BigARTM: библиотека с открытым кодом для тематического моделирования больших текстовых коллекций

© К. В. Воронцов © А. И. Фрей © П. А. Ромов © А. О. Янина
Московский Физико-Технический Институт (государственный университет),
Москва
voron@forecsys.ru sashafrey@gmail.com peter@romov.ru yanina.anastasia.mipt@gmail.com

© М. А. Суворова © М. А. Апишев
Московский Государственный Университет им. М. В. Ломоносова,
Москва
m.dudarenko@gmail.com great-mel@yandex.ru

Аннотация

Тематическое моделирование — это одно из современных направлений статистического анализа текстов, активно развивающееся последние 10–15 лет. Тематические модели выявляют латентные темы в коллекциях текстовых документов и используются для создания систем семантического поиска, категоризации, суммаризации, сегментации текстов. В данной работе представлена библиотека с открытым кодом BigARTM для вероятностного тематического моделирования больших текстовых коллекций. Приведены результаты тестирования производительности в сравнении с известными программными продуктами, показаны примеры применения на открытых текстовых коллекциях.

1 Введение

Вероятностная тематическая модель (probabilistic topic model) коллекции текстовых документов описывает каждую тему дискретным распределением на множестве терминов, каждый документ — дискретным распределением на множестве тем [8]. Одним из основных приложений тематического моделирования является информационный поиск [42]. Современные поисковые системы предназначены для поиска по коротким текстовым запросам. Они основаны на инвертированных индексах, в которых для каждого слова хранится список содержащих его документов [4]. Тематический поиск имеет другие цели и устроен по-другому. Если пользователь плохо ориентируется в терминологии или слабо представляет себе структуру предметной области, то его информационной потребностью является скорее получение «дорожной карты» предметной области, систематизация и визуализация релевантной информации по заданной теме. Тема запроса не формулируется словами, а задаётся текстовым фрагментом произвольный длины. Поисковая система строит тематическую модель запроса и определяет короткий список тем запроса. Затем для поиска документов схожей тематики применяются те же механизмы индексирования и поиска, только в роли слов выступают темы. Новые технологии информационного поиска на основе тематического моделирования активно разрабатываются в последние годы [39]. Тематические модели применяются также для выявления трендов в новостных потоках или научных публикациях [33], для многоязычного информационного поиска [38], для анализа данных социальных сетей [44], [35], [25], для классификации и категоризации документов [29], для тематической сегментации текстов [40], для анализа изображений и видеопотоков [20], [34], для тегирования веб-страниц [16], для обнаружения текстового спама [5], для рекомендательных систем [41], [39], [19], для анализа нуклеотидных последовательностей в биоинформатике [18].

Современные приложения предъявляют разнообразные требования к тематическим моделям. Они должны быть одновременно хорошо интерпретируемыми (автоматически строить темы, понятные конечным пользователям), мультимодальными (учитывать разнородные метаданные документов), динамическими (выявлять динамику тем во времени), иерархическими (автоматически разделять темы на подтемы), мультиграммными (использовать не только отдельные слова, но и ключевые фразы), и т.д.

За последние 15 лет в литературе были описаны сотни моделей с массой интересных приложений. Однако при решении практических задач зачастую используются лишь доступные реализации простых устаревших моделей PLSA (probabilistic latent semantic analysis) [13] и LDA (latent Dirichlet

Рис. 1. Мультимодальная тематическая модель наряду с распределениями вероятностей тем в документах и слов в темах оценивает распределения вероятностей других модальностей в каждой теме. Примерами модальностей являются элементы метаописания документов (авторы, моменты времени, источники, рубрики и т. д.), цитаты и ссылки между документами, объекты на изображениях, рекламные баннеры, пользователи и т. д.

allocation) [9]. Причина такого разрыва между теорией и практикой заключается в том, что тематическое моделирование в значительной степени остаётся исследовательской областью. Основой большинства моделей является теория байесовского обучения, в которой вывод каждой модели связан с решением отдельной математической задачи. Создание общего алгоритма оптимизации для широкого класса тематических моделей пока не представляется возможным в рамках байесовского подхода.

В [1] предложен альтернативный подход к тематическому моделированию, основанный на классической не-байесовской теории регуляризации некорректно поставленных задач по А.Н.Тихонову [6]. Показано, что он описывает широкий класс известных байесовских моделей [36]. Более того, *аддитивная регуляризация тематических моделей* (АРТМ) позволяет комбинировать регуляризаторы, тем самым комбинируя тематические модели. В отличие от байесовского подхода, оптимизация любых моделей и их комбинаций производится одним и тем же регуляризированным EM-алгоритмом. Для добавления регуляризатора в модель достаточно выписать его частные производные по параметрам модели.

Подчеркнём, что АРТМ — это не ещё одна тематическая модель, а общий подход к построению и комбинированию многих тематических моделей.

Следующим обобщением АРТМ являются *мультимодальные тематические модели*, которые описывают *метаданные* документов вместе с основным текстом, рис. 1.

Примерами метаданных являются: авторы [28], моменты времени [33], классы, жанры или категории [29], тэги [16] и именованные сущности (named entities) [23], цитируемые или цитирующие документы [11] или авторы [14], пользователи документов [39], графические элементы изображений [20], рекламные объявления на веб-страницах [24]. Метаданные помогают более точно определять тематику документов, и, наоборот, тематическая модель может использоваться для выявления семантики нетекстовых метаданных или для предсказания пропущенных метаданных. Различные типы метаданных принято называть *модальностями.* В коллекциях с параллельными текстами на нескольких языках модальностями являются языки [3], [22], [26], [17], [38].

