На правах рукописи

Карпович Сергей Николаевич

МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ВЕРОЯТНОСТНОГО ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ ПОТОКА ТЕКСТОВЫХ ДОКУМЕНТОВ

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

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Санкт-Петербург – 2017

Работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики» (Университет ИТМО)

Научный руководитель: Доктор технических наук, профессор
Смирнов Александр Викторович
Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН)
Заведующий лабораторией интегрированных систем автоматизации

Официальные оппоненты: Доктор технических наук, профессор
Хомоненко Анатолий Дмитриевич
Федеральное государственное бюджетное образовательное учреждение высшего образования «Петербургский государственный университет путей сообщения Императора Александра I»,
Заведующий кафедрой «Информационные и вычислительные системы»

Доктор технических наук, профессор
Водяхо Александр Иванович
Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)»,
Профессор кафедры «Вычислительная техника»

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого»

Защита диссертации состоится "___" ноября 2017 г. в __:__ часов на заседании совета по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук Д 002.199.01 при Федеральном государственном бюджетном учреждении науки Санкт-Петербургском институте информатики и автоматизации Российской академии наук (СПИИРАН) по адресу: 199178, Санкт-Петербург, 14-а линия В.О., 39, комн. 401. Факс: (812)-328-44-50 Тел: (812)-328-34-11.

С диссертацией и авторефератом можно ознакомиться на сайте Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации Российской академии наук (СПИИРАН): http://www.spiiras.nw.ru/dissovet/

Автореферат разослан "____"_____2017 г.

Ученый секретарь совета
Д 002.199.01

кандидат технических наук Зайцева А.А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. В связи с развитием цифровых технологий, постоянным ростом интернета, увеличением количества новостей, электронных писем, постов в блогах, растет потребность в алгоритмах для автоматической обработки текстов. Алгоритмы вероятностного тематического моделирования являются одним из перспективных направлений дистрибутивного анализа коллекций и потоков текстовых документов на естественном языке.

Вероятностное тематическое моделирование – это способ построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов. Вероятностные тематические модели (далее ВТМ) задают мягкую кластеризацию слов и документов по кластерам-темам, означающую, что слово или документ могут быть отнесены сразу к нескольким темам с различными вероятностями. ВТМ описывает каждую тему дискретным распределением на множестве терминов, каждый документ – дискретным распределением на множестве тем. В результате синонимы с большой вероятностью будут отнесены к одной теме, а омонимы попадут в разные. ВТМ, как правило, основаны на гипотезе «мешка слов» и «мешка документов», т.е. порядок слов в документе и порядок документов в коллекции не имеют значения.

ВТМ применяются для анализа потоков текстов. В данной работе под текстовым потоком понимается последовательность текстовых документов с определенным для каждого описанного события временем происшествия. Под обработкой потока текстовых документов понимается комплексная задача кластеризации поступающих документов и анализа эволюции тем этих документов. Для изучения особенностей алгоритмов вероятностного тематического моделирования при работе с русским языком необходимы русскоязычные текстовые корпуса.

Таким образом, подготовка данных для проведения исследований ВТМ, исследование свойств ВТМ и разработка методов вероятностного тематического моделирования для решения задач интеллектуального анализа текстов на естественном языке являются актуальными и востребованными задачами.

Степень разработанности темы. Появлению методологических основ ВТМ способствовала работа Х.Х. Пападимитриу, опубликованная в 1998 году. Развитие вероятностного тематического моделирования отражено в работах зарубежных ученых Т. Хофмана, Д. Блея, Э. Ына, М. Джордана и др. Вклад в развитие ВТМ внесли российские ученые Воронцов К.В., Потапенко А.А., Лукашевич Н.В., Нокель М.А., Коршунов А.В., Гомзин А.Г. Разработаны программные библиотеки для тематического моделирования, такие как Mallet, Gensim и BigArtm, позволяющие создавать ВТМ.

ВТМ успешно применяются в задачах информационного поиска, рекомендательных сервисах, методах разрешения морфологической

неоднозначности. В то же время остаются неизученные возможности ВТМ, например, применение тематического моделирования для многозначной классификации документов.

Целью диссертационной работы является разработка математического и программного обеспечения вероятностного тематического моделирования потока текстовых документов, позволяющего повысить доступность применения ВТМ за счет использования открытого программного обеспечения при решении прикладных задач информационного поиска, создании сервисов рекомендаций, анализе коллекции и потока текстовых документов.

Для достижения цели в работе поставлены следующие задачи:

1. Провести анализ современных методов вероятностного тематического моделирования для оценки ситуации в проблемной области и выявления путей повышения эффективности обработки текстовых данных.
2. Подготовить русскоязычный корпус текстов для тестирования алгоритмов вероятностного тематического моделирования, включающий помимо основного текста документа метатекстовую разметку о темах, к которым относится документ, его авторе и дате описанных событий, позволяющий эмулировать поток текстовых документов, исследовать динамические и темпоральные ВТМ.
3. Для анализа потока текстовых документов и отслеживания эволюции тем разработать алгоритм многозначной классификации текстовых документов с помощью вероятностного тематического моделирования.
4. Для пополнения словаря динамической ВТМ предложить метод определения тем «новых слов», отсутствующих в ВТМ на момент ее построения.
5. Апробировать предложенные метод и алгоритм путем создания прототипа программного комплекса для вероятностного тематического моделирования.

Методы исследования. При решении поставленных задач использовались методы системного анализа, математического и компьютерного моделирования, автоматической обработки естественного языка, теории вероятностей, математической статистики, прогнозирования временных рядов, теории машинного обучения и теории алгоритмов, разработки информационных систем и программирования.

Положения, выносимые на защиту:

1. Разработанный специальный русскоязычный корпус текстовых документов SCTM-ru позволяет исследовать алгоритмы вероятностного тематического моделирования.
2. Разработанный новый метод расчета матриц ВТМ на основе обучения с учителем (авторами документов) с учетом заданных связей между документами и темами упрощает построение ВТМ.
3. Разработанный оригинальный алгоритм классификации текстовых документов на базе ВТМ позволяет выполнять их многозначную классификацию.

