ЗАКЛЮЧЕНИЕ ДИССЕРТАЦИОННОГО СОВЕТА Д 002.199.01 НА БАЗЕ ФЕДЕРАЛЬНОГО ГОСУДАРСТВЕННОГО БЮДЖЕТНОГО УЧРЕЖДЕНИЯ НАУКИ САНКТ-ПЕТЕРБУРГСКОГО ИНСТИТУТА ИНФОРМАТИКИ И АВТОМАТИЗАЦИИ РОССИЙСКОЙ АКАДЕМИИ НАУК ПО ДИССЕРТАЦИИ НА СОИСКАНИЕ УЧЕНОЙ СТЕПЕНИ КАНДИДАТА НАУК

аттестационное дело № _____

решение диссертационного совета 16.11.2017 г. № 1

О присуждении Карповичу Сергею Николаевичу, гражданину Российской Федерации, ученой степени кандидата технических наук.

Диссертация «Математическое и программное обеспечение вероятностного тематического моделирования потока текстовых документов» по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей» принята к защите 14 сентября 2016 г., протокол № 1, диссертационным советом Д 002.199.01 на базе Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации Российской академии наук, 199178, Россия, Санкт-Петербург, 14 линия ВО, дом 39, утвержден приказом Рособрнадзора номер 2472-618 от 8 октября 2010 года.

Соискатель Карпович Сергей Николаевич, 1977 года рождения, в 2006 г. окончил Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина) по специальности «Автоматизированные системы обработки информации и управления» (диплом ВСГ №0242587), в настоящее время является аспирантом федерального автономного образовательного учреждения высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий механики и оптики» (Университет ИТМО) Министерства образования и науки Российской Федерации, по направлению/специальности «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей» (справка об обучении № 32/2017 от 22.06.2017). В настоящее время соискатель Карпович Сергей

Николаевич работает в акционерном обществе «Олимп» (учредитель – Правительство Москвы) руководителем группы развития поиска.

Диссертация «Математическое и программное обеспечение вероятностного тематического моделирования потока текстовых документов» выполнена в Федеральном государственном автономном образовательном учреждении высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики».

Научный руководитель – доктор технических наук, профессор СМИРНОВ Александр Викторович, основное место работы: Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН), заведующий лабораторией интегрированных систем автоматизации.

Официальные оппоненты:

ХОМОНЕНКО Анатолий Дмитриевич, доктор технических наук, профессор, Федеральное государственное бюджетное образовательное учреждение высшего образования «Петербургский государственный университет путей сообщения Императора Александра I», заведующий кафедрой «Информационные и вычислительные системы»;

ВОДЯХО Александр Иванович, доктор технических наук, профессор, Федеральное государственное автономное образовательное учреждение высшего образования Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), профессор кафедры вычислительной техники дали положительные отзывы на диссертацию.

Ведущая организация – Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого» в своем положительном отзыве, подписанном Поповым Сергеем Геннадьевичем, кандидатом технических наук, доцентом, доцентом кафедры «Телематика (при ЦНИИ РТК)», Курочкиным Михаилом Александровичем, кандидатом технических наук, доцентом, доцентом кафедры «Телематика (при ЦНИИ РТК)» и утвержденном Сергеевым Виталием Владимировичем, проректором по научной работе Федерального государственного

автономного образовательного учреждения высшего образования «Санкт-Петербургский политехнический университет Петра Великого», доктором технических наук, профессором, член-корреспондентом Российской академии наук, указала, что в целом диссертационная работа С.Н. Карповича представляет собой завершенную научно-исследовательскую работу, выполненную на актуальную тему, отличается научной новизной и практической значимостью полученных результатов. Автором в диссертации сформулирована и решена важная научно-техническая задача разработки математического и программного обеспечения вероятностного тематического моделирования коллекции и потока текстовых документов и создан корпус текстов пригодный для исследования алгоритмов обработки и анализа текстовых документов на естественном языке.