Простота и универсальность математического аппарата АРТМ позволяет создать модульную технологию построения мультимодальных и многоцелевых моделей на основе библиотеки регуляризаторов. Каждый регуляризатор обеспечивает определённое свойство тематической модели. Для решения прикладной задачи пользователь выбирает из библиотеки набор регуляризаторов, обеспечивающий построение модели с требуемыми свойствами. Таким образом, технология АРТМ целиком исключает из процесса разработки модели этап решения нетривиальных математических проблем и открывает доступ к достижениям современного тематического моделирования для прикладных аналитиков.

На основе АРТМ нами разработана библиотека тематического моделирования BigARTM с открытым кодом. В библиотеке реализован наиболее эффективный на сегодняшний день онлайновый параллельный EM-алгоритм для тематического моделирования больших коллекций. Имеется встроенная расширяемая библиотека регуляризаторов и метрик качества тематических моделей.

2 Теория и алгоритмы

2.1 Мультимодальные тематические модели

Обозначим через M множество модальностей. Каждая модальность $m \in M$ имеет словарь W^m, элементы которого называются *токенами*. Слова, составляющие основной текст документов, образуют первую модальность W^1. Объединение словарей всех модальностей обозначим через W.

Каждый документ — это последовательность токенов различных модальностей. Предполагается,

что тематика текста может быть выявлена по частотам токенов в документах, а порядок токенов не важен. Это предположение называют *гипотезой «мешка слов»*. Итак, коллекция задаётся частотами (числом вхождений) n_{dw} токенов w в документах d.

Вероятностная тематическая модель (ВТМ) описывает вероятность появления токенов w модальности m в документах d распределением $p(w|d) = \sum_t \phi_{wt}^m \theta_{td}$, где $\phi_{wt}^m = p(w|t)$ — представление темы t распределением вероятностей на множестве токенов W^m, $\theta_{td} = p(t|d)$ — представление документа d распределением вероятностей на множестве тем T, общее для всех модальностей. Матрицы $\Phi^m = \|\phi_{wt}^m\|$ и $\Theta = \|\theta_{td}\|$ являются искомыми параметрами модели. Обозначим через Φ матрицу, составленную из матриц Φ^m, поставленных в столбец друг на друга.

Для каждой модальности запишем задачу максимизации логарифма правдоподобия:

$$\mathrm{L_m}(\Phi^{\mathrm{m}}, \Theta) = \sum_{\mathrm{d,w}} \mathrm{n_{dw}} \log \sum_{\mathrm{t}} \phi_{\mathrm{wt}}^{\mathrm{m}} \theta_{\mathrm{td}} \rightarrow \max_{\Phi^{\mathrm{m}}, \Theta}.$$

Это невыпуклая задача оптимизации при ограничениях неотрицательности и нормировки столбцов матриц Φ^m и Θ.

Число тем $|T|$ обычно много меньше числа документов $|D|$ и объёма словаря $|W|$. Таким образом, построение тематической модели сводится к разложению матрицы нормированных частот $p(w|d) = \frac{n_{dw}}{n_d}$ в произведение двух матриц меньшего размера Φ и Θ. Данная задача имеет в общем случае бесконечно много решений, то есть является некорректно поставленной (недоопределённой).

2.2 Аддитивно регуляризованные тематические модели (APTM)

Согласно теории регуляризации А.Н.Тихонова, если задача недоопределена, то её решение можно сделать устойчивым, добавив к основному критерию дополнительный критерий-регуляризатор, учитывающий специфику задачи и знания предметной области [6].

Аддитивная регуляризация тематических моделей — это многокритериальный подход, в котором к основному критерию логарифма правдоподобия добавляется взвешенная сумма регуляризаторов $R(\Phi, \Theta) = \sum_{i=1}^n \tau_i R_i(\Phi, \Theta)$ [37].

Мультимодальная аддитивно регуляризованная тематическая модель строится путём максимизации взвешенной суммы логарифмов правдоподобия модальностей и регуляризаторов

$$\sum_m \tau_m L_m(\Phi^m, \Theta) + \sum_i \tau_i R_i(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta},$$

при ограничениях неотрицательности и нормировки столбцов матриц Φ^m и Θ. Веса τ_m и τ_i называются *коэффициентами регуляризации*.

Теорема. Если функции R_i гладкие и (Φ, Θ) — решение данной оптимизационной задачи, то оно удовлетворяет следующей системе уравнений со вспомогательными переменными p_{tdw}, n_{wt}, n_{td}:

$$p_{tdw} = \underset{t \in T}{\mathrm{norm}}(\phi_{wt}\theta_{td});$$

$$n_{wt} = \sum_{d \in D} \tau_{m(w)} n_{dw} p_{tdw};$$

$$n_{td} = \sum_{w \in d} \tau_{m(w)} n_{dw} p_{tdw};$$

$$\phi_{wt}^m = \underset{w \in W^m}{\mathrm{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right);$$

$$\theta_{td} = \underset{t \in T}{\mathrm{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right);$$

где $m(w)$ — модальность токена w, оператор $\underset{t}{\mathrm{norm}}(x_t)$ преобразует произвольный вектор (x_t) в дискретное распределение вероятностей путём обнуления отрицательных компонент вектора и последующей нормировки.

Интерпретация вспомогательных переменных следующая: $p_{tdw} = p(t|d, w)$ — это условное распределение тем для каждого токена w в каждом документе d; n_{wt} — счётчик числа употреблений токена w, связанных с темой t; n_{td} — счётчик числа употреблений токенов, связанных с темой t, в документе d.

