УТВЕРЖДАЮ
Директор
Федерального государственного бюджетного
учреждения науки
Институт проблем передачи информации им.
А.А. Харкевича Российской академии наук
д.ф.-м.наук, профессор РАН
_____А.Н. Соболевский

12 мая 2017 г.

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

о диссертационной работе Дмитрия Алексеевича Ильвовского
«Методы и алгоритмы обработки текстовых данных на основе графовых дискурсивных моделей», представленной к защите на соискание ученой степени кандидата технических наук по специальности 05.13.18 – математическое моделирование, численные методы и комплексы программ

Диссертация Д.А. Ильвовского посвящена обработке текстовых данных (на английском языке), в основу которой положены созданные автором графовые дискурсивные модели (т.е. модели, включающие сведения о том, как связаны фрагменты текста между собой – такие сведения могут содержаться в тексте эксплицитно или обнаруживаться особыми средствами).

Диссертация насчитывает 245 страниц и состоит из введения, пяти глав, заключения, списка литературы из 129 названий (в подавляющем большинстве случаев это англоязычные работы) и семи приложений с программным кодом.

Во Введении обосновывается актуальность работы, формулируются метод и предмет исследования, ставятся конкретные задачи исследования, характеризуется ее научная новизна, перечисляются теоретические и практические достоинства работы (при этом подчеркивается ориентированность работы на практические результаты, достигнутые в различных проектах, связанных с оптимизацией информационного поиска). Здесь же дается краткий синопсис диссертационного исследования, а также приводятся сведения об апробации работы и ее частей. Приводится, в частности, внушительный список из 12 конференций и симпозиумов, проводившихся в России и за границей с 2012 по 2015 годы, на которых диссертантом докладывались отдельные результаты, излагаемые в диссертации. К этому следует добавить, что ведущей организации – ИППИ РАН им. А.А. Харкевича – многие положения рецензируемого диссертационного исследования знакомы по выступлению его автора с докладом на тему «Модели и методы обработки текстовых абзацев» в феврале 2016 года на семинаре по теоретической семантике, руководимом академиком Ю.Д. Апресяном (http://iitp.ru/ru/science/seminars/1197.htm).

В первой главе обсуждаются теоретические основы моделирования текстовых данных. Под моделированием текстовых данных автор понимает основную задачу современной компьютерной лингвистики, которая чаще именуется “обработкой данных на естественном языке” (или, в англоязычной терминологии, natural language processing). Здесь кратко характеризуются различные теории и модели представления естественно-языкового текста (в частности, простейшие модели типа «мешок слов» (bag of words) и более продвинутые модели используемые в синтаксических компонентах лингвистических теорий основные подходы к отображению структуры

предложения - деревья составляющих и деревья зависимостей, иллюстрируемые на материале хорошо известных компьютерно-лингвистических ресурсов типа Penn Tree Bank), дается представление об уровнях лингвистического описания, излагается теория риторических структур В. Манна и его коллег, теория речевых актов Дж. Остина и Дж. Серля, теория представления дискурса (DRT) Х. Кампа и У. Рейле, теория «Смысл – Текст» И.А. Мельчука, некоторые менее известные и менее распространенные лингвистические и компьютерно-лингвистические теории, такие как семантика баз данных Р.Хауссера и др. В главе вводится несколько ключевых понятий: «чаща разбора» (в английской терминологии – parse thicket), «решётка замкнутых описаний», «ядерная функция», «прикладная онтология», при помощи которых затем, в последующих главах, строятся сами модели.

Вторая глава, озаглавленная «Модели и методы поиска ответов на сложные запросы», посвящена проблеме поиска в тексте ответов на нетривиальные запросы пользователя. Для решения задачи автор использует графовую модель представления текстовых абзацев (уже упоминаемую чащу разбора), в которой в дополнение к синтаксическим группам используются расширенные группы, построенные на основе риторических, кореферентных и коммуникативных связей между предложениями. Здесь существенным образом используются охарактеризованные в главе 1 теория риторических структур и теория речевых актов. Приводятся методы сравнения абзацев при помощи обобщения чащ разбора на основе решёток замкнутых описаний, а также обобщения **проекций** чащ разбора для приближённого быстрого сравнения. Результаты эксперимента показывают заметное улучшение релевантности поиска по длинным фразам (в среднем от 58% до 70%) при использовании чащ разбора. Как дополнение приводится алгоритм иерархической кластеризации, целью которого является группировка найденных текстов по похожести.

В третьей главе «Применение ядер для классификации коротких текстов» построенная модель используется для задачи классификации текстов ограниченного объема (несколько предложений). При помощи объединения синтаксических групп из разных деревьев (связанных, например, кореферентными, таксономичискими или коммуникативными отношениями) строятся т.н. расширенные деревья, которые не являются корректными синтаксическими деревьями, но образуют адекватное пространство признаков для сравнения текстов. На расширенных деревьях строится ядерная функция, которая численно выражает сходство между абзацами при помощи подсчёта общих подстурктур. Полученный алгоритм применяется к задаче классификации результатов поиска на релевантные и нерелевантные, а также к задаче классификации технических документов на «описание оригинальных разработок» и «инструкция по оформлению описаний оригинальных разработок». В обоих случаях предлагаемый алгоритм даёт заметный выигрыш в правильности классификации даже по сравнению с алгоритмом, использующим синтаксические деревья, не говоря уже об алгоритмах, основанных на «мешке слов».

