МИНОБРНАУКИ РОССИИ
федеральное государственное автономное образовательное учреждение высшего образования
«Санкт-Петербургский политехнический университет Петра Великого»
(ФГАОУ ВО «СПбПУ»)

ИНН 7804040077, ОГРН 1027802505279, ОКПО 02068574

Политехническая ул., 29, С.-Петербург, 195251
Телефон (812) 297-20-95, факс 552-60-80
E-mail: office@spbstu.ru

24.10.2017 № НК-Д-343

на № _____ от _____

УТВЕРЖДАЮ
Проректор по научной работе
Федерального государственного
автономного образовательного
учреждения высшего образования
«Санкт-Петербургский
политехнический университет Петра
Великого», доктор технических наук,
профессор, член-корреспондент
к
—
рович
2017г.

Отзыв

ведущей организации 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 диссертационной работы

Ежедневно появляется огромное количество новостей, сообщений в социальных сетях, электронных писем. Разработка автоматических и автоматизированных систем, позволяющих выделять из потока данных ценную информацию, является важной и актуальной задачей.

Анализ потока текстовых данных требует внимательного подхода как с точки зрения используемых алгоритмов, так и с точки зрения задействованных для решения методов. Особую ценность представляют алгоритмы,

позволяющие обрабатывать большие объемы данных, используя минимальные вычислительные ресурсы. Для облегчения принятия решения человеком, результаты анализа должны быть визуализированы. Поэтому для задач анализа потока текстовых документов, значимой считается подзадача визуализации результатов.

Содержание работы, соответствие паспорту специальности

Диссертационная работа состоит из введения, 4 глав, заключения, списка литературы и 2 приложений. Работа изложена на 153 страницах машинописного текста, содержит 14 таблиц и 51 рисунок. Библиографический список содержит 117 наименований.

Первая глава содержит обзор существующих исследований в области вероятностного тематического моделирования. Рассмотрены основные виды и методы оценки качества вероятностных тематических моделей. Уделено внимание практическому применению вероятностного тематического моделирования в задачах информационного поиска, при построении рекомендательных систем, в системах принятия решений. Обозначены существующие проблемы и направления исследований вероятностного тематического моделирования. Сформулирована содержательная постановка задачи исследования.

Вторая глава содержит обзор алгоритмов машинного обучения для анализа текстовых документов. В главе определены требования к разрабатываемому комплексу программных средств вероятностного тематического моделирования коллекций и потоков текстовых документов. Сформулированы требования к корпусу текстов, необходимому для проведения исследований вероятностных тематических моделей, представлен технологический процесс создания корпуса. Рассмотрены сценарии использования, на базе которых построена UML-диаграмма задач использования комплекса программных средств. Построена концептуальная схема программного комплекса, опирающаяся на событийно- и человеко-ориентированные подходы в разработке.

В третьей главе описан алгоритм многозначной классификации текстовых документов с использованием вероятностного тематического моделирования. Глава содержит метод определения темы «нового слова» через произведение Адамара тематических векторов документов, где оно встретилось. Предложенный алгоритм и метод используются для построения динамических вероятностных тематических моделей с бесконечным словарем на потоке текстовых документов.

Четвертая глава посвящена практической реализации математического и программного обеспечения комплекса программных средств на базе

микросервисной архитектуры. Комплекс программных средств включает микросервис создания корпуса текстов, построение вероятностной тематической модели на коллекции документов, построение динамической вероятностной тематической модели на потоке текстовых документов, микросервис визуализации результатов вероятностного тематического моделирования. Автор демонстрирует владение передовыми технологиями и подходами к разработке программного обеспечения.

Работа выполнена в соответствии с пунктами паспортов научных специальностей:

05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей:

- п. 3 «Модели, методы, алгоритмы, языки и программные инструменты для организации взаимодействия программ и программных систем»;
- п. 4 «Системы управления базами данных и знаний»;
- п. 7 «Человеко-машинные интерфейсы; модели, методы, алгоритмы и программные средства машинной графики, визуализации, обработки изображений, систем виртуальной реальности, мультимедийного общения».

