ОТЗЫВ

о диссертации Булатова Виктор Геннадьевича
«Методы оценивания качества и многокритериальной оптимизации тематических моделей в библиотеке TopicNet», представленной на соискание ученой степени кандидата технических наук по специальности 05.13.18 — Математическое моделирование, численные методы и комплексы программ.

Автор отзыва

ФИО: Зухба Анастасия Викторовна,
Учёная степень: кандидат физико-математических наук
Год присуждения ученой степени и научная специальность, по которой присуждена ученая степень: 2018 г., по специальности 01.01.09 — «Дискретная математика и математическая кибернетика»
Ученое звание: нет
Место работы: Московский физико-технический институт (национальный исследовательский университет), кафедра математических основ управления.
Должность: доцент
Контактная информация: e-mail: azukhba@gmail.com; a__l@mail.ru
Тел. : +7 9851993132

Актуальность работы

Тематическое моделирование (ТМ) находит применение в анализе текстовых и транзакционных данных, информационном поиске, рекомендательных системах и во многих других прикладных задачах интеллектуального анализа данных. Методы тематического моделирования используются для анализа текстов объявлений, энциклопедических статей, потоков новостей, форумов в социальных сетях, текстов песен и исходных кодов программ. Математически аппарат тематического моделирования опирается на теорию матричных разложений и теорию регуляризации.

Диссертационное исследование В.Г.Булатова выполнено в рамках подхода ARTM – аддитивной регуляризации тематических моделей. Это многокритериальный подход, позволяющий учитывать различные дополнительные требования при построении моделей. Некоторые вопросы практического применения ARTM до сих пор оставались открытыми. Прежде всего, это определение «стратегии регуляризации» – в какой последовательности применять регуляризаторы, как определять их весовые коэффициенты, а также число тем и другие гиперпараметры, и по каким критериям контролировать качество модели.

Указанный комплекс вопросов определяет актуальность темы исследования, а поиск практических ответов на них является целью работы.

Содержание работы

Построение работы традиционно. Диссертация состоит из введения, шести глав, заключения и списка научной литературы. Место решаемых задач в общей картине области тематического моделирования отражается в первой и последней главе.

Автор диссертации во введении обосновывает актуальность и новизну, формулирует цели и задачи, перечисляет выносимые на защиту положения, определяет теоретическую и практическую значимость осуществлённого исследования.

В первой главе автором обстоятельно представлен широкий спектр применений ТМ и комплекс проблем, связанных с его использованием в гуманитарных науках и препятствиями на пути их более широкого внедрения.

Вторая глава также носит обзорный характер. Диссертант рассматривает известные подходы к формализации понятия интерпретируемости тематических моделей и даёт обстоятельный обзор современной литературы по эмпирическому оцениванию качества тематических моделей.

В третьей главе критикуется распространённый критерий интерпретируемости ТМ, называемый когерентностью, и приводятся убедительные экспериментальные обоснования недостаточности этого критерия. В.Г.Булатовым предлагается новый критерий – внутритекстовая когерентность, обеспечивающий более полное покрытие текстового материала коллекции и отличающийся тем, что для его вычисления данные о частотах накапливаются по всем парам термов, а не только по «верхним» (наиболее частотным) термам тем.

В четвёртой главе описывается техника относительных коэффициентов регуляризации и предлагается новый регуляризатор, позволяющий строить тематические векторные представления документов максимально быстро, за один проход по всем словам документа. Приводятся результаты экспериментов, показывающие, что предложенный регуляризатор улучшает качество модели по критериям разреженности, различности тем и когерентности.

Пятая глава является центральной в данном диссертационном исследовании. В ней описывается новая библиотека тематического моделирования TopicNet, в разработке которой диссертант принимал непосредственное участие. Им была разработана архитектура библиотеки и основные инструментальные средства автоматизации экспериментов по построению аддитивно регуляризованных тематических моделей. Предложены концепции «кубов гиперпараметров», «дерева экспериментов» и «рецепта моделирования». Предложен универсальный рецепт моделирования, позволяющий строить тематические модели, превосходящие по качеству обычно используемые на практике модели LDA.

В шестой главе продемонстрировано практическое применение TopicNet при решении достаточно трудной прикладной задачи – кластеризации и классификации обращений клиентов в контактный центр для дальнейшей маршрутизации и обработки.

Обоснованность основных научных положений диссертационной работы подкреплена экспериментами на коллекциях текстовых документов. Разработанный программный комплекс предоставлен в открытый доступ на GitHub. Исходные коды экспериментов и наборы данных также предоставлены в открытый доступ, что обеспечивает воспроизводимость экспериментальных результатов работы. Диссертантом опубликованы три статьи в рецензируемых изданиях, две из которых проиндексированы Scopus. Имеются три свидетельства о регистрации программ для

ЭВМ. Публикации автора и автореферат достаточно полно отражают содержание диссертационной работы.

Диссертация соответствует паспорту специальности 05.13.18 — Математическое моделирование, численные методы и комплексы программ.

К диссертации имеются следующие замечания.

1. В обзоре недостаточно подробно описываются некоторые метрики качества, не проводится анализ их недостатков, не объясняются причины, по которым одни метрики были выбраны для реализации в TopicNet, а другие нет.

2. Стиль изложения недостаточно академичен. Изложение изобилует терминами, не переведёнными с английского языка (хотя их вполне можно было бы перевести) и «лабораторным жаргоном», что не способствует пониманию отдельных фрагментов работы.

Указанные недостатки не затрагивают сущности проведённого диссертантом исследования и не снижают теоретической и практической значимости результатов.

Заключение по работе

В диссертационной работе Булатова В. Г. получены новые научные результаты, которые изложены аргументировано, подтверждены экспериментами на реальных данных, в том числе путём сравнения предлагаемых решений с существующими альтернативами. Диссертант продемонстрировал высокую квалификацию и способность к самостоятельной научной работе. Основные результаты диссертации опубликованы своевременно и полно. В автореферате выполненные исследования и полученные результаты описаны достаточно полно. Диссертационная работа Булатова В. Г. вносит значительный вклад в практическое использование аддитивно регуляризованных тематических моделей для решения прикладных задач текстовой аналитики.

Диссертационная работа Булатова В. Г. «Методы оценивания качества и многокритериальной оптимизации тематических моделей в библиотеке TopicNet» является законченной научно-квалификационной работой и удовлетворяет критериям «Положения о присуждении ученых степеней кандидата наук, доктора наук в МФТИ», утвержденного приказом ректора МФТИ от 27.03.2020 г. № 600-1, предъявляемым к диссертациям на соискание ученой степени кандидата наук по специальности 05.13.18 — «Математическое моделирование, численные методы и комплексы программ». Автор работы Булатова Виктор Геннадьевича заслуживает присвоения ему учёной степени кандидата технических наук по данной специальности.

12 октября

/ А.В.Зухба

ПОДПИСЬ РУКИ
ЗАВЕРЯЮ:
ЗАВЕДУЮЩАЯ КАНЦЕЛЯРИЕЙ
АДМИНИСТРАТИВНОГО ОТДЕЛА
М.А.ГУСЕВА