Полнота, устойчивость и интерпретируемость вероятностных тематических моделей

Сухарева А.В.[1]

Интерпретируемость решения, возможность обучения без учителя, масштабируемость сделали тематическое моделирование одним из наиболее популярных инструментов статистического анализа текстов. Тематические модели позволяют снизить размерность пространства данных, так как описывают каждый документ как вероятностную смесь абстрактных тем, каждую тему как распределение над словами словаря коллекции. Переход из пространства слов в пространство тем приводит к естественному решению проблем синонимии и полисемии терминов. Однако есть и ряд недостатков, вызванных зависимостью решения от инициализации. Неустойчивость тематических моделей являются общеизвестным фактом, однако связанная с ней проблема полноты до сих пор в литературе не изучалась. Для решения этой задачи в статье исследуется новый алгоритм нахождения полного набора тем, основанный на построении выпуклой оболочки. Экспериментально подтверждается эффективность данного алгоритма. На практике полный набор тем использовался в качестве инициализации модели ARTM (additive regularization for topic modeling). По сравнению с рандомизированным начальным приближением, базис тем позволяет повысить устойчивость, перплексию на более 10%, когерентность в разы.

Ключевые слова: вероятностное тематическое моделирование, устойчивость тематических моделей, полный набор тем тематических моделей, латентное размещение Дирихле, LDA, регуляризация, ARTM, BigARTM.

[1] *Сухарева Анжелика Вячеславовна* — аспирантка каф. математических методов прогнозирования факультета вычислительной математики и кибернетики МГУ имени М.В.Ломоносова, e-mail: ang.sukhareva@gmail.com.

Sukhareva Anjelika Vyacheslavovna — graduate student, Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, Chair of Mathematical Methods of Forecasting.

1. Введение

Краткая характеристика содержания документов является стандартной задачей, решаемой в области информационного поиска, статистической обработки естественного языка и машинного обучения. Такое представление можно использовать для хранения, классификации или поиска по коллекции документов. Для получения векторного представления документов часто применяются модели «мешок слов», Doc2Vec и Context2Vec. В последнее время возрос интерес к вероятностным тематическим моделям, которые описывают каждый документ как вероятностную смесь абстрактных тем, каждую тему как распределение над словами словаря коллекции. Наиболее популяными тематическими моделями стали модели pLSA (probabilistic latent semantic analysis) [1] и LDA (latent Dirichlet allocation) [2].

Построение тематической модели сводится к решению некорректно поставленной задачи неотрицательного матричного разложения. Множество её решений в общем случае бесконечно. Это приводит к неустойчивости вычислительных методов и зависимости решения от случайного начального приближения. Многократное построение модели по одной и той же коллекции может приводить к нахождению новых тем и, как следствие, неоднозначному представлению документов. Несмотря на важность требования устойчивости в задачах компьютерной лингвистики и информационного поиска, проблема до сих пор относительно мало изучена. В литературе определено понятие устойчивости [3], предлагаются меры устойчивости [4, 5, 6]. Также в работах рассматриваются модели, повышающие устойчивость.

Модель Clustered LDA [7] основана на неслучайной инициализации LDA. Идея заключается в том, чтобы сначала обучить набор моделей LDA, отличающихся случайной инициализацией. Затем выполнить кластеризацию набора тем из разных моделей. Этот кластерный набор тем формирует основу новой инициализации для LDA, которая запускается для создания модели. Интуиция заключается в том, что кластеризация является естественным способом комбинирования похожих тем. В статье показано, что размер кластера (размер — количество тем в кластере) коррелирует с повторяемостью связной темы. Темы, которые с высокой вероятностью встречаются в документах, как правило, очень повторяемы в разных сериях и, следовательно, образуют большие кластеры. И наоборот, темы, встречающиеся с меньшей вероятностью

в документах в одной модели, менее склонны иметь аналоги в других моделях.

В гранулированной LDA (GLDA) [8] используется локальная регуляризация плотности: слова в локальном контекстном окне данного слова имеют более высокую вероятность получить ту же тему, что и это слово. Предполагается, что слова, характерные для одной и той же темы, часто размещаются внутри некоторого относительно небольшого окна.

Для повышения устойчивости и интерпретируемости моделей в работе [9] применяется новый многокритериальный подход — аддитивная регуляризация тематических моделей (additive regularization for topic modeling, ARTM). Вводятся регуляризаторы для повышения разреженности и различности тем. В экспериментах показывается, что комбинирование этих регуляризаторов улучшает разреженность, чистоту и контрастность тем без значимого ухудшения правдоподобия модели. Регуляризация может существенно влиять на результаты тематического моделирования. Она может как улучшать, так и свести к минимуму устойчивость тематического моделирования [8].

