в федеральном государственном автономном образовательном учреждении высшего профессионального образования «Национальный исследовательский университет «Высшая школа экономики».

Научный руководитель – доктор физико-математических наук, профессор Кузнецов Сергей Олегович, работает в федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский университет «Высшая школа экономики», руководитель Департамента анализа данных и искусственного интеллекта.

По итогам обсуждения принято следующее заключение:

Актуальность работы. Обработка текстов на естественном языке порождает значительное количество открытых проблем, связанных с созданием и реализацией эффективных алгоритмов, которые, с одной стороны, должны иметь теоретическое обоснование, а с другой стороны – учитывать особенности конкретной задачи. Большинство существующих алгоритмов не учитывают структурные особенности текста и не позволяют оперировать формальным описанием текста, обладающим достаточным уровнем абстракции.

Методы анализа формальных понятий (АФП), основанные на свойствах решеток замкнутых множеств и узорных структур, предоставляют удобный и эффективный математический аппарат для решения задач, связанных с обработкой текстовых данных. Эти методы сочетают в себе несколько качеств, делающих их пригодными для работы с текстами. Во-первых, АФП позволяет работать с формальными описаниями произвольной степени детализации. Во-вторых, позволяет абстрагироваться от конкретного смысла и значения этих описаний, после того как сформулированы несколько простых правил работы с ними (в общем случае достаточно лишь операции вычисления сходства, обладающей заданными свойствами). В-третьих, благодаря концепции так называемых замкнутых описаний, позволяет использовать мощный и интуитивно понятный аппарат теории решеток (частичных порядков с дополнительными свойствами). Решетка одновременно является и весьма удобной моделью представления знаний, допускающей различные уровни детализации, и весьма проработанным и развитым средством для работы с данными. Эти свойства делают решетки привлекательными в плане применения к задачам обработки текста, поскольку уже известны самые разные методы и модели, позволяющие построить формальное описание текста на синтаксическом и дискурсивном уровне.

Современные исследования в теории решеток замкнутых описаний развиваются в нескольких основных направлениях: анализ предпочтений (Объедков), ленивая классификация (Кузнецов), трикластеризация (Игнатов),

интеграция с дескриптивной логикой (Рудольф), анализ сложных последовательностей (Наполи, Бузмаков) и т.д. Также развивается применение методов данной области для обработки текстовых данных. В число приложений в этой области входит работа со структурным представлением текста, классификация и кластеризация текстов, повышение релевантности информационного поиска, извлечение знаний из текста и т.д.

Обоснованность научных положений подтверждена строгостью построенных математических моделей, экспериментальной проверкой результатов численных расчетов и практической эффективностью программных реализаций.

Личное участие соискателя ученой степени в получении результатов, изложенных в диссертации, заключается в следующем:

- Создание модели представления текстовых абзацев, использующей дискурсивные и синтаксические связи внутри текстового абзаца и развивающей принцип наименее общего обобщения;
- Применение построенной модели для решения задачи поиска по сложным запросам;
- Разработка метода, позволяющего повысить качество обучения на текстовых абзацах с помощью использования дискурсивной информации;
- Построение метода кластеризации текстов на основе решеток замкнутых описаний;
- Разработка метода выявления дискурсивных связей «та же сущность» в формальных описаниях, построенных на основе текстовых данных;
- Реализация разработанных моделей, методов и алгоритмов в виде программного комплекса.

Степень достоверности результатов проведенных соискателем ученой степени исследований подтверждается полученными рецензиями и отзывами ведущих специалистов на международных конференциях и тестированием на открытых данных в соответствии со стандартами воспроизводимости.

Научная новизна работы. В диссертации автором был получен ряд новых научных результатов:

1. Разработана графовая модель текстов, использующая и обобщающая модель структурного синтактико-дискурсивного представления 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: графовой модели текстов, основанной на деревьях синтаксического разбора, таксономическом представлении текстовых данных, модели и методе выявления тождественных денотатов для формальных описаний.

Практическая значимость результатов проведенных соискателем ученой степени исследований 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. Компания Zvents использовала алгоритм поиска с использованием разработанного представления текстовых абзацев при создании интернет-магазина. Компания Knowledge Trail 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.