Для решения системы уравнений используется EM-алгоритм [10], который в данном случае совпадает с методом простых итераций. Задаётся начальное приближение параметров модели ϕ_{wt}^m, θ_{td}, затем на каждой итерации вычисления производятся последовательно по приведённым в теореме формулам. Каждая итерация — это один проход всей коллекции. Вычислительная сложность EM-алгоритма линейна по объёму коллекции, по числу тем и по числу итераций. Он хорошо масштабируется, поскольку, число итераций, необходимых для сходимости процесса, как правило, невелико.

2.3 Онлайновый EM-алгоритм

Онлайновый EM-алгоритм реализует наиболее эффективную схему вычислений, при которой большие коллекции документов обрабатываются вообще за одну итерацию. Это возможно благодаря тому, что в матричном разложении $\Phi\Theta$ матрица Φ зависит от всей коллекции, тогда как в матрице Θ каждый столбец зависит от своего документа. На больших коллекциях темы неплохо определяются по небольшой доле документов. Поэтому матрица Φ успевает сойтись задолго до того, как заканчивается первая итерация.

Онлайновая версия EM-алгоритма предлагалась для моделей PLSA в [7] и LDA в [12] и [21]. В нашей работе она обобщается на случай мультимодальных регуляризованных моделей.

В онлайновом EM-алгоритме коллекция разделяется на пакеты документов. Каждый пакет обрабатывается при фиксированной матрице Φ, при этом для каждого документа из пакета производится несколько итераций до сходимости. На каждой

итерации вычисляются переменные p_{tdw}, n_{td} и θ_{td}. Переменные n_{wt} и ϕ_{wt}^{m} обновляются гораздо реже, после обработки каждого пакета.

Онлайновый алгоритм не держит в памяти всю коллекцию. Пакеты документов загружаются и выгружаются по необходимости. Это позволяет обрабатывать сколь угодно большие коллекции, которые не помещаются целиком ни в оперативную память, ни даже на диск одного компьютера. Эксперименты показывают, что качество модели и время обработки слабо зависят от размера пакетов, причём с ростом размера коллекции эта зависимость становится ещё менее заметной.

2.4 Параллельные реализации EM-алгоритма

При разработке параллельной архитектуры BigARTM учитывался опыт известных параллельных реализаций алгоритмов оптимизации тематических моделей.

В алгоритме Approximate Distributed LDA (AD-LDA) [31] коллекция распределяется по процессорам примерно в равных пропорциях, по окончании обработки данных всеми процессорами производится синхронизация, в результате которой обновляется глобальная матрица Ф. Недостатком этой реализации является то, что время, затрачиваемое на одну итерацию работы алгоритма, определяется самым медленным из процессоров. Сеть во время итераций простаивает, а во время синхронизаций перегружена. Кроме того, для каждого ядра хранится копия матрицы Ф, что приводит к необоснованному расходу памяти.

В алгоритме Y!LDA [30] коллекция распределяется по узлам, на каждом узле создаётся несколько потоков, занимающихся обработкой документов, и один поток, производящий обновление глобальных счётчиков n_{wt}. К этому потоку обработчики обращаются асинхронно по мере готовности новых обновлений. Параллелизм, основанный на многопоточности, позволяет хранить одну копию глобальных счётчиков для всех ядер на каждом узле. Синхронизация состояний всех узлов организована на основе архитектуры классной доски. Суть её в том, что глобальное состояние хранится в некоторой общей для всех узлов памяти, и поток синхронизации каждого узла асинхронно обращается к ней и производит обновление. Система производит параллельную обработку документов, эффективно используя имеющиеся вычислительные и сетевые ресурсы.

Алгоритм Mr.LDA [43] основан на технологиях MapReduce и Hadoop. На шаге Map производится оптимизация параметров, связанных с документами, на шаге Reduce — с темами. Технология MapReduce обеспечивает хорошую отказоустойчивость.

Библиотека Gensim предоставляет две реализации LDA на языке Python: LdaModel и LdaMulticore [27]. Первая из них почти в точности повторяет Online LDA из [12] и не является ни параллельной, ни распределённой. Тем не менее, LdaModel работает достаточно эффективно за счёт использования матричных вычислений в библиотеке NumPy. LdaMulticore является параллельным, архитектурно он имеет много общего с Y!LDA.

3 Проект BigARTM

BigARTM — это библиотека тематического моделирования с открытым кодом, в которой реализован онлайновый параллельный EM-алгоритм для мультимодальных аддитивно регуляризованных тематических моделей. BigARTM имеет встроенный набор регуляризаторов и набор метрик качества тематических моделей.

3.1 Параллельная архитектура

При разработке архитектуры библиотеки BigARTM мы исходили из требований асинхронной обработки данных, минимизации используемого объёма оперативной памяти, масштабируемости при увеличении количества ядер на узле, кроссплатформенности, возможности быстрой установки и использования на одной машине.

Для организации параллельной обработки данных на одном узле используется многопоточный параллелизм в пределах одного процесса. Это позволяет получить хорошую скорость обработки и хранить общую матрицу Ф для узла, а не для каждого ядра. Кроме того, обеспечивается возможность асинхронной работы с данными.

Библиотека BigARTM предназначена прежде всего для эффективной онлайновой параллельной обработки в пределах одной машины. Для достижения независимости объёма используемой оперативной памяти от размера обрабатываемой коллекции вся коллекция D делится на пакеты D_b, $b = 1, \ldots, B$, которые сохраняются на диск в отдельных файлах. В каждый момент времени в памяти находится только часть из них. Вся матрица Θ никогда не хранится. В каждый момент времени в памяти содержится только её фрагмент, соответствующий обрабатываемым документам, и после окончания обработки они удаляются.

Организация параллельной обработки в BigARTM, как в Y!LDA и Gensim, основана на многопоточности. Существует множество *обработчиков* (Processor) и один *поток слияния* (Merger), рис. 2.

