

ЗАКЛЮЧЕНИЕ ДИССЕРТАЦИОННОГО СОВЕТА

Д.002.073.04, созданного на базе Федерального исследовательского центра «Информатика и Управление» Российской академии наук (ФИЦ ИУ РАН), в соответствии с приказом № 747/нк от 22 июня 2016, действующим на момент защиты диссертации, по диссертации Черняк Екатерины Леонидовны на соискание ученой степени кандидата технических наук.

Диссертация «Разработка вычислительных методов анализа текстов с использованием аннотированных суффиксных деревьев» в виде рукописи по специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ» выполнена в федеральном государственном автономном учреждении высшего образования «Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), в Департаменте анализа данных и искусственного интеллекта.

Диссертация принята к защите 30 августа 2016 г., протокол № 1.

Соискатель Черняк Екатерина Леонидовна, гражданка Российской Федерации, занимает должности младшего научного сотрудника Международной научно-исследовательской лаборатории интеллектуальных систем и структурного анализа и старшего преподавателя Департамента анализа данных и искусственного интеллекта федерального государственного автономного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики». В 2010 году соискатель закончила бакалавриат по специальности «Прикладная математика и информатика», в 2012 году – магистратуру по специальности «Математическое моделирование» федерального государственного автономного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики». В 2015 году окончила очную аспирантуру при Департаменте анализа данных и искусственного интеллекта федерального государственного автономного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики» по специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ».

Научный руководитель – доктор технических наук, профессор Миркин Борис Григорьевич, профессор Департамента анализа данных и искусственного интеллекта федерального государственного автономного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики».

Официальные оппоненты:

Моттль Вадим Вячеславович, гражданин РФ, профессор, доктор технических наук, ведущий научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук.

Отзыв оппонента положительный, имеются замечания:

Не освещен вопрос оптимальной организации вычислений, например, параллелизации при очевидной низкой вычислительной эффективности метода аннотированного суффиксного дерева.

В диссертации приведены результаты трех групп экспериментов, иллюстрирующих три разные задачи анализа текстов – рубрикация научных

публикаций, пополнение таксономии научного знания и фильтрация обценной лексики. Очевидно, что язык текста (русский, английский, немецкий) оказывает существенное влияние на характер результата анализа. Однако в диссертации эксперименты с каждым из видов анализа текста проведены только на одном языке, русском либо английском, и фактически нет экспериментов по влиянию языка на результат анализа текста.

К сожалению диссертация не свободна от опечаток, например:

- стр. 15: «в корне *не верно* с точки зрения...»,
- стр. 54: «англоязычные статьи *в* из области... ».
- стр. 72: «подходы к построению или *пополнения*...».

Петровский Михаил Игоревич, гражданин РФ, кандидат физико-математических наук, доцент Кафедры автоматизации систем вычислительных комплексов Факультета вычислительной математики и кибернетики Московского государственного университета им. М. В. Ломоносова.

Отзыв оппонента положительный, имеется ряд замечаний:

В диссертационной работе приведен обзор популярных моделей представления текстовых данных, а также подходов и методов к определению меры сходства между текстами и к оценке релевантности строки тексту. В то же время в этом обзоре не отражены два таких важных современных направления, как использования модели представления типа word2vec, а также подходов к определению меры сходства или релевантности текстовых данных на основе ядерных функций (string kernels). Последнее особенно актуально, так как существуют методы построения ядерных функций для строк символов, позволяющих рассчитывать степень сходства на основе совпадения всех подстрок, что идейно близко к подходу, предложенному автором.

В работе не до конца освещены необходимость и полезность использования шкалирующих функций при задании меры релевантности, не обсуждается, как выбор функции шкалирования может повлиять на результат, и никак не ограничивается класс этих функций.

При представлении «строка–текст», предложенном автором, набор строк по сути генерируется шкалирующим окном из нескольких слов. Упоминается, что размер окна, и, судя по всему, степень перекрытия, должны подбираться эмпирически. При этом, вопрос выбора этих параметров в экспериментах практически не обсуждается, хотя, скорее всего, влияние этих параметров существенно.

В работе упоминается что «латентно-семантический анализ не является интерактивным методом, то есть, при появлении нового текста в коллекции необходимо заново формировать матрицу терм – текст и заново вычислять сингулярные матрицы и матрицу сингулярных значений». Это верно только отчасти. Некоторые методы LSA, такие как методы на основе неотрицательной матричной факторизации, действительно не позволяют представить новый документ в пространстве признаков на основе тематик коллекции без пересчета самой тематической модели. Но большинство методов LSA, в том числе, на основе главных компонент или латентного размещения Дирихле, позволяют представить новый документ в пространстве тематических признаков коллекции без пересчета тематической модели.

Также есть замечания по структуре работы. Главы с описанием практических задач не достаточно сбалансированы по уровню детализации, а также невелики по объему. Не приведена аргументация, почему именно эти задачи были выбраны в качестве демонстрационных. Заключительная глава с описанием программной

реализации очень короткая и собственно детали реализации, которые позволили бы оценить уровень трудозатрат автора, практически не освещены, а выводы по этой главе отсутствуют. Имело бы смысл объединить главы с третьей по пятую в одну общую главу с обоснованием выбора прикладных задач, а описание программной реализации вынести в приложение, добавив больше деталей и выводы.