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, 16-22 сентября (доклад на тему: «Выявление дубликатов объектов в прикладных онтологиях на основе методов анализа формальных понятий»);
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, 24 марта (доклад на тему «FCA-based Search for Duplicate Objects in Ontologies»);
3. 11-й Международной конференции по анализу формальных понятий (ICFCA-2013), Дрезден, Германия, 21-24 мая (доклад на тему «FCART: A New FCA-based System for Data Analysis and Knowledge Discovery»);
4. 8-й Международной конференции по компьютерной лингвистике ДИАЛОГ-2013, Москва, Россия, 29 мая – 2 июня (доклад на тему «Matching sets of parse trees for answering multi-sentence questions, in: Proceedings of the Recent Advances in Natural Language Processing»);
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, 3 августа (доклад на тему «Graph Structures for Knowledge Representation and Reasoning Third International Workshop»);
6. 7-й Международной конференции по компьютерной лингвистике RANLP-2013, Хисаря, Болгария, 7-14 сентября (доклад на тему «Matching sets of parse trees for answering multi-sentence questions»);
7. 8-й Международной конференции по компьютерной лингвистике RANLP-2015, Хисаря, Болгария, 5-11 сентября (доклад на тему «Text Classification into Abstract Classes Based on Discourse Structure»);
8. 16-й Международной конференции по интеллектуальному анализу данных AIMSА-2014, Варна, Болгария, 11-13 сентября (доклад на тему «Pattern Structure Projections for Learning Discourse Structures»);
9. 14-й Международной конференции по интеллектуальной обработке текста и компьютерной лингвистике CICLING-2014, Катманду, Непал, 6-12 апреля (доклад на тему «Extending tree kernels towards paragraphs»);
10. 15-й Международной конференции по интеллектуальной обработке текста и компьютерной лингвистике CICLING-2015, Каир, Египет, 14-20 апреля (доклад на тему «Text Integrity Assessment: Sentiment Profile vs Rhetoric Structure»);
11. 52-й Международной конференции Ассоциации компьютерной лингвистики ACL-2014, Балтимор, США, 22-27 июня (доклад на тему «Going beyond sentences when applying tree kernels»);
12. 53-й Международной конференции Ассоциации компьютерной лингвистики ACL-2015, Пекин, Китай, 26-31 июля (доклад на тему «Rhetoric Map of an Answer to Compound Queries»).

Полнота изложения результатов диссертации в публикациях. Основные результаты работы изложены в 15 научных статьях. Общий объем публикаций – 10.15 п.л. Личный вклад автора – 5 п.л.

Работы, опубликованные автором в ведущих рецензируемых научных журналах и изданиях, рекомендованных ВАК:

1. Ильвовский, Д. Выявление дубликатов объектов в прикладных онтологиях с помощью методов анализа формальных понятий / Д. Ильвовский, М. Климушкин // Научно-техническая информация. Серия 2. Информационные процессы и системы. – 2013. - № 1. - С.10-17. 0.75 п.л. (личный 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 / Д. Ильвовский // Научно-техническая информация. Серия 2. Информационные процессы и системы. - 2014. - № 2. - С.28-37, 0.9 п.л. (личный вклад автора 0.9 п.л.)
3. Ильвовский, Д.А. Системы автоматической обработки текстов / Д.А. Ильвовский, Е.Л. Черняк // Открытые системы. СУБД. - 2014. № 01. С. 51-53. 0.45 п.л. (личный вклад автора 0.3 п.л.)

Работы, опубликованные автором в рецензируемых научных изданиях, входящих в международные реферативные базы данных и системы цитирования Web of Science, Scopus:

4. Improving Text Retrieval Efficiency with Pattern Structures on Parse Thickets / S.O. Kuznetsov, F.V. Strok, D.A. Ilvovsky, B. Galitsky // Proceedings of the FCAIR. Vol. 977. M.: CEUR Workshop Proceeding. – 2013. P.6-21 (Scopus) 0.77 п.л. (личный вклад автора 0.25 п.л.)
5. Matching sets of parse trees for answering multi-sentence questions / B. Galitsky, D.A. Ilvovsky, S.O. Kuznetsov, F.V. Strok // Proceedings of the RANLP 2013. – INCOMA Ltd. – 2013. – P. 285–293 (Scopus) 0.8 п.л. (личный вклад автора 0.3 п.л.)
6. Finding Maximal Common Sub-parse Thickets for Multi-sentence Search / B. Galitsky, D.A. Ilvovsky, S.O. Kuznetsov, F.V. Strok // Graph Structures for Knowledge Representation and Reasoning. Springer. – 2014. – P. 39-57 (Scopus and Web of Science) 1.1 п.л. (личный вклад автора 0.3 п.л.)
7. Ilvovsky, D. Rhetoric map of an answer to compound queries / B. Galitsky, D.A. Ilvovsky, S.O. Kuznetsov // ACL-IJCNLP 2015.– Vol. 2: Short papers. Beijing: 2015.– P. 681-686 (Scopus) 0.6 п.л. (личный вклад автора 0.2 п.л.)
8. Ilvovsky, D. Text Classification into Abstract Classes Based on Discourse Structure / B. Galitsky, D.A. Ilvovsky, S.O. Kuznetsov // Proceedings of the