Обработчики производят параллельную обработку пакетов документов, вычисляют θ_{td} и накапливают счётчики n_{wt} для обновления матрицы Ф. Поток слияния асинхронно получает значения счётчиков от обработчиков n_{wt} и обновляет матрицу Ф. Для обработчиков существует *очередь заданий*, в которых содержатся подгруженные библиотекой в память пакеты. Также существует *очередь слияния*, в которую обработчики отправляют по мере готовности обновления матрицы Ф для потока слияния. Все данные хранятся в памяти статично и не копируются, перемещения производятся с помощью указателей. В каждый момент времени

Merger хранит две копии матрицы Φ, *базовую* и *активную*. Первая из них доступна для чтения обработчикам, вторая доступна для записи потоку слияния. Каждый обработчик перед началом работы захватывает указатель на текущую активную матрицу Φ и использует её для вычисления параметров θ_{td} своего пакета документов. Поток слияния, получая обновления из очереди слияния, обновляет базовую матрицу Φ. Как только все текущие обновления завершаются и обработчики отпускают указатели активной матрицы, поток слияния делает базовую матрицу активной и создаёт новую базовую матрицу. Обновление матрицы Φ (синхронизацию) можно инициировать в любой момент из пользовательского кода.

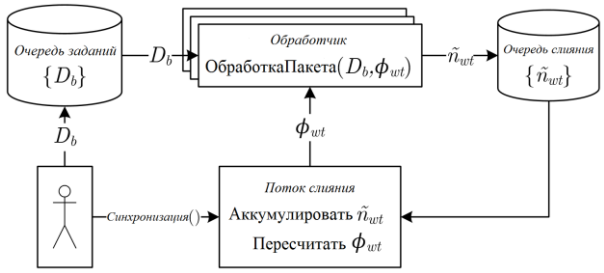

Рис. 2. Организация параллельной обработки в BigARTM.

3.2 Библиотека регуляризаторов

В текущей версии библиотеки BigARTM (на момент написания данной статьи v0.6.4) реализованы следующие регуляризаторы:

— обобщённый регуляризатор сглаживания, разреживания и частичного обучения для тем в матрице Φ^m (для любой модальности);

— обобщённый регуляризатор сглаживания, разреживания и частичного обучения для документов в матрице Θ;

— регуляризатор декоррелирования тем как столбцов матрицы Φ^m (для любой модальности);

— регуляризатор балансирования классов (label regularization) для задач классификации с сильно различающимися частотами классов.

Применение регуляризаторов разреживания и декоррелирования основных предметных тем совместно со сглаживанием фоновых тем общей лексики обеспечивает высокую разреженность матриц Φ, Θ и улучшает интерпретируемость тем, возможно, ценой несущественного ухудшения правдоподобия модели [2].

3.3 Библиотека метрик качества

В текущей версии библиотеки BigARTM (v0.6.4) реализованы следующие метрики качества:

— перплексия по каждой модальности:

$$\text{Perplexity} = \exp\left(-\frac{1}{n}L_m(\Phi^m,\Theta)\right);$$

— разреженность матриц Φ и Θ;

— чистота, контрастность и мощность лексических ядер тем [36]; ядром темы t считаются все слова, у которых $p(w|t) > 0.25$.

3.4 Автоматическое формирование словарей

Перед тематическим моделированием текстовых коллекций обычно выполняется предварительная обработка, включающая очистку текста от знаков препинания, переносов, нетекстовых данных и прочего «мусора»; лемматизацию; выделение ключевых фраз и именованных сущностей; удаление слишком частых слов (стоп-слов) и слишком редких слов. Все эти подготовительные операции выполняются с помощью сторонних средств за пределами BigARTM. Внутри библиотеки имеется возможность удалить редкие и частые слова во время предварительного прохода коллекции, при этом пороги максимальной и минимальной частоты задаются отдельно для каждой модальности.

3.5 Стратегии регуляризации

Решение разнообразных прикладных задач с большим объёмом сложно структурированных данных становится возможным в BigARTM благодаря свойству аддитивности регуляризаторов. Пользователь выбирает набор регуляризаторов, соответствующих целям моделирования, затем подбирает *стратегию регуляризации*, то есть определяет, по какому закону будут меняться весовые коэффициенты регуляризаторов.

Согласно теории регуляризации А.Н.Тихонова, коэффициенты регуляризации необходимо устремлять к нулю в ходе итераций для получения устойчивого решения исходной нерегуляризованной задачи. На практике одни регуляризаторы могут выполнять подготовительную работу для других, поэтому важна последовательность и постепенность включения и отключения регуляризаторов. Автоматический выбор стратегии регуляризации пока остаётся открытой теоретической проблемой.

В BigARTM используются *относительные коэффициенты регуляризации*, облегчающие перенос стратегий регуляризации с одной задачи на другую независимо от их размерных характеристик.

Для вычисления относительных коэффициентов регуляризации сначала оцениваются суммарные воздействия каждого регуляризатора R_i на каждую тему t в матрице Φ^m и на матрицу Φ^m в целом:

$$r_{it}^m = \sum_{w\in W}\left|\phi_{wt}^m\frac{\partial R_i}{\partial\phi_{wt}^m}\right|,\quad r_i^m = \sum_{t\in T} r_{it}^m,$$

а также на каждый документ d в матрице Θ и на матрицу Θ в целом:

$$r_{id} = \sum_{t\in T}\left|\theta_{td}\frac{\partial R_i}{\partial\theta_{td}}\right|,\quad r_i = \sum_{d\in D} r_{id}.$$