4. Разработанный метод определения кластеров-тем для слова с использованием произведения Адамара позволяет определить темы «нового слова» в потоке текстовых документов.
5. Комплекс программных средств, разработанный на основе микросервисной архитектуры для вероятностного тематического моделирования, обеспечивает создание персонифицированных приложений для интеллектуального анализа коллекций и потоков текстовых документов**.**

Научная новизна работы состоит в следующем:

1. Создан 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. Текст и метатекстовая разметка необходимы для построения различных видов BTM. Источником данных корпуса является сайт «Русские Викиновости».
2. Предложен метод расчета матриц BTM на основе обучения с учителем (авторами документов), учитывающий заданные связи между документами и темами, что позволяет упростить построение BTM за счет отсутствия итераций.
3. Предложен алгоритм многозначной классификации текстовых документов ml-PLSI, основанный на вероятностном тематическом моделировании, заключающийся в использовании матрицы «слово-тема» BTM для классификации документов, что позволяет определять темы «новых документов» при анализе потока текстовых документов в динамической тематической модели.
4. Предложен метод определения тем «нового слова», основанный на использовании произведения Адамара тематических векторов документов, содержащих это слово, позволяющий определять вектора тем для «новых слов» в потоке текстовых документов при построении динамической тематической модели с эффективностью, превосходящей существующие аналоги.
5. Разработан прототип комплекса программных средств для анализа потока текстовых документов с использованием вероятностного тематического моделирования, отличающийся использованием микросервисной архитектуры и позволяющий предоставить вариативность выбора подходящих способов решения конкретных практических задач, а также возможность визуализации промежуточных и конечных результатов вероятностного тематического моделирования.

Обоснованность и достоверность научных положений, основных выводов и результатов диссертационной работы обеспечиваются анализом состояния исследований в проблемной области, корректным использованием методов исследования, подтверждена результатами вычислительных экспериментов и

эффективностью алгоритмов (сложность, трудоемкость) и программного обеспечения (надежность) при внедрении, а также апробацией основных теоретических положений диссертации в печатных трудах и на конференциях.

Практическая ценность работы. Результаты диссертационной работы могут найти применение в задачах анализа текстов на естественном языке, информационном поиске и в сервисах рекомендаций. Разработанная система позволяет анализировать коллекции и потоки текстовых документов, строить ВТМ, анализировать изменение популярности тем во времени с помощью темпоральных ВТМ.

Реализация результатов работы. 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). Результаты, полученные в ходе исследования, применяются в системе анализа новостного потока принятой к использованию в АО «Олимп» (Правительство Москвы) и в сервисе многозначной классификации поисковых запросов пользователей, принятом к использованию в ООО «Rambler&Co», а также в учебном процессе по курсу «Управление знаниями» кафедры информационных систем Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики.

Апробация результатов работы. Результаты диссертационного исследования представлялись на международных научно-методических конференциях «Современное образование: содержание, технологии, качество» (Санкт-Петербург, 2013, 2015), международной конференции «Региональная информатика» (Санкт-Петербург, 2014), межрегиональной конференции «Информационная безопасность регионов России» (Санкт-Петербург, 2015), международной научной конференции научной «Корпусная лингвистика» (Санкт-Петербург, 2015, 2017), международной конференции ассоциации открытых инноваций FRUCT: FRUCT 20 (Санкт-Петербург, 2017). По разработанной системе было получено свидетельство о регистрации программы для ЭВМ «Система для анализа текстовых документов с использованием вероятностного тематического моделирования // Карпович С.Н.» №2017615118 от 3 мая 2017.

Публикации. По теме диссертационной работы 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.

Структура и объем работы. Диссертация объемом 153 машинописных страниц, содержит введение, четыре главы и заключение, список литературы (117 наименований), 14 таблиц, 51 рисунок, одно приложение с копиями актов внедрения.

СОДЕРЖАНИЕ РАБОТЫ

Во введении приводится обоснование актуальности работы, формируются основные цели работы, решаемые задачи, определяется научная новизна и указывается практическая ценность результатов работы, представлены выносимые на защиту научные положения.

Первая глава посвящена методам вероятностного тематического моделирования. Рассмотрены основные виды ВТМ и методы оценки качества вероятностного тематического моделирования. Глава содержит обзор существующих исследований в области вероятностного тематического моделирования, на основе которого выявлены существующие проблемы ВТМ. Уделено внимание применению вероятностного тематического моделирования в практических задачах.

ВТМ задает отношение между темами и документами в корпусе текстов. Переход из пространства терминов в пространство найденных тематик помогает разрешать синонимию и полисемию терминов, а также эффективнее решать задачи информационного поиска, классификации, суммаризации и аннотирования коллекций документов и новостных потоков. Тематическое моделирование как вид статистических моделей для нахождения скрытых тем, встреченных в коллекции документов, нашло своё применение в машинном обучении и обработке естественного языка.

Одна из самых распространенных ВТМ – латентное размещение Дирихле (LDA), эта модель является обобщением вероятностного семантического индексирования. Другие ВТМ, как правило, являются расширением LDA. На рисунке 1 представлена концептуальная модель построения ВТМ.

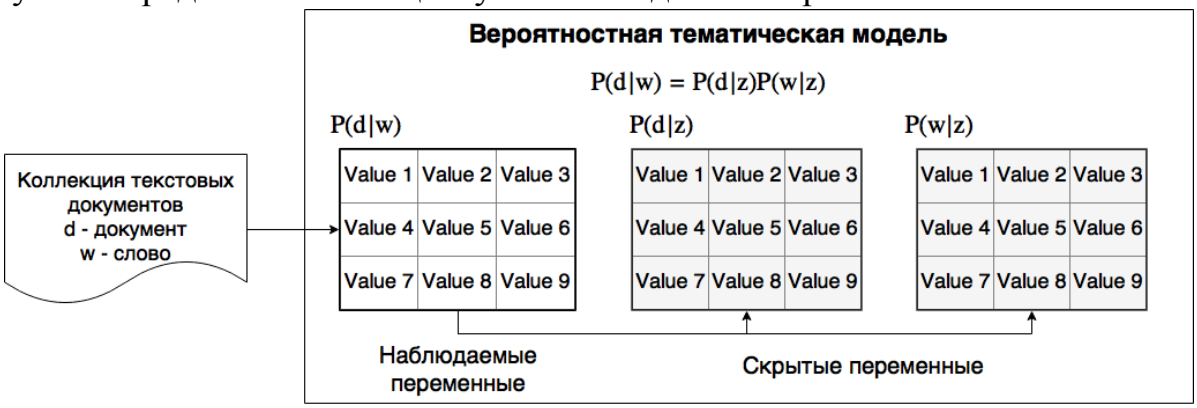

Рисунок 1 – Концептуальная модель вероятностного тематического моделирования: $P(w|z)$ – матрица искомых условных распределений слов по темам; $P(d|z)$ – матрица искомых условных распределений тем по документам; d – документ; w – слово; d, w – наблюдаемые переменные; z – тема (скрытая переменная)

Наиболее известным критерием оценки качества вероятностного тематического моделирования является перплексия – мера несоответствия модели словам, встречаемым в текстовых документах коллекции, и определяется через логарифм правдоподобия. Чем меньше перплексия, тем лучше модель обобщает данные.

