ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

доктора технических наук, профессора

Хомоненко Анатолия Дмитриевича

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 диссертации

С развитием информационных технологий остро встает вопрос автоматической обработки создаваемых и накапливаемых цифровых данных, в том числе текстовых документов. Растет потребность в инструментах анализа коллекций и потоков текстовых документов для систем принятия решения, информационно-поисковых систем, сервисов рекомендаций. Учитывая скорость появления новостей, сообщений электронной почты, публикаций в социальных сетях, можно утверждать, что разработка математического и программного обеспечения для автоматизации обработки текстовых документов является важной практической и научной задачей.

Вероятностное тематическое моделирование и нейронные сети относятся к передовым интеллектуальным средствам обработки естественного языка. В современной научной литературе решение задачи разработки эффективных методов многозначной классификации текстовых документов и их анализа на основе вероятностного тематического моделирования не получила исчерпывающего решения. Все это подтверждает ***актуальность*** темы диссертационной работы Карповича С.Н., посвященной разработке математического и программного обеспечения для автоматизации обработки текстовых документов на основе

методов многозначной классификации и вероятностного тематического моделирования.

Степень обоснованности научных положений, выводов и рекомендаций

В диссертационной работе выполнен обзор современных подходов к построению вероятностных тематических моделей, выявлены проблемы в предметной области, определены направления для развития вероятностного тематического моделирования. Предложенный в работе алгоритм многозначной классификации текстовых документов решает важную практическую задачу анализа текстовых документов, а также расширяет возможности применения вероятностного тематического моделирования. Вероятностные тематические модели (ВТМ) используются, в первую очередь, для решения задач кластеризации и классификации, что составляет важное функциональное расширение направления вероятностного тематического моделирования.

Диссертационная работа Карповича С.Н., состоящая из введения, четырех глав, заключения, библиографического списка, изложена на 153 страницах машинописного текста, включает пятьдесят один рисунок, четырнадцать таблиц. Введение соответствует всем необходимым формальным требованиям и содержит краткое описание сути решаемой научной задачи, цель и основные направления исследования.

Первая глава содержит подробный обзор существующих исследований в области вероятностного тематического моделирования, на основе которого выявлены существующие проблема ВТМ. Рассмотрены основные виды ВТМ и методы оценки качества построенных ВТМ. В главе уделено внимание применению вероятностного тематического моделирования в практических задачах, таких как информационной поиск, рекомендательные системы, системы принятия решения. Сформулирована содержательная постановка задачи исследования.

Вторая глава посвящена определению требований к разрабатываемому комплексу программных средств вероятностного тематического моделирования

потока текстовых документов. Определены и сформулированы требования к корпусу текстов, необходимому для проведения исследований ВТМ, предложен технологический процесс создания корпуса. Рассмотрены сценарии использования, на основе которых построена UML-диаграмма задач использования комплекса программных средств. Построена концептуальная схема программного комплекса, учитывающая необходимость разработки человеко- и событийно-ориентированных подходов к его разработке.

Третья глава описывает предложенный алгоритм многозначной классификации и метод определения тем «новых слов» в ВТМ для пополнения словаря в динамической вероятностной тематической модели. Содержит обзоры существующих подходов к обучению ВТМ и построению динамических вероятностных тематических моделей, демонстрирует преимущества предложенного алгоритма и метода.

Четвертая глава посвящена практической реализации математического и программного обеспечения комплекса программных средств. Включает создание корпуса текстовых документов, источником которого стал сайт Русские Викиновости, разработку предложенных алгоритма и метода, а также визуализацию промежуточных и конечных результатов вероятностного тематического моделирования. При построении комплекса программных средств использовалась микросервисная архитектура, позволяющая использовать каждый микросервис автономно от программного комплекса, что удобно для использования только нужного микросервиса при решении практических задач. Автор продемонстрировал уверенное владение современными технологиями и инструментальными средствами разработки программного обеспечения.

Все сформулированные в диссертационной работе научные положения, выводы и рекомендации являются вполне обоснованными.

