УДК 519.711.3 *Научная статья* DOI: 10.18287/2223-9537-2022-12-2-186-199

Аддитивная регуляризация при тематическом моделировании текстов сообществ онлайновых социальных сетей

© 2022, И.О. Датьев ✉, А.М. Федоров

Институт информатики и математического моделирования ФИЦ КНЦ РАН, Апатиты, Россия

Аннотация

Задача моделирования сообществ (групп) пользователей в социальных медиа является актуальной в рамках информационной поддержки принятия решений на разных уровнях государственного управления. Для автоматизированного извлечения смысла текстовой и сопутствующей информации используются методы тематического моделирования. В статье представлен опыт улучшения результатов тематического моделирования сообществ онлайновых социальных сетей с помощью аддитивной регуляризации тематических моделей. Улучшение результатов достигается посредством применения базовых регуляризаторов, доступных в программной библиотеке с открытым исходным кодом *BigARTM*. Тематические модели, полученные с использованием регуляризатора, сравниваются с тематическими моделями, полученными методами латентного размещения Дирихле и вероятностного латентно-семантического анализа. На подготовленном датасете, содержащем предварительно обработанные тексты постов сообществ онлайновой социальной сети проведены эксперименты по сравнению качества тематических моделей по метрикам когерентности, чистоты тем, разреженности матриц распределения. Обсуждаются недостатки метрик когерентности для оценки качества тематических моделей, полученных с помощью метода аддитивной регуляризации. Предложены дополнительные метрики, которые могут быть полезны для оценки качества тематических моделей. Сделаны выводы о применимости предложенного подхода для моделирования сообществ онлайновых социальных сетей. Результаты работы могут быть применены при разработке информационно-аналитических систем поддержки управления региональным развитием.

Ключевые слова: *управление региональным развитием, информационно-аналитические системы, сообщества социальных сетей, методы тематического моделирования, метрики когерентности.*

Цитирование: *Датьев И.О., Фёдоров А.М.* Аддитивная регуляризация при тематическом моделировании текстов сообществ онлайновых социальных сетей // Онтология проектирования. 2022. Т. 12, №2(44). С.186-199. DOI:10.18287/2223-9537-2022-12-2-186-199.

Финансирование: Исследование выполнено в рамках государственного задания ИИММ ФИЦ КНЦ РАН. Тема НИР «Методология создания информационно-аналитических систем поддержки управления региональным развитием, основанных на формирующем искусственном интеллекте и больших данных» (регистрационный номер темы НИР: 122022800551-0).

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение

Задачи управления региональным развитием, стоящие перед органами региональной власти, имеют сложный и разнородный характер. Для их качественного решения помимо представителей власти требуется привлечение специалистов из различных предметных областей (ПрО) [1]. В проведении решений при управлении региональным развитием заинтересованы и участвуют представители трёх сфер: власть, бизнес и гражданское общество.

Для современных процессов всеобщей цифровизации актуальной представляется задача разработки комплексных информационно-аналитических систем (ИАС) для поддержки управления региональным развитием. Основными характеристиками таких систем являются

интеллектуализированная обработка больших объёмов данных, распределённость, динамично изменяющаяся структура, модульные решения разнородных прикладных задач. Проектирование и разработка ИАС, отвечающих этим требованиям, должны основываться на системе формализованных знаний о ПрО [2].

В задачах регионального управления одной из важных ПрО является общество, которое определяется как отдельными индивидами (жителями региона), так и общественными организациями, объединениями и сообществами, действующими в регионе. Формализованные знания об обществе составляют базовую онтологию для создания модуля ИАС «Общество» (рисунок 1).

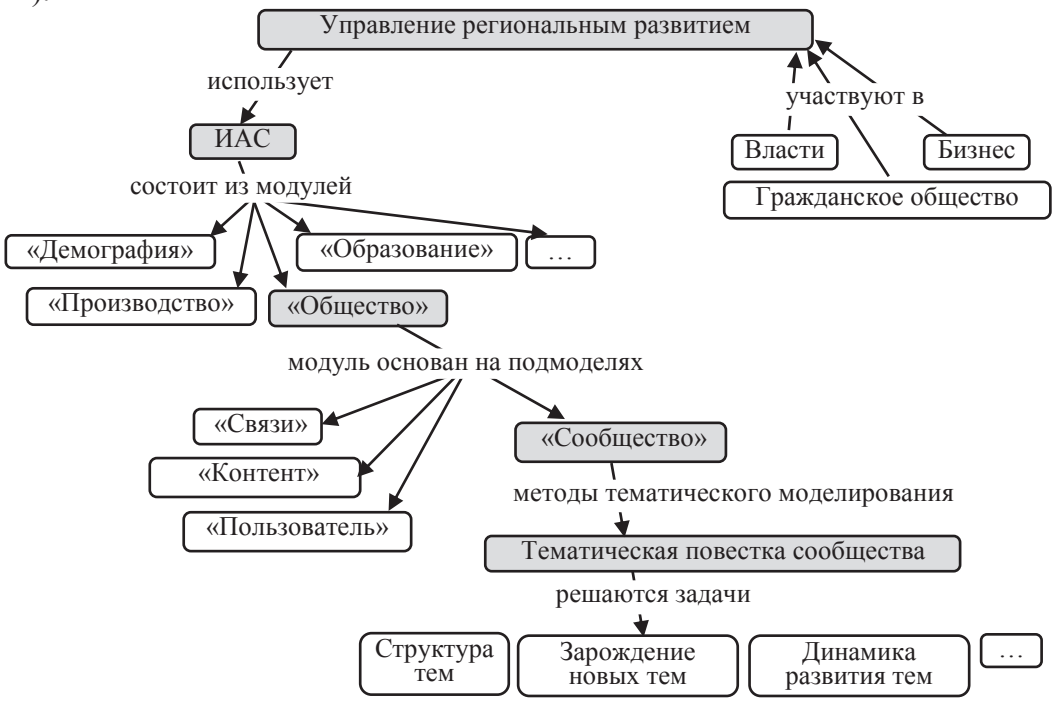

Рисунок 1 — Модуль «Общество» ИАС в структуре управления региональным развитием

Один из современных методов моделирования общества основан на исследовании социальных сетей (СС). СС можно рассматривать как образ общества в информационном онлайновом пространстве. Существование такого пространства открывает возможности автоматизации получения данных, проведения мониторинга, оценки, анализа и построения прогнозов.

Для задач управления региональным развитием важной характеристикой общества является *тематическая повестка*. В информационном пространстве СС тематическая повестка определяется совокупностью публикаций (постов), комментариев, мультимедийных приложений, которые размещаются на персональных страницах пользователей, а также в организуемых ими сообществах (группах). Для автоматизированного анализа текстов постов и комментариев с целью выявления в них *тематических кластеров* используются методы *тематического моделирования*.

Данная статья посвящена применению тематического моделирования для построения моделей сообществ СС и улучшению его результатов.