В главе рассмотрены основные виды ВТМ: классические ВТМ – PLSA, LDA, ARTM; автор-тематические модели; обучаемые с учителем; учитывающие зависимости между словами документа; темпоральные ВТМ; динамические и онлайн тематические модели; многоязычные ВТМ; непараметрические ВТМ. Рассмотрены подходы к визуализации результатов вероятностного тематического моделирования. Целью данной главы является анализ актуальных проблем вероятностного тематического моделирования.

Проведенный анализ позволил выявить следующие проблемы:

1. Отсутствие русскоязычных текстовых корпусов, пригодных для исследования алгоритмов вероятностного тематического моделирования и включающих помимо основного текста документа дату, необходимую для построения темпоральных ВТМ: информацию об авторе документа, необходимую для построения автор-тематических моделей; информацию о том к каким темам относится каждый документ, необходимую для построения ВТМ методом обучения с учителем. Наличие русскоязычных текстовых корпусов, распространяемых по открытой лицензии, позволит специалистам, исследующим особенности ВТМ, сосредоточить свои усилия на исследованиях алгоритмов обработки текстов на естественном языке и сделает более доступным изучение особенностей русского языка, а именно синонимию, полисемию, омонимию.
2. Несоответствие существующих программных решений с использованием вероятностного тематического моделирования практическим потребностям, таким как обучение ВТМ, визуализация результатов ВТМ, экспорт ВТМ, анализ потока текстовых документов. Большинство алгоритмов вероятностного тематического моделирования используют статичный набор данных, в то время как в реальных задачах востребованы методы постоянного пополнения ВТМ новыми документами, словами и темами. При использовании классических алгоритмов ВТМ для потока текстовых документов, необходимо постоянно переобучать ВТМ при получении новых документов, так как словарь модели не пополняется новыми словами, и неперестроенная модель теряет свою актуальность со временем.
3. Для анализа потока текстовых документов с помощью ВТМ необходимо построить ВТМ на начальном наборе данных, а затем решать задачу многозначной классификации для определения к каким темам относится новый документ. Поэтому существует потребность в методах применения вероятностного тематического моделирования для решения задач многозначной классификации, а также в методах обучения ВТМ с учителем.
4. Для анализа потока текстовых документов не решен вопрос определения тем «нового слова» в ВТМ. Под «новым словом» в данной работе подразумевается слово, отсутствующее в словаре ВТМ на момент ее построения.

Решение обозначенных проблем позволит эффективнее использовать вероятностное тематическое моделирование в задачах обработки текста на естественном языке.

Во второй главе определены требования к разрабатываемому комплексу программных средств вероятностного тематического моделирования потока текстовых документов. Предложена концептуальная схема программного комплекса (рисунок 2). Определены требования к корпусу текстов для построения BTM. Предложен технологический процесс создания корпуса.

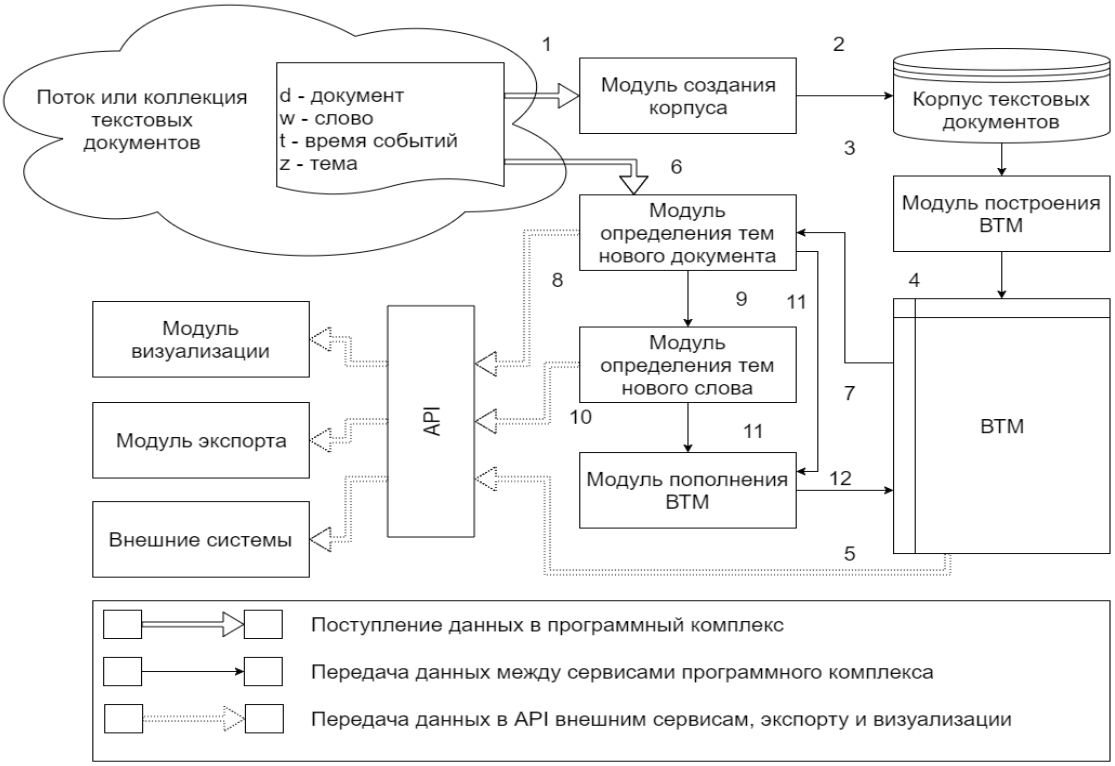

Рисунок 2 – Концептуальная схема программного комплекса

Для решения практических задач вероятностного тематического моделирования коллекций и потоков текстовых документов необходимо наличие открытых человеко- и событийно-ориентированных комплексов программных средств, в которых специалист, решающий свою прикладную задачу, сможет самостоятельно выбрать и задействовать подходящий алгоритм или программную библиотеку. Большинство существующих систем нацелено на решение одной задачи с заданным форматом входных данных, не предоставляют возможности пользователю вносить изменения в порядок их работы.