Относительные коэффициенты регуляризации $\tilde{\tau}_i$ обычно принимают значения в отрезке [0,1] и показывают относительную силу воздействия регуляризатора. Коэффициенты регуляризации τ_i выражаются через них по формулам

$$\tau_i = \tilde{\tau}_i \left(\gamma_i \frac{n_{it}^m}{r_{it}^m} + (1 - \gamma_i) \frac{n^m}{r_i^m} \right),$$

$$\tau_i = \tilde{\tau}_i \left(\gamma_i \frac{n_d}{r_{id}} + (1 - \gamma_i) \frac{n}{r_i} \right),$$

где n_d — длина документа, n — длина коллекции, n_d^m и n^m— длина документа и длина коллекции в модальности m; параметр γ_i называется *степенью индивидуализации* регуляризатора. При $\gamma_i = 1$ воздействия регуляризатора максимально различны по темам (или по документам), при этом воздействие на малые темы (или на короткие документы) меньше, на большие темы (на длинные документы) — больше. При $\gamma_i = 0$ воздействия регуляризатора не различаются по документам.

3.6 Открытый код

Исходный код проекта BigARTM доступен через https://github.com/bigartm. Он имеет лицензию New BSD License, допускающую как некоммерческое, так и коммерческое применение. Документация доступна по адресу http://bigartm.org.

Для разработки ядра библиотеки был выбран язык C++ из соображений кроссплатформенности и скорости вычислений.

Имеется API (интерфейс программирования приложений) для C++ и Python, а также приложение командной строки. Для обмена данными использована технология Google Protocol Buffers.

Библиотека является кросс-платформенной, поддерживаются операционные системы Linux, Windows и OS X в 32- и 64-битных конфигурациях.

4 Эксперименты

Целью экспериментов является сравнение библиотеки BigARTM с другими доступными средствами тематического моделирования по производительности и качеству моделей.

4.1 Тесты производительности

Для проверки линейной масштабируемости по числу процессоров был проведён эксперимент на четырёх открытых коллекциях, из которых pubmed и wiki уже могут считаться большими, см. Таблицу 1. Результаты показаны на рис. 3.

В Таблице 2 приведены результаты сравнения производительности BigARTM с другими популярными средствами тематического моделирования, Vowpal Wabbit и Gensim. Все три библиотеки реализуют онлайновый алгоритм, который запускался на одной и той же коллекции после одинаковой предварительной обработки, при одинаковом размере пакетов в 10 тысяч документов, и в одинаковой вычислительной среде Amazon AWS c3.8xlarge с 32 виртуальными ядрами. Реализация LDA в Vowpal Wabbit не является параллельной, поэтому она запускалась только на одном ядре.

4.2 Задача категоризации текстов

Известно, что метод опорных векторов (SVM), в котором частоты слов используются в качестве признаков, является одним из лучших методов категоризации текстов. Однако в [29] было показано, что тематические модели ещё лучше решают задачи категоризации в случаях большого числа несбалансированных, взаимозависимых, пересекающихся категорий.

Цель данного эксперимента — показать, что мультимодальные регуляризованные тематические модели в BigARTM справляются с задачей не хуже, чем алгоритм Dependency LDA из [29].

Таблица 1. Объёмные характеристики коллекций.

коллекция	$\lvert W \rvert, 10^3$	$\lvert D \rvert, 10^6$	$n, 10^6$	объём, GB
enron	28	0.04	6.4	0.07
nytimes	103	0.3	100	0.13
pubmed	141	8.2	738	1.0
wiki	100	3.7	1009	1.2

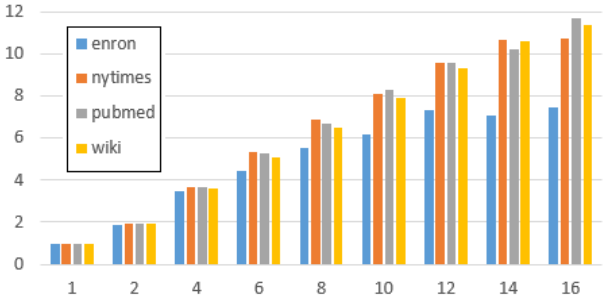

Рис. 3. Зависимость ускорения от числа процессоров.

Таблица 2. Сравнение алгоритмов тематического моделирования при запусках с разным числом параллельных процессов по времени обучения модели, по времени моделирования тестовой коллекции из 100 тысяч документов и по перплексии (чем меньше перплексия, тем лучше).

алгоритм	число проц.	обучение, мин.	тест, с.	перплек-сия
BigARTM	**1**	**35**	**72**	**4000**
Gensim.LdaModel	1	369	395	4161
VowpalWabbit.LDA	1	73	120	4108
BigARTM	**4**	**9**	**20**	**4061**
Gensim.LdaMulticore	4	60	222	4111
BigARTM	**8**	**4,5**	**14**	**4304**
Gensim.LdaMulticore	8	57	224	4455

Коллекция EUR-lex, использовавшаяся в [29], уже разбита на обучающую и тестовую части, что позволяет добиться воспроизводимости результатов. Коллекция содержит около 20 тысяч документов, словарь составляет 190 тысяч слов. Предобработка, предложенная в [29], удаляет все слова, встретившиеся в коллекции менее 20 раз. После этого словарь сокращается до 20 тысяч слов. Число категорий равно 3950. При этом каждый документ

может относиться ко многим категориям. В ходе предобработки были удалены метки классов, которые встретились во всей коллекции один раз (таких оказалось около 700).

Как для Dependency LDA, так и для аддитивно регуляризованной модели, использовался регуляризатор балансирования классов (label regularization) [29].

В эксперименте вычислялись метрики качества по отложенной выборке: AUC-PR — площадь под кривой precision-recall; AUC — площадь под ROC-кривой; OneErr — доля документов, у которых самая вероятная категория неверная; IsErr — доля документов, у которых множество категорий не совпадает с идеальным.

