

УТВЕРЖДАЮ

Проректор Национального
исследовательского университета
«Высшая школа экономики»

к.э.н., доцент Рощин Сергей Юрьевич
« 18 » июня 2016 г.

ЗАКЛЮЧЕНИЕ

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

Диссертация Черняк Екатерины Леонидовны на тему: «Разработка вычислительных методов анализа текстов с использованием аннотированных суффиксных деревьев» выполнена в Департаменте анализа данных и искусственного интеллекта факультета компьютерных наук федерального государственного автономного образовательного учреждения высшего профессионального образования «Национальный исследовательский университет «Высшая школа экономики».

В период подготовки диссертации соискатель ученой степени Черняк Екатерина Леонидовна работала в Международной лаборатории интеллектуальных систем и структурного анализа федерального государственного автономного образовательного учреждения высшего профессионального образования «Национальный исследовательский университет «Высшая школа экономики» в должности младшего научного сотрудника.

В 2012 г. окончила магистратуру федерального государственного автономного образовательного учреждения высшего профессионального образования «Национальный исследовательский университет «Высшая школа экономики» по направлению «Прикладная математика и информатика».

В 2015 г. окончила очную аспирантуру федерального государственного автономного образовательного учреждения высшего профессионального образования «Национальный исследовательский университет «Высшая школа экономики» по научной специальности 05.13.18 «Математическое моделирование, численные методы и комплексы программ».

Справка об обучении с результатами сдачи кандидатских экзаменов выдана в 2015 г. в федеральном государственном автономном образовательном учреждении высшего профессионального образования «Национальный исследовательский университет «Высшая школа экономики».

Научный руководитель – д.т.н., с.н.с. Борис Григорьевич Миркин – работает в федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский университет «Высшая школа экономики» в должности профессора Департамента анализа данных и искусственного интеллекта факультета компьютерных наук.

По итогам обсуждения принято следующее заключение:

Актуальность темы

Проникновение вычислительной техники во все сферы частной, государственной и общественной жизни привело к необходимости разработки методов автоматического семантического анализа текстовых документов. Особое внимание в международной литературе уделяется задачам поиска и извлечения информации, категоризации текстов, извлечения ключевых словосочетаний, извлечение фактов и др. В большинстве методов решения таких задач используются знания о морфологической структуре, синтаксисе и грамматике языка, на котором написаны тексты (Natural Language Processing). Однако существует явная потребность в разработке более универсальных методов, не использующих знания о языке и лингвистические инструменты.

До некоторой степени подобный универсализм может быть достигнут за счет использования моделей представления коллекций текстовых документов, не учитывающих ни морфологию, ни синтаксис языка, на котором написаны тексты, и не требующих морфологического и синтаксического анализа. В данном исследовании в качестве такой модели рассматривается аппарат аннотированных суффиксных деревьев, представляющий собой модификацию стандартного аппарата суффиксных деревьев, заимствованных из биоинформатики. Помимо того, что модель представления текста, основанная на аннотированных суффиксных деревьях является независимой от языка, она обладает и другим очевидным преимуществом: она естественным образом порождает такую меру релевантности строк тексту, которая учитывает нечеткие совпадения между строкой и текстом, за счет чего и достигается требуемый уровень точности.

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, проявляется в непосредственном участии соискателя в постановке экспериментов и подготовке публикаций и выступлений на научных семинарах и конференциях, то есть, в самостоятельной апробации работы.

Личный вклад автора:

1. Создание модели нормированного аннотированного суффиксного дерева для представления коллекций текстов;
2. Создание меры релевантности строка-текст СУВСС, основанной на нормированном суффиксном дереве;
3. Разработка метода рубрикации научных статей с использованием СУВСС;
4. Разработка метода пополнения таксономии предметной области на основе ресурсов Интернет-энциклопедии Википедии с использованием СУВСС;
5. Реализация разработанных моделей, методов и алгоритмов в виде программных комплексов.

