На правах рукописи

Булатов Виктор Геннадьевич

Методы оценивания качества и многокритериальной оптимизации тематических моделей в библиотеке TopicNet

Специальность 05.13.18 —
«Математическое моделирование, численные методы и комплексы программ»

Автореферат
диссертации на соискание учёной степени
кандидата технических наук

Москва — 2020

Работа прошла апробацию на кафедре «Анализ данных» Федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)».

Научный руководитель: доктор физико-математических наук, профессор РАН, руководитель лаборатории машинного интеллекта МФТИ.
Воронцов Константин Вячеславович

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт системного программирования им. В.П. Иванникова Российской академии наук.

Защита состоится 28 декабря 2020 г. в 16:00 на заседании диссертационного совета ФПМИ.05.13.18.013 по адресу: 141701, Московская область, г. Долгопрудный, Институтский переулок, д. 9..

С диссертацией можно ознакомиться в библиотеке и на сайте Московского физико-технического института (национального исследовательского университета) https://mipt.ru/education/post-graduate/soiskateli-tekhnicheskie-nauki.php.

Работа представлена «15» октября 2020 г. в Аттестационную комиссию федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)» для рассмотрения советом по защите диссертаций на соискание ученой степени кандидата наук, доктора наук в соответствии с п.3.1 ст. 4 Федерального закона «О науке и государственной научно-технической политике».

Общая характеристика работы

Актуальность темы исследования. Тематическое моделирование — это обширное направление исследований в области автоматической обработки текстов. Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ и из каких слов состоит каждая тема. В отличие от обычных методов кластеризации, тематическая модель может относить документ не к одному кластеру-теме, а к нескольким, то есть она производит «мягкую кластеризацию», причём не только документов, но и слов.

Обычно тематические модели относят к методам машинного обучения «без учителя», поскольку они не требуют размеченных обучающих выборок. Это позволяет использовать тематическое моделирование в тех случаях, когда никаких дополнительных данных, кроме собственно текстовой коллекции, не имеется, например, для информационного поиска в больших текстовых массивах, для анализа специализированных текстов или текстов на редких языках, для анализа больших массивов текстоподобных данных, таких как программный код, тексты песен, банковские транзакции, географические данные, музыкальные произведения.

Результатом вероятностного тематического моделирования является конечное множество *тем*, каждая из которых описывается вероятностным распределением на множестве слов. Важным свойством темы является её интерпретируемость. Слова, имеющие большую вероятность в данной теме, должны относиться к одной предметной области и быть семантически связанными. Тема считается интерпретируемой, если, рассматривая наиболее частотные слова темы, эксперт может сказать, о чём эта тема, и дать ей определённое название [1]. Если все темы (или почти все) интерпретируемые, то о такой модели говорят, что она в целом является интерпретируемой. В таком случае модель может быть полезна для понимания тематической структуры коллекции. Интерпретируемость является трудно формализуемой характеристикой. Существуют различные экспертные и вычислительные методики её количественного оценивания [2].

Развитие вероятностного тематического моделирования началось с работы Т.Хофманна [3], в которой была предложена модель вероятностного латентного семантического анализа (Probabilistic Latent Semantic Analysis, PLSA). Построение тематической модели является некорректно поставленной задачей стохастического матричного разложения, которая имеет бесконечное множество решений. Для доопределения постановки задачи и выбора наиболее подходящего решения необходимо вводить дополнительные ограничения на модель. Следующей важной вехой стала модель латентного размещения Дирихле (Latent Dirichlet Allocation, LDA) [4], основанная на байесовской регуляризации искомых дискретных распределений с помощью априорных распределений Дирихле. В последующие

годы на основе PLSA и LDA были разработаны сотни специализированных моделей, отличающихся способами регуляризации, структурой исходных данных и матричного разложения [5—7].

Аддитивная регуляризация тематических моделей (ARTM) позволяет комбинировать регуляризаторы для создания моделей с заданными свойствами [8; 9]. Это многокритериальный подход, основанный на оптимизации взвешенной суммы основного критерия (логарифма правдоподобия) и некоторого количества дополнительных критериев-регуляризаторов. Многокритериальный подход является ответом на практическую потребность строить модели, обладающие целым рядом необходимых свойств одновременно [10]. Соответственно, в каждой конкретной задаче тематического моделирования может быть много не только оптимизационных критериев, но и метрик качества, с помощью которых валидируется (оценивается) построенная модель. В частности, в [9] было показано, что комбинирование регуляризаторов сглаживания, разреживания, декоррелирования и отбора тем позволяет одновременно улучшить несколько метрик качества (разреженность, различность и интерпретируемость тем) без заметного ухудшения основного критерия правдоподобия (или перплексии) модели. Тематические модели для разведочного информационного поиска, в дополнение к этим свойствам, должны быть также мультимодальными (учитывать не только слова, но и биграммы, теги, категории, авторство документов) и иерархическими (разделять крупные темы на более мелкие подтемы) [11]. Построение таких моделей требует не только выбора множества модальностей и регуляризаторов, но и подбора различных гиперпараметров.

На основе теории аддитивной регуляризации тематических моделей была разработана библиотека тематического моделирования с открытым кодом BigARTM [12; 13]. Свойство аддитивности регуляризаторов позволило реализовать в BigARTM модульный подход, когда пользователь выбирает из библиотеки нужный ему набор регуляризаторов для построения модели с требуемыми свойствами. Возможность конструирования новых композитных моделей из «готовых блоков» существенно отличает BigARTM от других средств тематического моделирования, основанных на теории байесовского обучения, в которой каждая новая модель требует проведения уникальных математических выкладок (байесовского вывода) и, как следствие, разработки нового программного кода.