Результаты приведены в таблице 3. По трём из четырёх метрик качества BigARTM обходит как Dependency LDA, так и лучший результат SVM.

Таблица 3. Сравнение моделей классификации ARTM, Dependency LDA и SVM. Наилучшие результаты выделены жирным шрифтом. T_{opt} — оптимальное число тем для тематической модели.

алгоритм	T_{opt}	AUC-PR	AUC	OneErr	IsErr
BigARTM	10000	**0.513**	0.980	**29.1**	**95.5**
DepLDA	200	0.492	**0.982**	32.0	97.2
SVM	–	0.435	0.975	31.6	98.1

Интересно отметить, что при увеличении числа тем качество категоризации в BigARTM только улучшается, и оптимальные значения достигаются при 10 тысячах тем, тогда как для Dependency LDA оптимальным оказывается 200 тем.

4.3 Задача кросс-язычного поиска

Целью следующего эксперимента является проверка того, что мультимодальная тематическая модель может использоваться как мультиязычная, если модальностями считать языки параллельных текстов. Эксперимент проводился на коллекции EuroParl [15] протоколов заседаний Европейского парламента. Были выбраны протоколы на английском и испанском языках, так как именно эта языковая пара часто используется для сравнения мультиязычных тематических моделей. Как и в работах [22], [26], [17], отдельным документом считалась речь одного выступающего на одном заседании.

Измерялось качество кросс-язычного поиска, когда для документа–запроса q в одном языке требуется найти его документ–перевод d в другом языке. Документы сравнивались по расстоянию Хеллингера между их тематическими профилями:

$$H(d,q) = \frac{1}{2}\sum_{t\in T}\left(\sqrt{\theta_{td}} - \sqrt{\theta_{tq}}\right)^2.$$

Качество кросс-язычного поиска измерялось критерием *точности* — долей документов, для которых перевод оказался самым близким.

В качестве обучающей выборки были взяты заседания за период с 1996 по 1999 года, а также за 2001–2002 года, в качестве тестовой выборки — заседания за 2000 год и первые 9 месяцев 2003 года. Такое же разбиение использовалось в работах [26] и [17]. Кроме того, так же, как и в [22], [17], в тестовую выборку отбирались документы длиной не менее 100 слов. Число документов в обучающей выборке составило 67379, в тестовой — 16068.

Использовалась встроенная в BigARTM возможность фильтрации словаря: отбрасывались все редкие слова, которые встречаются менее, чем в 20 документах коллекции и стоп-слова, которые встречаются более, чем в 50% документов.

В таблице 4 приведено сравнение с моделями из работ [22], [26], [17]. Для первых двух моделей авторы сообщают точность поиска только для 50 тем. BigARTM немного уступает модели JPLSA, однако заметим, что одна итерация BigARTM занимает 30 секунд для 50 тем и 40 секунд для 100 тем, тогда как одна итерация JPLSA — 31 минуту [26]. По сравнению с моделями из [17] BigARTM показывает несколько лучшие результаты.

Таблица 2. Сравнение точности кросс-язычного поиска для разных моделей. Жирным шрифтом выделено лучшее значение в каждом столбце.

	Число тем T			
Модель	50	100	200	500
PLTM [22]	0.812	—	—	—
JPLSA [26]	**0.989**	—	—	—
PLTM-He [17]	0.9430	0.9845	0.9940	0.9931
PLTM-He kd-trees [17]	0.9486	0.9890	0.9950	0.9961
BigARTM	0.9721	**0.9900**	**0.9956**	**0.9967**

4.4 Задача рекомендации статей habrahabr.ru

Тематические модели часто используются для выявления тематических интересов пользователей в рекомендательных системах и социальных сетях [41], [39], [19]. Такие модели могут основываться на данных о частотах слов в документах (content-based), либо на данных о посещениях документов пользователями (usage-based). В обоих случаях для каждого пользователя и для каждого документа строятся векторы латентных тем или *интересов*. Сравнение и ранжирование этих векторов позволяет давать рекомендации. Модели, основанные на объединении данных о контенте и посещениях, называются *гибридными*. Мультимодальные тематические модели являются естественным инструментом для построения гибридных моделей, так как пользователи и слова являются равноправными модальностями, связанными с документами. Гибридные модели обычно оказываются более точными, поскольку они объединяют два принципиально разных источника информации.

Для эксперимента с несколькими модальностями была выбрана коллекция коллективного блога habrahabr.ru из 132157 статей. Документы содержали пять модальностей: слова, авторы, комментаторы, теги, хабы. Сравнивались три модели: первая использовала только модальность слов, вторая — только модальность комментаторов, третья — все пять модальностей. Целью данного эксперимента было показать, что совместное использование большего числа модальностей позволяет улучшить качество рекомендаций.

Для оценивания качества рекомендаций коллекция разбивалась 1:1 на обучающую и тестовую. Для построения рекомендаций во всех трёх моделях оценивались тематические профили пользователей $p(t|u)$ по обучающей выборке. Затем по тестовой выборке оценивались распределения их предпочтений $p(d|u) = \sum_t p(d|t)p(t|u)$, строился ранжированный список предпочтений, и по первым k элементам списка вычислялись стандартные критерии качества Precision@k (доля релевантных документов среди найденных) и Recall@k (доля найденных документов среди релевантных). Релевантными рекомендациями считались документы, которые данный пользователь комментировал. Поскольку коллекция большая, а пользователи в среднем комментируют лишь малую долю потенциально интересных для них документов, более адекватным критерием качества рекомендаций является Recall@k.

Таблица 5. Результаты сравнения качества рекомендаций по критерию Recall@k для унимодальных моделей с модальностью слов, с модальностью комментаторов и со всеми модальностями.