TopicMapping [10] основан на неслучайной инициализации pLSA или LDA. Авторы экспериментально показали, что алгоритмы, разработанные для обнаружения сообществ в сетях, могут улучшать устойчивость тем.

В данной статье также ставится проблема полноты тематических моделей. Экспериментально исследуется взаимосвязь между полнотой, устойчивостью, интерпретируемостью и правдоподобием тематических моделей. Проблема полноты в литературе вообще не рассматривалась. Актуальность проблемы обусловлена большим числом прикладных задач анализа текстов, в которых требуется как можно полнее определить тематический состав коллекции документов.

Статья организована следующим образом. В разделе 2 мы вводим основные обозначения и терминологию, кратко описываем проблему устойчивости и полноты моделей. В разделе 3 и 4 будут исследованы устойчивость, полнота и интерпретируемость моделей. В разделе 3 вводится понятие устойчивости тем, рассматриваются различные меры устойчивости и интерпретируемости моделей. В разделе 4 вводится понятие полноты моделей. Для построения полного набора тем в разделе 4 предлагается алгоритм на основе выпуклой оболочки. Эмпирические результаты влияния полноты на устойчивость, интерпретируемость и правдоподобие моделей обсуждаются в разделе 5. Наконец, в заключении представлены наши выводы.

2. Постановка задачи

Введем следующие обозначения: $D = \{d_1, \ldots, d_{|D|}\}$ — коллекция текстовых документов, $W = \{w_1, \ldots, w_{|W|}\}$ — словарь термов, встретившихся в них, $T = \{t_1, \ldots, t_{|T|}\}$ — конечное множество тем. В качестве термов могут использоваться слова, n-граммы, коллокации, именованные сущности и т.д. Число тем является гиперпараметром и задается заранее ($|T| \ll |D|$). Каждое вхождение терма $w \in W$ в документ $d \in D$ связано с некоторой темой $t \in T$. Термы и документы являются наблюдаемыми переменными, темы — латентными (скрытыми).

Тематическая модель появления слов в документах выглядит следующим образом:

$$p(d, w) = p(d)p(w|d) = p(d)\sum_{t \in T} p(w|t)p(t|d) = p(d)\sum_{t \in T} \phi_{wt}\theta_{td}, \qquad (1)$$

где w — слово, t — тема, d — документ коллекции.

При построение тематической модели (1) требуется оценить по известной коллекции D параметры модели $\Phi = (\phi_{wt})_{W \times T}$ и $\Theta = (\theta_{td})_{T \times D}$. Построение вероятностной тематической модели является некорректно поставленной задачей стохастического матричного разложения. Множество ее решений в общем случае бесконечно:

$$F \approx \Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta', \qquad (2)$$
$$\Phi \neq \Phi', \Theta \neq \Theta',$$

где $F = \big(p(d, w)\big)_{W \times D}$ — известная матрица частот, S — произвольная невырожденная матрица, при условии, что все матрицы стохастические.

Решение будем искать с помощью максимизации логарифма правдоподобия:

$$L(\Phi, \Theta) = \sum_{d \in D}\sum_{w \in W} n_{dw} \log p(d, w) = \sum_{d \in D}\sum_{w \in W} n_{dw} \log \sum_{t \in T} \phi_{wt}\theta_{td} \to \max_{\Phi, \Theta} \qquad (3)$$

$$\sum_{w \in W} \phi_{wt} = 1, \phi_{wt} \geq 0, \quad \sum_{t \in T} \theta_{td} = 1, \theta_{td} \geq 0,$$

где $\Phi = (\phi_{wt})_{W \times T}$, $\Theta = (\theta_{td})_{T \times D}$, n_{dw} — число вхождений терма w в документ d.

Каждый раз модель находит локальный максимум правдопобия (3). Локальных экстремумов экспоненциально много и они различаются по глубине. Однако из локальных максимумов можно построить базис векторов тем тематических моделей, состоящий из линейно независимых, хорошо интерпретируемых тем. Базис будем считать *полным набором тем*, описывающим весь корпус текстов.

