ОТЗЫВ

официального оппонента
кандидата технических наук Краснова Федора Владимировича
на диссертационную работу Апишева Мурата Азаматовича
«Эффективная реализация алгоритмов тематического моделирования с аддитивной регуляризацией»,
представленную на соискание ученой степени кандидата технических наук по специальности
05.13.17 – «Теоретические основы информатики»

Актуальность темы

Методы тематического моделирования являются одним из направлений исследований в обработке естественного языка. Они активно развиваются последние 20 лет и используются в разных задачах текстовой аналитики, таких как кластеризация и классификация текстов, автореферирование, построение векторных представлений слов и документов, информационный поиск. С учётом постоянного роста объемов обрабатываемых данных, крайне актуальной становится проблема повышения эффективности обучающих алгоритмов.

Теоретический аппарат аддитивно регуляризованных тематических моделей (ARTM) даёт возможность гибко строить разнообразные модели с различными свойствами, однако на сегодняшний день более популярным является старый, относительно громоздкий байесовский подход (LDA). В связи с этим в научной литературе публикуется множество работ, предлагающих более производительные реализации алгоритмов обучения LDA. Для обучения же моделей ARTM существует только один полноценный инструмент – библиотека с открытым кодом BigARTM. Диссертационная работа М.А. Апишева посвящена исследованию методов повышения эффективности алгоритмов обучения моделей ARTM в этой библиотеке, а также возможностей использования аддитивно регуляризованных тематических моделей для решения прикладных задач анализа данных.

Содержание

Диссертационная работа состоит из введения, двух обзорных глав, трех глав, посвященных результатам исследований, заключения и списка литературы, содержащего 77 наименований. Общий объем диссертации – 124 страницы.

Во **введении** описываются цель и основные задачи исследования, демонстрируются их актуальность, новизна и практическая значимость, формулируются основные положения, выносимые на защиту.

Первая глава посвящена формальной постановке задачи тематического моделирования. Представлены три основных подхода: вероятностный латентный семантический анализ (PLSA), латентное размещение Дирихле (LDA) и аддитивная регуляризация тематических моделей (ARTM), в рамках которого выполнена настоящая диссертационная работа. Вводятся необходимые понятия и обозначения.

Вторая глава носит обзорный характер и посвящена описанию и анализу предлагаемых в научной литературе методов повышения производительности алгоритмов обучения тематических моделей LDA.

В **третьей главе** рассматриваются различные параллельные EM-алгоритмы для обучения аддитивно регуляризованных моделей, используемые в библиотеке BigARTM. Демонстрируются недостатки синхронных алгоритмов и асинхронного онлайнового алгоритма Async. Описывается предлагаемый автором алгоритм DetAsync, реализующий обучение модели с запаздывающим на один шаг обновлением параметров и общей для параллельных потоков матрицей счётчиков. Экспериментально демонстрируется более высокая скорость сходимости DetAsync по сравнению с предшествующими алгоритмами. Дополнительно производится сравнение по скорости работы с конкурирующими инструментами для обучения модели LDA, показывается существенное превосходство BigARTM с точки зрения скорости работы и масштабируемости по числу параллельных потоков. Предлагается метод хранения и обработки данных для случая существенно разреженных моделей. В серии экспериментов показываются преимущества использования нового подхода для всех реализованных в BigARTM алгоритмов.

Четвертая глава вводит теорию обучения тематических моделей с произвольной функцией потерь. Одним из её следствий является возможность обучения с использованием быстрого E-шага. Это модификация обычного E-шага, которая не требует подсчета нормировочной константы. Предлагаются различные стратегии комбинирования обычных и быстрых E-шагов. В серии экспериментов выявляются наилучшие стратегии для оффлайнового и онлайновых алгоритмов, определяются условия применимости этих стратегий для повышения скорости и качества моделирования.

В **пятой главе** рассматриваются алгоритмические расширения и приложения аддитивно регуляризованных тематических моделей. В первой части главы вводится теория мультимодальных моделей. Описывается предлагаемая автором модель для решения задачи выявления из корпуса текстов тематик специфической направленности, описываемых небольшим набором ключевых слов. В ходе многочисленных экспериментов показывается превосходство предлагаемого подхода над более простыми моделями. Также исследуются методы настройки и повторного использования разработанной модели. Во второй части главы описывается теория гиперграфовых (транзакционных) тематических моделей T-ARTM. Описываются детали реализации алгоритма обучения таких моделей в BigARTM. Экспериментально демонстрируется преимущество использования моделей T-ARTM при обработке транзакционных текстовых данных.