В работе [3] показано использование ряда методов вероятностного тематического моделирования на примерах постов сообществ онлайновой СС ВКонтакте. Эти эксперименты показали, что подход к аддитивной регуляризации тематических моделей (*Additive Regularization for Topic Modeling, ARTM*) является одним из самых быстрых по времени обучения тематической модели. В качестве другого преимущества отмечена мультимодальность, значимая для моделирования сообществ СС, поскольку в них присутствует сопут-

ствующая информация, которая не всегда является обычным текстом. Для автоматической оценки качества тем, получаемых с помощью различных методов тематического моделирования, выбрана метрика когерентности *UMass*, поскольку ей присуща наибольшая корреляция с человеческими оценками семантической интерпретируемости тем [4, 5]. В роли дополнительного критерия качества использовалась сумма вероятностей верхних десяти токенов темы (т.н. чистота темы) [6]. Метод латентного размещения Дирихле *LDA* существенно превзошел *ARTM* и другие методы по значению когерентности *UMass*. Результаты объясняются использованием *ARTM* без регуляризаторов, поскольку в таком режиме подход *ARTM* является обычным методом вероятностного латентно-семантического анализа *PLSA* [7].

В данной статье приводятся результаты экспериментов с настройкой базовых регуляризаторов, доступных в библиотеке *BigARTM* [8], и с варьированием гиперпараметрами метода *LDA*. В качестве опорного для сравнения используются метод *LDA* и метод *ARTM* без регуляризации.

1 Особенности подхода ARTM

В проведённом исследовании использованы открытая библиотека *BigARTM* и программно реализованный в ней подход *ARTM* к построению и комбинированию тематических моделей [9, 10]. *ARTM* является применением классической теории регуляризации некорректно поставленных задач [11] к тематическому моделированию. Построение тематической модели сводится к задаче стохастического матричного разложения, которая в общем случае имеет бесконечно много решений, т.е. является некорректно поставленной.

ARTM имеет два принципиальных отличия [6] от байесовского подхода. Во-первых, не ставится задача построения чисто вероятностной модели порождения текста. Многие ограничения (в том числе лингвистические) формализуются с помощью оптимизационных критериев, а не через априорные распределения. При этом регуляризаторы могут не иметь вероятностную интерпретацию. Наиболее распространённый регуляризатор — распределение Дирихле — может быть заменён одновременно несколькими проблемно-ориентированным регуляризаторами [12-15]. Во-вторых, вместо байесовского вывода используется регуляризованный EM-алгоритм [16]. Благодаря аддитивности регуляризаторов добавление регуляризатора в модель требует лишь небольшой модификации M-шага в готовом EM-подобном алгоритме, что позволяет без особых вычислительных затрат использовать не один, а комбинацию регуляризаторов. В [17] утверждается, что *ARTM* — это обобщённый подход к тематическому моделированию как к задаче многокритериальной оптимизации.

Возможно также построение с помощью *ARTM* мультимодальных моделей, позволяющих учитывать дополнительную (сопутствующую), зачастую нетекстовую, информацию, содержащуюся в коллекциях документов [18, 19]. Примерами такой информации применительно к СС могут служить: время, автор, источник публикации, гиперссылки, хэштеги, именованные сущности, названия сообществ пользователей, эмотиконы и др.

Дальнейшее развитие подход *ARTM* получил в работах [20-22], в направлении совершенствования регуляризаторов и исследования сходимости алгоритма *ARTM* [23].

В работах [6, 17, 24] показано, что комбинирование регуляризаторов сглаживания, разреживания, декоррелирования повышает интерпретируемость тем и образует базовый набор регуляризаторов, достаточный для большинства задач тематического моделирования. В данной работе проведены эксперименты с тремя основными регуляризаторами *ARTM*: сглаживание, разреживание, декорреляция. Для метода *LDA* [25], производится настройка гиперпараметров *alpha* и *beta*, которые определяют степень аппроксимации матриц *Phi, Theta*, к распределению Дирихле. В случае малых (меньше единицы) значений *alpha* (или *beta*) получа-

ются разреженные распределения, в которых почти все вероятности равны или близки к нулю, и только небольшая часть – существенно ненулевые. Для тематического моделирования это соответствует предположению, что в документе присутствует небольшое число тем (это соответствует разреженной матрице *Theta*), и тему можно чётко определить небольшим количеством слов (разреженная матрица *Phi*) [26]. Таким образом, распределение Дирихле выполняет в тематическом моделировании (метод *LDA*) только роль регуляризатора «разрежива-ния-сглаживания», поскольку «не имеет лингвистических обоснований, не является моделью какого-либо языкового явления, и его применение продиктовано исключительно удобством аналитического интегрирования в байесовском выводе» [16].

2 Настройка регуляризаторов и некоторые метрики качества

Оценивание тематических моделей проводилось на реальном датасете, содержащем открытые данные нескольких региональных сообществ популярной российской СС ВКонтакте. Датасет состоит из 15754 уникальных токенов, 9084 лемматизированных постов [3]. В процессе экспериментов производилось обучение серий тематических моделей. Всего было получено девять серий моделей с диапазоном количества тем от одной до четырёхсот для каждой серии. Во всех моделях, кроме *LDA_auto* и *ARTM* без регуляризации (*PLSA*), значения коэффициентов регуляризации задавались вручную на основании результатов нескольких экспериментов.

Наряду с настройкой базовых регуляризаторов *ARTM* предпринимались попытки найти значения параметров метода *LDA*, способствующие построению наилучших тематических моделей. Для этого применялись несколько программных библиотек *BigARTM*[1], *Gensim*[2], *Gensim/Mallet*, в которых существует возможность настройки двух основных (за исключением *Gensim/Mallet*, где был найден только один параметр — *alpha*) гиперпараметров *LDA* — *alpha* и *beta*.

На рисунках 2-9 представлены результаты проведённых экспериментов и приняты следующие обозначения:

1) *LDA_auto* — модель *LDA* из пакета *Gensim* с автоматической инициализацией параметров *alpha* и *beta*;
2) *LDA*_a=e=10-4 — модель *LDA* из пакета *Gensim* с инициализацией параметров *alpha* и *beta* значениями 10^{-4};
3) *LDA_mlt*_a=10-4 — модель *LDA* из пакета *Gensim* с модификацией *Mallet*[3] с инициализацией параметра *alpha* значением 10^{-4};
4) ARTM_LDA_a=b=10-4 — модель *LDA* из пакета *BigARTM* с инициализацией параметров alpha и beta значениями 10^{-4};
5) *ARTM* — модель *ARTM* без регуляризации (*PLSA*);
6) *ARTM_phi*-2 — модель *ARTM* с разреживанием матрицы *Phi, tau*=-2;
7) *ARTM_theta*-2 — модель *ARTM* с разреживанием матрицы *Theta, tau*=-2;
8) *ARTM_decor*-2 — модель *ARTM* с декоррелированием тем, *tau*=-2;
9) *ARTM_plan*-2 — модель *ARTM* с применением последовательности разреживающих регуляризаторов при *tau*=-2 (по 10 итераций) для матриц *Phi, Theta* и декоррелирующего регуляризатора.

[1] https://bigartm.org/.

[2] https://pythonru.com/biblioteki/gensim.

[3] Python wrapper for Latent Dirichlet Allocation (LDA) from MALLET, the Java topic modelling toolkit - https://radimrehurek.com/gensim_3.8.3/models/wrappers/ldamallet.html

Цель экспериментов - отыскание значений коэффициентов регуляризации для подхода *ARTM*. Общее количество итераций (проходов по коллекции документов) для каждой модели равно 30.

На рисунках 2-5 представлены значения различных метрик - когерентность *UMass*, сумма вероятностей верхних десяти токенов темы матрицы *Phi*, количество околонулевых (находящихся в интервале $[0, 10^{-4}]$) значений элементов матриц *Phi* и *Theta* в зависимости от количества тем для тематических моделей, полученных с помощью метода *LDA* и подхода *ARTM*. Метрики рассчитываются для каждой темы отдельно. На графиках представлены усреднённые (медианные) значения для каждой тематической модели.