Несмотря на модульность, гибкость, масштабируемость и высокую производительность [10], практическое применение библиотеки BigARTM наталкивается на ряд трудностей. Пользователь должен хорошо разбираться в теории ARTM, чтобы грамотно выбрать стратегию регуляризации, то есть последовательность включения регуляризаторов, затем подобрать число тем, коэффициенты регуляризации для каждого регуляризатора и

другие гиперпараметры. Эта работа связана с проведением серий вычислительных экспериментов, которые требуют тщательного планирования, журнализации, валидации, визуализации и критического осмысления промежуточных результатов. Таким образом, BigARTM перекладывает на пользователя значительный объём работы, требующей пристального внимания и высокой квалификации.

Актуальной задачей является создание технических средств для автоматизации экспериментов по построению аддитивно регуляризованных тематических моделей, их валидации и выбора лучшей модели по заданной совокупности метрик качества.

Степень разработанности темы исследования. Тематическое моделирование является полезным инструментом в цифровых гуманитарных исследованиях (digital humanities) [14; 15]. Однако процессы построения, валидации и выбора тематических моделей практически не алгоритмизированы [15]. Большое прикладное значение имеет разработка новых методов визуализации, автоматической валидации и выбора моделей [16].

Также исследователи сталкиваются с неустойчивостью [17] и неинтерпретируемостью тем [18]. Известно, что настройка гиперпараметров модели может повысить её устойчивость [19], однако на практике она выполняется редко, и для неё нет единой принятой методологии [19; 20].

Настройка гиперпараметров может помочь и с интерпретируемостью тем, особенно в рамках подхода ARTM. Известные регуляризаторы, такие как регуляризатор декоррелирования, способствуют повышению различности и интерпретируемости тем [1]; также можно использовать регуляризатор, напрямую оптимизирующий заданный критерий, как это было сделано в [21] для критерия средней когерентности тем.

К сожалению, интерпретируемости трудно дать формальное определение. Попытки определить плохие темы через расстояние до известных «мусорных» тем или через низкие значения когерентности дают лишь срез проблематичных тем и имеют ограниченную область применимости [18]. Системный подход к измерению интерпретируемости предполагает оценивание каждой темы по нескольким критериям качества [22].

Таким образом, принятые методологии перекладывают ответственность за подбор гиперпараметров на исследователя; при этом процедура подбора остаётся нерегламентированной, что создаёт высокий барьер входа для неспециалистов. В обзорной монографии [7] подчёркивается важность снижения порога входа и более жёсткой регламентации процесса моделирования: «первоочередная исследовательская задача в тематическом моделировании... сделать его более доступным».

Мы видим, что важными нерешёнными проблемами являются: обеспечение интерпретируемости моделей, подбор гиперпараметров и стандартизация процесса построения тематической модели для широкого класса пользовательских прикладных задач.

Целью данного диссертационного исследования является разработка и реализация технологии построения интерпретируемых аддитивно регуляризованных тематических моделей, применимых для решения широкого класса задач тематического моделирования.

Для достижения поставленной цели решаются следующие **задачи**.

1. Реализация, эмпирическое исследование и улучшение автоматически вычисляемых критериев интерпретируемости тематических моделей, в том числе нового критерия внутритекстовой когерентности.
2. Разработка методологии и средств автоматизации проведения экспериментов по подбору стратегии регуляризации и выбору гиперпараметров тематической модели.
3. Проектирование архитектуры библиотеки TopicNet с открытым кодом на GitHub для реализации данной методологии. Разработка и реализация интерфейсов, обеспечивающих создание пользовательских регуляризаторов и метрик качества в TopicNet.
4. Поиск универсального «рецепта» построения аддитивно регуляризованных тематических моделей, превосходящих LDA по совокупности критериев качества, применение которого не требовало бы от пользователя знания теории ARTM.
5. Решение прикладных задач с использованием разработанной библиотеки TopicNet, в частности, задачи кластеризации интентов в текстовой коллекции обращений клиентов в контактный центр.

Научная новизна. Предложена новая методология многокритериального выбора моделей на основе концепций «дерева экспериментов», «кубов гиперпараметров» и «рецептов моделирования» в рамках теории аддитивной регуляризации тематических моделей (ARTM). Разработан универсальный «рецепт» построения аддитивно регуляризованных тематических моделей, превосходящих LDA по совокупности критериев качества. Предложен новый способ построения иерархических тематических моделей с разными весами модальностей на разных уровнях иерархии.

Теоретическая значимость. Работа вносит вклад в развитие теории аддитивной регуляризации тематических моделей (ARTM), предоставляя исследователям удобную инструментальную среду, позволяющую накопить эмпирический материал для изучения стратегий регуляризации и их влияния на качество тематических моделей при многокритериальном оценивании. Вводятся понятия внутритекстовой когерентности, относительных и абсолютных коэффициентов регуляризации, фактора

балансировки, дерева экспериментов, куба гиперпараметров, рецепта моделирования.

Практическая значимость. Предложенные подходы и методы реализованы в библиотеке тематического моделирования с открытым кодом TopicNet, которая может быть использована и уже используется для решения различных прикладных задач анализа текстовых и транзакционных данных. Реализованные в библиотеке концепции дерева экспериментов, куба гиперпараметров и рецепта моделирования позволяют находить, сохранять и распространять в сообществе исследователей удачные приёмы решения прикладных задач тематического моделирования.