В работе содержатся опечатки, не препятствующие общему пониманию работы, но иногда вводящие в заблуждение, например:

- На стр. 6 полная релевантность строки тексту определяется как сумма вероятностей максимальных совпадений, а на стр. 41 тот же термин определяется как усредненная сумма.
- На стр. 27 в тексте для описания рисунка 1.1 используются обозначения вершин, не совпадающие с обозначениями на рисунке (не u и v , а u и w).
- На стр. 43 упоминается три типа шкалирующих функций, в то время как в работе используется семь разных шкалирующих функций.
- Также присутствует ряд технических опечаток. Например: на стр. 16 – тире в начале строчки «за отсеивание слов»; стр. 21 – «скрытое переменной, тема состоит набора»; стр. 32 – «Предварительна подготовка»; стр. 39 – «текстовый текст»; стр. 45 – отсутствует точка в последней строчке и другие.

Ведущая организация – Федеральное государственное учреждение «Федеральный исследовательский центр Институт прикладной математики им. М. В. Келдыша» Российской академии наук, г. Москва дала положительное заключение (заключение составлено Ройтбергом Михаилом Абрамовичем, доктором физико-математических наук, заведующим лабораторией прикладной математики ФГУ ФИЦ ИПМ им. Келдыша). Отзыв положительный, не имеет замечаний по существу работы.

На диссертацию и автореферат поступило 3 отзыва, которые подписали:

1. Зобнин Алексей Игоревич, кандидат физико-математических наук, руководитель группы морфологии ООО «Яндекс».

Отзыв положительный, содержит 4 замечания:

- Имеется ряд опечаток: например, вместо «каждая строка помечена непустой подстрой строки S », очевидно, должно быть «каждое ребро помечено непустой подстрокой строки S » (стр. 7).
- Для нестрогих неравенств следовало бы использовать символ “ $\leq$ ”, а не “ $<=$ ” (стр. 4).
- При сравнении разных методов решения задач обработки текста стоило бы сравнивать не только вычислительную сложность, но и затраты потребляемой памяти.
- n -граммные языковые модели со сглаживанием названы «вычислительно неэффективными» (стр. 4), хотя на самом деле их эффективность зависит от используемых структур данных.

2. Браславский Павел Исаакович, кандидат технических наук, старший научный сотрудник Института математики и компьютерных наук Уральского федерального университета им. первого Президента России Б. Н. Ельцина.

Отзыв положительный, содержит 7 замечаний:

- На наш взгляд, *модель релевантности* – не очень удачный термин, в зависимости от приложения лучше использовать *модель информационного поиска* (*information retrieval model*) или *мера близости текста* (*text similarity*).

measure). *Релевантность* – это соответствие документа информационной потребности пользователя (а не текста запросу).

- Автор обосновывает актуальность предложенного метода недостатками классических моделей информационного поиска (BM24, языкового моделирования). Было бы логично и последовательно провести оценку предлагаемого метода по сравнению с другими моделями на задаче поиска документа по запросу, благо есть достаточное количество размеченных коллекций.

- Текст автореферата не содержит важных деталей оценки предложенного метода в задаче рубрикации аннотаций научных статей. Например, непонятно, почему были выбраны метрики MAP и nDCG, предназначенные для оценки ранжированного поиска, или, почему не были использованы стандартные коллекции для оценки методов классификации текстов (например, RCV1), которые позволили бы сравнить эффективность разработанного метода с более широким спектром аналогов.

- Наряду с определением аннотированного суффиксного дерева (АСД), в автореферате дается 2 свойства АСД. При этом нет каких-либо пояснений о важности этих свойств и их применении в работе.

- Не указана оценка сложности наивного алгоритма построения АСД и, соответственно, не приводятся сравнения по сложности модификации алгоритма Укконена и наивного алгоритма.

- Из описания четвертой главы неясно, почему согласованность разных частей исследования измеряется по-разному. Логично было бы в обоих случаях приводить и процент согласия, и каппу Коэна. Значение каппы Коэна 0,319 говорит о низкой согласованности экспертов.

- Результаты сравнения методов в описании пятой главы было бы логично оформить в виде таблицы. Не очень понятно, что являлось единицей оценки – отдельное слово или весь документ?

3. Козеренко Елена Борисовна, кандидат филологических наук, заведующая Лабораторией компьютерной лингвистики и когнитивных технологий обработки текстов Института проблем информатики Российской академии наук Федерального исследовательского центра «Информатика и управление» Российской академии наук. Отзыв положительный и не содержит замечаний.