Научная новизна полученных результатов

1. Разработан русскоязычный корпус текстов, позволяющий исследовать возможности алгоритмов вероятностного тематического моделирования. Предложенный подход к разработке корпуса позволяет создавать коллекции документов и корпуса для решения прикладных задач, проведения экспериментов. Корпус может быть использован для исследования различных алгоритмов обработки естественного языка, классификации и кластеризации текстовых документов.
2. Предложенный метод построения BTM на основе обучения с учителем позволяет расширить круг задач в которых применяется вероятностное тематическое моделирование, позволяет использовать BTM в задачах многозначной классификации.
3. Алгоритм многозначной классификации ml-PLSI демонстрирует мощность концепции вероятностного тематического моделирования в задачах классификации. Удобный вероятностный математический аппарат позволяет интуитивно понимать механизмы определения классов.
4. При анализе потока текстовых документов большое значение имеют характеристики «новых слов». Предложенный метод определения тем «новых слов» в BTM отличается высокой вычислительной эффективностью в сравнении с ранее описанными подходами.

5. Микросервисная архитектура, являясь трендом разработки на 2017 год, позволяет создавать масштабируемые приложения, состоящие из отдельных, полностью автономных блоков – микросервисов.

Достоверность и обоснованность полученных результатов

Достоверность и обоснованность основных результатов подтверждается анализом состояния исследований в области обработки текстов на естественном языке и в смежных областях знаний. Ознакомление с источниками, представленными автором, демонстрирует его осведомленность о текущем состоянии исследований и глубину проработки темы. Теоретические выводы, представленные в диссертационном исследовании, подтверждаются проведенными экспериментами над подготовленным корпусом текстов. Также в работе представлены акты внедрения, подтверждающие практическую апробацию результатов диссертационной работы и указывающие на их практическую значимость для различных вариантов использования.

Значимость полученных результатов для науки и практики

Теоретическая значимость

Теоретическая значимость результатов работы заключается в формализации алгоритма ml-PLSI с использованием математического аппарата вероятностного тематического моделирования, которая демонстрирует возможность использования вероятностного тематического моделирования в задачах классификации. Созданный корпус текстов SCTM-ru позволяет проводить эксперименты с алгоритмами обработки текстов на естественном языке.

Практическая значимость

Результаты, представленные в работе, имеют большую практическую значимость для задач анализа коллекций и потоков текстовых документов. Они могут быть использованы при решении задач классификации, кластеризации. Микросервисная архитектура комплекса программных средств позволяет использовать отдельные микросервисы для решения практических задач. Представленные результаты могут найти свое применение в информационном поиске и рекомендательных системах.

Публикации, апробация работы и личное участие автора в получении результатов диссертации

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.

В качестве подтверждения возможности практического применения результатов диссертационного исследования автором приводятся акты внедрения. Акт от ООО «Олимп» подтверждает использование результатов диссертационной работы в системе анализа новостного потока, разработанном для Правительства Москвы. Комплекс программных средств позволяет строить темпоральную вероятностную тематическую модель для отслеживания эволюции тем. Построенная BTM используется для анализа и планирования действий новостной редакции, позволяет выделять наиболее популярные темы новостей.

Акт от ООО «Rambler&Co» подтверждает возможность использования вероятностного тематического моделирования для классификации коротких текстов, которыми являются поисковые запросы. Следует отметить сложность автоматической классификации коротких текстов, необходимость тщательного отбора данных для обучения алгоритмов классификации, трудности, связанные с обучением алгоритмов многозначной классификации, при этом предложенный в работе алгоритм опережает альтернативные по качеству результатов классификации.

Возможность и необходимость использования результатов в учебном процессе подтверждается актом от Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики.

Рекомендации по использованию результатов и выводов диссертационной работы.