Критерий	Слова	комм.	все
Recall@5	0.79	0.62	**0.80**
Recall@10	0.64	0.59	**0.71**
Recall@15	0.63	0.60	**0.74**
Recall@20	0.68	0.65	**0.69**

Таблица 6. Рекомендованные и прочитанные статьи случайно выбранного пользователя X.

Статьи, рекомендованные пользователю X:
Мобильные устройства — друзья или враги?
Баг или фича с related ссылками?
Чтение RSS-потоков
Борьба с комментариями-дубликатами
Кредитная карта + мобильный телефон
Статьи, которые пользователь X комментировал:
Кредитная карта + мобильный телефон
Мобильные устройства — друзья или враги?
Старейшему сетевому изданию о музыке 10 лет
Оптимизация стоимости при работе с Amazon S3
Чтение RSS-потоков

В таблице 5 представлены результаты сравнения трёх моделей. Они показывают, что гибридная (мультимодальная) модель действительно превосходит унимодальные.

В таблице 6 показан пример рекомендаций случайно выбранному пользователю X в сравнении со списком статей, которые он комментировал.

4 Выводы

BigARTM — это программное обеспечение с открытым кодом, предназначенное для построения мультимодальных регуляризованных тематических моделей больших коллекций текстовых документов. Основой BigARTM является классическая (не-байесовская) теория регуляризации некорректно поставленных задач. Аддитивная регуляризация позволяет комбинировать регуляризаторы и модальности и строить модели с заданными свойствами. Эксперименты показывают, что BigARTM превосходит известные аналоги по скорости и по качеству. Дальнейшее развитие BigARTM предполагает расширение библиотеки регуляризаторов, библиотеки метрик качества и отказ от модели «мешка слов» для обработки последовательного текста.

Работа выполнена при финансовой поддержке РФФИ, гранты 14-07-00847, 14-07-00908, 14-07-31176 и программы ОМН РАН «Алгебраические и комбинаторные методы математической кибернетики и информационные системы нового поколения».

Литература

[1] Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. 2014. Т. 456, № 3. С. 268–271.

[2] Воронцов К.В., Потапенко А.А. Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции Диалог (Бекасово, 4–8 июня 2014 г.). Вып. 13 (20). М: Изд-во РГГУ, 2014. С. 676–687.

[3] Дударенко М. А. Регуляризация многоязычных тематических моделей // Вычислительные методы и программирование. 2015. Т. 16. С. 26–38.

[4] Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск. Вильямс, 2011.

[5] Павлов А. С., Добров Б. В. Метод обнаружения массово порожденных неестественных текстов на основе анализа тематической структуры // Вычислительные методы и программирование: новые вычислительные технологии. 2011. Т. 12. С. 58–72.

[6] Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач. М.: Наука, 1986.

[7] Bassiou N., Kotropoulos C. Online PLSA: Batch updating techniques including out-of-vocabulary words // Neural Networks and Learning Systems, IEEE

Transactions on. Nov 2014. Vol. 25, no. 11. Pp. 1953–1966.

[8] Blei D. M. Probabilistic topic models // Communications of the ACM. 2012. Vol. 55, no. 4. Pp. 77–84.

[9] Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. Vol. 3. Pp. 993–1022.

[10] Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM algorithm // J. of the Royal Statistical Society, Series B. 1977. no. 34. Pp. 1–38.

[11] Dietz L., Bickel S., Scheffer T. Unsupervised prediction of citation influences // Proceedings of the 24th international conference on Machine learning. ICML '07. New York, NY, USA: ACM, 2007. Pp. 233–240.

[12] Hoffman M. D., Blei D. M., Bach F. R. Online learning for latent dirichlet allocation // NIPS. Curran Associates, Inc., 2010. Pp. 856–864.

[13] Hofmann T. Probabilistic latent semantic indexing // Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 1999. Pp. 50–57.

[14] Kataria S., Mitra P., Caragea C., Giles C. L. Context sensitive topic models for author influence in document networks // Proceedings of the Twenty-Second international joint conference on Artificial Intelligence Volume 3. IJCAI'11. AAAI Press, 2011. Pp. 2274–2280.

[15] Koehn P. Europarl: A Parallel Corpus for Statistical Machine Translation. In Conference Proceedings: the tenth Machine Translation Summit, pages 79–86, Phuket, Thailand, 2005. AAMT, AAMT.

[16] Krestel R., Fankhauser P., Nejdl W. Latent Dirichlet allocation for tag recommendation // Proceedings of the third ACM conference on Recommender systems. ACM, 2009. Pp. 61–68.

[17] Krstovski K., Smith D. A. Online polylingual topic models for fast document translation detection. In Proceedings of the Eighth Workshop on Statistical Machine Translation, WMT'13, pages 252–261, 2013.

[18] La Rosa M., Fiannaca A., Rizzo R., Urso A. Probabilistic topic modeling for the analysis and classification of genomic sequences // BMC Bioinformatics. 2015. Vol. 16, no. Suppl 6. P. S2.

[19] Lee S. S., Chung T., McLeod D. Dynamic item recommendation by topic modeling for social networks // Information Technology: New Generations (ITNG), 2011 Eighth International Conference on. IEEE, 2011. Pp. 884–889.

[20] Li X.-X., Sun C.-B., Lu P., Wang X.-J., Zhong Y.-X. Simultaneous image classification and annotation based on probabilistic model // The Journal of China Universities of Posts and Telecommunications. 2012. Vol. 19, no. 2. Pp. 107–115.

[21] Mimno D., Hoffman M., Blei D. Sparse stochastic inference for latent Dirichlet allocation // Proceedings of the 29th International Conference on Machine Learning. 2012. Pp. 1599–1606

[22] Mimno D., Wallach H. M., Naradowsky J., Smith D. A., McCallum A. Polylingual topic models // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2. EMNLP '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. Pp. 880–889.