В дискуссии приняли участие члены диссертационного совета:

1. Арлазаров В. Л., д. т. н., профессор, член-корреспондент Российской академии наук

2. Попков Ю. С., д. т. н., профессор, академик Российской академии наук
Соискатель имеет 21 опубликованную работу, из них по теме диссертации опубликовано – 13 научных работ общим объемом 5.5 печатных листов, в том числе 3 статьи в научных журналах и изданиях, которые включены в перечень российских рецензируемых научных журналов и изданий для опубликования основных научных результатов диссертаций, а также 1 публикация в зарубежных научных изданиях.

Соискателем опубликовано 9 работ в материалах всероссийских и международных конференций и симпозиумов.

Наиболее значимые работы по теме диссертации:

1. Черняк Е.Л. Меры релевантности строка-текст в проблеме рубрикации научных статей / Черняк Е.Л., Миркин Б.Г. // Бизнес-информатика. 2014. № 2. С. 51–62.

2. Черняк Е.Л. Метод аннотированного суффиксного дерева для оценки

степени вхождения строк в текстовые документы / Черняк Е.Л., Миркин Б.Г., Чугунова О.Н. // Бизнес-информатика. 2012. № 2. С. 31–41.

3. Черняк Е.Л. Системы автоматической обработки текстов / Черняк Е.Л., Ильвовский Д.А. // Открытые системы. СУБД. 2014. № 1. С. 51–43.

4. Chernyak E.L. Mirkin B.G. An AST method for scoring string-to-text similarity in semantic text analysis. Springer. 2014. Vol. 92. P. 92 – 96

5. Chernyak E.L. An approach to the problem of annotation of research publications / Chernyak E.L. // Proceedings of 8th ACM International Conference On Web Search and Data Mining. ACM. 2014. Vol. 58. P. 429-434

Диссертационный совет отмечает, что на основании выполненных соискателем исследований:

- Разработана новая теоретико-множественная модель совокупности «строка – текст»;
- Предложена адаптация метода аннотированного-суффиксного дерева для вычисления параметров теоретико-множественной модель, адаптация алгоритма Укконена для построения аннотированного суффиксного дерева, а также новые методы: рубрикации научных статей; пополнения таксономии предметной области; фильтрации обценной лексики;
- Проведены эксперименты, в которых продемонстрирована эффективность предложенных методов: метода рубрикации научных статей, метода пополнения таксономии предметной области, метода фильтрации обценной лексики;
- Разработаны комплексы программ, реализующие метод построения аннотированного суффиксного дерева и метод вычисления оценок релевантности на его основе.

Теоретическая значимость работы обоснована тем, что: разработана новая теоретико-множественная модель совокупности «строка – текст», расширена модель нормированного аннотированного суффиксного дерева с критерием релевантности СУВСС, предложен новый метода построения таблиц релевантности «строка – текст» (РСТ) для применения в конкретных задачах, проведен сравнительный анализ моделей представлений текстов и их применимости в различных задачах автоматического анализа и обработки коллекций текстовых документов; проведена модернизация аппарата суффиксных деревьев для использования в задачах автоматического анализа и обработки текстов.

Практическая значимость подтверждается тем, что разработан и внедрен метод фильтрации обценной лексики для анализа и определения тональности текстов в социальных сетях в системе FORS Media ГК «ФОРС»; разработан и внедрен метод рубрикации проектной документации ЗАО «ЕС-Лизинг»; определены пределы и перспективы использования теоретико-множественный моделей для автоматического анализа и обработки коллекций текстовых документов; создан комплекс программ компьютерного моделирования коллекций текстовых документов на базе разработанных подходов.

Оценка достоверности результатов исследования выявила: что идея диссертация базируется на использовании соискателем классического аппарата дискретной математики, на анализе международной литературы по компьютерной лингвистике и автоматической обработке текстов и обобщении передового опыта по разработки алгоритмического и программного обеспечения; что установлено качественное превосходство теоретико-множественного моделирования по сравнению с известными подходами.

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

Личный вклад автора заключается в:

1. разработке модели нормированного аннотированного суффиксного дерева для представления коллекций текстов;
2. создании меры релевантности строка-текст СУВСС, основанной на нормированном суффиксном дереве;
3. разработке метода рубрикации научных статей с использованием СУВСС;
4. разработке метода пополнения таксономии предметной области на основе ресурсов Интернет-энциклопедии Википедии с использованием СУВСС;
5. разработке метода фильтрации обценной лексики;
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.

Диссертационный совет пришел к выводу о том, что: диссертация представляет собой научно-квалификационную работу, которая соответствует критериям, установленным Положением о присуждении ученых степеней, и принял решение присудить Черняк Екатерине Леонидовне ученую степень кандидата технических наук.

При проведении тайного голосования диссертационный совет в количестве 36 человек, из них 14 докторов наук по специальности защищаемой диссертации, участвовавших в заседании, из 42 человек, входящих в состав совета, проголосовали: за присуждение ученой степени 34, против присуждения ученой степени 1, недействительных бюллетеней 1.

Председатель диссертационного совета Д 002.073.04

Чл.-корр. РАН



Ю.С. Попков

Ученый секретарь диссертационного совета Д 002.073.04

Д.т.н., профессор

В.Н. Крутько