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 соискателем получен ряд новых научных результатов:

1. Разработана модель нормированного аннотированного суффиксного дерева для представления текстов с естественно интерпретируемой мерой релевантности строка-текст СУВСС и таблицей релевантности строка-текст (РСТ);
2. Предложена методика рубрикации научных статей и проведены вычислительные эксперименты, показавшие эффективность меры СУВСС в сравнении с мерами релевантности, основанными на векторных пространствах и вероятностных моделях для представления текстовых документов, традиционно используемых в международных публикациях;
3. Разработана методика использования структуры и состава русскоязычной Википедии для пополнения предметных таксономий. Методика апробирована при решении задачи пополнения таксономии чистой и прикладной математики;
4. Разработано программное обеспечение, реализующее методику построения РСТ на основе нормированного аннотированного суффиксного дерева и меры релевантности СУВСС, а также обслуживающее способы решения задач, перечисленных выше.

Теоретическая значимость работы заключается в разработке принципиально новых конструкций: модели нормированного аннотированного суффиксного дерева с функцией релевантности СУВСС,

метода построения таблиц релевантности строка–текст, метода пополнения таксономии с использованием структуры и состава русскоязычной Википедии.

Практическая значимость результатов проведенных соискателем ученой степени исследований подтверждена экспериментами по сравнительной оценке использования мер релевантности для рубрикации научных статей и результатами расчетов по пополнению таксономии. Все разработанные методы реализованы в виде программных комплексов, предназначенных для решения исследовательских и прикладных задач. Разработанные методы и алгоритмы были успешно применены в реальных проектах компании ФОРС и проектах, выполнявшихся по грантам ВШЭ в 2010 – 2015 гг., а также в преподавательской деятельности Департамента анализа данных и искусственного интеллекта.

Апробация работы

Материалы диссертационного исследования были представлены на следующих научных мероприятиях:

1. 1-ой и 2-ой Всероссийских научных конференциях «Анализ изображений, сетей и текстов» (АИСТ-2012, АИСТ-2013), 16 – 18 мая 2012, 4 – 6 апреля 2013, Екатеринбург, Россия; темы докладов: «Автоматизация использования таксономий для аннотирования текстовых документов», «Использование ресурсов интернета для построения таксономии»
2. 1-ом Семинаре по кластерам, деревьям и порядкам (COT-2013), 12 – 13 декабря 2012, Москва, Россия; тема доклада: «An AST method for scoring string-to-text similarity in semantic text analysis»
3. 8-ой Международной конференции «Диалог» (Диалог-2013), 29 мая – 2 июня, Бекасово, Россия; тема доклада: «Computational refining of Russian-language taxonomy using Wikipedia»
4. 3-ей Международной научной конференции «Анализ изображений, сетей и текстов» (АИСТ-2014), 10 – 12 апреля, Екатеринбург, Россия; тема доклада: «Conceptual maps: construction over a text collection and analysis»
5. 2-ой Международной конференции «Информационные технологии и количественный менеджмент» (ITQM-2014), 3 – 5 июня 2014, Москва, Россия; тема доклада: «A method for refining a taxonomy by using annotated suffix trees and Wikipedia resources»
6. 3-ей Всероссийской конференции «Искусственный интеллект и естественный язык» (AINL-2014), 12 – 13 сентября 2014, Москва, Россия; тема доклада: «Создание и визуализация газетного интернет-корпуса»
7. 8-ой Международной конференции «Веб-поиск и майнинг данных» (WSDM-2015), 31 января – 6 февраля, Шанхай, КНР; тема доклада: «An approach to the problem of annotation of research publications»;
8. 2-ом Международном семинаре по майнингу данных и автоматической обработке текстов (DMNLP-2015), 7 сентября, Порту, Португалия; тема доклада: «Some thoughts on using annotated suffix trees for NLP tasks»

Полнота изложения материалов диссертации в публикациях

Основные результаты диссертационного исследования опубликованы в 10 работах общим объемом 4.59 п.л.; личный вклад автора составляет 2.9 п.л.

