

**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ**

федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский университет ИТМО»
(Университет ИТМО)

Кронверкский пр-т, д. 49, лит. А,
Санкт-Петербург, Россия, 197101
Тел.: (812) 232-97-04 | Факс: (812) 232-23-07
od@itmo.ru | itmo.ru

ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА
кандидата физико-математических наук
Фильченкова Андрея Александровича
на диссертационную работу Ирхина Ильи Александровича
«Единственность матричного разложения и сходимостъ регуляризованных алгоритмов в
вероятностном тематическом моделировании»,
представленную на соискание ученой степени кандидата физико-математических наук
по специальности 05.13.17 – «Теоретические основы информатики»

Актуальность работы

Представленная работа рассматривает задачу тематического моделирования, относящуюся к специальной области машинного обучения и заключающуюся в построении вероятностных моделей порождения текстов по неразмеченной коллекции документов. Одним из способов построения тематических моделей является подход аддитивной регуляризации тематических моделей. Он состоит в сведении задачи к построению тематической модели, минимизирующей функционал правдоподобия с дополнительными ограничениями в виде регуляризаторов, с помощью EM-подобного алгоритма. Однако введение дополнительных регуляризаторов затрудняет теоретическое исследование алгоритма их построения и особенностей его работы, в частности, об условиях сходимости и единственности получаемого решения.

Диссертация Ирхина И.А. посвящена исследованию теоретических свойств указанного алгоритма построения, и призвана расширить понимание его свойств и особенностей работы. В частности, в работе доказываются теоремы о сходимости алгоритма и единственности получаемого в ходе его работы матричного разложения. Помимо этого в диссертации предлагаются различные теоретически обоснованные модификации алгоритма, позволяющие достичь более высокого качества получаемых тематических моделей.

Поскольку тематическое моделирование предлагает актуальный инструментарий для решения различных задач в обработке естественного языка и информационном поиске, а также в ряде других областей, включая рекомендательные системы и анализ речи, а сама проблемы, поднимаемые в диссертационной работе, являются значимыми для теоретических и алгоритмических основ тематического моделирования, актуальность работы не вызывает сомнений.

Содержание работы

В диссертации можно выделить следующие части: введение, пять глав и заключение.

Во введении обоснована актуальность работы, определены цели и задачи исследования.

В главе 1 описывается задача тематического моделирования и подход аддитивной регуляризации тематических моделей, формулируется итерационный алгоритм, используемый для получения тематических моделей.

В главе 2 рассматривается проблема сходимости данного итерационного алгоритма. Доказывается теорема о достаточных условиях сходимости и выполнение предложенных условий проверяется в эксперименте на текстовой коллекции. Доказывается теорема об изменении значения регуляризатора на итерации алгоритма, которая используется в предлагаемой далее модификации алгоритма. В эксперименте сравниваются новая и исходная версия алгоритма и делается вывод об улучшении от использования модификации.

В главе 3 рассматривается проблема единственности матричного разложения, получаемого в качестве результата работы алгоритма. Доказывается теорема о достаточных условиях единственности стохастического матричного разложения. Приводится экспериментальное подтверждение выполнения предложенных условий на текстовой коллекции. Делается вывод о причинах неединственности решения в задачах тематического моделирования.

В главе 4 рассматривается задача повышения разреженности тематических моделей. Доказываются теоремы об изменении правдоподобия при занулении параметров модели и об аппроксимации этих изменений. Предлагается альтернативный способ разреживания тематических моделей в подходе аддитивной регуляризации. В эксперименте сравниваются предложенный алгоритм разреживания и стандартный способ с использованием регуляризатора разреживания. Заключается, что предложенный способ позволяет получить тот же уровень разреженности модели при большем значении правдоподобия.