3. Устойчивость и интерпретируемость модели

Понятие *устойчивости тем* (stability topics) было введено в работе М. Стэйверса и Т. Гриффитса [3] в 2007 году. В статье замечено, что некоторые темы являются устойчивыми и появляются во всех моделях, другие же специфичны и встречаются только в данном решении. Авторы построили две вероятностные тематические модели, отличающиеся инициализацией. Затем измерялось различие между темами t и s с помощью симметричного расстояния Кульбака-Лейблера (Kullback-Leibler):

$$KL(t,s) = \frac{1}{2}\sum_{w\in W}\phi'_{wt}\log_2\frac{\phi'_{wt}}{\phi''_{ws}} + \frac{1}{2}\sum_{w\in W}\phi''_{ws}\log_2\frac{\phi''_{ws}}{\phi'_{wt}}, \qquad (4)$$

где ϕ', ϕ'' — оценки распределений слов по темам для двух моделей, отличающихся только случайным начальным приближением.

После чего векторы тем второй модели сопоставлялись венгерским алгоритмом [11] наиболее близким темам из первой модели. На основании экспериментов авторы сделали вывод, что по-разному проинициализированные модели дают разные решения, но многие темы устойчивы во всех запусках. Стоит отметить, что инициализация матрицы Φ сильно влияет на результат моделирования, в то время как инициализация матрицы Θ при фиксированной Φ — незначительно. Объясняется это тем, что правдоподобие (3) зависит больше от Φ, чем от Θ. На практике обычно матрица Φ инициализируется случайно. Лучшим алгоритмом инициализации матрицы Φ считается алгоритм Ароры [12].

Аналогично, в [4] предлагается рассматривать расстояние между векторами распределений тем по документам d и c двух моделей с различной случайной инициализацией:

$$KL(d,c) = \sum_{t\in T}\theta'_{td}\log_2\frac{\theta'_{td}}{\theta''_{tc}}.$$

Несмотря на широкое применение KL-дивергенции (4) как меры близости между двумя распределениями, она имеет ряд недостатков при измерении сходства тем. Во-первых, в величине KL-дивергенции доминирует длинный хвост слов с низкой вероятностью. Во-вторых, KL-дивергенция зависит от размера словаря. Это означает, что различные словари и пары тем будут давать широкий диапазон значений KL-дивергенции для двух разных тем. Чтобы сделать расстояние между двумя различными темами равным 1, авторы статьи [5] вводят нормированную KL-дивергенцию:

$$NKLS(t,s) = 1 - \frac{KL(t,s)}{\max\limits_{t',s'} KL(t',s')}.$$

Под *устойчивостью тематической модели* будем понимать способность модели сохранять построенные темы в разных запусках. Устойчивость модели может применяться как внутренний критерий качества вероятностных тематических моделей [4], наряду с перплексией. В отличие от перплексии устойчивость не зависит от словаря. Она желательна по нескольким причинам [7]:

1) высокая вариативность тем, возникающих в результате различных инициализаций, может означать, что любая отдельная модель может пропустить некоторые полезные темы;

2) существует корреляция между повторяемостью и интерпретируемостью темы, интерпретируемые темы, как правило, повторяются чаще;

3) улучшение устойчивости темы — один из возможных способов устранения зашумленных тем.

Устойчивость модели можно определить, вычислив все попарные KL-дивергенции тем (или расстояние Жаккара, Хеллингера и т.д.) и усреднив их. В табл. 1 сравниваются различные инициализации модели ARTM по основным критериям, в том числе и устойчивости.

В статье [6] предлагается другой подход к измерению устойчивости модели. Здесь темы рассматриваются как множества отранжированных слов. Для определения сходства между двумя темами вычисляется среднее расстояние Жаккара (Jaccard):

$$AJ(\phi', \phi'') = \frac{1}{n}\sum_{d=1}^{n} \gamma_d(\phi', \phi''), \quad \gamma_d = \frac{|\phi'_d \cap \phi''_d|}{|\phi'_d \cup \phi''_d|},$$

где ϕ'_d — первые d топ-слов темы ϕ', n — общее число топ-слов.

Авторы предлагают меру близости тем, основанную на расстоянии Жаккара:

$$agree(\Phi_x, \Phi_y) = \frac{1}{K} \sum_{i=1}^{K} AJ\big(\phi_{xt}, \pi(\phi_{xt})\big),$$

где $\pi(\phi_{xt})$ — отранжированный список тем Φ_y, поставленных в соответствие темам Φ_x перестановкой π. По аналогии с [3] темы сопоставляются венгерским алгоритмом.

