ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

доктора технических наук, профессора Водяхо Александра Ивановича

На диссертационную работу *Карповича Сергея Николаевича* по теме *«Математическое и программное обеспечение вероятностного тематического моделирования потока текстовых документов»*, представленную на соискание ученой степени кандидата технических наук по специальности *05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».*

Актуальность темы исследования. Согласно недавним оценкам, в мире ежедневно генерируется более 2,5 квинтиллионов байт данных. Более 90% всех мировых данных создано за последнее десятилетие, при этом около 80% из них не структурированы или слабо структурированы. Обработка такого объема данных лежит за пределами возможностей стандартных аналитических методов, а сами данные слишком обширны, чтобы человек был в силах их воспринять. Активно ведутся исследования алгоритмов машинного обучения, позволяющих обрабатывать данные большого объема (Big Data). Одним из перспективных направлений автоматической обработки текстовых данных большого объема является вероятностное тематическое моделирование. Поэтому исследования методов построения вероятностных тематических моделей на коллекциях и потоках текстовых документов, являются актуальными. Диссертационная работа Карповича Сергея Николаевича направлена на создания общедоступного инструмента для построения вероятностных тематических моделей коллекций и потоков текстовых документов.

Диссертационная работа Карповича С.Н., объемом 153 машинописных страниц, содержит введение, четыре главы и заключение, список литературы (117 наименований), 14 таблиц, 51 рисунок, одно приложение с копиями актов внедрения. Введение соответствует всем формальным требованиям и содержит краткое описание сути проблемы, цель и основные направления исследования.

Основные результаты. В ходе диссертационного исследования были получены следующие основные результаты:

1. Разработан русскоязычный корпус текстов, позволяющий исследовать возможности алгоритмов вероятностного тематического моделирования. Корпус содержит помимо текста документа, мета параметры: дату описанных событий, автора, категории. Распространяется по свободной лицензии. Корпус позволяет создавать классические

тематические модели, автор тематические модели, темпоральные тематические модели, использовать авторскую разметку документов по категориям для создания классификатора документов. Корпус может быть использован для исследования различных алгоритмов обработки естественного языка, классификации и кластеризации текстовых документов.

2. Предложен алгоритм многозначной классификации ml-PLSI использующий математический аппарат вероятностного тематического моделирования для классификации текстовых документов. Алгоритм многозначной классификации расширяет круг задач, в которых успешно применяются вероятностные тематические модели. Позволяет определять темы «новых документов» в динамических тематических моделях.
3. Предложен метод определения тем «новых слов» через произведение Адамара векторов тем документов, где встретилось «новое слово». Предложенный метод отличается высокой вычислительной эффективностью в сравнении с альтернативными подходами. Применяется для определения тем «новых слова» в динамических тематических моделях.
4. Разработан комплекс программных средств на основе микросервисной архитектуры, предоставляющий набор алгоритмов и методов для анализа коллекций и поток текстовых документов. Каждый микросервис предназначен для решения конкретной задачи и может быть использован отдельно. Комплекс программных средств распространяется по открытой лицензии. Предоставляет возможность настраивать параметры и изменять программный код каждого микросервиса, включает настраиваемый интерфейс визуализации результатов тематического моделирования.

Степень обоснованности научных положений, выводов и рекомендаций. Тема диссертационной работы и полученные в ходе исследования результаты соответствуют пунктам 3 и 4 паспорта специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей». Основные научные положения, приведенные в автореферате и вынесенные на защиту, а также выводы и результаты диссертационного исследования обоснованы и аргументированы. Обоснованность и достоверность научных положений, основных выводов и результатов диссертации обеспечивается анализом современного состояния исследований в проблемной области, корректностью предложенных методов и технологий. Работа в достаточной степени

структурирована, в конце каждой главы присутствуют выводы, которые отражают суть проводимых исследований и разработок.

Научная новизна и достоверность результатов. Научная новизна полученных результатов состоит в усовершенствовании вероятностного тематического моделирования, в расширении функциональных возможностей вероятностных тематических моделей за счет использования ранее не рассматриваемых свойств тематических моделей.