Показано, что использование относительных коэффициентов регуляризации обеспечивает возможность переноса стратегии обучения тематической модели на другие текстовые коллекции: один и тот же набор значений относительных коэффициентов регуляризации и/или весов модальностей может быть использован для различных прикладных задач. В случае, когда непосредственный перенос численных значений нецелесообразен из-за специфики новой коллекции, относительные коэффициенты облегчают подбор оптимальных значений, поскольку они находятся в диапазоне $[0, 1]$ и интерпретируются как степень воздействия регуляризатора на модель в сравнении с основным критерием логарифмированного правдоподобия.

Предложенные в данной работе и реализованные в TopicNet механизмы были успешно применены для решения ряда прикладных задач: для кластеризации интентов в текстовой коллекции обращений клиентов в контактный центр [1], для анализа банковских транзакционных данных [23] и других.

Методология и методы исследования. В работе использованы методы теории вероятностей, численной оптимизации, автоматической обработки текстов, машинного обучения, вероятностного тематического моделирования. Экспериментальное исследование проводится на языке Python; опубликованная на GitHub библиотека TopicNet, подытоживающая результаты исследования, открыта для свободного использования и удовлетворяет принципам воспроизводимости результатов.

Основные положения, выносимые на защиту:

1. Разработана методология построения аддитивно регуляризованных тематических моделей, обеспечивающая формирование «рецептов моделирования» с автоматизированным подбором гиперпараметров по множеству критериев и отличающаяся использованием относительных коэффициентов регуляризации и кубов гиперпараметров.
2. Выстроена архитектура библиотеки TopicNet, обеспечивающая программную реализацию данной методологии и отличающаяся использованием удобного языка описания кубов гиперпараметров

и возможностью создания пользовательских регуляризаторов и метрик качества на языке Python.
3. Создан универсальный рецепт моделирования, обеспечивающий многокритериальный выбор тематических моделей для широкого класса задач, отличающийся предварительной настройкой куба гиперпараметров по набору разнородных задач тематического моделирования.
4. Выполнена программная реализация нового критерия когерентности, обеспечивающая его эффективное вычисление и отличающаяся более полным использованием данных о сочетаемости слов внутри текстовых документов.

Достоверность полученных результатов обеспечивается вычислительными экспериментами на реальных текстовых коллекциях. Методика и результаты подробно описаны в тексте работы. Разработанный код библиотеки TopicNet и проведённых экспериментов находится в открытом доступе, что обеспечивает воспроизводимость результатов. Достоверность также подтверждается тремя свидетельствами о регистрации программы для ЭВМ (№2019661840, №2019662102 и №2020613851).

Апробация работы. Основные результаты диссертации докладывались на следующих конференциях и семинарах:

- Международная конференция по компьютерной лингвистике «Диалог», Москва, 1 июня 2018.
- International Conference Recent Advances in Natural Language Processing (RANLP), Варна, 3 сентября 2019.
- Открытая лекция в рамках образовательного проекта Физтех.Рост, Долгопрудный, 18 октября 2019.
- Открытый научный семинар «Методы анализа текстов», Москва, 28 марта 2018.
- Открытый научный семинар «Презентация TopicNet», Москва, 10 августа 2019.
- OpenTalks.AI — ведущая открытая конференция по искусственному интеллекту, Москва, 20 февраля 2020 года.
- International Conference on Language Resources and Evaluation (LREC), Марсель (должна была состояться в мае 2020).

Личный вклад. Личный вклад диссертанта в работы, выполненные с соавторами, заключается в следующем:

- В [2] предложен метод генерации полусинтетической выборки, проведены эксперименты по анализу репрезентативности высокочастотных слов в темах.
- В [1] выполнена реализация иерархической тематической модели средствами библиотеки TopicNet; предложен метод разделения

слов и n-грам по функциональному назначению; предложены методы анализа ошибок моделирования; выполнены эксперименты с относительными коэффициентами регуляризации.
- В [3] описана архитектура библиотеки TopicNet, концепция кубов гиперпараметров и дерева эксперимента, методы отбора моделей и связанный с ними специализированный язык описания кубов; проведена часть экспериментов, связанная с GenSim и с различностью тем.
- В [4] предложена адаптация псевдорегуляризатора для библиотеки TopicNet и выполнена его программная реализация; проведена связанная с этим часть экспериментов.

Публикации. Основные результаты по теме диссертации изложены в трёх рецензируемых публикациях, две из которых проиндексированы Scopus. Статья [4] принята к публикации в 2020 году (ВАК и Scopus). Также получены три свидетельства о государственной регистрации программы для ЭВМ (№2019661840, №2019662102 и №2020613851).

Объем и структура работы. Диссертация состоит из введения, двух обзорных глав, четырёх глав с результатами проведенного исследования, заключения, библиографии и приложения. Полный объём диссертации составляет 147 страниц, включая 19 рисунков и 14 таблиц. Список литературы содержит 159 наименований.

Содержание работы

Во **введении** отражается актуальность исследований, проводимых в рамках данной диссертационной работы, приводится обзор научной литературы по изучаемой проблеме, формулируется цель, ставятся задачи работы, определяется научная новизна, практическая и теоретическая значимости представляемой работы. Приводится список публикаций автора по теме диссертации и формулируются положения, вносимые на защиту.

Первая глава посвящена постановке задачи тематического моделирования. Рассмотрены основные подходы к тематическому моделированию и подробно описан математический аппарат подхода ARTM.

Задача тематического моделирования заключается в нахождении матрицы Φ, содержащей дискретные распределения вероятности слов для каждой темы, и матрицы Θ, содержащей дискретные распределения вероятности тем для каждого документа. Основное требование к модели заключается в том, что произведение этих двух матриц должно приближать распределения вероятности слов для каждого документа. Подход ARTM позволяет вводить различные дополнительные требования к матрицам Φ и Θ.