Пусть Φ_0 — матрица слова-темы тематической модели с K темами, обученной на всей выборке документов, а $\{\Phi_1, \ldots, \Phi_\tau\}$ — матрицы слова-темы той же модели, полученные на различных подвыборках этих документов. Подвыборки генерируются случайно, без возвращения документов, доля выбранных документов среди всех документов коллекции может быть любой. Тогда устойчивость тематической модели определяется как:

$$stability(K) = \frac{1}{\tau} \sum_{i=1}^{\tau} agree(\Phi_0, \Phi_i). \tag{5}$$

Данный подход может использоваться для выбора числа тем K из некоторого диапазона $[K_{\min}, K_{\max}]$. Выбирается то значение, при котором модель будет наиболее устойчива к возмущениям в данных.

Устойчивость тесно связана с таким важнейшим свойством тематических моделей как интерпретируемость. Требование интерпретируемости тем плохо формализуется. На практике часто привлекаются асессоры, которые оценивают интерпретируемость тем по топу термов с помощью различных методик, например, можно ли по набору терминов понять, о чем речь в теме или добавляется лишнее слово из другой темы и экспертам предлагается его найти и т.д. Автоматической мерой интерпретируемости принято считать когерентность [13, 14], которая оценивает совместную встречаемость главных термов темы в одних и тех же контекстах. В данной статье для оценки интерпретируемости модели дополнительно к когерентности применяются критерии частоты и контрастности тем, предложенные в работе [9]. Они основаны на гипотезе о том, что в интерпретируемой теме должно хорошо выделяться семантическое ядро (множество слов, характерных для предметной области).

Терм w относится к ядру W_t темы t, если

$$p(w|t) > \frac{1}{|W|}. \tag{6}$$

Чистота темы вводится следующим образом:

$$\sum_{w \in W_t} p(w|t). \tag{7}$$

Контрастность темы вычисляется как:

$$\frac{1}{|W_t|} \sum_{w \in W_t} p(t|w). \tag{8}$$

Чем выше чистота и контрастность тем, тем лучше.

4. Алгоритм построения полного набора тем

Из неустойчивости тем (2) следует проблема полноты набора тем, найденного тематической моделью. Возникают следующие вопросы:

- действительно ли темы новые или это комбинации предыдущих;
- можно ли найти все темы корпуса, сколько моделей для этого нужно построить.

Для ответов на эти вопросы рассмотрим алгоритм построения базиса тем моделей [15], отличающихся инициализацией. Все описанные выше модели — pLSA, ARTM и LDA — находят темы в пространстве распределений над словами. Каждое такое распределение можно рассматривать как точку в единичном $(|W|-1)$-симплексе слов $\triangle$.

Определение. *Множество $V = \{v_1, \ldots, v_m\} \subset \triangle$ — базис тем множества матриц тематических моделей $\Phi_1, \ldots, \Phi_n \subset \triangle$, если $\forall \phi \in \Phi_j$ выполняется:*

$$\min_{v \in conv V} \rho(\phi, v) \leq \varepsilon, \tag{9}$$

$$conv V = \Big\{ v = \sum_{i=1}^{m} \alpha_i v_i \big| v_i \in V, \quad \sum_i \alpha_i = 1, \quad \alpha_i \geq 0 \Big\},$$

где $\rho(\phi, v)$ — функция расстояния.

Из определения следует, что базис V состоит из векторов тем $\phi \in \Phi_j, \quad j = \overline{1,n}$, для которых $\min\limits_{v \in convV} \rho(\phi, v) > \varepsilon$, именно они линейно независимы. Кроме того, решение каждой отдельной тематической модели локально оптимально, а значит, его можно улучшить с помощью замены тем на близкие им в случае повышения правдоподобия (3). На этом и основан данный жадный алгоритм для нахождения базиса тем.

Алгоритм состоит из чередования двух этапов. На первом этапе происходит замена тем с целью расширения имеющейся системы векторов тем и максимизации правдоподобия. Алгоритм итеративный, на каждой итерации для тем набора находятся схожие с некоторым порогом γ и выполняется замена, если она улучшает правдоподобие всего набора тем. На втором этапе происходит поиск темы для добавления в набор:

- включаются линейно независимые темы;
- исключаются выбросы (выбросами считались темы, которые имеют более двух коэффициентов $\alpha_i > 0.15$ в (9), что соответствовало несогласованным темам, зависимым нелинейно от тем полного набора);
- выбирается тема, максимально повышающая правдоподобие набора тем.