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

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. – 1.15 п.л. (личный вклад автора – 0.5 п.л.)
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. – 1 п.л. (личный вклад автора – 0.45 п.л.)
3. Черняк Е.Л. Системы автоматической обработки текстов / Черняк Е.Л., Ильвовский Д.А. // Открытые системы. СУБД. 2014. № 1. С. 51–43. – 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, Scopus Astrophysics, Data System, PubMed, MathSciNet, zbMATH, Chemical Abstracts, Springer, Agris, GeoRef:

4. Chernyak E.L. A Method for Refining a Taxonomy by Using Annotated Suffix Trees and Wikipedia Resources / Chernyak E.L., Mirkin B.G. // 2nd International Conference On Information Technology and Quantitative Management ITQM 2014. Procedia Computer Science. 2014. Vol. 31. P. 193 – 200 (Scopus, WoS) – 0.4 п.л. (личный вклад автора – 0.2 п.л.)
5. Chernyak E.L. An AST method for scoring string-to-text similarity in semantic text analysis / Chernyak E.L., Mirkin B.G. // Springer. 2014. Vol. 92. P. 92 – 96 (WoS) – 0.35 п.л. (личный вклад автора – 0.2 п.л.)
6. 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 (Scopus, WoS) – 0.4 п.л. (личный вклад автора – 0.4 п.л.)
7. Chernyak E.L. Conceptual maps: construction over a text collection and analysis / Morenko E.N., Chernyak E. L., Mirkin B. G. // Springer International Publishing. 2014. Vol. 439. P. 163-169 (Scopus, WoS) – 0.5 п.л. (личный вклад автора – 0.2 п.л.)

Работы, опубликованные автором в других изданиях:

8. Chernyak E.L. Conceptual maps: construction over a text collection and analysis / Chernyak E. L., Mirkin B. G. // Материалы ежегодной конференции Диалог. 2013. Т. 2. Стр. 177 – 185 – 0.5 п.л. (личный вклад автора – 0.25 п.л.)
9. Черняк Е.Л. Аннотированные суффиксные деревья: особенности реализации / Дубов М. С., Черняк Е. Л. // Использование ресурсов Интернета для построения таксономии // Доклады всероссийской научной конференции АИСТ'2013. 2013. Стр. 49 – 57 – 0.45 п.л. (личный вклад автора – 0.2 п.л.)
10. Черняк Е.Л. Использование ресурсов Интернета для построения таксономии / Черняк Е. Л., Миркин Б. Г. // Использование ресурсов Интернета для построения таксономии // Доклады всероссийской научной конференции АИСТ'2013. 2013. Стр. 49 – 57 – 0.35 п.л. (личный вклад автора – 0.2 п.л.)

Ценность научных работ соискателя ученой степени

Публикации полностью соответствуют теме диссертационного исследования и раскрывают её 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. В научных работах представлен метод аннотированного суффиксного дерева и его различные приложения к задачам автоматической обработки текстов. Метод АСД используется для решения задачи категоризации текстов, построения матрицы строка–текст и ее последующей категоризации, для достраивания таксономии. В работах проводится аналогия между мерой сходства, порождаемой АСД – СУВСС, и понятием релевантности и проводятся сравнения между популярными мерами релевантности и СУВСС, показывается эффективность СУВСС по точности, MAP и nDCG, объясняемая нечетким характером меры.

Диссертация Черняк Екатерины Леонидовны на тему: «Разработка вычислительных методов анализа текстов с использованием аннотированных суффиксных деревьев» – 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 п. 3 Разработка, обоснование и тестирование эффективных вычислительных

методов с применением современных компьютерных технологий по паспорту специальности.

Диссертация Черняк Екатерины Леонидовны на тему: «Разработка вычислительных методов анализа текстов с использованием аннотированных суффиксных деревьев» рекомендуется к защите на соискание ученой степени кандидата технических наук по специальности 05.13.18 «Математическое моделирование, численные методы и комплексы программ» (технические науки).

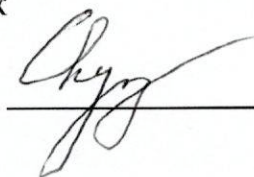
Заключение принято на заседании Департамента анализа данных и искусственного интеллекта факультета компьютерных наук федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики» 30 мая 2016 г., протокол № 6.

На заседании присутствовало 8 человек. Результаты голосования: «за» – 7 человек, «против» – 0 человек, «воздержалось» – 1 человека.

Руководитель

департамента анализа данных и искусственного интеллекта
факультета компьютерных наук
НИУ ВШЭ,

д.ф.-м.н., профессор



Кузнецов Сергей Олегович

Подпись заверяю

Специалист по кадрам

КОНЕВСКАЯ Л.В.

«03» июня 2016