В рамках ARTM тематические модели строятся при помощи *EM-алгоритма*, в котором текущие значения матриц Φ и Θ итеративно

обновляются по определённым формулам. Введение каждого дополнительного регуляризатора приводит к появлению в этих формулах новых слагаемых, не меняющих общую структуру EM-алгоритма. Его программная реализация может быть выполнена один раз в самом общем виде, при этом каждый регуляризатор реализуется независимо от остальных в виде отдельного модуля.

Область применения, уникальная для тематических моделей — описание коллекции, дающее общее представление о тематической кластерной структуре больших объёмов данных. Зачастую исследователь ищет ответы на вопросы о структуре и природе коллекции, а функция тематической модели как модели языка, способной предсказывать слова в тексте, интересует его меньше.

Приложения тематического моделирования в различных областях наталкиваются на проблемы плохой интерпретируемости тем, дублирующих, мусорных и вводящих в заблуждение тем, неустойчивости результатов моделирования.

Большую часть этих проблем можно разрешить при помощи настройки гиперпараметров, однако в практико-ориентированной литературе не хватает хорошо проработанной систематической методологии настройки гиперпараметров. В большинстве работ ограничиваются подбором числа тем и ещё одного-двух гиперпараметров по грубым сеткам значений.

Сложность подбора гиперпараметров у многокритериальных моделей вызывает необходимость как измерения, так и оптимизации качества тематических моделей одновременно по множеству критериев.

Во второй главе рассматриваются критерии качества тематических моделей, используемые в литературе. Проведена категоризация распространённых подходов к измерению качества. Особое внимание уделяется мерам качества, связанными с анализом верхних (наиболее частотных, вероятных, «топовых», top-10) токенов в темах.

Третья глава посвящена анализу общих недостатков этих мер качества. Большинство используемых в литературе подходов к оценке интерпретируемости тем укладываются в следующую схему:

1. Для каждой темы выбирается какой-то небольшой набор характеризующих её токенов (как правило, это 10 верхних токенов).
2. Этот набор анализируется одним из двух способов:
   - качество тем оценивается экспертом визуально по этим токенам;
   - качество тем оценивается путём автоматического вычисления определённых статистик, в частности, парной сочетаемости верхних токенов в текстовой коллекции.

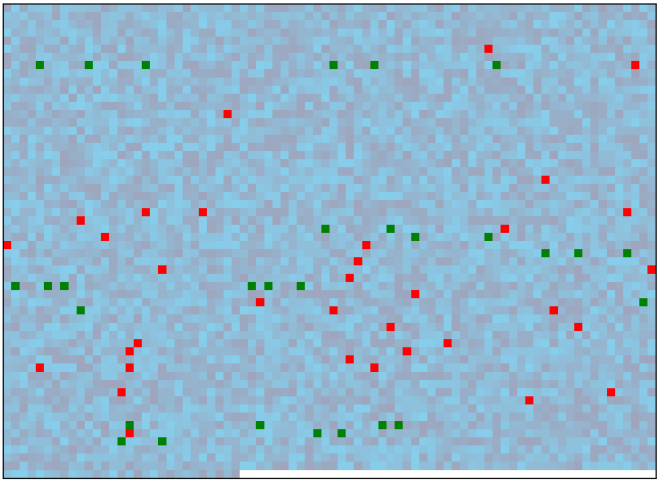

Рис. 1 — Демонстрация доли текста, покрытой верхними словами, на примере одного документа. Словопозиции обозначены серо-синим цветом, словопозиции верхних слов показаны красным цветом, зелёным цветом показаны словопозиции, имеющие ненулевой вклад в расчёт когерентности (т.е. попадающие в скользящее окно вместе с другим верхним словом).

В пункте (1) вышеописанной схемы тема фактически подменяется небольшим списком её верхних токенов. Каким бы образом ни проводился дальнейший анализ, обоснование качества тематической модели после такой подмены представляется проблематичным.

Рисунок 1 демонстрирует то, что верхние токены покрывают исчезающе малую часть коллекции, и ситуация ещё более усложняется наличием дополнительного требования парной сочетаемости токенов в окне. Численные расчёты подтверждают, что когерентность отдельно взятой темы, в большинстве случаев, учитывает менее тысячной доли всего корпуса текста.

Также в этой главе предлагаются несколько мер качества, основанных на идее *внутритекстовой когерентности*. Традиционные меры когерентности сначала выделяют небольшое множество верхних слов в заданной теме по их ϕ_{wt} и затем анализируют, каким образом эти слова встречаются в тексте (анализ *от темы к тексту*). В предлагаемом же методе сначала выделяются все соседние слова текста, распределение ϕ_{wt} которых затем анализируется (*от текста к теме*).

Предполагается, что этот метод будет лучше характеризовать интерпретируемость тем, нежели традиционные меры когерентности. Эксперимент на полусинтетической коллекции показывает, что предлагаемый подход действительно отличается большей чувствительностью.

В четвёртой главе рассматриваются способы увеличения интерпретируемости тематических моделей. Предложен метод подбора коэффициентов сглаживания, коэффициентов разреживания и весов дополнительных модальностей при помощи математического аппарата *относительных коэффициентов*.

Эта техника облегчает построение тематических моделей, все темы которых различны и не «загрязнены» большим числом неинформативных токенов (слов общей лексики). Также она позволяет перенести имеющуюся стратегию обучения тематической модели на другие текстовые коллекции схожей структуры.