[23] Newman D., Chemudugunta C., Smyth P. Statistical entity-topic models // Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '06. New York, NY, USA: ACM, 2006. Pp. 680–686.

[24] Phuong D. V., Phuong T. M. A keyword-topic model for contextual advertising // Proceedings of the Third Symposium on Information and Communication Technology. SoICT '12. New York, NY, USA: ACM, 2012. Pp. 63–70.

[25] Pinto J. C. L., Chahed T. Modeling multi-topic information diffusion in social networks using latent Dirichlet allocation and Hawkes processes // Tenth International Conference on Signal-Image Technology & Internet-Based Systems. 2014. Pp. 339–346.

[26] Platt J. C., Toutanova K., Yih W.-T. Translingual document representations from discriminative projections. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10, pages 251–261, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[27] Rehurek R., Sojka P. Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45-50, 2010.

[28] Rosen-Zvi M., Griffiths T., Steyvers M., Smyth P. The author-topic model for authors and documents // Proceedings of the 20th conference on Uncertainty in artificial intelligence. UAI '04. Arlington, Virginia, United States: AUAI Press, 2004. Pp. 487–494.

[29] Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine Learning. 2012. Vol. 88, no. 1-2. Pp. 157–208.

[30] Smola A., Narayanamurthy S. An Architecture for Parallel Topic Models. Proceedings of the VLDB Endowment, vol. 3 (1–2), pp. 703-710, 2010.

[31] Smyth P., Newman D., Asuncion A., Welling M. Distributed Algorithms for Topic Models. Neural Information Processing Systems — The Journal of Machine Learning Research, Vol. 10, pp. 1801-1828, 2009.

[32] Steyvers M., Griffiths T. Finding scientific topics // Proceedings of the National Academy of Sciences. 2004. Vol. 101, no. Suppl. 1. Pp. 5228–5235.

[33] Cui W., Liu S., Tan L., Shi C., Song Y., Gao Z., Qu H., Tong X. TextFlow: Towards better understanding of evolving topics in text // IEEE transactions on visualization and computer graphics. 2011. Vol. 17, no. 12. Pp. 2412–2421.

[34] Varadarajan J., Emonet R., Odobez J.-M. A sparsity constraint for topic models application to temporal activity mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions. 2010.

[35] Varshney D., Kumar S., Gupta V. Modeling information diffusion in social networks using latent topic information // Intelligent Computing Theory / Ed. by D.-S. Huang, V. Bevilacqua, P. Premaratne. Springer International Publishing, 2014. Vol. 8588 of Lecture Notes in Computer Science. Pp. 137–148.

[36] Vorontsov K. V., Potapenko A. A. Additive regularization of topic models // Machine Learning,

Special Issue on Data Analysis and Intelligent Optimization. 2014.

[37] Vorontsov K. V., Potapenko A. A. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization // AIST'2014, Analysis of Images, Social networks and Texts. Vol. 436. Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS), 2014. Pp. 29–46. 31

[38] Vulic I., De Smet W., Tang J., Moens M.-F. Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications // Information Processing & Management. 2015. Vol. 51, no. 1. Pp. 111–147.

[39] Wang C., Blei D. M. Collaborative topic modeling for recommending scientific articles // Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2011. Pp. 448–456.

[40] Wang H., Zhang D., Zhai C. Structural topic model for latent topical structure analysis // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol. 1. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. Pp. 1526–1535.

[41] Yeh J.-h., Wu M.-l. Recommendation based on latent topics and social network analysis // Proceedings of the 2010 Second International Conference on Computer Engineering and Applications. Vol. 1. IEEE Computer Society, 2010. Pp. 209–213.

[42] Yi X., Allan J. A comparative study of utilizing topic models for information retrieval // Advances in Information Retrieval. Springer Berlin Heidelberg, 2009. Vol. 5478 of Lecture Notes in Computer Science. Pp. 29–41.

[43] Zhai K., Boyd-Graber J., Asadi N., Alkhouja M. Mr.LDA: A Flexible Large Scale Topic Modeling Package using Variational Inference in MapReduce. Proceedings of the 21st international conference on World Wide Web, pp. 879-888, 2012.

[44] Zhao X. W., Wang J., He Y., Nie J.-Y., Li X. Originator or propagator?: Incorporating social role theory into topic models for Twitter content analysis // Proceedings of the 22Nd ACM International Conference on Conference on Information and Knowledge Management. CIKM '13. New York, NY, USA: ACM, 2013. Pp. 1649–1654.

BigARTM: Open Source Library for Topic Modeling of Large Text Collections

Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, Marina Dudarenko, Anastasia Yanina

Topic modeling is a rapidly developing branch of statistical text analysis last 10–15 years. Topic model uncovers a hidden thematic structure of a text collection and finds a highly compressed representation of each document by a set of its topics. The topical representation of a document captures the most important information about its semantics and therefore is useful for many applications including information retrieval, classification, categorization, summarization and segmentation of texts. Additive Regularization for Topic Modeling (ARTM) is a recent semi-probabilistic approach, which provides a much simpler inference for many models previously studied in the Bayesian settings. Additive regularization makes topic models easier to design, to infer, to combine, and to explain, thus reducing barriers to entry into topic modeling research field. In this paper we present the BigARTM open source project (http://bigartm.org) for regularized multimodal topic modeling of large collections. Experiments on Wikipedia corpus show that BigARTM performs faster and gives better perplexity comparing to other popular packages, such as Vowpal Wabbit and Gensim. We also demonstrate how BigARTM can be used for text categorization, cross-language search, and recommendations for social network users.