В ходе изучения сценариев работы специалистов по анализу данных были выделены базовые задачи, которые легли в основу прецедентов использования и позволили определить основные требования к комплексу программных средств. На рисунке 3 представлена диаграмма задач использования программного комплекса.

Основные требования к разрабатываемому комплексу программных средств для вероятностного тематического моделирования: управляемая предварительная обработка данных; выбор метода построения ВТМ; использование вероятностного тематического моделирования для многозначной классификации текстовых документов; построение ВТМ на коллекции и потоке текстовых документов. Концептуальная схема программного комплекса для вероятностного тематического моделирования представлена на рисунке 2.

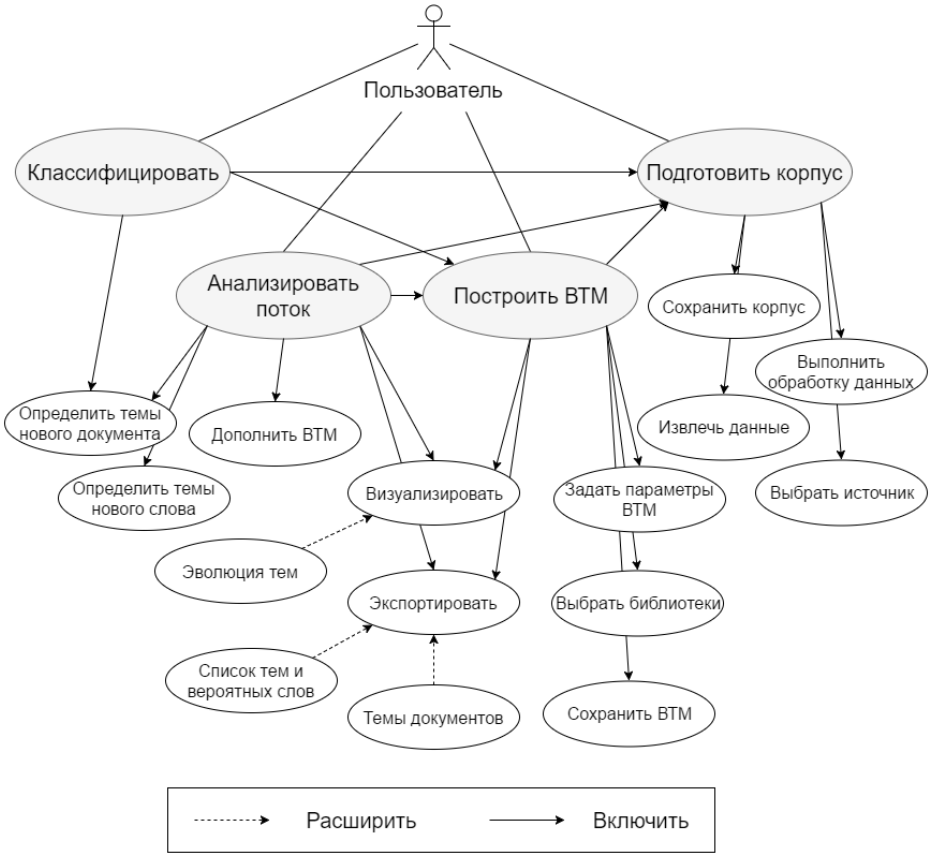

Рисунок 3 – UML-диаграмма задач использования программного комплекса

Определено, что разрабатываемый корпус должен распространяться по свободной лицензии, количество документов должно быть достаточным для проведения исследований с учетом особенностей языка корпуса, а именно синонимии, полисемии, омонимии. В результате анализа существующих текстовых корпусов и наборов данных, предложен технологический процесс создания корпуса, который состоит из следующих шагов: определение источника, предварительная обработка текстов документов корпуса, извлечение данных и метатекстовая разметка параметров каждого документа в корпусе, обеспечение доступа к корпусу. В качестве источника данных корпуса использован международный новостной сайт «Русские Викиновости» (Викиновости), тексты статей которого распространяются по свободной лицензии Creative Commons Attribution 2.5 Generic.

В третьей главе проведен обзор существующих методов обучения BTM и алгоритмов многозначной классификации. Предложен алгоритм многозначной классификации текстовых документов с использованием вероятностного тематического моделирования ml-PLSI. Выполнен обзор существующих подходов к расширению словаря BTM. Реализован метод определения тем «нового слова», в котором тематический вектор «нового слова» рассчитывается через произведение Адамара тематических векторов документов, где это слово встретилось. Разработан алгоритм, позволяющий расширять словарь BTM.

Схема построения BTM представлена на рисунке 4а. Если отождествить понятие темы BTM и тема документа в Викиновостях, а также учесть, что задача построения BTM имеет бесконечно много решений, то можно построить BTM (1), обучившись на размеченном корпусе.

$$P(w|d) = P(w|z)P(z|d), \qquad (1)$$

где: $P(w|d)$ – вероятность отношения «слово-документ»; $P(w|z)$ – вероятность отношения «слово-тема»; $P(z|d)$ – вероятность отношения «тема-документ», схема обучения представлена на рисунке 4б.

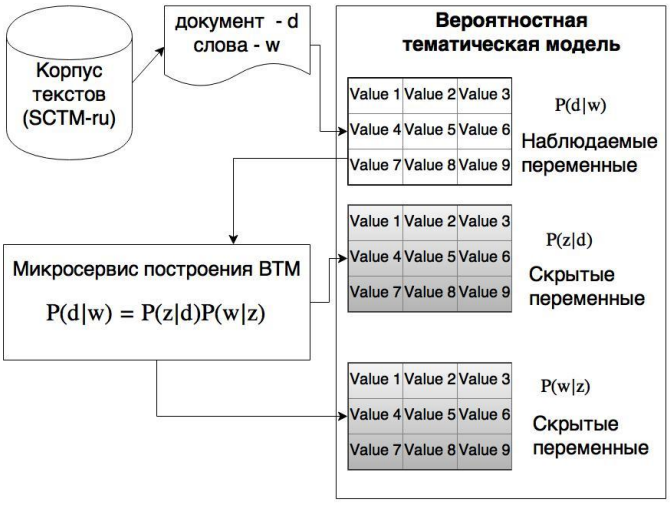

Рисунок 4а – Схема построения вероятностной тематической модели по корпусу SCTM-ru