Четвёртая глава несколько выбивается из тематики диссертационного исследования. Она посвящена поиску тождественных денотатов в онтологиях, автоматически сгенерированных из предварительно обработанных текстовых данных.

Поставленная задача решается методом фильтрации решёток формальных понятий. Для этого онтология преобразуется в формальный контекст, где каждому объекту сопоставлено множество его бинарных признаков. При помощи кластеризации объектов в формальном контексте формируются формальные понятия. Затем на основе специально разработанных числовых критериев фильтрации определяется, какие кластеры объединяют тождественные объекты. Приводится сравнение результатов работы алгоритма с альтернативами – методами на основе экстенсиональной

устойчивости понятия, на основе меры абсолютного сходство и на основе расстояния Хэмминга. Продемонстрировано улучшение, достигаемое за счёт использования нового метода. Выдающимся результатом является небольшое падение точности алгоритма (до 90%) при росте полноты вплоть до 70%.

В пятой главе приводятся описание программных комплексов, используемых для построения чащ разбора, для работы с решётками замкнутых описаний, для анализа формальных понятий и др.

Диссертация Д.А. Ильвовского обладает целым рядом особенностей, которые ярко выделяют её на фоне других исследований, посвященных автоматической обработке текста.

Во-первых, в работе органично сочетаются структурные методы анализа текста (синтаксические деревья, кореферентные связи) со статистическими (ядерные функции, метод опорных векторов, машинное обучение).

Во-вторых, автор не ограничивается анализом связей внутри предложения, как это бывает в большинстве работ, а в дополнение к ним использует дискурсивные связи между предложениями абзаца, что является по существу центральной темой диссертации,. Автор убедительно показывает, что такой подход позволяет добиться лучших результатов в ряде задач, связанных с поиском, классификацией и кластеризацией текстов.

В-третьих, несмотря на использование разнородного программного обеспечения (Stanford NLP для построения синтаксических деревьев и кореферентных связей, парсер Joty для построения риторических связей, Bing API и Apache SOLR для взаимодействия с поисковыми системами, собственный алгоритм для построения коммуникативных связей и др.) их грамотное взаимодействие при построении чащ разбора для абзацев на основе решёток замкнутых описаний позволяет добиться синергетического эффекта.

Практическая ценность работы подтверждается ее внедрением в реальных проектах в компаниях «ФОРС-Центр разработки» и Авикомп.

К недостаткам работы можно отнести следующие.

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 «чащи 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.

Указанные замечания не носят принципиального характера и не оказывают влияния на положительную характеристику диссертационной работы в целом.

Результаты рецензируемого научного исследования соответствуют паспорту специальности 05.13.18 «Математическое моделирование, численные методы и комплексы программ» в пунктах: 1 – «Разработка новых математических методов моделирования объектов и явлений», 2 – «Развитие качественных и приближенных аналитических методов исследования математических моделей», 3 – «Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий», 4 – «Реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента».

Диссертационное исследование Дмитрия Алексеевича Ильвовского представляет собой законченную научно-квалификационную работу, в которой решены актуальные задачи

разработки структурной модели текстов на естественном языке, ориентированной на поиск, классификацию и кластеризацию текстов и использующей синтаксические и дискурсивные связи внутри текста;

применения построенной модели в задаче **поиска сходства текстов** с целью улучшения релевантности информационного поиска по сложным запросам пользователя;

применения построенной модели в задаче **классификации текстов** с целью повышения качества существующих методов за счет использования дискурсивной информации;

построения на основе разработанной модели таксономического представления текстовых документов с использованием решеток замкнутых структурных описаний и применения представления в задаче **кластеризации текстов**;

разработки математической модели и метода для определения связи типа «тождественная сущность» в построенных на основе текстовых данных формальных описаниях и эффективной алгоритмической реализации данной модели;

реализации разработанных моделей, методов и алгоритмов в виде программного комплекса.

Диссертация содержит важные результаты, обладающие научной новизной, теоретической значимостью и практической ценностью.

Таким образом, диссертационная работа Д.А. Ильвовского “Методы и алгоритмы обработки текстовых данных на основе графовых дискурсивных моделей”, отвечает

требованиям пунктов 9-14 Положения о защите ученых степеней, утвержденного Постановлением Правительства Российской Федерации от 24.09.2013 г. № 842 и требованиям ВАК, Минобрнауки РФ, предъявляемым к диссертациям на соискание ученой степени кандидата наук по специальности 05.13.18 - «Математическое моделирование, численные методы и комплексы программ», а ее автор, Д.А.Ильвовский, заслуживает присвоения ему ученой степени кандидата технических наук по указанной специальности.

11 мая 2017 г., г Москва.

Вед. научный сотрудник Лаборатории компьютерной лингвистики
ИППИ РАН им. А.А. Харкевича,
к.ф.-м.н.

Л.Л.Цинман

Отзыв утвержден на заседании лаборатории компьютерной лингвистики (№ 15) Института проблем передачи информации им. А.А.Харкевича РАН 11 мая 2017 г., протокол № 1 от 11 мая 2017 г.

И.о. зав. лабораторией компьютерной лингвистики
ИППИ РАН им. А.А. Харкевича,
к.ф.н.

Л.Л.Иомдин

Подписи Л.Л. Цинман и Л.Л. Иомдин
УДОСТОВЕРЯЮ
Зав. канцелярией

Сведения о ведущей организации: Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН).
Адрес: Россия, 103051, Москва, Большой Каретный пер., д. 19, строение 1. Телефон: +7 (495) 650-42-25.
E-mail: director@iitp.ru