Таким образом, мы дополняем линейно независимую систему векторов до базиса. Описанные шаги повторяются до сходимости. Так как алгоритм итеративный, то после того, как алгоритм сошелся, нужно выполнить процедуру замены близких тем, используя обученные модели в обратном порядке.

Для решения оптимизационной задачи (9) использовался алгоритм SLSQP (Sequential Least SQuares Programming) [16], который реализован во многих популярных математических пакетах, в том числе и SciPy.

5. Эмпирические результаты

В этом разделе мы приводим результаты работы алгоритмов на реальных данных коллекции ПостНаука[1]. Коллекция ПостНаука — небольшой корпус текстов интернет-журнала ПостНаука, состоящий из

[1] https://postnauka.ru/

научно-популярных статей о современной фундаментальной науке и учёных, которые её создают.

В экспериментах, описанных ниже, использовались теги к документам коллекции ПостНаука, размеченные экспертами. Всего было 930 тегов. Они обозначали общие и частные понятия, например, биология, клетка, эволюционная биология, ген, эукариоты, квантовая физика, Россия, человек и т.д. Каждый документ в среднем описывался 6 тегами. Данные были случайно разделены на обучающую и контрольную выборки в отношении 80% к 20%.

5.1. Моделирование документов

Для построения полного набора использовались модели LDA (Gibbs sampling) с 20 темами. LDA[2] с параметрами $\eta = 0.01$, $\alpha = \frac{1}{20}$ обучалась 3500 итераций. Всего было построено около 2000 моделей. В полный набор Φ_0 были отобраны 23 темы.

Затем для изучения полноты, устойчивости и интерпретируемости тематических моделей обучалась модель ARTM с 23 темами. В экспериментах модель была проинициализирована двумя способами: случайно и базисом тем, который был получен, как описано выше. Во втором случае перплексия повышалась при добавлении шумовой компоненты:

$$\phi_\alpha = \alpha\phi_t + (1-\alpha)\phi'_t,$$

где $\phi_t \in \Phi_0$, $\phi'_t \sim \mathrm{Dir}(\beta)$ — шум.

В экспериментах на тегах коллекции ПостНаука брались $\alpha = 0.5$ и $\beta = 10$ для сглаживания Φ_0. Встряхивание весов позволило избежать переобучения.

При построении ARTM применялся регуляризатор разреживания матрицы Θ с различными коэффициентами. Модель ARTM строилась с помощью библиотеки BigARTM[3].

5.2. Влияние полноты на устойчивость и интерпретируемость тематических моделей

Для изучения связи между полнотой, устойчивостью и интерпретируемостью тематических моделей мы провели такой эксперимент:

[2]https://github.com/lda-project/lda
[3]https://github.com/bigartm/bigartm

- построили базис тем для моделей LDA с 20 темами, алгоритм отобрал 23 темы;
- сравнили качество моделирования модели ARTM при фиксированной матрице Φ, матрица Θ бралась случайной. Рассмотрели два случая. В первом случае матрица Φ была случайной из равномерного распределения, во втором — проинициализирована полным набором тем.

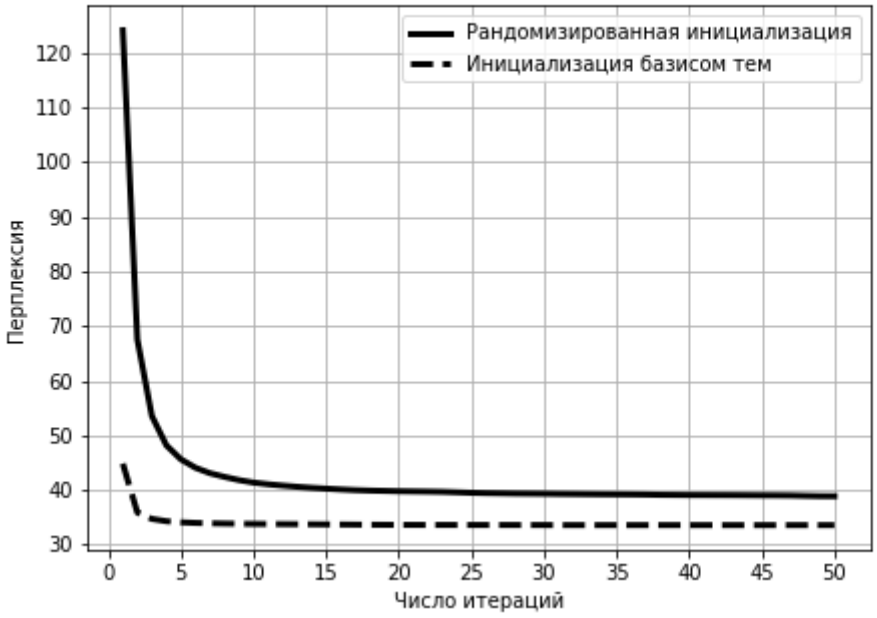

Рис. 1. Обучение модели ARTM, проинициализированной случайно и полным набором тем.