Показано, что любому абсолютному коэффициенту сглаживания/разреживания Φ и Θ соответствует какой-то относительный коэффициент (и наоборот). Приводятся следующие формулы преобразований из λ в τ:

$$\tau = \frac{\lambda \sum_{d \in D} n_d}{(1-\lambda)|D| \cdot |T|} \tag{1}$$

$$\tau = \frac{n}{|T| \cdot |W|} \frac{\lambda}{(1-\lambda)} \tag{2}$$

Регуляризатор сглаживания Θ, воздействующий на темы T и документы D, коэффициент τ которого вычислен по формуле (1), можно проинтерпретировать как нахождение «компромисса»: результирующая Θ_{td} будет на λ состоять из априорного распределения $\frac{1}{|T|}$ и на $(1-\lambda)$ из оценки максимума правдоподобия $\frac{n_{td}}{n_d}$.

Аналогично интерпретируется и формула (2) для регуляризатора сглаживания Φ, воздействующего на темы T и токены W. Здесь задаётся взвешенная комбинация из априорного распределения $\frac{1}{|W|}$ и оценки максимума правдоподобия $\frac{n_{wt}}{n_t}$.

Это означает, что любую существующую модель можно переформулировать в терминах относительного сглаживания и разреживания без каких-либо потерь. Практическая значимость этого результата заключается в вытекающей из него возможности переносить существующую схему сглаживания/разреживания на другую коллекцию схожей структуры.

Оставшаяся часть главы посвящена изучению роли матриц Φ и Θ в тематическом моделировании. Вопрос рассматривается с двух позиций: с точки зрения математической постановки задачи и с точки зрения построения и использования тематических моделей на практике. В терминах вероятностного смысла задачи обе матрицы являются равноправными.

При этом существует ряд причин, по которым естественно считать матрицу Θ второстепенной по отношению к Φ. Особое значение для рассматриваемой работы имеет аргумент об интерпретируемости: описанный в предыдущей главе процесс оценки тем на интерпретируемость обычно состоит из выбора небольшого набора верхних слов для каждой темы и представления этого набора эксперту-человеку [24]. В этом процессе используется только матрица Φ. Таким образом, в контексте интерпретации найденных тем экспертами качество матрицы Φ оказывается более важным, чем качество матрицы Θ.

В работе [4] доказывается, что введение функциональной зависимости $\Theta = f(\Phi)$ требует модификации EM-алгоритма. Эту модификацию можно истолковать, как добавление псевдорегуляризатора, что приводит к алгоритму быстрой векторизации документов на основании одной лишь матрицы Φ. Именно в таком виде данный псевдорегуляризатор был интегрирован в открытую библиотеку TopicNet[1].

Оставшаяся часть четвёртой главы посвящена изучению свойств этого псевдоргеуляризатора. Поведение модели с предложенным псевдоргеуляризатором исследуется на реальной текстовой коллекции. Результаты показывают, что этот псевдорегуляризатор действительно повышает ряд критериев качества, связанных с интерпретируемостью и успешно комбинируется с другими регуляризаторами.

Пятая глава посвящена разработке библиотеке TopicNet. TopicNet — открытая надстройка над библиотекой BigARTM, предоставляющая более удобные возможности для подбора гиперпараметров, для работы с пользовательскими регуляризаторами и для визуализации тематических моделей. Описанная библиотека доступна онлайн на GitHub.

Главная мотивация TopicNet — создать инструмент, удобный как для новичков, так и для продвинутых пользователей. Большое внимание уделяется удобству работу и наличию «рецептов моделирования», показывающих хорошее качество без трудоёмкой настройки гиперпараметров. Численный эксперимент показывает, что TopicNet с настройками «по умолчанию» превосходит модель LDA из открыто доступной библиотеки GenSim, также с настройками «по умолчанию», по критериям различности, когерентности и информативности тем.

Библиотека TopicNet состоит из двух больших модулей: `Viewers` и `Cooking Machine`.

Модуль `Viewers` содержит различные инструменты визуализации. Дизайн придерживается философии Unix: каждый вьюер имеет ограниченную область ответственности и способен возвращать результат операции в JSON-подобном виде.

[1] https://github.com/machine-intelligence-laboratory/TopicNet/blob/master/topicnet/demos/Topic-Thetaless-Regularizer.ipynb

Это даёт возможность комбинировать содержащиеся в модуле вьюверы, не теряя при этом удобные для конечного пользователя методы, возвращающие `pandas.DataFrame`, строку сформированного HTML, или отображающие результат сразу в ячейке вывода Jupyter Notebook.

Модуль `Cooking Machine` содержит различные инструменты для моделирования, расположенные в иерархии основных классов. Эти классы отвечают за построение и обучение модели заданной структуры, за отбор моделей согласно заданным пользователем ограничениям, а также за сохранение, загрузку и журналирование происходящего процесса.

Процесс построения тематической модели представим в виде дерева. Каждый узел дерева содержит в себе тематическую модель, а ориентированные рёбра хранят информацию об отношениях «предок-потомок» вида «модель Y была получена из модели X при помощи преобразования T_{XY}». Не все деревья эксперимента являются допустимыми. Мы накладываем ограничения на допустимые преобразования: требуем, чтобы все рёбра одного уровня описывали преобразования из одного и того же семейства, различающиеся лишь набором параметров.

Одним из примеров таких преобразований является «Применить к модели регуляризатор с произвольными параметрами».

Класс `Experiment` отвечает за хранение, журналирование и актуализацию этой структуры.