Научная новизна, достоверность и практическая значимость

Научная новизна полученных результатов заключается в разработке методов для построения вероятностной тематической модели в потоке текстовых документов. В частности, разработан метод построения ВТМ на основе обучения с учителем, в котором учитываются связи между документами и темами при расчете матрицы ВТМ. Кроме того, разработан алгоритм многозначной классификации текстовых документов, основанный на математическом аппарате ВТМ. Этот алгоритм играет важную роль при анализе потока текстовых документов в динамических вероятностных тематических моделях. Для исследований алгоритмов обработки текстов, в том числе и указанных алгоритмов, разработан русскоязычный корпус текстов SCTM-ru, содержащий метатекстовую разметку, такую как: автор, время описанных событий, тема. В работе предложен эффективный метод определения тем для «новых слов» в потоке текстовых документов, основанный на использовании произведения Адамара тематических векторов документов, содержащих это слово, и разработан алгоритм пополнения словаря динамической вероятностной тематической модели.

Практическая значимость состоит в разработанной программной инфраструктуре и программном обеспечении для построения вероятностных тематических моделей коллекций и потоков текстовых документов. Предложенная микросервисная архитектура позволяет использовать комплекс программных средств целиком либо его отдельные автономные микросервисы для решения практических задач обработки и анализа текстовых документов.

Достоверность полученных результатов подтверждается результатами экспериментальных исследований, подтверждающими вычислительную эффективность и повышенную точность построения ВТМ, опубликованными автором материалами исследований, а также внедрением полученных результатов.

Замечания по диссертационной работе

При высоком в целом уровне выполнения диссертационной работы в ней имеются следующие недостатки.

1. Говоря об оценке качества языковых моделей, автор в первой главе называет Перплексию и как критерий, и как меру несоответствия модели $P(w|d)$ словам w, встречаемым в текстовых документах. На мой взгляд, здесь достаточно ограничиться мерой, а в качестве критерия использовать одно из условий, определяющих критерий пригодности, превосходства или оптимальности.
2. Во второй главе рассмотрены прецеденты использования разрабатываемого комплекса программных средств, при этом приведена UML-диаграмма задач использования программного комплекса, и не представлена диаграмма сценариев использования.
3. В работе упоминается наличие категорий определенных авторами документов, попавших в корпус текстов, при этом в автореферате эти же категории называются темами.
4. Упомянутые в четвертой главе способы практического применения вероятностного тематического моделирования в задачах информационного поиска и рекомендательных сервисах описаны недостаточно подробно.
5. По оформлению работы можно отметить: отсутствие списка сокращений, наличие отдельных неточностей, например, «гиперпараментры»; автор указывает как листинги спецификацию кода, тогда как сами листинги не приводятся; рис. 7 в диссертации плохо виден, в автореферате этот недостаток устранен.

Заключение по диссертации в целом

Диссертация Карповича С.Н. представляет собой завершенную научно-квалификационную работу, в которой решена актуальная научная задача разработки методов и алгоритма анализа потока текстовых документов, имеющая

значение для развития теории и практики обработки текстов на естественном языке. Выводы и рекомендации в работе научно обоснованы и достоверны. Считаю, что представленная диссертация является законченной научной работой, полностью удовлетворяет требованиям, предъявляемым к диссертациям на соискание ученой степени кандидата технических наук, и соответствует пункту 9 действующего положения о присуждении ученых степеней, утвержденного постановлением Правительства РФ от 24.09.2013 № 842, а ее автор, Карпович Сергей Николаевич, заслуживает присуждения ему ученой степени кандидата технических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Официальный оппонент:
Заведующий кафедрой «Информационные
и вычислительные системы»
ФГБОУ ВО «Петербургский
государственный университет путей
сообщения Императора Александра I»

Доктор технических наук,
профессор Хомоненко Анатолий Дмитриевич

«23» октября 2

Сведения о составителе отзыва:
ФИО: Хомоненко Анатолий Дмитриевич
Ученая степень: доктор технических наук
Ученое звание: профессор
Место работы: Федеральное государственное бюджетное образовательное учреждение высшего образования «Петербургский государственный университет путей сообщения Императора Александра I»
Должность: Заведующий кафедрой «Информационные и вычислительные системы»
Почтовый адрес: 190031, Санкт-Петербург, Московский пр., 9.
Телефон: (812) 457-80-23
Адрес э

Ученый
Колодк , доцент