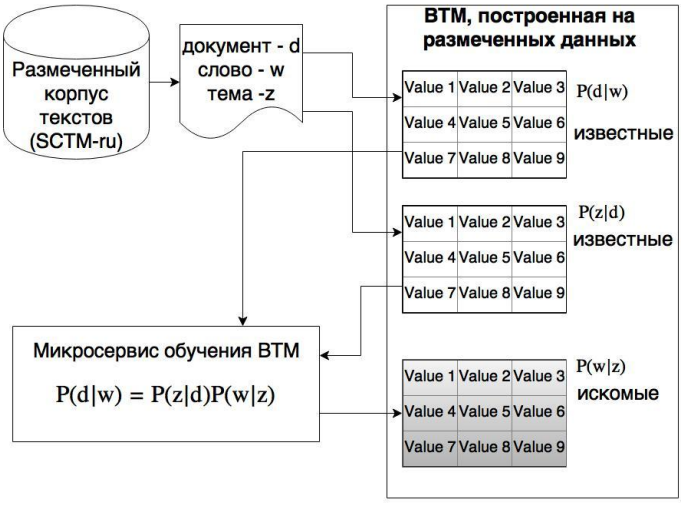

Рисунок 4б – Схема обучения BTM по размеченным данным корпуса SCTM-ru

Коллекция документов или корпус текстов, содержащие помимо текста документов информацию о том, к каким темам относится каждый документ, далее называется размеченными данными. Обучить BTM по размеченным данным – это значит рассчитать матрицу «слово-тема» по известным значениям матриц «слово-документ», «документ-тема» для каждого слова и каждого документа из корпуса текстов. Для большинства BTM не важна последовательность слов в тексте, каждый документ представлен неупорядоченным набором слов, называемым «мешок-слов». Алгоритм многозначной классификации ml-PLSI представлен в листинге 1.

Важной задачей для динамических ВТМ является определение тем «новых слов». Существуют методы построения динамических ВТМ в которых вклад «новых слов» в тему «нового документа» не учитывается, словарь таких моделей не меняется со временем, что снижает их качество. Существуют алгоритмы вероятностного тематического моделирования, в которых вектор тем «нового слова» берется из равномерного распределения Дирихле либо из процесса Дирихле и словарь ВТМ изменяется со временем. Однако не рассмотрены методы определения тем «новых слов» по тематическим векторам документов, содержащих это слово.

Листинг 1. Многозначная классификация на базе вероятностного тематического моделирования, ml-PLSI.

Вход: ВТМ, матрицы слово-тема, тема-документ и документ-слово, новый документ d_{new}

Выход: список предсказанных для документа тем с вероятностями.

1. Для всех $w \in d_{new}$ $\sum_{w \in d} p(w|z)$
2. Список тем документа, отсортированных по убыванию вероятности отнесения к теме.

Метод определения тем «нового слова» через сумму векторов «документ-темы», содержащих это слово, вычислительно эффективен, однако может приводить к ошибочным оценкам тематической принадлежности. Если документ содержит «новое слово» по ошибке автора документа, то метод суммы векторов не обнаружит ошибку. Устойчивым к ошибкам подобного рода является метод последовательного покомпонентного перемножения векторов документов так называемого произведения Адамара (2), содержащих «новое слово». Произведение Адамара обнуляет значение вероятности для непересекающихся векторов тем. Для расширения словаря ВТМ предлагается использовать алгоритм, схема которого представлена на рисунке 5.

$$p_{new}(w|z) = p_{i=1}(d|z) \circ p_{i+1}(d|z) \circ \ldots \circ p_n(d|z) \qquad (2)$$

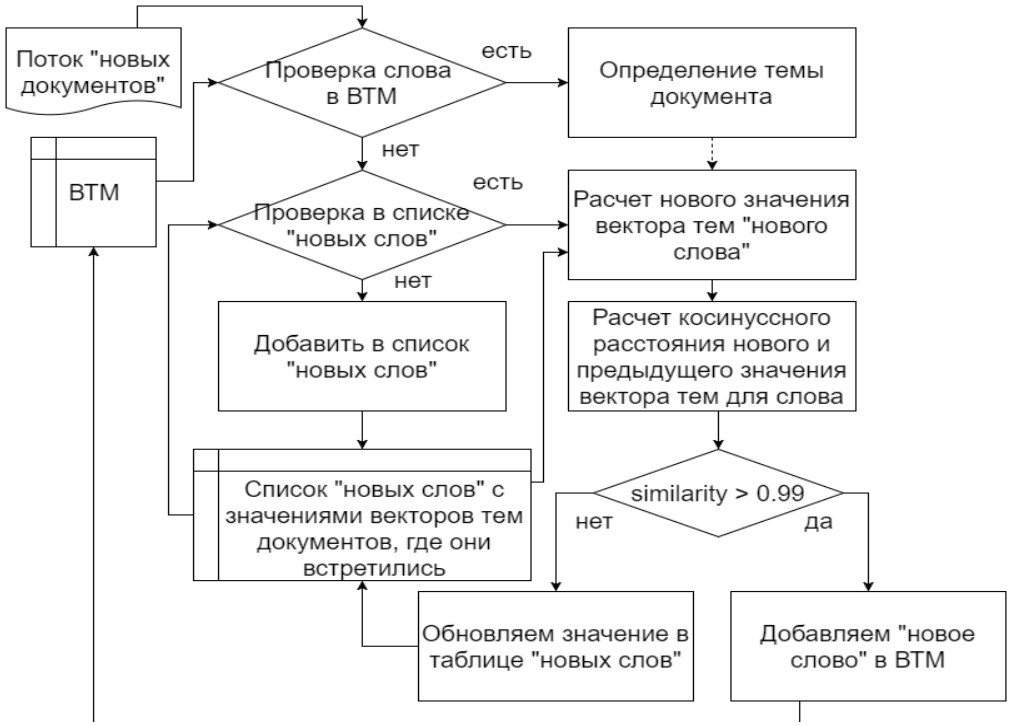

Рисунок 5 – Схема алгоритма определения тем «нового слова» и добавление его в ВТМ

Построение и регулярный пересчет ВТМ на корпусе текстов вычислительно сложная задача, поэтому при работе с динамическими и онлайн ВТМ актуально дополнение модели новыми словами и документами. Произведение Адамара – это бинарная операция над двумя векторами, сложность вычисления линейно зависит от длины векторов. Поэтому предложенный алгоритм вычислительно эффективнее и может быть использован для решения практических задач.