Соискателем создан русскоязычный корпус текстов SCTM-ru, позволяющий исследовать алгоритмы обработки текстов на естественном языке, в том числе алгоритмы вероятностного тематического моделирования. Предложен метод расчета вероятностной тематической модели на основе связей документов и тем установленных авторами документов, а также метод определения темы «нового слова» использующий произведение Адамара векторов тем документов, где это слово встретилось. Предложен алгоритм многозначной классификации текстовых документов ml-PLSI, основанный на математическом аппарате вероятностного тематического моделирования. Текст автореферата полностью соответствует содержанию диссертации. Диссертационное исследование «Математическое и программное обеспечение вероятностного тематического моделирования потока текстовых документов» является научно-квалификационной работой и соответствует критериям, изложенным в п. 9 “Положения о присуждении ученых степеней”, утвержденного постановлением Правительства Российской Федерации № 842 от 24 сентября 2013 г., предъявляемых к кандидатским диссертациям, а его автор Карпович Сергей Николаевич заслуживает присуждения ученой степени кандидата технических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Соискатель имеет 12 опубликованных работ, в том числе по теме диссертации 12 работ, опубликованных в рецензируемых научных изданиях – 4 работы, из них опубликованных в изданиях, рекомендуемых ВАК РФ – 3.

Основные научные результаты опубликованы в 12 научных трудах общим объемом 4,66 п.л., из которых 1 статья объемом 0,44 п.л., выполнены в соавторстве, а 11 статей объемом 4,22 п.л. – лично. Наиболее значимые работы по теме диссертации:

1. **Карпович С.Н.** Русскоязычный корпус текстов SCTM-ru для построения тематических моделей // Труды СПИИРАН. – СПб., 2015. – № 39. С. 123-142. УДК 004.912 (ВАК). *Личный вклад соискателя – 100%*
2. **Карпович С.Н.** Многозначная классификация текстовых документов с использованием вероятностного тематического моделирования ml-PLSI // Труды СПИИРАН. – СПб., 2016. – Т. 4. – №. 47. – С. 92-104 (ВАК, Scopus). *Личный вклад соискателя – 100%*
3. **Карпович С.Н.** Тематическая модель с бесконечным словарем // Информационно-управляющие системы. 2016. №6 С. 43-49. doi:10.15217/issn1684-8853.2016.6.43 (ВАК). *Личный вклад соискателя – 100%*
4. Smirnov A. Topic model visualization with iPython // A. Smirnov, N. Teslya, **S. Karpovich**, A. Grigorev // Open Innovations Assoc. FRUCT, Proc. of 20th Conf. – 2017. – Pp. 131-137 (Web of Science, Scopus). *Личный вклад соискателя – 25%*

Оригинальность содержания диссертации составляет не менее 95% от общего объёма текста; цитирование оформлено корректно; заимствованного материала, использованного в диссертации без ссылки на автора либо источник заимствования, не обнаружено; научных работ, выполненных соискателем учёной степени в соавторстве без ссылок на соавторов не выявлено. Недостоверные сведения об опубликованных соискателем ученой степени работах в диссертации отсутствуют.

На автореферат диссертации поступило 6 отзывов, все отзывы положительны:

1) ФГБОУ ВО «Петрозаводский государственный университет». Отзыв составил доцент ФГБОУ ВО «Петрозаводский государственный университет», к.ф.-м.н., доцент Корзун Дмитрий Жоржевич. Замечания: в тексте автореферата не представлены система параметров, характеризующая размеры и другие свойства получаемого корпуса текстов, имеющие значимое влияние на сложность предлагаемых алгоритмов для автоматической обработки текстов; не приведены результаты по теоретическому исследованию (например, в виде теорем) таких типовых свойств предлагаемых алгоритмов, как корректность, точность, временная и емкостная сложность.

2) ФГБУН Институт информатики и математического моделирования технологических процессов Кольского научного центра РАН. Отзыв составил врио директора ФГБУН Институт информатики и математического моделирования технологических процессов Кольского научного центра РАН, д.т.н. Олейник Андрей Григорьевич. Замечания: в тексте автореферата отсутствуют пояснения к указанным на рисунках 2 и 6 числовым обозначениям, определяющим, вероятно, порядок выполнения операций или информационного взаимодействия модулей разработанного программного комплекса; характеризуя научную новизну полученных результатов диссертационного исследования (п.5, стр.5), автор отмечает обеспечение в рамках созданного прототипа комплекса программных средств возможности варьирования способов решения конкретных практических задач анализа потока текстовых документов; на стр. 13 автореферата (абз.3) также указано, что любой микросервис, входящий в состав комплекса, может быть изменен без нарушения целостности системы, однако в тексте автореферата не представлено, каким образом выбирается «подходящий» способ решения задачи, кто и как может изменять микросервисы; в тексте автореферата неоднократно (стр. 5, 13) упоминается эффективность алгоритмов, но не приводятся ее количественные характеристики.

