«Утверждаю»
Врио директора Федерального
государственного бюджетного
учреждения науки
Институт системного
программирования
им. В.П. Иванникова РАН
академик РАН, д.ф.-м.н.

_____ А.И. Аветисян

«18» ноября 2020

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ
Федерального государственного бюджетного учреждения науки «Институт системного программирования им. В.П. Иванникова»
Российской академии наук
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

Вероятностное тематическое моделирование – это один из инструментов текстовой аналитики, позволяющий выделять из корпусов текстов интерпретируемые тематические кластеры, которые полезны при решении задач визуализации данных, категоризации, информационного поиска, автореферирования. Теория аддитивно регуляризованных тематических моделей (ARTM) дает возможность строить модели, удовлетворяющие разнообразным требованиям, формализуемым в виде регуляризаторов. Такой подход является существенно более простым и гибким, чем наиболее распространённый на текущий момент аппарат байесовских моделей на основе латентного размещения Дирихле (LDA). Тем не менее, в отличие от LDA, для модели ARTM существует только один полноценный инструмент – библиотека с открытым кодом BigARTM. С учетом постоянного роста объемов обрабатываемых данных, существенной становится проблема производительности алгоритмов обучения тематических моделей в BigARTM. Данная работа посвящена разработке и реализации более эффективных алгоритмов обучения моделей ARTM, а также их использованию для решения прикладной задач анализа данных.

Содержание

Диссертационная работа состоит из введения, пяти глав и заключения.

Во введении обоснована актуальность работы, установлены цели и задачи исследования, сформулированы основные положения, выносимые на защиту,

обоснованы научная новизна и значимость работы, достоверность результатов, перечислены основные публикации по теме диссертации.

В первой главе поставлена задача вероятностного тематического моделирования, рассмотрены и изложены в единой нотации различные алгоритмы построения таких моделей. Для модели LDA приведены несколько методов оценивания параметров, показана идейная схожесть этих методов с друг с другом и с EM-алгоритмом для обучения модели ARTM.

Во второй главе приведен обзор предложенных в научной литературе реализаций эффективных алгоритмов для обучения модели LDA. Выделены, описаны и проанализированы различные методы повышения производительности, используемые в этих реализациях.

В третьей главе описаны имеющиеся в BigARTM варианты EM-алгоритмов для модели ARTM и предложены новые, более эффективные. Разработан онлайновый асинхронный алгоритм DetAsync, а также новый метод хранения и обработки разреженных тематических моделей. Проведены эксперименты, демонстрирующие преимущества новых алгоритмов перед старыми, а также их значительно более высокую скорость работы по сравнению с конкурирующими реализациями в равных условиях.

В четвертой главе описана теория обучения тематических моделей с произвольной функцией потерь, следствием которой является возможность использования EM-алгоритма с E-шагом, не требующим подсчёта нормировочной константы (быстрый E-шаг). Предложен набор стратегий комбинирования обычных и быстрых E-шагов. Экспериментально выявлена одна лучшая стратегия для оффлайнового и онлайновых алгоритмов, выработаны рекомендации относительно условий их использования.

В пятой главе рассмотрены алгоритмические расширения аддитивно регуляризованных тематических моделей: мультимодальные тематические модели (M-ARTM) и гиперграфовые (транзакционные) тематические модели (T-ARTM). В рамках аппарата M-ARTM автором предложена модель для решения задачи выявления из корпуса текстов тематик специфической направленности, описываемых небольшим набором ключевых слов. В проведенных экспериментах демонстрируется превосходство предложенной модели над более простыми конкурентами. Описаны детали реализации алгоритма обучения моделей T-ARTM в BigARTM, экспериментально показано их преимущество при обучении моделей на данных транзакционной природы.

Основные результаты и их новизна

В рамках диссертационной работе М.А. Апишев получены следующие новые результаты:

1. Разработан алгоритм параллельного асинхронного онлайнового обучения регуляризованных мультимодальных тематических моделей.
2. Разработан алгоритм обучения регуляризованных мультимодальных тематических моделей с разреженным хранением параметров.
3. Разработана модификация EM-алгоритма с ускоренным E-шагом без нормировки и стратегии ее применения для оффлайнового и онлайновых алгоритмов.