Доля околонулевых элементов в матрице *Phi* для различных моделей показывает сходное поведение (рисунок 4). Наблюдается резкий рост общей разреженности и выход на плато около значения 1.0, что соответствует полному «занулению» матрицы *Phi*.

Судить о преимуществе одной из моделей лишь на основании метрик на рисунках 2-5 не представляется возможным. Действительно, с ростом количества тем разреженность матриц *Phi* и *Theta* повышается, в матрицах увеличивается количество нулевых значений в случае применения подхода *ARTM* и околонулевых значений в случае применения метода *LDA*. Изменение доли околонулевых элементов во всей матрице в зависимости от количества тем показано на рисунках 4 и 5. Разреженность рассчитывается как отношение количества околонулевых элементов матрицы к общему количеству элементов матрицы. Следует отметить, что на рисунках 4-7 графики модели *ARTM_decor* и модели *ARTM* совпадают.

Интересно оценить разреженность отдельных тем. Тема представляется столбцом в матрице *Phi*. Разреженность темы определяется количеством околонулевых значений в соответствующем столбце матрицы *Phi*. На рисунке 6 показано изменение разреженности по темам в зависимости от их количества в модели.

Для наглядности на рисунке 6 продублирована доля околонулевых элементов в матрице *Phi* тематических моделей (показана на рисунке 4), которые в легенде рисунка 6 отмечены префиксом «*ttl_*». Остальные обозначения легенды соответствуют обозначениям на рисунке 2. Разреженность по темам рассчитывается как отношение тем, в которых все значения околонулевые (лежащие в интервале $[0, 10^{-4}]$), к общему количеству тем в модели.

Разреженность по темам демонстрирует более сложное по сравнению с общей разреженностью по матрицам поведение для разных моделей. Только одна из моделей (*ARTM_theta*-2) выявила тенденцию к полному «занулению» отдельных тем.

Разреженность по темам в части их верхних токенов (рисунок 7) рассчитывается как отношение количества тем, в которых имеется менее 10 неоколонулевых (не лежащих в интервале $[0, 10^{-4}]$) значений, к общему числу тем в модели. Можно предположить, что темы представленные небольшим числом токенов с ненулевым значением *Phi*, отражают долю потенциально пригодных для интерпретации тем в модели.

Существенной метрикой тематической модели является полнота охвата и использование в темах всех токенов словаря. В матрице *Phi* этот факт представляется отсутствием нулевых строк. Наличие пустых строк в матрице *Phi* свидетельствует о разреженности матрицы *Phi* по токенам. Разреженность матрицы *Phi* по токенам рассчитывается как отношение количества строк, заполненных нулевыми (или околонулевыми) элементами, к общему количеству строк. На рисунке 8 показана зависимость разреженности матрицы *Phi* по токенам от количества тем в модели. Для оценки разреженности используются не абсолютно нулевые значения, а интервал $[0, 10^{-4}]$. С учетом этого, разреженность матрицы *Phi* по строкам характеризует долю токенов словаря, которые почти никогда не используются в темах, т.е. имеют крайне низкую вероятность встретиться в любой теме тематической модели.

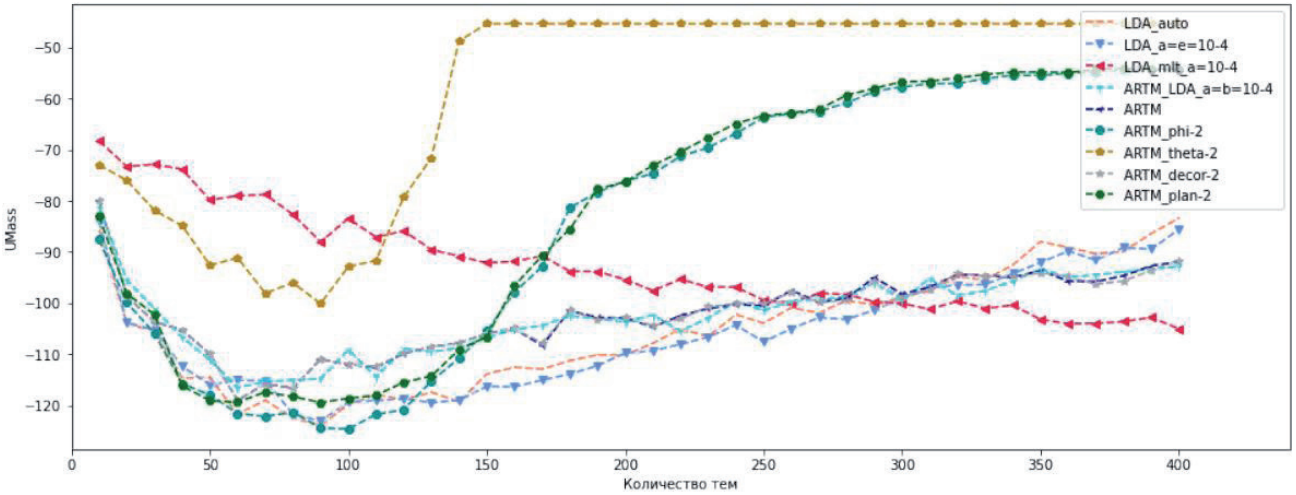

Рисунок 2 — Значения метрики когерентности *UMass* тематических моделей, полученных методами *LDA* и *ARTM* в зависимости от количества тем

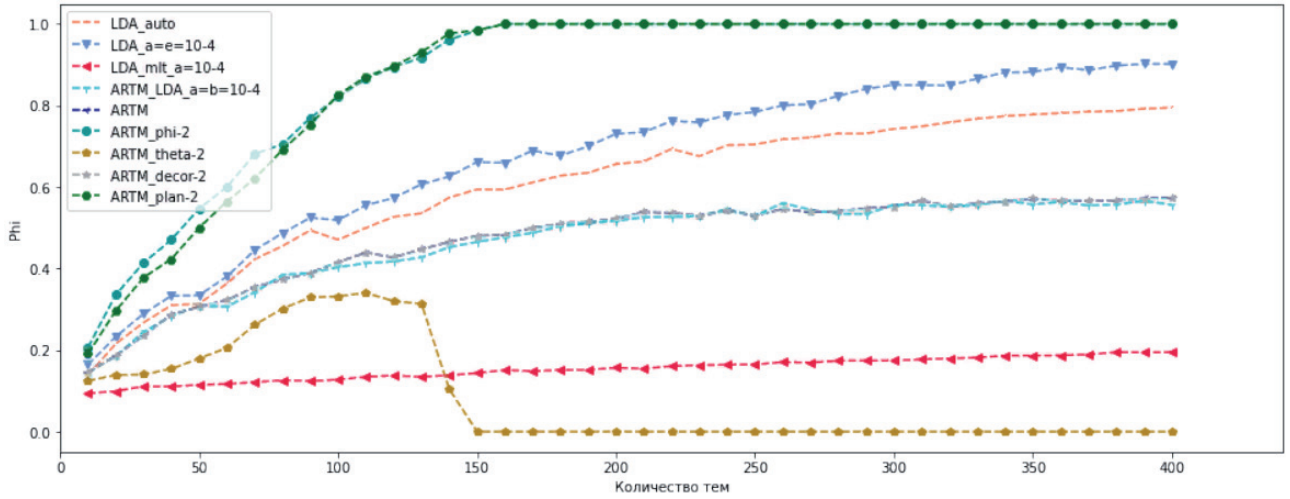

Рисунок 3 — Медианы сумм вероятностей верхних десяти токенов темы (*Phi*) в зависимости от количества тем