Теоретическая база и результаты рекомендуются к использованию в учебном процессе по направлениям машинное обучение, обработка естественного языка, компьютерная лингвистика, например, в Университете ИТМО. Практические результаты рекомендуются к использованию в системах информационного поиска, рекомендательных сервисах, системах принятия решения, для анализа коллекций и потоков текстовых документов, например, в организациях, которым требуется анализ текстовых документов, среди которых ООО «Олимп», ООО «Rambler&Co».

Замечания по диссертационной работе

Несмотря на высокий уровень представленных результатов, отметим следующие недостатки, выявленные в результате анализа текста диссертационной работы.

1. В диссертационной работе используется не техническая лексика, что затрудняет оценку проделанной автором работы.
2. В работе приведен обзор существующих алгоритмов вероятностного тематического моделирования, выявлены проблемы, над которыми работал

автор, но недостаточно полно представлен перечень существующих проблем в предметной области, не уделено внимание направлениям, над которыми ведутся работы другими учеными. Например, особенности построения вероятностных тематических моделей на коротких текстах, которыми являются сообщения в социальной сети Твиттер.

3. В работе уделено внимание тематической близости документов и отдельных слов, но не раскрыто понятие семантической близости. Упоминается технология дистрибутивной семантики word2vec, но не раскрыты сходства и различия между технологией word2vec и вероятностным тематическим моделированием.
4. В диссертационной работе слабо отражены перспективы дальнейшего исследования свойств вероятностных тематических моделей. В том числе возможность применения разработанных алгоритма многозначной классификации и метода определения темы «нового слова» при решении практических задач информационного поиска и при разработке рекомендательных сервисов.

Заключение

Диссертация Карповича С.Н. на соискание ученой степени кандидата технических наук является законченной научно-исследовательской работой, в которой решена практически значимая научная задача – разработка математического и программного обеспечения вероятностного тематического моделирования коллекции и потока текстовых документов и создан корпус текстов пригодный для исследований алгоритмов обработки и анализа текстовых документов на естественном языке. Диссертация написана грамотным научно-техническим языком с соблюдением установленных требований, имеет логически правильное построение и оформлена согласно действующим государственным стандартам, регулирующим оформление текста диссертации и ее дополнительных элементов. По каждому разделу представлены аргументированные выводы. Представленные замечания не снижают общую ценность и значимость результатов работы и не влияют на положительный вывод о качестве представленной диссертации.

Диссертационная работа Карповича С.Н. полностью удовлетворяет требованиям, установленным п. 9 положения о присуждении ученых степеней, утвержденного постановлением №842 Правительства РФ от 24 сентября 2013, 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».

Диссертационная работа и отзыв рассмотрены и одобрены на заседании кафедры «Телематика (при ЦНИИ РТК)» Федерального государственного

автономного образовательного учреждения высшего образования «Санкт-Петербургский политехнический университет Петра Великого», присутствовало 10 человек, протокол №3 от 10.10.2017 г.

Отзыв составили:

Доц. кафедры «Телематика (при Ц
Доцент, к.т.н.

Доц. кафедры «Телематика (при Ц
Доцент, к.т.н.

Сведения о составителях отзыва:

ФИО: Попов Сергей Геннадьевич
Ученая степень: кандидат технических нау
Ученое звание: доцент
Место работы: Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого»
Должность: доцент кафедры «Телематика (при ЦНИИ РТК)»
Почтовый адрес: Политехническая ул., 29-4-306, Санкт-Петербург, 195251
Телефон: +79219613493
Адрес электронной почты: popovserge@spbstu.ru

ФИО: Курочкин Михаил Александрович
Ученая степень: кандидат технических наук
Ученое звание: доцент
Место работы: Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого»
Должность: доцент кафедры «Телематика (при ЦНИИ РТК)»
Почтовый адрес: Политехническая ул., 29-4-306, Санкт-Петербург, 195251
Телефон: +78125526521
Адрес электронной почты: kurochkin.m@gmail.com