Все преобразования связаны с экземпляром класса `Cube`. Каждый `Cube` играет роль чертежа, задающего все преобразования на текущем уровне эксперимента. Таким образом, процесс обучения можно представить как цепочку кубов, последовательно соединённых друг с другом.

Куб выполняет две важные функции. Первая — это *спецификация*: во время инициализации куб преобразует заданные пользователем параметры в многомерное пространство поиска. Вторая функция — *применение*: получив точку в пространстве поиска и тематическую модель, куб изменяет заданное множество параметров и/или гиперпараметров модели. Таким образом, он играет роль инкубатора для моделей, что отражено в названии класса. На Рис. 2 приведена схема обучения, состоящая из двух кубов, применённых к одной модели.

Классы `Experiment` и `Cube` позволяют сделать сложные стратегии обучения и журналирование экспериментов более сжатыми и доступными. Модуль `config_parser` делает ещё один шаг в сторону облегчения конфигурируемости: стратегию обучения можно задать при помощи текстового конфигурационного файла в формате YAML.

В реальных экспериментах не у каждой модели есть потомки; большинство моделей отбрасывается в соответствии с каким-то критерием. Самый естественный, но в то же время самый трудозатратный способ анализа тематической модели — это ручное изучение списков верхних токенов и верхних документов в темах, на основании которого пользователь

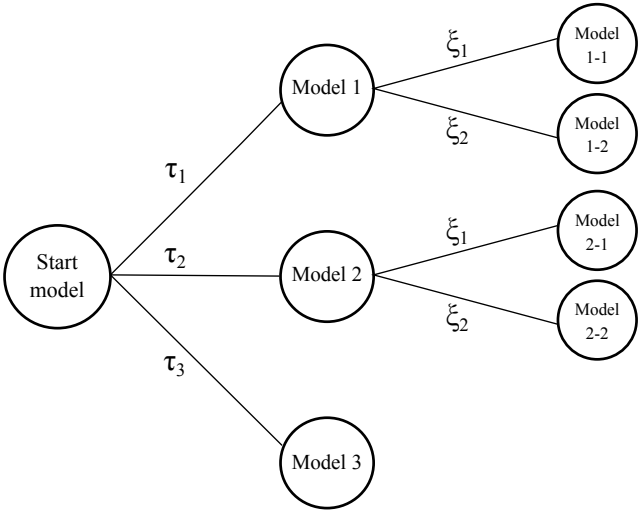

Рис. 2 — Пример двухэтапной схемы эксперимента. На первом этапе применяется регуляризатор с коэффициентом τ, принимающим значения из некоторого множества $\{\tau_1, \tau_2, \tau_3\}$. Лучшими моделями после первого этапа являются *Model 1* и *Model 2*, поэтому *Model 3* больше не участвует в процессе обучения. Второй этап связан с другим регуляризатором с коэффициентом ξ, принимающим значения из множества $\{\xi_1, \xi_2\}$. В результате этого этапа у каждой из ранее отобранных моделей появляется два потомка.

принимает решение, какие из моделей являются «удовлетворительными». Другой подход заключается в сравнении численных показателей; обычно используется перплексия и когерентность. Библиотека BigARTM добавляет к их числу дополнительные метрики разреженности, различности, чистоты и контрастности тем [9], и другие.

Библиотека TopicNet также поддерживает пользовательские критерии качества. Значительная часть мер, описанных во второй главе, реализована на платформе TopicNet.

Для того чтобы облегчить ручной анализ моделей, мы реализовали простой специализированный (domain-specific) язык для отбора моделей (пример приведён на Рис. 3). Использование этого языка делает процесс многокритериального отбора моделей более простым и прозрачным.

В шестой главе рассматривается задача создания таксономии коллекции диалогов контактного центра без наличия разметки. Предлагается регуляризованная тематическая модель, играющая роль первого приближения к структуре коллекции.

```
TopicKernel@word.average_contrast > 0.95 * MAXIMUM(
        TopicKernel@word.average_contrast)
   and PerplexityScore@all < 1.1 * MINIMUM(
        PerplexityScore@all)
   and SparsityPhiScore@word -> max
   COLLECT 3
```

Рис. 3 — Пример строки, задающей критерий отбора моделей. Здесь в качестве критериев отбора участвуют перплексия, контраст лексического ядра модальности `@word` и разреженность матрицы Ф. Результатом будут три модели, контраст которых не более чем на 5% отличается от наилучшего достигнутого контраста, имеют допустимую перплексию и как можно более разреженны.

Был использован усложнённый вариант распространённой техники, улучшающей интерпретируемость — использования информативных n-грам (коллокаций) в качестве дополнительной модальности. Для того чтобы извлечь информативные n-грамы, был использован алгоритм TopMine [25], основанный на статистике парной сочетаемости слов в текстовой коллекции.

Для нашей задачи имело смысл внести ряд правок в алгоритм TopMine. Во-первых, логика подсчёта статистик парной сочетаемости была модифицирована таким образом, чтобы в ней использовались мультимножества слов вместо последовательностей слов.

Во-вторых, TopMine не выделяет пересекающиеся коллокации. Это приводит к тому, что похожие предложения («записать ребёнка в детский сад» и «записать ребёнка в детский садик») могут вовсе не содержать общих коллокаций. Примером служат предложение «получение паспорта РФ» (выделится коллокация `получение_паспорт_РФ`) и предложение «паспорт РФ был утерян» (выделится коллокация `паспорт_РФ`). Это следует из процесса поиска коллокаций алгоритмом: на каждом шаге обработки соседние коллокации-кандидаты сливаются, если их объединение удовлетворяет критерию информативности. Для того чтобы устранить вышеописанную проблему, достаточно изменить процесс итеративного слияния фраз так, чтобы при успешном слиянии коллокаций они не удалялись из множества кандидатов. Данная модификация увеличивает потребление памяти алгоритмом, однако делает процесс поиска менее «жадным».