В главе 5 рассматривается возможность изменения оптимизационной задачи тематического моделирования с целью повышения интерпретируемости получаемых моделей. Формулируется новая оптимизационная задача, в которой разные параметры тематической модели находятся в функциональной зависимости. Выводится итерационный алгоритм для оптимизации и доказывается его сходимость. В эксперименте демонстрируется увеличение согласованности и интерпретируемости получаемых тем.

В заключении делается вывод о том, что задачи исследования выполнены и цель достигнута.

Основные результаты работы

В качестве основных результатов работы выносятся следующие:

- Теорема о достаточных условиях сходимости алгоритма аддитивной регуляризации тематических моделей.
- Теорема о достаточных условиях единственности стохастического матричного разложения.
- Модификация алгоритма аддитивной регуляризации тематических моделей, ускоряющая сходимость итерационного процесса.
- Метод разреживания тематической модели.

Достоверность полученных результатов

Научные положения, выводы и рекомендации, сформулированные в диссертации, являются обоснованными, они сформулированы в виде математических утверждений и подкреплены математическими доказательствами. Также достоверность подтверждается публикацией результатов в научных журналах и докладами на научных конференциях и семинарах.

Новизна и научная значимость работы

Теоретическая значимость работы заключается в доказательстве теорем о сходимости итерационного алгоритма аддитивной регуляризации тематических моделей, доказательстве теоремы о достаточных условиях единственности стохастического матричного разложения. Предложенные доказательства являются новыми.

Практическая значимость заключается в применении предложенных модификаций на текстовых коллекциях и улучшении метрик качества получаемых тематических моделей. Часть предложенных модификаций реализована в библиотеке с открытым исходным кодом TopicNet.

Замечания

1. Предложенные достаточные условия единственности матричного разложения в главе 2 являются общими и не используют того, что разложение матрицы рассматривается в точке сходимости итерационного алгоритма. Стоит учесть при дальнейшем исследовании условий единственности.
2. В главе 4 в экспериментальном исследовании сравнение проводится только с разреженным LDA (sparse LDA) и автор не обосновывает, почему не рассматриваются другие базовые алгоритмы построения тематических моделей.
3. В главе 5 при сравнении разных алгоритмов на текстовых коллекциях и при сравнении реализации в TopicNet с другими алгоритмами используются разные наборы метрик качества тематических моделей, что затрудняет их сравнение.
4. Публикации по теме диссертационной работы выполнены только на русском языке, что ограничивает доступность результатов для мировой науки. Одновременно с этим в работе эксперименты проводятся только на англоязычных корпусах, исследование поведения предлагаемых алгоритмов для русскоязычных текстов вызывает дополнительный интерес.
5. В работе содержатся ряд орфографических и грамматических ошибок, а также ошибок в оформлении. В частности, в формуле 2.2 допущена опечатка: в определении KL-дивергенции пропущен логарифм, однако, дальнейший вывод формул является верным.

Заключительная оценка

В диссертационной работе Ирхина И.А. представлено теоретическое обоснование свойств алгоритма аддитивной регуляризации тематических моделей. Предложенные теоремы и методы имеют важное как теоретическое, так и практическое значение. Работа выполнена на хорошем математическом уровне, текст последователен и структурирован. Доказанные математические результаты дополняются экспериментальным исследованием предложенными на их основе алгоритмами.

Автореферат соответствует диссертации, отражает её содержание и даёт представление об актуальности темы, целях, задачах, методах исследования, научной новизне, теоретической и практической ценности, кратком содержании и основных результатах работы.

Диссертационная работа Ирхина И.А. «Единственность матричного разложения и сходимость регуляризованных алгоритмов в вероятностном тематическом моделировании» является законченной самостоятельной научно-исследовательской работой, посвященной актуальной проблеме и содержащей новые научные результаты.

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».

07.12.2020

Официальный оппонент
Фильченков Андрей Александрович
кандидат физико-математических наук,
доцент факультета ИТиП
afilchenkov@itmo.ru



Фильченков А.А.
А.А. Фильченков