Достоверность результатов. Достоверность результатов подтверждается проведенными экспериментами, опубликованными автором работами по результатам исследований, апробацией результатов работы на научно-технических конференциях и семинарах, а также актами внедрения.

Теоретическая и практическая значимость результатов. Теоретическая значимость результатов работы заключается в формализации алгоритма многозначной классификации с использованием математического аппарата вероятностного тематического моделирования и формализации метода определения тем «нового слова» при построении динамической вероятностной тематической модели.

Практическая значимость заключается в разработанном комплексе программных средств вероятностного тематического моделирования коллекций и потоков текстовых документов. Предложенная микросервисная архитектура комплекса программных средств позволяет использовать отдельные микросервисы в других программных системах для решения практических задач. Разработанный русскоязычный корпус текстов SCTM-ru пригоден для исследования алгоритмов обработки текстовых документов на естественном языке.

Внедрение результатов. Результаты исследований внедрены в ООО «Rambler&Co» в качестве сервиса многозначной классификации поисковых запросов, задаваемых в поисковую систему пользователями, в ООО «Олимп» для анализа новостного потока сайта Правительства Москвы. В Университете ИТМО результаты исследований используются в учебном процессе по курсу «Управление знаниями».

Полнота публикаций научных результатов. По материалам диссертационного исследования автором лично и в соавторстве опубликовано 12 печатных работ, включая 3

работы в журналах из списка ВАК («Труды СПИИРАН», «Информационно-управляющие системы») и 1 работа в международном издании, индексирующимся в реферативных базах Web of Science и Scopus.

Замечания. Несмотря на высокий в целом уровень диссертационной работы, в ней присутствует ряд недостатков.

1. Автору следовало более четко определить сферу применения предложенного комплекса программных средств.

2. В диссертационной работе автором выносятся на защиту пять положений. При этом одно из положений «метод расчета матриц вероятностной тематической модели на основе обучения с учителем…» в автореферате представлен недостаточно подробно, в части описания существующих альтернативных подходов к обучению вероятностных тематических моделей.

3. Следовало более четко определить связь между критериями оценки качества вероятностной тематической модели и бизнес целями практических задач, решаемых с помощью тематического моделирования.

4. Предложенная микросервисная архитектура недостаточно описана в автореферате, что затрудняет возможность практического использования полученных результатов.

Заключение. В диссертационной работе Карповича С.Н. изложены решения задачи по разработке комплекса математических и программных средств интеллектуального анализа потока текстовых документов, являющиеся актуальными для проблемной области вероятностного тематического моделирования. Научные результаты, полученные в ходе выполнения работы, являются новыми и практически значимыми, что доказано предоставленными актами внедрения. Отмеченные выше недостатки не влияют на общую положительную оценку результатов работы.

Рукопись диссертационного исследования и автореферат представлены в соответствии с требованиями. Личный вклад автора по решению поставленных задач не вызывает сомнения. Автореферат адекватно отражает содержание диссертационного исследования.

С учетом содержания диссертации и автореферата считаю, что представленная диссертация является законченной научно-квалификационной работой, отвечающей требованиям, установленным в п. 9 Положения о присуждении ученых степеней, утвержденного постановлением Правительства Российской Федерации от 24 сентября 2013 № 842, предъявляемым к кандидатским диссертациям, а ее автор заслуживает присуждения

ученой степени кандидата технических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Официальный оппонент:

Профессор кафедры вычислительной техники

ФГБОУ ВО «Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)»

Доктор технических наук,

Профессор Водяхо Александр Иванович

«16» октября 2017 г.

Сведения о составителе отзыва:

ФИО: Водяхо Александр Иванович

Ученая степень: доктор технических наук

Ученое звание: профессор

Место работы: Федеральное государственное бюджетное образовательное учреждение высшего образования «Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)»

Должность: профессор кафедры вычислительной техники

Почтовый адрес: 197376, Россия, Санкт-Петербург, ул. Профессора Попова, дом 5.

Телефон: (812) 234-25-03

Адрес электронной почты: aivodyaho@mail.ru

Подпись А.И. Водяхо заверя

Начальник отдела диссертац Русяева Т.Л.

16.10 2017 г.