Четвертая глава посвящена проектированию комплекса программных средств вероятностного тематического моделирования для анализа текстовых документов и рассмотрению методов применения системы для решения прикладных задач интеллектуального анализа текстов.

На основе анализа требований к современным системам анализа текстовых документов для реализации программного комплекса предложена микросервисная архитектура (Microservice Architecture), представленная на рисунке 6. Программный комплекс строится как набор небольших микросервисов, каждый из которых работает в собственном процессе и коммуницирует с остальными, используя HTTP. Специалист управляет работой программного комплекса, инициирует процессы, выполняемые в микросервисах. Любой микросервис может быть изменен и это изменение не нарушит целостности системы.

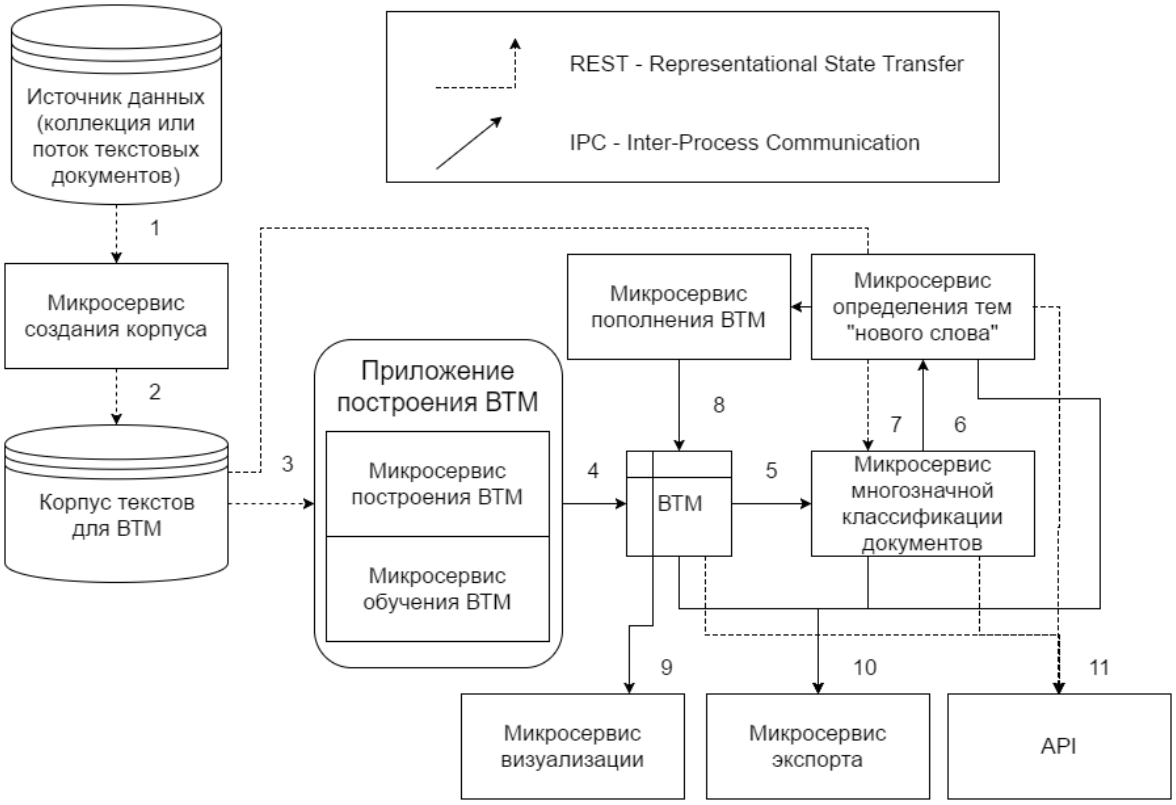

Рисунок 6 – Микросервисная архитектура комплекса программных средств вероятностного тематического моделирования

Для реализации комплекса были выбраны следующие программные средства: языковая платформа Python, интерактивная оболочка iPython, дистрибутив python Anaconda, библиотеки gensim, Pandas, Numpy, Scikit-learn, pyLDAvis, Matplotlib. Вместе выбранные программные средства и библиотеки обеспечивают необходимую гибкость и функциональность для реализации комплекса программных средств.

Реализован микросервис предварительной обработки Википедий для удаления невостребованных при построении BTM данных и извлечения необходимых. Предложена XML схема разметки корпуса. Подготовлен корпус текстов SCTM-ru, пригодный для исследования алгоритмов вероятностного тематического моделирования. Корпус SCTM-ru состоит из 12 тыс. документов, 320 авторов. События, описанные в документах, распределены по датам с ноября 2005 года по январь 2017 года. В корпусе SCTM-ru более 2,5 млн словоупотреблений, состоящих только из русских букв. Словарный состав корпуса – 262 тыс. уникальных словоформ.

Реализован микросервис многозначной классификации с использованием BTM. Для оценки эффективности предложенного алгоритма было проведено сравнение трех алгоритмов многозначной классификации: ml-PLSI, Random forest, Logistic regression. Корпус текстов SCTM-ru разделен на два непересекающихся множества документов: D – корпус текстов (SCTM-ru), D_1– документы корпуса для обучения,

D_2– документы корпуса для обучения, $D_1 \cap D_2 = \emptyset, D_1 + D_2 = D$. Оценка делалась на 100 новостях, для сравнения брались первые 50 предсказанных тем. Результат представлен в Таблице 1, алгоритм ml-PLSI точнее предсказывает первую наиболее вероятную тему и показывает лучшую полноту. Высокое значение функции потерь Хэмминга связано с большим количеством предсказываемых тем.

Таблица 1 – Сравнение качества алгоритмов многозначной классификации

Метрика качества	ml-PLSI	Random forest	Logistic regression
Функция потерь Хэмминга	0,70	0,96	0,95
Первая предсказанная тема совпала с одной из тем документа	62%	11%	11%
Процент верных предсказаний из 50 первых тем	96%	53%	84%

Для анализа потока текстовых документов необходимо располагать средствами определения тем «новых документов» и тем «новых слов», встретившихся в этих документах. Реализован микросервис, в котором через произведение Адамара рассчитывается вектор тем для «новых слов», встретившихся в потоке текстовых документов, на основе векторов тем документов, содержащих «новое слово».