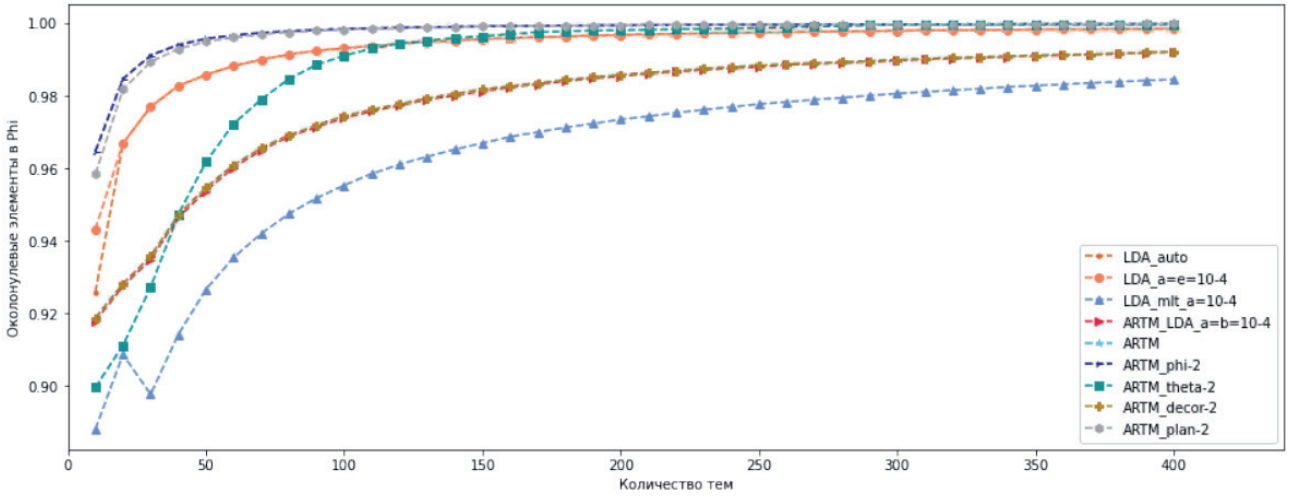

Рисунок 4 — Доля околонулевых элементов в матрице *Phi*

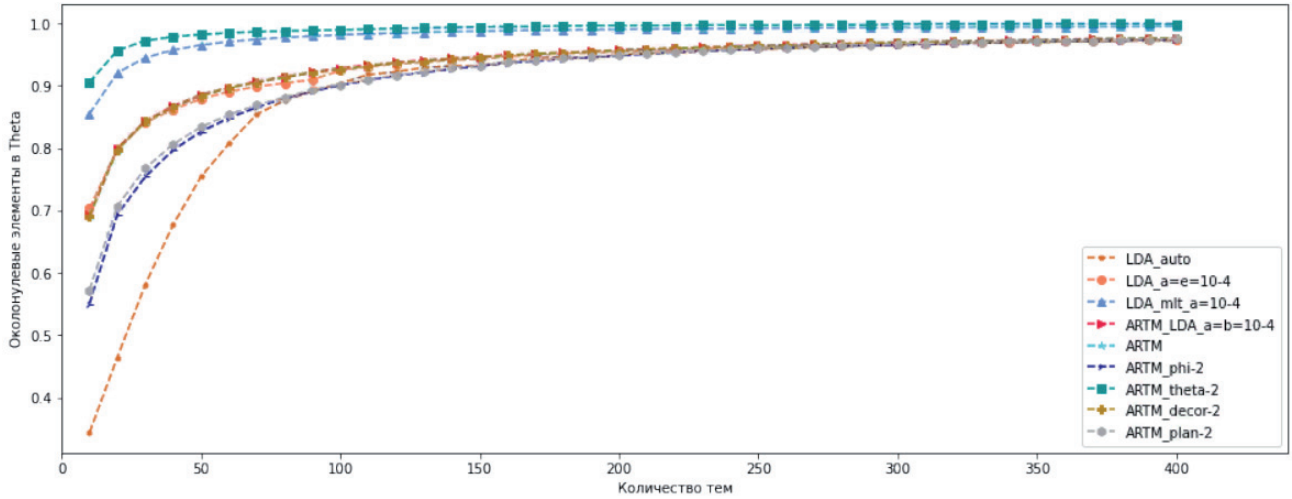

Рисунок 5 — Доля околонулевых элементов в матрице *Theta*

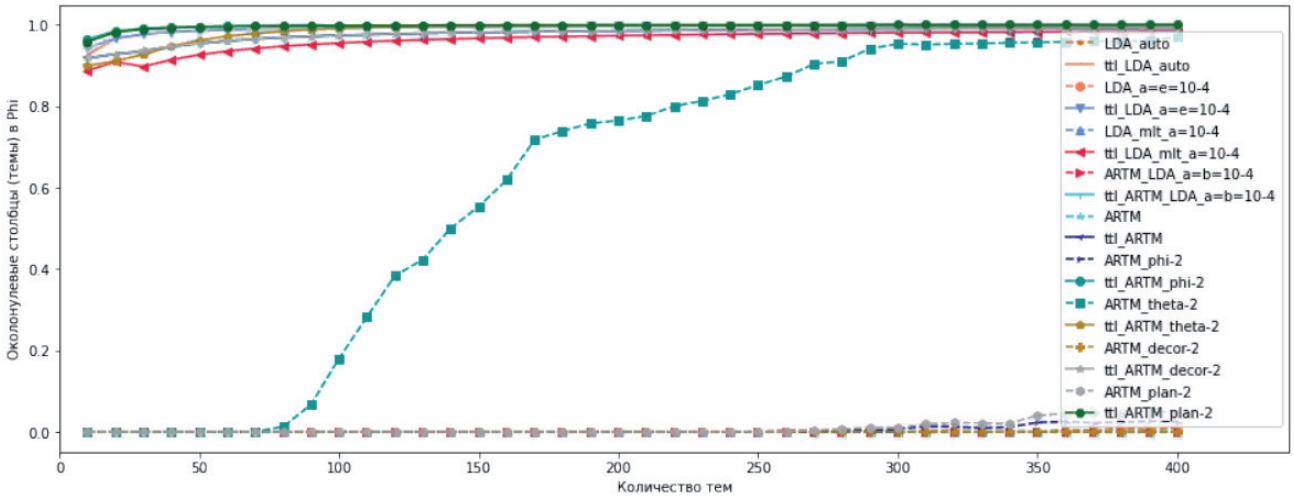

Рисунок 6 — Доля столбцов в матрице *Phi*, в которых все элементы околонулевые

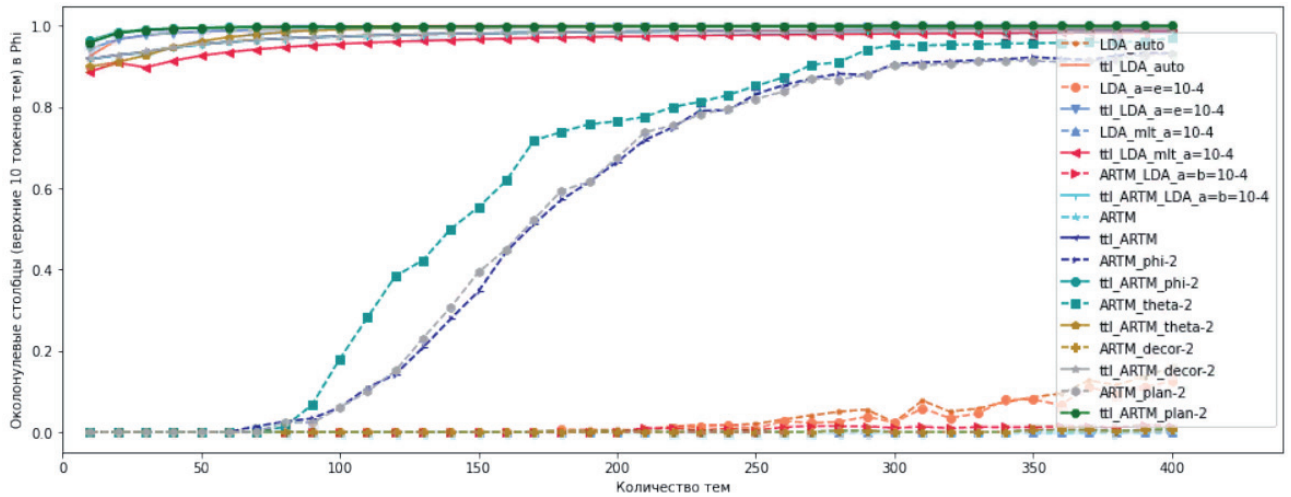

Рисунок 7 — Доля столбцов в матрице *Phi*, десять верхних токенов которых содержат околонулевые элементы

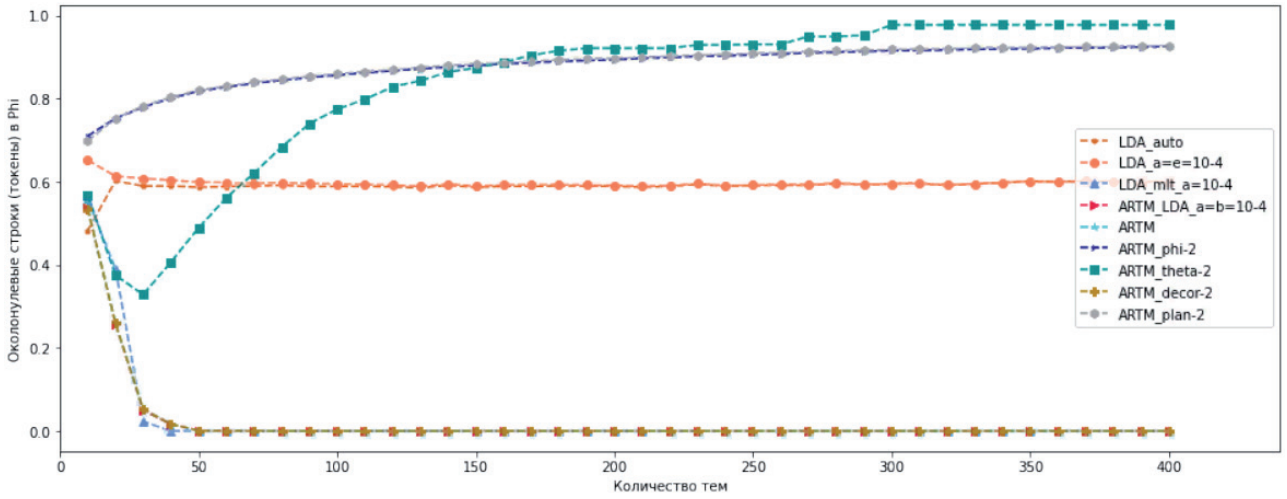

Рисунок 8 — Доля околонулевых строк в матрице *Phi*
(т.е. доля токенов словаря, которые не участвуют ни в одной из тем модели)

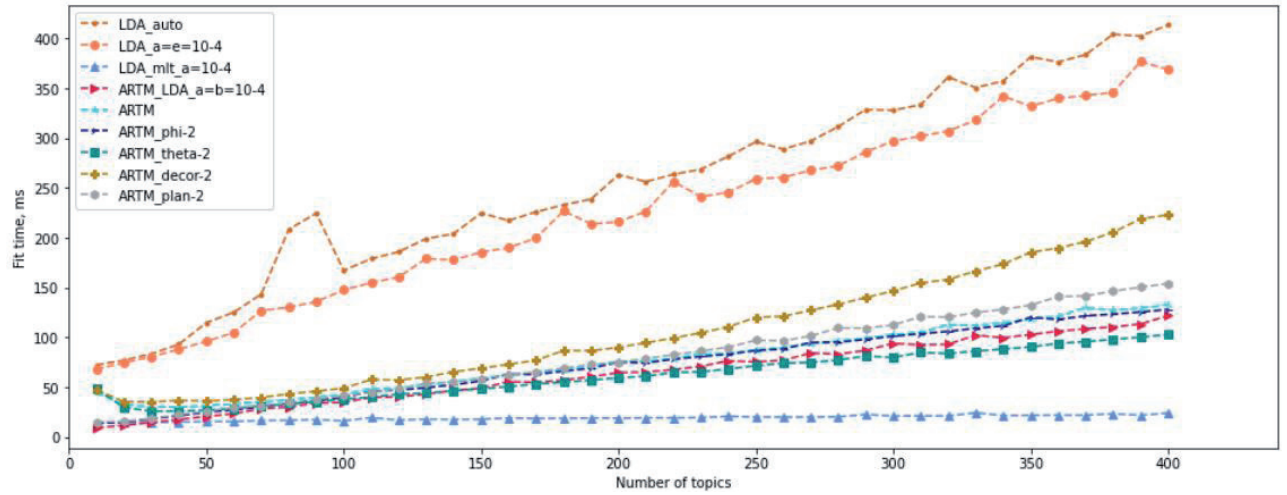

Рисунок 9 — Время обучения модели для разных методов в зависимости от количества тем

В работе был проведён тест на время обучения модели. В отличие от работы [1] здесь использовано большее количество моделей, присутствует библиотека *Gensim* (*Mallet*, обозначена *LDA_mallet*), максимальное количество тем составляет 400 (рисунок 9). В таких условиях подход *ARTM* находится на втором месте, однако время обучения сопоставимо с лидером и не является критично завышенным. Время обучения *ARTM* зависит от количества проходов по коллекции (в данном случае 30), т.е. результаты можно улучшить.

3 Анализ результатов

Цель проведённых экспериментов состояла в поиске средств и методов улучшения результатов тематического моделирования. Найденные и протестированные закономерности показали, что классические метрики, используемые для оценки качества тематических моделей, полученных с помощью *ARTM*, имеют ряд особенностей, которые необходимо учитывать при выборе параметров моделирования.

В результате экспериментов выявлено, что хотя когерентность *UMass* при применении *ARTM* с регуляризаторами возрастает, матрицы, а самое главное, темы или верхние токены тем могут получаться нулевыми. Такие темы вряд ли можно считать корректными или по-

лезными. Вероятнее всего, они требуют дополнительной обработки (или исключения), а, возможно, введения новых штрафных слагаемых в метрику когерентности. В таких условиях полагаться только на среднюю когерентность всех тем тематической модели при оценке качества темы становится невозможным.