4. Предложена стратегия комбинирования регуляризаторов для выделения специфических тем по заданному словарю с приложением к анализу этно-релевантных тем в текстах социальной сети.
5. Разработана реализация алгоритма обучения гиперграфовых тематических моделей транзакционных данных.

Теоретическая и практическая значимость

В работе М.А. Апишева изложены и систематизированы различные подходы к повышению эффективности работы алгоритмов обучения тематических моделей LDA. Предложены и реализованы новые алгоритмы обучения моделей ARTM.

В рамках подхода аддитивной регуляризации тематических моделей предложена комбинация регуляризаторов, позволяющая улучшить качество решения задачи извлечения из текстов тематик специфической направленности по сравнению с решениями на основе более простых тематических моделей.

В экспериментах на трех текстовых корпусах (статьи Википедии, аннотации Pubmed и посты LiveJournal) продемонстрирована практическая значимость полученных результатов. Результаты работы М.А. Апишева существенно модернизировали библиотеку BigARTM, которая используется различными компаниями и научными группами в проектах по обработке естественного языка.

Достоверность результатов

Предлагаемые алгоритмы основаны на исследовании большого объема современных научных работ по теме, реализации их выполнены в библиотеке с открытым кодом и являются общедоступными. Основные результаты работы получены в ходе экспериментов, проведенных с применением стандартных для данной области программных инструментов и удовлетворяющих принципам воспроизводимости.

Результаты работы докладывались автором на конференциях FRUCT, AIST и «Ломоносов», материалы диссертации использовались в рамках учебных курсов «Машинное обучение на больших данных» (ШАД) и «Машинное обучение и большие данные» (ФИВТ МФТИ).

Замечания

1. В главе 2 проводится подробный обзор реализаций алгоритмов обучения тематического моделирования, но при этом из обзора исключены реализации, осуществляющие обучение на GPU. Исключение из обзора моделей, обучающихся на GPU выглядит недостаточно обоснованным, в похожей задаче обучения искуственных нейронных сетей использование GPU способно ускорить обучение в десятки раз.
2. В главе 3 в начале приводятся диаграммы Ганта, очевидно полученные в ходе эксперимента. При этом харрактеристики вычислительного узла приводятся только в середине главы. Это нарушает логику повествования.
3. В главе 3 формулируется теорема 1, но не приводится ни доказательство этой теоремы ни ссылка на работу с доказательством теоремы.

Заключительная оценка

Диссертационная работа Апишева Мурата Азаматовича «Эффективная реализация алгоритмов тематического моделирования с аддитивной регуляризацией», выполненная под руководством д.ф.-м.н., профессора К.В. Воронцова, является законченной научно-квалификационной работой, содержащей новые научные результаты в областях алгоритмов обучения тематических моделей и методов использования аддитивной регуляризации для анализа данных. Все результаты, выносимые на защиту, строго обоснованы и подтверждены в ходе вычислительных экспериментов.

Результаты работы опубликованы в 11 печатных трудах, 8 из которых – в изданиях, рекомендованных ВАК. Автореферат диссертации, опубликованные статьи и тезисы докладов достаточно полно отражают основное содержание работы. Отмеченные недостатки работы не снижают общей положительной оценки.

Диссертация соответствует всем критериям, установленным Положением о присуждении ученых степеней, предъявляемым к диссертациям на соискание ученой степени кандидата технических наук по специальности 05.13.17 – «Теоретические основы информатики», а ее автор – Апишев Мурат Азаматович заслуживает присуждения ученой степени кандидата технических наук по указанной специальности.

Настоящий отзыв обсуждался и был одобрен на заседании отдела информационных систем Федерального государственного бюджетного учреждения науки Института системного программирования им. В.П. Иванникова РАН 17.11.2020 г. протокол № 1.

Заведующий отделом
информационных систем ИСП РАН
к.ф.-м.н. Д.Ю. Турдаков