На рисунке 7а представлен интерфейс микросервиса визуализации, отображены изменения в темах построенной BTM на временной шкале. На рисунке 7б представлен интерактивный интерфейс визуализации BTM, который позволяет оценить близость тем и получить представление о словарном составе каждой темы.

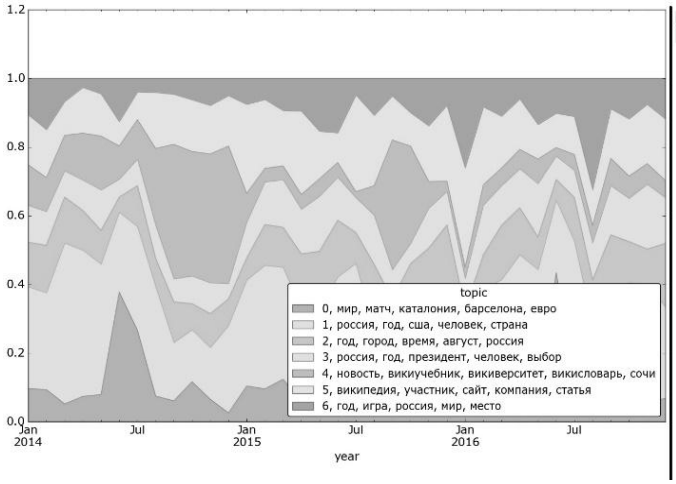

Рисунок 7а – Темпоральная BTM

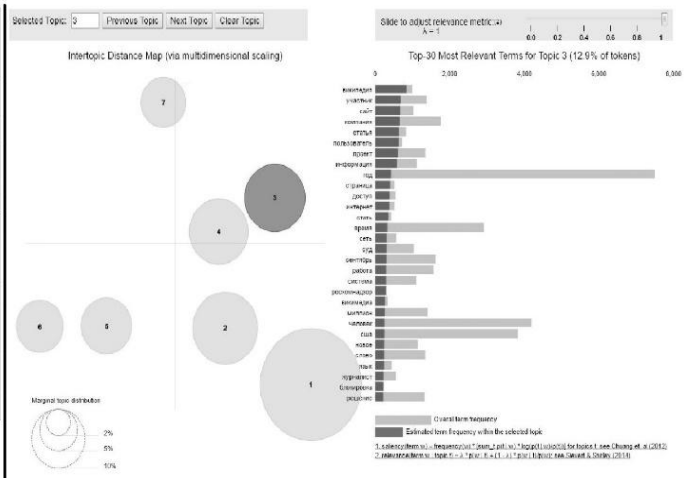

Рисунок 7б – Интерактивная визуализация BTM

На рисунке 8 приведён сценарий применения BTM для решения подзадачи классификации при информационном поиске. На рисунке 9 представлен сценарий применения BTM в рекомендательном сервисе. Преимущество предложенной схемы заключается в использовании алгоритма дополнения BTM новыми словами вместо перестроения при появлении новых документов в базе данных.