У метода *LDA* значения элементов матриц уменьшаются значительно медленнее, *LDA* не допускает нулевых значений в матрицах — значения становятся близкими к нулю. Чтобы отразить долю малозначимых элементов матриц на рисунках 3-7, эмпирическим путём был установлен порог близости к нулю равный 10^{-4}.

ARTM - сравнительно новый подход и не все метрики когерентности пригодны для оценки тематических моделей, полученных с его помощью. Эта непригодность может быть объяснена тем, что метрики когерентности типа *UMass* и *UCi* изначально не рассчитаны на наличие нулевых вероятностей в матрицах.

Выделены следующие возможные причины некорректной работы метрик когерентности.

- По расчётной формуле значение метрики *Umass* нулевой темы может быть больше, чем ненулевой темы, что вносит неверный вклад в метрику.
- Особенность документов СС — слишком короткие документы (посты), — поэтому может сокращаться взаимная встречаемость терминов.
- Метрика когерентности темы вполне корректна для расчёта когерентности одной темы. Когда имеется несколько тем, в процессе регуляризации некоторые темы «зануляются». Т.е. метрика подходит для отдельных тем, но когда необходимо оценить тематическую модель в целом (множество тем для сообществ СС) более высокие значениями когерентности могут получить плохие нулевые темы.
- В [27] отмечено, что метрики когерентности, подобные *UMass* и *UCi*, как правило берут лишь 10 верхних токенов для темы, которые в корпусе текстовой коллекции составляют 1-2 %, причём доля текста, покрываемая верхними токенами, никак не контролируется. Возможно, что в коротких текстах СС эта доля ещё меньше.

Названные причины были учтены при разработке метрик качества тематических моделей, рассмотренных в экспериментальной части данной статьи.

4 Выводы

Данная работа направлена на обоснование выбора подхода *ARTM* для моделирования сообществ СС. Ранее, используя *ARTM* без настройки регуляризаторов, авторам данной работы не удалось получить тематические модели, превосходящие по значениям автоматизированных метрик когерентности тематические модели, полученные с помощью метода *LDA*. Поскольку *ARTM* обладает ключевым преимуществом для моделирования сообществ СС — мультимодальностью, было решено провести исследование с использованием базовых регуляризаторов *ARTM*. Для сравнения в качестве опорного использован метод *LDA* и подход *ARTM* без регуляризации (т.е. *PLSA*).

Исследования показали, что подход *ARTM* с регуляризацией сопоставим с другими методами тематического моделирования по значениям метрики когерентности *UMass* и сумме вероятностей верхних десяти токенов темы. Было обнаружено, что метрика когерентности *UMass* не всегда подходит для автоматизированной оценки качества тематических моделей, полученных с помощью метода *ARTM* при использовании регуляризаторов. Поэтому предложено несколько дополнительных метрик, которые могут быть полезными при оценивании качества тематической модели.

Результаты данной работы позволяют надеяться на улучшение качества тематических моделей при использовании в работе ИАС поддержки управления региональным развитием.

Список источников

[1] ***Боргест Н.М.*** Границы онтологии проектирования // Онтология проектирования. 2017. Т. 7, №1(23). С. 7-33. – DOI: 10.18287/2223-9537-2017-7-1-7-33.

[2] ***Смирнов С.В.*** Онтологическое моделирование в ситуационном управлении // Онтология проектирования. 2012. №2. С. 16-24.

[3] ***Fedorov A.M., Datyev I.O. Shchur A.L.*** Social Media Communities Topic Modeling // In: Silhavy R., Silhavy P., Prokopova Z. (eds.): Data Science and Intelligent Systems. CoMeSySo 2021. Lecture Notes in Networks and Systems. Vol. 231. Springer, Cham, 2021. P. 605-614. https://doi.org/10.1007/978-3-030-90321-3_50.

[4] ***Mimno D. Wallach H., Talley Ed., Leenders M., McCallum A.*** Optimizing semantic coherence in topic models // In: Proc. of the 2011 Conf. on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK. - Association of Computational Linguistics, 2011. P.262-272.

[5] ***Newman D. Lau J.H., Grieser K., Baldwin T.*** Automatic evaluation of topic coherence // In: Human Language Technologies: The 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics (HLT 2010). - Association for Computational Linguistics, Stroudsburg, 2010. P.100-108.

[6] ***Vorontsov K., Potapenko A.*** Additive regularization of topic models. // Mach Learn 101. 2015. P. 303-323. https://doi.org/10.1007/s10994-014-5476-6.

[7] ***Hofmann T.*** Probabilistic latent semantic indexing // In: Proc. of the 22nd annual international ACM SIGIR conf. on Research and development in information retrieval (SIGIR '99). - Association for Computing Machinery, New York, NY, USA, 1999. P.50–57. https://doi.org/10.1145/312624.312649.

[8] ***Datyev I.O., Fedorov A.M., Shchur A.L.*** Framework for civic engagement analysis based on open social media data // In: Silhavy R. (ed.): CSOC 2020. AISC. Vol. 1225. Springer, Cham, 2020. P. 586-597. https://doi.org/10.1007/978-3-030-51971-1_48.

[9] ***Kochedykov D. Apishev M., Golitsyn L., Vorontsov K.*** Fast and Modular Regularized Topic Modelling // In: 21st Conf. of Open Innovations Association (FRUCT). - FRUCT Oy, Helsinki, Uusimaa, Finland, 2017. P. 182-193 https://doi.org/10.23919/FRUCT.2017.8250181.

[10] ***Vorontsov K.V.*** Additive regularization for topic models of text collections. *Doklady Mathematics*. 2014. 3(89). P. 301–304. https://doi.org/10.1134/S1064562414020185.

[11] ***Tikhonov A.N., Arsenin V.Y.***: Solution of ill-posed problems. - Winston, Washington DC, 1977.

[12] ***Khalifa O., Corne D.W., Chantler M., Halley F.*** Multi-objective topic modeling // In: Purshouse R.C., Fleming P.J., Fonseca C.M., Greco S., Shaw J. (eds.): Evolutionary Multi-Criterion Optimization (EMO 2013). LNCS. Vol 7811. Springer, Heidelberg, 2013. P. 51-65. https://doi.org/10.1007/978-3-642-37140-0_8.

[13] ***Si L., Jin R.*** Adjusting mixture weights of gaussian mixture model via regularized probabilistic latent semantic analysis // In: Ho T.B., Cheung D.W.-L., Liu H. (eds.): Proc. of the Ninth Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD). LNCS. Vol. 3518. Springer, 2005. P. 622– 631.

[14] ***Chien J.-T., Wu M.-S.*** Adaptive bayesian latent semantic analysis // IEEE Transactions on Audio, Speech, and Language Processing. 2008. Vol. 1(16). P. 198–207.

[15] ***Larsson M.O., Ugander J.*** A concave regularization technique for sparse mixture models // In: Shawe Taylor J., Zemel R., Bartlett P., Pereira F., Weinberger K .(eds.): Advances in Neural Information Processing Systems 24 (NIPS 2011), 2011. P. 1890–1898.

[16] ***Воронцов К.В., Потапенко А.А.*** Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных. 2013. Т. 1, № 6. С. 657-686.