Предлагаемая модель является двухуровневой, то есть состоит из двух «обычных» тематических моделей. Модель первого уровня и модель второго ориентируются на различные признаки. Это связано с тем, что первый уровень иерархии предназначен для определения предмета диалога, а цель второго уровня иерархии — нахождение действий, о которых говорит пользователь.

В контексте поставленной задачи было предпринято разделение признаков по их функциональному назначению. Из всех токенов (слов и n-грам) были выделены две группы на основании их частеречного состава: «тематическая» и «функциональная». «Функциональная» группа состоит из одиночных глаголов и n-грам, содержащих хотя бы один глагол. «Тематическая» группа состоит из одиночных существительных, одиночных прилагательных и n-грамм, включающих в себя хотя бы одно существительное и не имеющих в своём составе глаголов.

Таким образом, предлагаемая тематическая модель использует пять модальностей: `@lemmatized` (просто слова), `@verb_lemmatized` (слова-глаголы), `@noun_lemmatized` (слова-существительные и слова-прилагательные), `@theme_ngramms` (n-грамы с существительными и без глаголов), `@verb_ngramms` (n-грамы с глаголами).

Одна и та же стратегия обучения успешно применяется к двум разным коллекциям диалогов. Первая коллекция состоит из диалогов клиентов с представителями различными государственных организаций, а вторая представляет собой логи технической поддержки провайдера. Механизм относительных коэффициентов позволил успешно перенести веса модальностей, подобранные на первой коллекции, на вторую коллекцию.

В **заключении** приведены основные результаты работы, которые заключаются в следующем:

1. Методология построения аддитивно регуляризованных тематических моделей, обеспечивающая формирование «рецептов моделирования» с автоматизированным подбором гиперпараметров по множеству критериев и отличающаяся использованием относительных коэффициентов регуляризации и кубов гиперпараметров.
2. Архитектура библиотеки TopicNet, обеспечивающая программную реализацию данной методологии и отличающаяся использованием удобного языка описания кубов гиперпараметров и возможностью создания пользовательских регуляризаторов и метрик качества на языке Python.
3. Универсальный рецепт моделирования, обеспечивающий многокритериальный выбор тематических моделей для широкого класса задач, отличающийся предварительной настройкой куба гиперпараметров по набору разнородных задач тематического моделирования.
4. Программная реализация нового критерия когерентности, обеспечивающая его эффективное вычисление и отличающаяся более полным использованием данных о сочетаемости слов внутри текстовых документов.

Публикации автора по теме диссертации

В изданиях, входящих в международную базу цитирования Scopus

1. Unsupervised dialogue intent detection via hierarchical topic model [Текст] / A. Popov, V. Bulatov, D. Polyudova, E. Veselova // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). — 2019. — С. 932—938.
2. *Alekseev, V.* Intra-text coherence as a measure of topic models' interpretability [Текст] / V. Alekseev, V. Bulatov, K. Vorontsov // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue. — 2018. — С. 1—13.
3. TopicNet: Making Additive Regularisation for Topic Modelling Accessible [Текст] / V. Bulatov, V. Alekseev, K. Vorontsov, D. Polyudova, E. Veselova, A. Goncharov, E. Egorov // Proceedings of The 12th Language Resources and Evaluation Conference. — 2020. — С. 6745—6752.
4. *Ирхин, И. А.* Аддитивная регуляризация тематических моделей с быстрой векторизацией текста [Текст] / И. А. Ирхин, В. Г. Булатов, К. В. Воронцов. — 2020.

Зарегистрированные программы для ЭВМ

5. *Свидетельство о гос. регистрации программы для ЭВМ.* Topic Net Cooking Machine [Текст] / Г. А. Владимирович, Б. В. Геннадьевич, В. К. Вячеславович ; Ф. государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт (национальный исследовательский университет)». — № 2019660705 ; заявл. 30.08.2019 ; опубл. 17.09.2019, 2019662102 (Рос. Федерация).
6. *Свидетельство о гос. регистрации программы для ЭВМ.* Система создания таксономии текстовой коллекции диалогового контактного центра [Текст] / Г. А. Владимирович, Е. Е. Олегович, В. Е. Романовна, Б. В. Геннадьевич ; Ф. государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт (национальный исследовательский университет)». — № 2020612931 ; заявл. 23.03.2020 ; опубл. 17.03.2020, 2020613851 (Рос. Федерация).

7. *Свидетельство о гос. регистрации программы для ЭВМ.* Topic Net Viewers [Текст] / Г. А. Владимирович, Б. В. Геннадьевич, В. К. Вячеславович ; Ф. государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт (национальный исследовательский университет)». — № 2019660707 ; заявл. 30.08.2019 ; опубл. 10.09.2019, 2019661840 (Рос. Федерация).

