«Утверждаю»
Директор Федерального
государственного бюджетного
учреждения науки
Институт системного
программирования
им. В.П. Иванникова РАН
академик РАН, д.ф.-м.н.
_____ А.И. Аветисян
«30» ноября 2020

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ
Федерального государственного бюджетного учреждения науки «Институт системного программирования им. В.П. Иванникова»
Российской академии наук
на диссертационную работу Булатова Виктора Геннадьевича
«Методы оценивания качества и многокритериальной оптимизации тематических моделей в библиотеке TopicNet», представленную на соискание ученой степени кандидата технических наук по специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ»

Актуальность темы диссертации

Вероятностное тематическое моделирование – это одно из направлений анализа текстовой информации. Тематическая модель выявляет тематику коллекции текстовых документов, описывая каждый документ посредством составляющих его тем, и связывая каждую тему с набором характерных для неё слов. Полученные таким образом тематические кластеры могут быть полезны при решении широкого спектра прикладных задач, в том числе для визуализации и интерпретации данных.

Построение полезных тематических моделей требует настройки гиперпараметров. Особенно актуальным этот вопрос является для аддитивно регуляризованных тематических моделей (ARTM), чей математический аппарат позволяет строить модели, удовлетворяющие разнообразным требованиям, формализуемым в виде регуляризаторов. Обратной стороной гибкости этого подхода является необходимость определения нужных регуляризаторов и подбора их параметров --- задача, также являющаяся частным случаем задачи настройки гиперпараметров.

Данная работа посвящена разработке методологии, облегчающей настройку гиперпараметров тематических моделей.

Содержание работы

Диссертационная работа состоит из введения, шести 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.

Первая глава определяет основные термины, формулирует задачу вероятностного тематического моделирования, излагает важные для дальнейшего изложения приложения тематических моделей. Описываются нерешённые проблемы области тематического моделирования и обосновывается, почему они представляются важными.

Во второй главе приведен обзор предложенных в научной литературе мер качества тематических моделей. Автором предложена категоризация различных подходов к оценке качества, что позволило сгруппировать их в несколько рубрик. Приведены аргументы в пользу того, что хорошая тематическая модель должна удовлетворять многим требованиям сразу, т.е. задача оптимизации является многокритериальной.

В третьей главе приводятся экспериментальные обоснования, почему известные способы измерения интерпретируемости, основанные на когерентности, недостаточно информативны. Предлагается альтернативный критерий внутритекстовой когерентности и метод измерения его чувствительности.

В четвертой главе рассматриваются два приёма, способствующих увеличению интерпретируемости и переносимости стратегий тематического моделирования, применимых в широком диапазоне ситуаций. Во-первых, рассматриваются относительные коэффициенты регуляризации, позволяющие использовать одни и те же численные значения параметров для анализа различных коллекций с предсказуемым результатом. Во-вторых, изучается влияние на модель введения явной зависимости одной из двух матриц модели от другой; показано, что использование данного псевдорегуляризатора, улучшает качество модели (разреженность, различность, когерентность). Приведено эвристическое объяснение данного эффекта.

В пятой главе вводится ряд понятий, формализующих принятые практики тематического моделирования. На основе этого терминологического аппарата разрабатывается архитектура библиотеки TopicNet, обеспечивающая планирование и журналирование экспериментов, а также делающая более удобной процедуру многокритериального отбора моделей.

В шестой главе описывается применение библиотеки TopicNet (в частности, относительных весов модальностей) в прикладной задаче построения тематической иерархии над обращениями пользователей в контакт-центр.

Основные результаты и их новизна

В рамках диссертационной работы В.Г. Булатова получены следующие новые результаты:

1) Впервые разработана методология построения аддитивно регуляризованных тематических моделей, позволяющая организовать автоматизированный подбор гиперпараметров по множеству критериев.

2) Выстроена архитектура библиотеки TopicNet, обеспечивающая программную реализацию данной методологии, использующая относительные коэффициенты регуляризации, позволяющая описывать стратегии обучения моделей посредством удобного язык и поддерживающая возможность создания пользовательских регуляризаторов и метрик качества на языке Python.

3) Предложен универсальный "рецепт моделирования", производящий многокритериальный выбор тематических моделей для широкого класса разнородных задач тематического моделирования.

4) Эмпирически изучено, в какой степени традиционные меры когерентности используют данные о сочетаемости слов внутри текстовых документов. Предложен ряд новых критериев когерентности, более полно использующих эту информацию.

Теоретическая и практическая значимость

На текущий момент полноценных инструментов для настройки гиперпараметров ТМ не существует. Часть имеющихся решений не являются открытыми, часть поддерживают только ограниченное число моделей и/или критериев качества. При этом ряд авторов отмечает важность создания инструментария, позволяющего прозрачным образом подстраивать модель под нужды пользователя и доступного для неспециалистов.

В данной диссертационной работе представлена открытая библиотека TopicNet, предназначенная облегчить настройку гиперпараметров тематических моделей. Поддержка пользовательских регуляризаторов, мер качества, стратегий поиска и методов визуализации обеспечивает возможность дальнейшего расширения библиотеки. Результаты работы В.Г. Булатова дополняют библиотеку BigARTM, которая используется различными компаниями и научными группами в проектах по обработке естественного языка.

Достоверность результатов

Достоверность полученных результатов обеспечивается вычислительными экспериментами на реальных текстовых коллекциях, а также апробацией на научных конференциях и в открытой печати. Разработанный код библиотеки TopicNet и проведённых экспериментов находится в открытом доступе, что обеспечивает воспроизводимость результатов. Достоверность и значимость работы также подтверждается рядом свидетельств о государственной регистрации программы для ЭВМ.

Замечания

Можно отметить ряд недостатков данной работы.

1) Наблюдается некоторая непоследовательность изложения эмпирических данных. Часть результатов представлена на графиках, содержащих доверительные интервалы (но без численных характеристик этих доверительных интервалов). Часть результатов представлена посредством таблиц без указания выборочного стандартного отклонения, притом, что для ряда задач эта вариабельность может быть существенной.

2) В главе 5 постулируется многостадийная схема экспериментов. Естественность этой схемы следовало бы обосновать более тщательно; в частности, в обзоре имеющихся публикаций не описываются работы, в которых применяется схожий принцип обучения тематических моделей, хотя и даётся ссылка на необходимые публикации

3) В главе 5 описывается новый куб, обеспечивающий гибкую настройку коэффициентов регуляризации во время обучающих итерации. В этот раздел желательно бы было

включить более подробное обсуждение вопроса о том, для каких проблем практического характера применим данный подход.

Указанные замечания не снижают общей положительной оценки диссертационной работы.

Заключительная оценка

Диссертационная работа Булатова Виктора Геннадьевича «Методы оценивания качества и многокритериальной оптимизации тематических моделей в библиотеке TopicNet», выполненная под руководством д.ф.-м.н., профессора К.В. Воронцова, является законченной научно-квалификационной работой, содержащей новые научные результаты, которые изложены аргументировано, подтверждены экспериментами на реальных данных, в том числе путём сравнения предлагаемых решений с существующими альтернативами.

Результаты работы опубликованы в рецензируемых научных изданиях, индексируемых базой SCOPUS, в том числе в трудах конференции LREC, входящей в число ведущих международных конференций по компьютерной лингвистике.

Диссертация соответствует всем критериям, установленным Положением о присуждении ученых степеней, предъявляемым к диссертациям на соискание ученой степени кандидата технических наук по специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ», а ее автор – Булатов Виктор Геннадьевич 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 16.11.2020 г. протокол № 3.

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

Д.Ю. Турдаков