[17] ***Vorontsov K. Potapenko A., Plavin A.*** Additive Regularization of Topic Models for Topic Selection and Sparse Factorization // In: Gammerman A., Vovk V., Papadopoulos H. (eds.): Statistical Learning and Data Sciences (SLDS 2015). LNCS. Vol. 9047. Springer Cham, 2015. P.193-202. https://doi.org/10.1007/978-3-319-17091-6_14.

[18] ***Chirkova N.A., Vorontsov K.V.*** Additive Regularization for Hierarchical Multimodal Topic Modeling. *Machine Learning and Data Analysis*. 2016. Vol. 2. Issue 2. P. 187–200. https://doi.org/10.21469/22233792.2.2.05.

[19] ***Янина А.О., Воронцов К.В.*** Мультимодальные тематические модели для разведочного поиска в коллективном блоге // Машинное обучение и анализ данных. 2016. №2(2). С. 173–186. https://doi.org/10.21469/22233792.2.2.04.

[20] ***Apishev M., Koltcov S., Koltsova O., Nikolenko S., Vorontsov K.*** Additive Regularization for Topic Modeling in Sociological Studies of User-Generated Texts // In: Sidorov G., Herrera-Alcántara O. (eds.): Advances in Computational Intelligence (MICAI 2016). LNCS. Vol. 10061. SpringerCham, 2017. P. 169-184. https://doi.org/10.1007/978-3-319-62434-1_14.

[21] Bulatov V., Alekseev V., Vorontsov K., Polyudova D., Veselova E., Goncharov A., Egorov E**.** TopicNet: Making Additive Regularisation for Topic Modelling Accessible // In: Proc. of the 12th Language Resources and Evaluation Conf. - European Language Resources Association, Marseille, France, 2020. P. 6745–6752. https://aclanthology.org/2020.lrec-1.833.pdf.

[22] ***Veselova E., Vorontsov K.*** Topic Balancing with Additive Regularization of Topic Models // In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Online. - Association for Computational Linguistics, 2020. P. 59–65. https://doi.org/10.18653/v1/2020.acl-srw.9

[23] ***Ирхин И.А., Воронцов К.В.*** Сходимость алгоритма аддитивной регуляризации тематических моделей // Труды института математики и механики УРО РАН. 2020. №3(26). С. 56–68. https://doi.org/10.21538/0134-4889-2020-26-3-56-68

[24] ***Сухарева А.В., Воронцов К.В.*** Построение полного набора тем вероятностных тематических моделей // Интеллектуальные системы. Теория и приложения. 2019. Т. 23, № 4. С. 7-23.

[25] ***Blei D.M., Ng A.Y., Jordan M.I.*** Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. No. 3. P. 993-1022.

[26] ***Wallach H.M., Mimno D.M., McCallum A.*** Rethinking lda: Why priors matter. // In: NIPS. Vol. 22. 2009. P. 1973–1981.

[27] ***Alekseev V.A., Bulatov V.G., Vorontsov K.V.*** Intra-text coherence as a measure of topic models' interpretability // In: Computational Linguistics and Intellectual Technologies: Proc. of the Int. Conf. “Dialogue 2018” (Moscow, May 30 - June 2, 2018). P. 1-13. https://www.dialog-21.ru/media/4281/alekseevva.pdf

Сведения об авторах

Датьев Игорь Олегович 1981 г. рождения. Окончил Кольский филиал Петрозаводского государственного университета (2004). К.т.н. (2011). Учёный секретарь Института информатики и математического моделирования ФИЦ КНЦ РАН. Автор более 100 научных работ в области разработки моделей и технологий для региональных информационно-коммуникационных систем. Author ID (РИНЦ): 180256; Author ID (Scopus): 56070103900; Researcher ID (WoS): J-1839-2018. *datyev@iimm.ru*. ✉

Федоров Андрей Михайлович 1978 г. рождения. Окончил Кольский филиал Петрозаводского государственного университета (2000). К.т.н. (2005). Заместитель директора по научной работе Института информатики и математического моделирования ФИЦ КНЦ РАН. С 2000 г. преподаёт в филиале Мурманского арктического государственного университета в городе Апатиты, доцент кафедры информатики и вычислительной техники. Область научных интересов - разработка моделей и технологий информационной поддержки для регионального управления. Author ID (RSCI): 4285-9780; Author ID (Scopus): 57203929412; Researcher ID (WoS): D-5859-2016. *fedorov@iimm.ru*.

Поступила в редакцию 11.04.2022, после рецензирования 12.05.2022. Принята к публикации 06.06.2022.

Scientific article DOI: 10.18287/2223-9537-2022-12-2-186-199

Additive regularization for topic modeling of social media communities

© 2022, I.O. Datyev, A.M. Fedorov

Institute for Informatics and Mathematical Modeling, Kola Science Centre of the Russian Academy of Sciences, Apatity, Russia

Abstract

The task of modeling communities (groups) of users in social media is relevant in the framework of information support for decision-making at different levels of government. For automated extraction of the meaning of textual and related information, topic modeling methods are used. This article presents the experience of improving the results of social networks communities topic modeling using the Additive Regularization for Topic Modeling (ARTM). The improvement of the results is achieved through the use of basic regularizers available in the open-source software BigARTM. Topic models obtained using regularized ARTM are compared with topic models obtained by Latent Dirichlet Allocation and Probabilistic Latent Semantic Analysis. The experiments were carried out on a dataset, containing preprocessed texts of posts from communities of an online social network. In particular, the quality of topic models in terms of coherence, the purity of topics, and the sparsity of the distribution matrices are compared. Disadvantages of coherence metrics for assessing the quality of topic models obtained using the ARTM method are discussed. Additional metrics are proposed that can be used for assessing the quality of topic models. Conclusions are drawn about the suitability of the ARTM approach for modeling communities of online social networks. The results of this work can be applied in the development of information and analytical systems for supporting the management of regional development.

***Key words**: regional development management, information and analytical systems, social network communities, topic modeling methods, coherence metrics.*

***For citation:** Datyev IO, Fedorov AM.* Additive regularization for topic modeling of social media communities [In Russian]. *Ontology of designing*. 2022; 12(2): 186-199. DOI:10.18287/2223-9537-2022-12-2-186-199.

Financial Support: The work is supported by the Ministry of Science and Higher Education of the Russian Federation. Topic title: Methodology for creating information and analytical systems to support the management of regional development based on formative artificial intelligence and big data (reg. no. 122022800551-0).

Conflict of interest: The author declares no conflict of interest.

List of Figures

References

[1] ***Borgest NM.*** Boundaries of the ontology of designing [In Russian]. *Ontology of designing*. 2017; 7(1): 7-33. DOI: 10.18287/2223-9537-2017-7-1-7-33.

[2] ***Smirnov SV.*** Ontological modeling in situational management [In Russian]. *Ontology of designing*. 2012; No. 2: 16-24.

[3] ***Fedorov AM, Datyev IO, Shchur AL.*** Social Media Communities Topic Modeling. In: Silhavy R., Silhavy P., Prokopova Z. (eds.): Data Science and Intelligent Systems. CoMeSySo 2021. Lecture Notes in Networks and Systems. Vol. 231. Springer, Cham, 2021: 605-614. https://doi.org/10.1007/978-3-030-90321-3_50.

[4] ***Mimno D, Wallach H, Talley Ed, Leenders M, McCallum A.*** Optimizing semantic coherence in topic models. In: Proc. of the 2011 Conf. on Empirical Methods in Natural Language Processing (Edinburgh, Scotland, UK). Association of Computational Linguistics, 2011: 262–272.