Список литературы

1. Reading tea leaves: How humans interpret topic models [Текст] / J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, D. M. Blei // Advances in neural information processing systems. — 2009. — С. 288—296.
2. Automatic Evaluation of Topic Coherence [Текст] / D. Newman, J. H. Lau, K. Grieser, T. Baldwin // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. — Los Angeles, California : Association for Computational Linguistics, 2010. — С. 100—108. — (HLT '10). — URL: http://dl.acm.org/citation.cfm?id=1857999.1858011.
3. *Hoffman, T.* Probabilistic latent semantic indexing [Текст] / T. Hoffman // Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. — New York : ACM Press, 1999. — С. 50—57.
4. *Blei, D. M.* Latent dirichlet allocation [Текст] / D. M. Blei, A. Y. Ng, M. I. Jordan // Journal of machine Learning research. — 2003. — Т. 3, Jan. — С. 993—1022.
5. Knowledge discovery through directed probabilistic topic models: a survey [Текст] / A. Daud, J. Li, L. Zhou, F. Muhammad // Frontiers of Computer Science in China. — 2010. — Т. 4, № 2. — С. 280—301.
6. *Blei, D. M.* Probabilistic topic models [Текст] / D. M. Blei // Commun. ACM. — 2012. — Т. 55, № 4. — С. 77—84. — URL: http://doi.acm.org/10.1145/2133806.2133826.
7. Applications of topic models [Текст] / J. Boyd-Graber, Y. Hu, D. Mimno [и др.] // Foundations and Trends® in Information Retrieval. — 2017. — Т. 11, № 2/3. — С. 143—296.
8. *Vorontsov, K.* Additive regularization for topic models of text collections [Текст] / K. Vorontsov // Doklady Mathematics. Т. 89. — Citeseer. Pleiades Publisher, 2014. — С. 301—304.

9. *Vorontsov, K. V.* Additive Regularization of Topic Models [Текст] / K. V. Vorontsov, A. A. Potapenko // Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications. — 2015. — Т. 101, № 1—3. — С. 303—323. — URL: http://dx.doi.org/10.1007/s10994-014-5476-6.

10. Fast and modular regularized topic modelling [Текст] / D. Kochedykov, M. Apishev, L. Golitsyn, K. Vorontsov // 2017 21st Conference of Open Innovations Association (FRUCT). — IEEE. 2017. — С. 182—193.

11. *Ianina, A.* Regularized multimodal hierarchical topic model for document-by-document exploratory search [Текст] / A. Ianina, K. Vorontsov // 2019 25th Conference of Open Innovations Association (FRUCT). — IEEE. 2019. — С. 131—138.

12. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections [Текст] / K. Vorontsov, O. Frei, M. Apishev, P. Romov, M. Dudarenko // Analysis of Images, Social Networks and Texts - 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9-11, 2015, Revised Selected Papers. Т. 542 / под ред. M. Y. Khachay, N. Konstantinova, A. Panchenko, D. I. Ignatov, V. G. Labunets. — Springer, 2015. — С. 370—381. — (Communications in Computer and Information Science). — URL: http://dx.doi.org/10.1007/978-3-319-26123-2.

13. *Frei, O.* Parallel non-blocking deterministic algorithm for online topic modeling [Текст] / O. Frei, M. Apishev // International Conference on Analysis of Images, Social Networks and Texts. — Springer. 2016. — С. 132—144.

14. *Grimmer, J.* Text as data: The promise and pitfalls of automatic content analysis methods for political texts [Текст] / J. Grimmer, B. M. Stewart // Political analysis. — 2013. — Т. 21, № 3. — С. 267—297.

15. *Pääkkönen, J.* Humanistic interpretation and machine learning [Текст] / J. Pääkkönen, P. Ylikoski // Synthese. — 2020. — С. 1—37.

16. *Schmidt, B. M.* Words alone: Dismantling topic models in the humanities [Текст] / B. M. Schmidt // Journal of Digital Humanities. — 2012. — Т. 2, № 1. — С. 49—65.

17. *Mantyla, M. V.* Measuring LDA topic stability from clusters of replicated runs [Текст] / M. V. Mantyla, M. Claes, U. Farooq // Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. — 2018. — С. 1—4.

18. *Boyd-Graber, J.* Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements [Текст] / J. Boyd-Graber, D. Mimno, D. Newman // Handbook of Mixed Membership Models and Their Applications. —.

19. *Agrawal, A.* What is wrong with topic modeling? and how to fix it using search-based software engineering [Текст] / A. Agrawal, W. Fu, T. Menzies // Information and Software Technology. — 2018. — Т. 98. — С. 74—88.

20. *Chen, T.-H.* A survey on the use of topic models when mining software repositories [Текст] / T.-H. Chen, S. W. Thomas, A. E. Hassan // Empirical Software Engineering. — 2016. — Т. 21, № 5. — С. 1843—1919.

21. *Mavrin, A.* Four Keys to Topic Interpretability in Topic Modeling [Текст] / A. Mavrin, A. Filchenkov, S. Koltcov // Conference on Artificial Intelligence and Natural Language. — Springer. 2018. — С. 117—129.

22. *Fan, A.* Assessing topic model relevance: Evaluation and informative priors [Текст] / A. Fan, F. Doshi-Velez, L. Miratrix // Statistical Analysis and Data Mining: The ASA Data Science Journal. — 2019. — Т. 12, № 3. — С. 210—222.

23. Topic Modelling for Extracting Behavioral Patterns from Transactions Data [Текст] / E. Egorov, F. Nikitin, V. Alekseev, A. Goncharov, K. Vorontsov // 2019 International Conference on Artificial Intelligence: Applications and Innovations (IC-AIAI). — IEEE. 2019. — С. 44—49.

24. *Röder, M.* Exploring the space of topic coherence measures [Текст] / M. Röder, A. Both, A. Hinneburg // Proceedings of the eighth ACM international conference on Web search and data mining. — ACM. 2015. — С. 399—408.

25. Scalable topical phrase mining from text corpora [Текст] / A. El-Kishky, Y. Song, C. Wang, C. R. Voss, J. Han // Proceedings of the VLDB Endowment. — 2014. — Т. 8, № 3. — С. 305—316.