Основные результаты и их новизна

В своей диссертационной работе М.А. Апишев получил следующие новые результаты:

1. алгоритм параллельного асинхронного онлайнового обучения регуляризованных мультимодальных тематических моделей;
2. алгоритм обучения регуляризованных мультимодальных тематических моделей с разреженным хранением параметров;
3. модификация EM-алгоритма с ускоренным E-шагом без нормировки и стратегии ее применения для оффлайнового и онлайновых алгоритмов;
4. стратегия комбинирования регуляризаторов для выделения специфических тем по заданному словарю с приложением к анализу этно-релевантных тем в текстах социальной сети;

5. реализация алгоритма обучения гиперграфовых тематических моделей транзакционных данных.

Достоверность результатов

Достоверность представленных результатов обеспечивается большим объемом исследуемого материала, широкой теоретической базой, а также применением стандартных программных инструментов и общепринятых методик проведения эксперимента. Все проделанные эксперименты удовлетворяют принципам воспроизводимости, разработанный программный код является общедоступным.

Значимость результатов

Практическая значимость диссертационного исследования заключается в разработке и реализации высокоэффективных алгоритмов обучения тематических моделей в библиотеке BigARTM, что открывает исследователям возможность существенно быстрее строить разнообразные модели ARTM на больших данных. Наиболее значимым теоретическим результатом является предложенная в работе новая аддитивно регуляризованная модель для извлечения из текстов тематик специфической направленности.

Замечания

1. В разделе 4.1 (стр.63) кажется не самым подходящим выбор обозначений: F для полного набора обычных итераций и N для полного набора быстрых итераций. Вероятно, более логичным было бы обозначить их наоборот: F — быстрые (fast) итерации, N — обычные (normal).
2. В разделе 4.2. указано, что в экспериментах используются наилучшие значения коэффициентов регуляризации. Возможно, имело смысл перечислить, какие ещё коэффициенты были опробованы и к каким результатам они привели.
3. В разделе 5.3 (стр. 83) указано, что в экспериментах используется 25 итераций обработки каждого документа, что кажется избыточным, не вполне ясно, по какой причине было выбрано именно это число.
4. В разделе 4.2 (стр. 64) написано: «Число k наиболее вероятных слов во всех экспериментах принято равным 10». Необходимо пояснить такой выбор числа k для подсчета когерентности (4.2), так как, например, если k = W, то значение когерентности становится константой относительно Ф.
5. В разделе 4.2 (стр. 65) целесообразно уточнить, что речь идет о df-когерентности.
6. В разделе 5.3 (стр. 86) в таблице 5.1 приведены средние значения когерентности. Вызывает сомнение применение усреднения по распределению когерентности для каждой тематики, так как характер распределения не известен и нет оснований полагать распределение нормальным.
7. Оформление: согласно ГОСТ Р 7.0.11-2011 «СИБИД. Диссертация и автореферат диссертации. Структура и правила оформления» со ссылкой на ГОСТ 2.105-95 «Единая система конструкторской документации (ЕСКД). Общие требования к текстовым документам» ГОСТ название таблиц следует помещать над таблицей (п. 4.4.1).

Следует отметить, что указанные замечания не снижают значимости полученных результатов и не влияют на общую положительную оценку диссертационного исследования Апишева М. А.

Заключение

Целью диссертационной работы М.А. Апишева являлась модернизация библиотеки тематического моделирования для построения моделей больших текстовых коллекций с возможностью гибкой настройки процесса обучения. Разработанные алгоритмы позволяют существенно ускорить обучение моделей ARTM на большом числе документов и с большим числом тем. Эти алгоритмы были реализованы в программном проекте с открытым кодом BigARTM. М.А. Апишев проделал большой объем исследовательской и экспериментальной работы: было изучено значительное количество научных работ, посвященных проблемам повышения эффективности обучения тематических моделей, были реализованы и протестированы новые алгоритмы, разработана и исследована новая аддитивно регуляризованная тематическая модель.

По теме диссертации опубликовано 11 печатных работ, 8 из них – в изданиях, входящих в перечень ВАК, 7 из них индексируются в базе Scopus.

Диссертационная работа М.А. Апишева «Эффективная реализация алгоритмов тематического моделирования с аддитивной регуляризацией» является законченной самостоятельной научно-исследовательской работой, содержащей новые научные результаты.

Работа удовлетворяет всем требованиям, предъявляемым ВАК к диссертациям, представленным на соискание ученой степени кандидата технических наук по специальности 05.13.17 – «Теоретические основы информатики». Диссертант М.А. Апишев заслуживает присвоения ему ученой степени кандидата технических наук.

Официальный оппонент,
кандидат технических наук,
директор департамента информационных
систем управления компании NAUMEN (АО «Наумен»)
Адрес: 109147, Москва, ул. Воронцовская, 35Б,
корп. 3, бизнес-центр «Time Center», 5 этаж
Телефон: +7-981-781-48-47
E-mail: fkrasnov@naumen.ru

Краснов Федор Владимирович

Дата 26.11.2020

Подпись удостоверяю
Генеральный директор
АО «Наумен»
Атаманов В.Ю.

Акционерное общество «НАУМЕН» Российская Федерация г. Екатеринбург ОГРН 1026605241807