[5] ***Newman D, Lau JH, Grieser K, Baldwin T.*** Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics (HLT 2010, Stroudsburg). Association for Computational Linguistics, 2010: 100–108.

[6] ***Vorontsov K, Potapenko A.*** Additive regularization of topic models. Mach Learn. 2015; 101: 303–323. https://doi.org/10.1007/s10994-014-5476-6.

[7] ***Hofmann T.*** Probabilistic latent semantic indexing. In: Proc. of the 22nd annual international ACM SIGIR conf. on Research and development in information retrieval (SIGIR '99). Association for Computing Machinery, New York, NY, USA, 1999: 50–57. https://doi.org/10.1145/312624.312649.

[8] ***Datyev IO, Fedorov AM, Shchur AL.*** Framework for civic engagement analysis based on open social media data. In: Silhavy R (ed.) CSOC 2020. AISC, vol. 1225. Springer, Cham, 2020: 586–597. https://doi.org/10.1007/978-3-030-51971-1_48.

[9] ***Kochedykov D, Apishev ., Golitsyn L, Vorontsov K.*** Fast and Modular Regularized Topic Modelling. In: 21st Conf. of Open Innovations Association (FRUCT). FRUCT Oy, Helsinki, Uusimaa, Finland, 2017: 182-193 https://doi.org/10.23919/FRUCT.2017.8250181.

[10] ***Vorontsov KV.*** Additive regularization for topic models of text collections. Doklady Mathematics. 2014; 3(89): 301–304. https://doi.org/10.1134/S1064562414020185.

[11] ***Tikhonov AN, Arsenin VY.*** Solution of ill-posed problems. Winston, Washington DC, 1977.

[12] ***Khalifa O, Corne DW, Chantler M, Halley F.*** Multi-objective topic modeling. In: Purshouse RC, Fleming PJ, Fonseca CM, Greco S, Shaw J (eds.): Evolutionary Multi-Criterion Optimization (EMO 2013). LNCS. Vol. 7811. Springer, Heidelberg, 2013: 51–65. https://doi.org/10.1007/978-3-642-37140-0_8.

[13] ***Si L, Jin R.*** Adjusting mixture weights of gaussian mixture model via regularized probabilistic latent semantic analysis. In: Ho TB, Cheung D.W-L, Liu H. (eds.): Proc. of the Ninth Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD). LNCS. Vol. 3518. Springer, 2005: 622– 631.

[14] ***Chien J-T, Wu M-S.*** Adaptive bayesian latent semantic analysis. IEEE Transactions on Audio, Speech, and Language Processing. 2008; 1(16): 198–207.

[15] ***Larsson MO, Ugander J.*** A concave regularization technique for sparse mixture models. In: Shawe Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger K (eds.): Advances in Neural Information Processing Systems 24 (NIPS 2011). 2011: 1890–1898.

[16] ***Vorontsov KV, Potapenko AA.*** EM-like algorithms modification for probabilistic topic modeling [In Russian]. Machine learning and data analysis. 2013; vol. 1, No. 6: 657–686.

[17] ***Vorontsov K, Potapenko A, Plavin A.*** Additive Regularization of Topic Models for Topic Selection and Sparse Factorization. In: Gammerman A, Vovk V, Papadopoulos H (eds.): Statistical Learning and Data Sciences (SLDS 2015). LNCS. Vol. 9047. Springer, Cham, 2015: 193-202. https://doi.org/10.1007/978-3-319-17091-6_14.

[18] ***Chirkova NA, Vorontsov KV.*** Additive Regularization for Hierarchical Multimodal Topic Modeling. Machine Learning and Data Analysis. 2016; 2(2): 187–200. https://doi.org/10.21469/22233792.2.2.05.

[19] ***Ianina A, Vorontsov KV.*** Multimodal topic modeling for exploratory search in collective blog [In Russian]. Machine Learning and Data Analysis. 2016; 2(2): 173–186. https://doi.org/10.21469/22233792.2.2.04.

[20] ***Apishev M, Koltcov S, Koltsova O, Nikolenko S, Vorontsov K.*** Additive Regularization for Topic Modeling in Sociological Studies of User-Generated Texts. In: Sidorov G, Herrera-Alcántara O (eds.): Advances in Computational Intelligence. MICAI 2016. LNCS. Vol. 10061. Springer, Cham, 2017: 169-184. https://doi.org/10.1007/978-3-319-62434-1_14.

[21] ***Bulatov V, Alekseev V, Vorontsov K, Polyudova D, Veselova E, Goncharov A, Egorov E***. TopicNet: Making Additive Regularisation for Topic Modelling Accessible. In: Proc. of the 12th Language Resources and Evaluation Conf. European Language Resources Association, Marseille, France, 2020: 6745–6752 https://aclanthology.org/2020.lrec-1.833.pdf.

[22] ***Veselova E, Vorontsov K.*** Topic Balancing with Additive Regularization of Topic Models. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Online. Association for Computational Linguistics, 2020: 59–65. https://doi.org/10.18653/v1/2020.acl-srw.9.

[23] ***Irkhin IA, Vorontsov KV.*** Convergence of the algorithm of additive regularization of topic models. [In Russian]. Trudy Instituta Matematiki i Mekhaniki URO RAN. 2020; 3(26): 56–68. https://doi.org/10.21538/0134-4889-2020-26-3-56-68.

[24] ***Sukhareva AV, Vorontsov KV.*** Building a complete set of topics for probabilistic topic models [In Russian]. Intelligent systems. Theory and Applications. 2019; 4(23): 7-23.

[25] ***Blei DM, Ng AY, Jordan MI.*** Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 2003: 993-1022.

[26] ***Wallach HM, Mimno DM, McCallum A.*** Rethinking lda: Why priors matter. In NIPS, vol. 22, 2009: 1973–1981.

[27] ***Alekseev VA, Bulatov VG, Vorontsov KV.*** Intra-text coherence as a measure of topic models' interpretability. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018” (Moscow, May 30 - June 2, 2018). 2018: 1-13. https://www.dialog-21.ru/media/4281/alekseevva.pdf.

About the authors

Igor Olegovich Datyev (b. 1981) graduated from the Kola branch of Petrozavodsk State University in 2004. Cand. Sci. Eng. Scientific secretary at the Institute for Informatics and Mathematical Modeling of the KSC RAS. The author of more than 100 scientific papers in the field of development of models and technologies for regional information and communication systems. Author ID (RSCI): 180256; Author ID (Scopus): 56070103900; Researcher ID (WoS): J-1839-2018. *datyev@iimm.ru*. ✉

Andrei Mikhailovich Fedorov (b. 1978). graduated from the Kola branch of the Petrozavodsk State University (2000). Cand. Sci. Eng. (2005). A deputy director for research at the Institute for Informatics and Mathematical Modeling of the KSC RAS. Since 2000, he has been teaching at the Murmansk Arctic State University branch in the Apatity, associate professor of the Department of Informatics and Computer Engineering. The area of scientific interests is currently focused on the development of models and technologies for information support for regional management. Author ID (RSCI): 4285-9780; Author ID (Scopus): 57203929412; Researcher ID (WoS): D-5859-2016. *fedorov@iimm.ru.*

Received April 11, 2022. Revised May 12, 2022. Accepted June 6, 2022.