«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

«23» ноября 2020

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

Федерального государственного бюджетного учреждения науки «Институт системного программирования им. В.П. Иванникова»
Российской академии наук
на диссертационную работу Ирхина Ильи Александровича
«Единственность матричного разложения и сходимость регуляризованных алгоритмов в вероятностном тематическом моделировании», представленную на соискание ученой степени кандидата физико-математических наук
по специальности 05.13.17 – «Теоретические основы информатики»

Актуальность темы

Тематическое моделирование – одно из приложений машинного обучения к анализу текстов. Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова образуют каждую тему. Полученные таким образом вектора могут быть использованы для классификации и категоризации документов или визуализации данных. Этот подход используется не только в анализе текстов, но и в задачах информационного поиска, рекомендаций, анализа изображений и других.

Одним из методов построения тематических моделей является метод аддитивной регуляризации тематических моделей. Данный подход позволяет объединять различные требования к тематической модели в виде функционалов-регуляризаторов и учитывать их при решении оптимизационной задачи. В данной работе доказываются теоретические свойства указанного подхода такие как сходимость оптимизационного алгоритма и единственность решения, а также разрабатываются улучшения алгоритма.

Содержание

Диссертационная работа состоит из введения, пяти глав и заключения.

Во введении обоснована актуальность работы, установлены цели и задачи исследования, сформулированы основные выносимые на защиту положения, установлена научная новизна, обоснованы теоретическая и практическая значимость, достоверность результатов, перечислены основные публикации по теме диссертации.

В первой главе сформулирована задача тематического моделирования, описан метод аддитивной регуляризации тематических моделей (ARTM) и введено обобщение оптимизационной задачи максимизации регуляризованного правдоподобия на случай

произвольной монотонной функции потерь. Приведён используемый в подходе регуляризованный EM-подобный итерационный алгоритм.

Во второй главе доказываются теоремы о достаточных условиях сходимости итерационного алгоритма ARTM. Данный алгоритм рассматривается как GEM-алгоритм и используются известные результаты о сходимости GEM-алгоритмов. На основе теоретических результатов предлагается модификация итерационного алгоритма и в эксперименте подтверждается улучшение качества тематических моделей при использовании этой модификации.

В третьей главе рассматривается проблема единственности стохастического матричного разложения в точке сходимости итерационного алгоритма ARTM. Доказывается теорема о достаточных условиях единственности такого разложения. Выполнение предложенных условий подтверждается в экспериментах. Описывается связь задачи тематического моделирования и разложения матрицы. По итогу приводится вывод, что неединственность решения задачи тематического моделирования вызвана неоднозначностью выбора приближения исходной матрицы, а не неединственностью точного стохастического матричного разложения выбранного приближения.

В четвертой главе предлагается метод разреживания тематических моделей. Приводятся теоремы, оценивающие изменение максимизируемого функционала при зануления параметров модели. На основе этих теорем приводится алгоритм разреживания, который в экспериментах сравнивается со стандартным способом разреживания подхода ARTM. Эксперимент показывает, что предложенный алгоритм даёт те же значения разреженности при больших значения оптимизируемого функционала.

В пятой главе формулируется изменение оптимизационной задачи ARTM на случай, когда матрицы параметров модели находятся в функциональной зависимости. Для данного случая выводится итерационный алгоритм, который оказывается совместимым с алгоритмом ARTM. В экспериментах показывается, что предложенный алгоритм позволяет получить более согласованные, различные и интерпретируемые темы.

Основные результаты и их новизна

В рамках диссертационной работы Ирхина И.А. получены следующие новые результаты:

1. Доказана теорема о достаточных условиях сходимости алгоритма аддитивной регуляризации тематических моделей. Предложенные условия являются интерпретируемыми и проверяемыми на практике.
2. Доказана теорема о достаточных условиях единственности стохастического матричного разложения. Сформулированы причины неединственности решения для задач тематического моделирования.
3. Предложена и теоретически обоснована модификация алгоритма ARTM, ускоряющая сходимость итерационного процесса.
4. Разработан метод разреживания тематической модели, не увеличивающий перплексию получаемой модели.

Теоретическая и практическая значимость

Полученные в диссертации результаты могут использоваться в научных исследованиях в области тематического моделирования.

Работа Ирхина И.А. носит преимущественно теоретический характер. Основной её ценностью являются теоретические обоснования наблюдаемых на практике свойств алгоритма аддитивной регуляризации тематических моделей и предложенные на их основе

модификации алгоритма, которые позволяют увеличить качество получаемых тематических моделей.

Часть предложенных модификаций реализована в библиотеке с открытым кодом TopicNet.

Достоверность результатов

Математические результаты диссертационной работы Ирхина И.А. сформулированы в виде утверждений и теорем с корректными доказательствами. Проведена экспериментальная проверка предложенных методов на открытых коллекциях текстовых документов. Исходный код и данные всех экспериментов находятся в открытом доступе. Результаты работы докладывались на конференциях и научных семинарах.

Замечания

1. В главе 2 при проверке эффекта от предложенной модификации проводятся эксперименты только с регуляризатором декоррелирования. Стоит также провести эксперименты и с другими регуляризаторами.
2. Алгоритмы глав 3 и 4 реализованы только в экспериментальной реализации алгоритма ARTM в репозитории автора. Для возможности практического использования результата стоит перенести алгоритмы в общеиспользуемую библиотеку.
3. В работе имеются грамматические ошибки и опечатки (Например, в Определении 2, формуле 2.2)

Отмеченные недостатки не оказывают влияния на общую положительную характеристику работы

Заключительная оценка

Диссертационная работа Ирхина Ильи Александровича «Единственность матричного разложения и сходимость регуляризованных алгоритмов в вероятностном тематическом моделировании», выполненная под руководством д.ф.-м.н., профессора Воронцова К.В., является законченной научно-квалификационной работой, содержащей новые научные результаты в области семантического анализа текстов. Все результаты, выносимые на защиту, обоснованы и подтверждены в ходе вычислительных экспериментов.

Работа удовлетворяет всем требованиям, предъявляемым ВАК к диссертациям, представленным на соискание учёной степени кандидата физико-математических наук, а её автор, Ирхин И.А. заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 05.13.17 – «Теоретические основы информатики».

Работа была заслушана на научном семинаре ИСП РАН под руководством академика РАН А.И. Аветисяна от 30.10.2020. Настоящий отзыв обсуждался и был одобрен на заседании отдела информационных систем Федерального государственного бюджетного учреждения науки Института системного программирования им. В.П. Иванникова РАН 16.11.2020 г. протокол №3.

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