- RANLP 2015. Hissar.— 2015.— P. 200-207 (Scopus) 0.8 п.л. (вклад автора 0.3 п.л.)
9. Ilvovsky, D. Text integrity assessment: Sentiment profile vs rhetoric structure / B. Galitsky, D.A. Ilvovsky, S.O. Kuznetsov // 16th International Conference, CICLing 2015.— Cairo.—Proceedings.— Part II.— Vol. 9042.— Springer International Publishing.— 2015.— P.126-139 (Scopus and Web of Science) 0.7 п.л. (личный вклад автора 0.3 п.л.)
 10. Ilvovsky, D. Pattern structures for news clustering. / T.N. Makhalova, D. Ilvovsky, B. Galitsky // Proceedings of the International Workshop FCA4AI at IJCAI 2015.— Buenos Aires.— 2015.— Vol. 1430. P.35-42 (Scopus) 0.6 п.л. (личный вклад автора 0.2 п.л.)
 11. Ilvovsky, D. Pattern Structure Projections for Learning Discourse Structures. / F.V. Strok, B. Galitsky, D.A. Ilvovsky // 16th International Conference, AIMS 2014, Varna, Bulgaria, 2014. Proceedings. Vol. 8722. L., NY, Dordrecht, Heidelberg, Springer.— 2014.— P.254-260 (Scopus and Web of Science) 0.5 п.л. (личный вклад автора 0.2 п.л.)

Работы, опубликованные автором в других изданиях:

12. Ilvovsky, D.A. FCART: A New FCA-based System for Data Analysis and Knowledge Discovery / A. Neznanov, D.A. Ilvovsky, S. Kuznetsov // Contributions to the 11th International Conference on Formal Concept Analysis. Dresden: Qucoza.— 2013.— P.31-44. 0.6 п.л. (личный вклад автора 0.2 п.л.)
13. Parse thicket representations of text paragraphs / B. Galitsky, D.A. Ilvovsky, S.O. Kuznetsov, F.V. Strok // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Т.1. Вып. 12 (19). М.: РГГУ.— 2013.— С. 134-145. 0.73 п.л. (личный вклад автора 0.2 п.л.)
14. Ильвовский, Д.А. Выявление дубликатов объектов в прикладных онтологиях на основе методов анализа формальных понятий / Д.А. Ильвовский, М.А. Климушкин // Сборник докладов 9-й Международной конференции ИОИ-2012. Торус Пресс.— 2012.— С.625-628. 0.2 п.л. (личный вклад автора 0.1 п.л.)
15. Ilvovsky, D. Going beyond sentences when applying tree kernels / D. Ilvovsky // Proceedings of the Student Research Workshop.— ACL 2014.— P. 56-63. 0.75 п.л.

Публикации полностью соответствуют теме диссертационного исследования и раскрывают ее основные положения. В публикациях подробно изложены все основные результаты и положения диссертационного исследования. Научные работы были представлены на нескольких ведущих международных конференциях и получили одобрение рецензентов.

Ценность научных работ соискателя ученой степени

В научных работах представлена модель представления текстового абзаца и её приложения к задачам автоматической обработки текстов. Модель

используется для поиска по сложным запросам, классификации и кластеризации текстов. В работах демонстрируется эффективность модели в предложенных задачах и её преимущество по сравнению с популярными альтернативными моделями по точности и полноте.

Диссертационная работа Ильвовского Дмитрия Алексеевича на тему: «Методы и алгоритмы для обработки текстовых данных на основе графовых дискурсивных моделей» – это законченная научно-квалификационная работа, которая соответствует требованиям пунктов 9, 10, 14 Положения о присуждении ученых степеней, утвержденного постановлением Правительства Российской Федерации от 24 сентября 2013 г. № 842, а также Паспорту научной специальности 05.13.18 «Математическое моделирование, численные методы и комплексы программ» (технические науки), а именно пунктам:

1. Разработка новых математических методов моделирования объектов и явлений;

2. Развитие качественных и приближенных аналитических методов исследования математических моделей;

4. Комплексные исследования научных и технических проблем с применением современной технологии математического моделирования и вычислительного эксперимента;

5. Разработка новых математических методов и алгоритмов проверки адекватности математических моделей объектов на основе данных натурального эксперимента.

Диссертация Ильвовского Дмитрия Алексеевича на тему: «Методы и алгоритмы для обработки текстовых данных на основе графовых дискурсивных моделей» рекомендуется к защите на соискание ученой степени кандидата технических наук по специальности 05.13.18 «Математическое моделирование, численные методы и комплексы программ» (технические науки).

Заключение принято на заседании Департамента анализа данных и искусственного интеллекта факультета компьютерных наук федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики». Присутствовало на заседании - 5 чел. Результаты голосования: «за» - 5 чел., «против» - 0 чел., «воздержалось» - 0 чел., протокол № 2 от 03 ноября 2016 г.

Доцент Департамента
анализа данных и искусственного интеллекта
факультета компьютерных наук
к.т.н,

Объедков Сергей Александрович

Подпись заверяю

специалист по кадрам
1 категории Коневская

21.11.2016