3) Национальный исследовательский университет Высшая школа экономики (Нижний Новгород). Отзыв составил профессор кафедры информационных систем и технологий, д.т.н., доцент Бабкин Эдуард Александрович. Замечания: на рисунке 1 отображена концептуальная схема вероятностного тематического моделирования. В каждой из трех матриц вероятностной тематической модели в качестве значений ячеек указаны одни и те же выражения: Value 1, Value 2 … В этом случае остается непонятно – это одни и те же значения в каждой матрице или значения этих матриц различаются (то же относится к рис.4); на схеме алгоритма определения тем «нового слова» (рис.5) допущена грамматическая ошибка: необходимо заменить слово «косинуссного» на «косинусного».

4) ФГБУН Институт проблем управления им В.А. Трапезникова РАН. Отзыв составил ведущий научный сотрудник лаборатории №68, доктор технических наук, профессор Ульянов Михаил Васильевич. Замечания: автором разработан новый русскоязычный корпус текстовых документов SCTM-ru, но в тексте автореферата нет

данных о сравнении этого корпуса с существующими корпусами, что не дает возможность в полной мере оценить значимость этой разработки; применение произведения Адамара очевидно сокращает трудоемкость вычислений, что важно при потоковой обработке, однако автор не указывает на возможные ошибки тематической классификации, порождаемые применением этого аппарата; на стр. 3 автореферата при описании текстового потока автор пишет «… время происшествия», автор отзыва считает, что более корректно – время регистрации документа.

5) ФГБУН Институт проблем управления сложными системами Российской академии наук (ИПУСС РАН). Отзыв составил заместитель директора по научной работе, доктор технических наук Смирнов Сергей Викторович. Замечания: автор выбирает для исследования вероятностное моделирование тем, не позиционируя эту модель среди других конкурентных подходов к малоразмерному представлению текстов – неотрицательному матричному разложению (NMF), моделям word2vec, t-sne и т.п.; в тексте автореферата приведены количественные оценки качества разработанного метода, однако важна скорость работы соответствующих алгоритмов; на рисунках 2 и 6 связи блоков имеют цифровые обозначения, которые не комментируются в тексте автореферата.

6) ФГБУН Институт систем информатики им А.П. Ершова (ИСИ им. А.П. Ершова). Отзыв подготовил заведующий лабораторией искусственного интеллекта ИСИ им А.П. Ершова, кандидат технических наук О.А. Загорулько. Замечания: недостаточно полно отражены существующие направления исследований в области вероятностного тематического моделирования; очень кратко описаны сценарии применения вероятностного тематического моделирования в практических задачах информационного поиска и рекомендательных системах; на стр. 15 ошибочно указано, что D – документы корпуса для обучения. На мой взгляд, должно быть сказано, что D – это документы корпуса для тестирования алгоритма.

Выбор официальных оппонентов и ведущей организации обосновывается тем, что д.т.н., профессор Хомоненко А.Д. является известным ученым в области численных методов теории массового обслуживания, программирования, компьютерной безопасности, баз данных и информационных систем, систем искусственного интеллекта.; д.т.н., профессор, Водяхо А.И. – известный ученый в области

архитектуры и проектирования информационных систем; ведущая организация, Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого», занимающий ведущие позиции в рейтинге технических вузов России, играющий заметную роль в научно-образовательном сообществе страны и мира, выполняющий подготовку кандидатов и докторов наук по 92 научным специальностям включающим информатику и вычислительную технику, в котором получены значимые теоретические результаты в области обработки текстов на естественном языке.