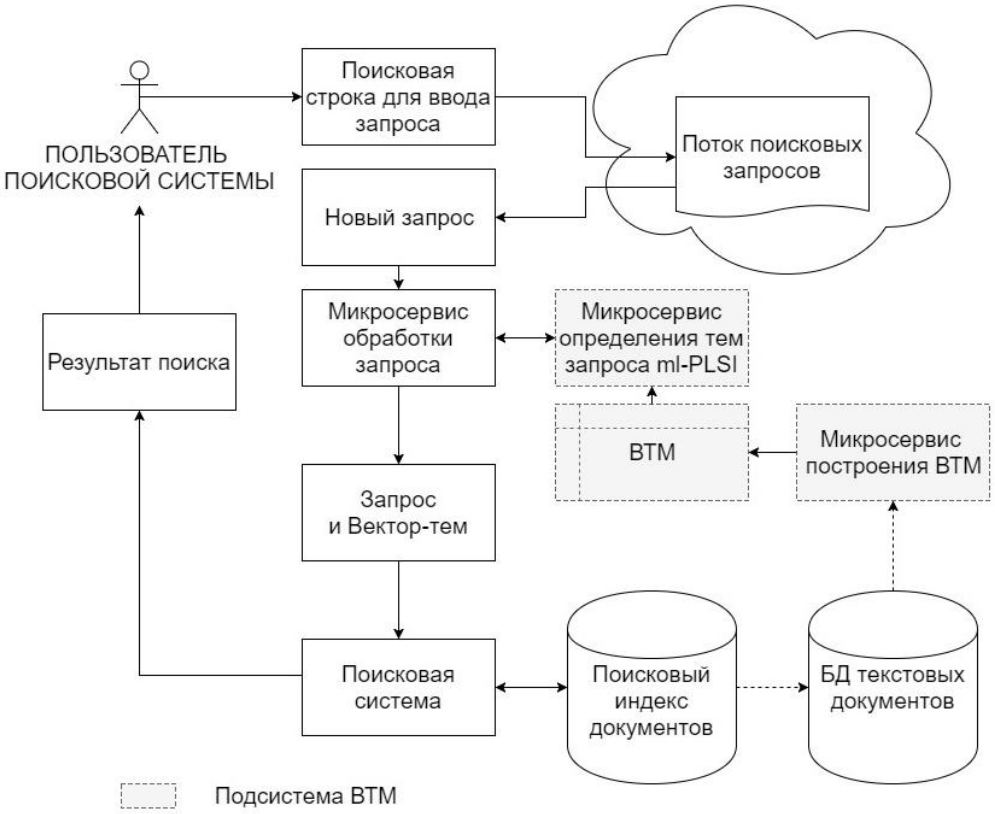

Рисунок 8 – Сценарий применения BTM для решения задач информационного поиска

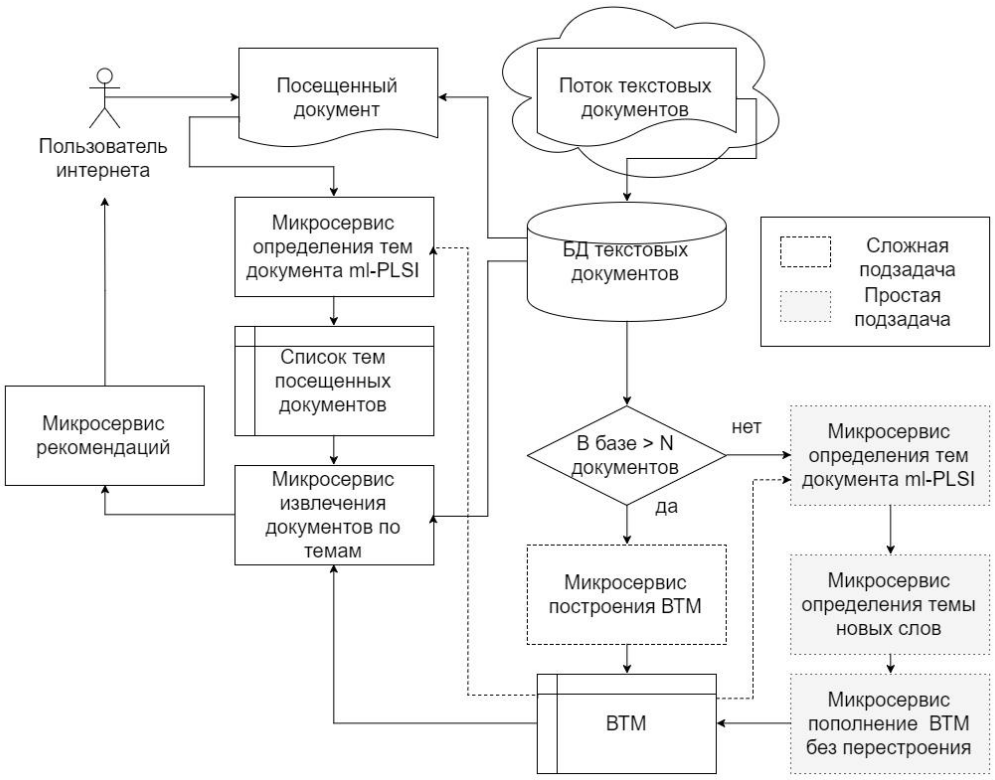

Рисунок 9 – Применение BTM в рекомендательном сервисе

ЗАКЛЮЧЕНИЕ

В диссертационной работе предложено решение актуальной научно-технической задачи по разработке комплекса математических и программных средств интеллектуального анализа потока текстовых документов с использованием вероятностного тематического моделирования, основанного на микросервисной архитектуре и позволяющего обеспечить специалиста необходимыми средствами анализа, возможностью выбирать источники данных, задавать параметры вероятностного тематического моделирования. В процессе решения данной задачи были получены следующие результаты:

1. Создан 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. Текст и метатекстовая разметка необходимы для построения BTM различных видов. Источником данных корпуса является сайт «Русские Викиновости».
2. Разработан метод расчета матриц BTM на основе обучения с учителем (авторами документов), учитывающий заданные связи между документами и темами, позволяющий упростить построение BTM за счет отсутствия итераций.
3. Разработан алгоритм многозначной классификации текстовых документов ml-PLSI, основанный на вероятностном тематическом моделировании, заключающийся в использовании матрицы «слово-тема» BTM для классификации документов, что позволяет определять темы «новых документов» при анализе потока текстовых документов в динамической тематической модели.
4. Разработан метод определения тем «нового слова», основанный на использовании произведения Адамара тематических векторов документов, содержащих это слово, позволяющий определять вектора тем для «новых слов» в потоке текстовых документов при построении динамической тематической модели с эффективностью, превосходящей существующие аналоги.
5. Разработан прототип комплекса программных средств для анализа потока текстовых документов с использованием вероятностного тематического моделирования, отличающийся использованием микросервисной архитектуры и позволяющий предоставить вариативность выбора подходящих способов решения конкретных практических задач, а также возможность визуализации промежуточных и конечных результатов вероятностного тематического моделирования.

Разработанный прототип комплекса программных средств рекомендован к использованию для построения и визуализации BTM. Дальнейшее исследование вероятностного тематического моделирования потока текстовых документов

возможно за счет улучшения функциональности комплекса программных средств. Перспективным направлением исследования является использование ВТМ в задачах ассоциативной классификации в комбинации с другими классификаторами, а также использование ВТМ как решателя в алгоритмах коллективного распознавания.

Полученные результаты соответствуют п. 3 «Модели, методы, алгоритмы, языки и программные инструменты для организации взаимодействия программ и программных систем», п. 4 «Системы управления базами данных и знаний» паспорта специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

ОСНОВНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

В рецензируемых журналах из списка ВАК и изданиях, приравненных к ним:

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**)
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**)
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**)
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**)

Зарегистрированное программное обеспечение и базы данных:

5. Система для анализа текстовых документов с использованием вероятностного тематического моделирования // Карпович С.Н. №2017615118 от 3 мая 2017

В других изданиях:

6. Карпович С.Н. Проблемы базового образования и профессиональной подготовки в сфере информационных технологий // Современное образование: содержание, технологии, качество: XX междунар. научно-метод. конф. СПб., 23 апр.2013. Материалы конф. Т.1. – С.66-68.
7. Карпович С.Н. Создание алгоритмов обработки неструктурированных данных большого объема // Материалы конференции. Региональная информатика (РИ-2014. X1У Санкт-Петербургская международная конференция «Региональная информатика (РИ-2014)», 29-31 окт. 2014. – С.39.
8. Карпович С.Н. Универсальный потоковый анализ и комбинирование тематических моделей в обработке неструктурированных данных большого объёма // 68-я Научно-техническая конференция профессорско-преподавательского состава СПбГЭТУ "ЛЭТИ" Санкт-Петербург, 28 января-5 февраля 2015 г. Сборник докладов студентов, аспиратнотов и молодых ученых

9. Карпович С.Н. Автоматическая обработка текстов в тематическом моделировании // Современное образование: содержание, технологии, качество: XX междунар. науч.-метод. конф.: Изд-во СПбГЭТУ «ЛЭТИ», СПб, 22 апр. 2015 г., Т.2 – С. 167-168.
10. Карпович С.Н. Русскоязычный корпус текстов СКТМ-ру для построения тематических моделей // Международная конференция «Корпусная лингвистика». 22-26 июня 2015. Сборник докладов конференции (РИНЦ)
11. Карпович С.Н. Информационно-психологическая безопасность тематической модели с непрерывным временем для контроля публикаций в сети интернет // Информационная безопасность регионов России (ИБРР-2015). IX Санкт-Петербургская межрегиональная конференция. Санкт-Петербург, 28-30 октября 2015 г.: Материалы конференции / СПОИСУ. - СПб., 2015. - С. 309-310.
12. Карпович С.Н. Визуализация тематических моделей с помощью iPython // Международная конференция «Корпусная лингвистика». 2017. Сборник докладов конференции (РИНЦ)