Диссертационный совет отмечает, что на основании выполненных соискателем исследований:

разработаны специальный русскоязычный корпус текстовых документов для исследования алгоритмов вероятностного тематического моделирования, комплекс программных средств для построения и визуализации вероятностных тематических моделей;

предложены:

метод расчета матриц ВТМ на основе обучения с учителем (авторами документов), учитывающий заданные связи между документами и темами, отличающийся вычислительной эффективностью за счет отсутствия итераций, позволяющий создавать ВТМ на различных наборах, размеченных данных. Обучение ВТМ может найти свое применение в анализе новостного потока, где темы — это теги, проставленные авторами документов, а также в анализе записей в блогах;

алгоритм многозначной классификации текстовых документов ml-PLSI, основанный на математическом аппарате вероятностного тематического моделирования, заключающийся в использовании известной матрицы «слово-тема» ВТМ для классификации документов, что позволило определять темы «новых документов» при анализе потока текстовых документов в динамической тематической модели. По сравнению с логистической регрессией и деревьями решений, предложенный алгоритм лучше справляется с ситуацией, когда у одного документа большое количество меток, обучение ВТМ происходит быстрее чем обучение альтернативных алгоритмов. Предложенный алгоритм многозначной классификации

точнее определяет классы, даже в том случае если в классифицируемом документе встречаются незнакомые вероятностной тематической модели слова, логистическая регрессия при встрече неизвестного слова, случайным образом присваивает ему тему, что зачастую приводит к ошибкам;

метод определения тем «нового слова», основанный на использовании произведения Адамара тематических векторов документов, содержащих это слово, позволяющий определять вектора тем «новых слов» в потоке текстовых документов при построении динамической тематической модели с эффективностью, превосходящей альтернативные методы определения тем «новых слов» основанные на распределении и процессе Дирихле.

микросервисная архитектура комплекса программных средств, обеспечивающего реализацию предложенных в диссертационной работе методов и алгоритма.

доказана перспективность использования алгоритма многозначной классификации при построении динамических тематических моделей, а также перспективность использования произведения Адамара для определения тем «нового слова» через вектора тем документов, где оно встретилось.

введены:

- Требования математическому и программному обеспечению, необходимому для построения вероятностных тематических моделей.

Теоретическая значимость исследования обоснована тем, что:

доказаны сформулированные в работе теоретические утверждения результатами проведенных экспериментов с использованием разработанного комплекса программных средств, реализующего предложенные методы и алгоритм. Эти утверждения относятся к расчету матриц BTM на основе обучения с учителем, определению тем «нового слова» через произведение Адамара и использованию вероятностного тематического моделирования для классификации текстовых документов.

применительно к проблематике диссертации результативно (эффективно, то есть с получением обладающих новизной результатов)

использованы методы системного анализа, математического и компьютерного моделирования, автоматической обработки естественного языка, теории вероятности, математической статистики, прогнозирования временных рядов, теории машинного обучения и теории алгоритмов, разработки информационных систем и программирования;

изложены методологические и методические основы построения вероятностной тематической модели потока текстовых документов;

раскрыты

проблемные аспекты применения имеющихся подходов в области вероятностного тематического моделирования;

основные вопросы, связанные с анализом потока текстовых документов, проблемы определения тем «новых» документов и слов, проблемы пополняемости словаря вероятностной тематической модели;

изучены существующие методы построения вероятностных тематических моделей, большая часть которых предназначена для работы со статическими коллекциями документов, уделено внимание визуализации результатов вероятностного тематического моделирования;

проведена модернизация существующих методов пополнения словаря вероятностной тематической модели, при построении динамической тематической модели, предложенный метод использует произведение Адамара векторов тем документов, в которых встретилось новое слово.

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

разработаны и внедрены следующие результаты диссертационной работы:

- Многозначная классификация ml-PLSI;
- Вероятностное тематическое моделирование;
- Создание темпоральной вероятностной тематической модели;
- Представление результатов тематического моделирования;

внедрены в системе анализа новостного потока принятой к использованию в АО «Олимп», а также в сервисе классификации поисковых запросов принятом к использованию в ООО «Rambler&Co». На основе приведенных результатов были разработаны: сервис анализа потока новостей, позволяющий обрабатывать большие объемы данных, автоматически группировать схожие новости, отслеживать динамику изменения популярности тем во времени, визуализировать результаты тематического моделирования для последующего представления этих результатов в отчетах; также был разработан сервис классификации поисковых запросов, позволяющий определять наиболее вероятные тематические группы очень коротких текстов, таких как запрос в поисковую систему.

- русскоязычный корпус текстов SCTM-ru, позволяющий исследовать алгоритмы вероятностного тематического моделирования и отличающийся от других корпусов наличием оригинального текста документа и метатекстовой разметки: автор, время описанных событий, тема;
- метод расчета матриц вероятностей тематической модели на основе обучения с учителем (авторами документов) с учетом заданных связей документов и тем;
- алгоритм многозначной классификации текстовых документов ml-PLSI, основанный на вероятностном тематическом моделировании и заключающийся в использовании матрицы «слово-тема» вероятностной тематической модели для классификации документов, позволяющий определять темы «новых-документов» при анализе потока текстовых документов в динамической тематической модели;

внедрены в учебном процессе при изучении дисциплины «Управление знаниями» магистерской программы по специальности 38.04.05 – информационные системы бизнеса, кафедры информационных систем «Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики». Использование указанных результатов в учебном процессе позволило предоставить студентам актуальные знания о методах построения вероятностных тематических моделей. В материалах курса демонстрируются современные методы расчета матриц

«слово-тема» и «документ-тема» вероятностной тематической модели, включая разработанный в диссертации метод на основе обучения с учителем. Подробно описывается алгоритм для многозначной классификации текстовых документов, позволяющий классифицировать по темам ранее не встречающиеся документы из непрерывного потока документов.

Исследования, отраженные в диссертации, проведены в рамках НИР № 714630 «Разработка теоретических и технологических основ социо-киберфизических систем», проводимой в Университете ИТМО (государственная программа поддержки ведущих университетов РФ, субсидия 074-U01).

определены возможности и перспективы практического использования полученных результатов диссертации в задачах анализа текстов на естественном языке, информационном поиске и в сервисах рекомендаций;

создан комплекс программных средств для вероятностного тематического моделирования коллекций и потоков текстовых документов, обладающий возможностями настройки работы отдельных модулей для решения конкретных практических задач, позволяющий визуализировать промежуточные и конечные результаты тематического моделирования;

представлены предложения и направления для дальнейших научных исследований, в основу которых могут быть положены разработанные методы и алгоритм.

Оценка достоверности результатов исследования выявила:

достоверность полученных результатов подтверждена проведением всестороннего анализа работ по исследуемой проблеме, корректным применением научно-методического аппарата в виде использованных методов и теорий, апробацией основных результатов диссертации в печатных трудах и докладах на международных и всероссийских конференциях, положительными итогами практической реализации результатов работы;

теория построена на известных принципах, проверенных данных и фактах с использованием современных известных и апробированных методов исследования, согласуется с опубликованными частными результатами других исследователей;

идея базируется на анализе работ отечественных и зарубежных исследователей в области вероятностного тематического моделирования, основанного на автоматической обработке текстов на естественном языке;

использованы методы математического и компьютерного моделирования;

установлено качественное и количественное соответствие результатов решения задачи многозначной классификации текстовых документов, при этом подтверждено преимущество предложенного алгоритма перед существующими алгоритмами Логистической регрессии и Деревья решений.

Личный вклад соискателя состоит в:

- исследовании существующих методов и алгоритмов вероятностного тематического моделирования;
- создании русскоязычного корпуса текстов пригодного для исследования алгоритмов вероятностного тематического моделирования;
- исследовании и разработке методов построения вероятностных тематических моделей через обучение на размеченных данных;
- разработке алгоритма многозначной классификации текстовых документов на базе вероятностного тематического моделирования;
- исследовании и разработке метода определения темы «нового слова» при анализе потока текстовых документов для построения динамических вероятностных тематических моделей с пополняемым словарем;
- разработке комплекса программных средств для вероятностного тематического моделирования коллекций и потоков текстовых документов с визуализацией результата моделирования;
- подготовке основных публикаций по выполненной работе.

Диссертационный совет считает, что Карпович С.Н. в своей диссертационной работе решил научную задачу разработки математического и программного обеспечения вероятностного тематического моделирования потока текстовых документов, имеющую важное значение для построения вероятностных тематических моделей с пополняемым словарем, а также для применения подобных моделей в задачах классификации.

На заседании 16.11.2017 г. диссертационный совет принял решение присудить Карповичу С.Н. ученую степень кандидата технических наук.

При проведении тайного голосования диссертационный совет в количестве 22 человек, из них 8 докторов наук по специальности рассматриваемой диссертации, участвовавших в заседании, из 26 человек, входящих в состав совета, проголосовали: за 22, против нет, недействительных бюллетеней нет.

Председатель диссерта
доктор технических на
член-корреспондент Р. супов Рафаэль Мидхатович

Ученый секретарь дис
кандидат технических ева Александра Алексеевна
16.11.2017 г.