Процесс обучения модели ARTM проиллюстрирован на рис. 1. На графике видно, что модель сошлась по перплексии. Инициализация модели полным набором позволила повысить перплексию на 13%.

Критерий	ARTM (случайно)	ARTM (базис тем)
Перплексия на обучении	38.44	33.53
Перплексия на контроле	15.73	14.20
Устойчивость тем по Θ	$1.59 \cdot 10^{-8}$	$7.77 \cdot 10^{-9}$
Устойчивость тем по Φ	$6.32 \cdot 10^{-8}$	$1.83 \cdot 10^{-8}$
Полнота	0.4	1
Когерентность	-8.21	-3.32
Средний размер ядра	40.47	39.78
Средняя контрастность	0.95	0.93
Средняя чистота	0.9	0.85
Разреженность Φ	0.94	0.94
Разреженность Θ	0.89	0.91

Таблица 1. Сравнение результатов работы тематической модели подхода ARTM, проинициализированной случайно (первый столбец) и полным набором (второй столбец). Качество оценивалось по основным критериям, связанным с полнотой, устойчивостью и интерпретируемостью моделей. В качестве меры близости тем бралось расстояние Хеллингера. Темы считались близкими, если расстояние Хеллингера меньше 0.5.

В табл. 1 приводятся критерии, по которым велось сравнение случайной инициализации с базисом тем. Критерий устойчивости тем по матрице Φ (документов по матрице Θ) определялся как среднее расстояние Хеллингера между соответствующими распределениями тем (документов). Видно, что при фиксированной матрице Φ модель устойчива как по матрице Φ, так и по матрице Θ. Также способствует повышению устойчивости сильная разреженность матриц Φ и Θ. Однако полный набор позволяет строить решение на порядок более устойчивое, чем случайное начальное приближение.

Полнота тематической модели определялась как доля тем близких по расстоянию Хеллингера к темам базиса. Эксперимент показал, что около 40% тем модели похожи на темы базиса, остальные являются их комбинациями. Часто смешиваются абсолютно разные тематики. Это приводит к несогласованности тем, например: биология, эволюция, микробиология, экология, микробы, палеонтология, бактерии, геология,

география, земля. Следует отметить, что все темы базиса удалось найти за 20 моделей ARTM с 23 темами.

Полный набор тем позволяет строить не только более устойчивые модели, но и повышает в разы интерпретируемость (когерентность) тем. При этом наблюдается незначительное уменьшение размера ядра (6), чистоты (7) и контрастности тем (8). Темы полного набора более крупные и общие, чем при случайной инициализации. Более того, содержание тем полного набора не изменилось при моделировании.

6. Заключение

Данная работа посвящена изучению взаимосвязи устойчивости, полноты и интерпретируемости тематических моделей. Под устойчивостью тематической модели будем понимать способность модели сохранять построенные темы в разных запусках. Полнота тематической модели определялась как доля тем близких по расстоянию Хеллингера к темам базиса. Базис строился с помощью алгоритма на основе выпуклой оболочки из тем моделей, отличающихся инициализацией матрицы Ф. Свойство интерпретируемости тематических моделей плохо формализуется. На практике часто привлекаются эксперты для оценки того, насколько понятна тема людям и можно ли дать ей название. Оценки экспертов хорошо коррелируют с основной мерой интерпретируемости — когерентностью.

Неустойчивость тематических моделей являются общеизвестным фактом, однако связанная с ней проблема полноты до сих пор в литературе не изучалась. Для решения этой задачи в статье рассматривается алгоритм нахождения полного набора тем, основанный на построении выпуклой оболочки векторов тем тематических моделей, отличающихся только инициализацией матрицы Ф. Алгоритм состоит из двух процедур — замены близких тем и добавления новой темы — которые выполняются по очереди до сходимости алгоритма. Таким образом, происходит приближенное дополнение линейно независимой системы до базиса. По построению в базис добавляются только те новые темы, в δ-окрестности которых нет выпуклых комбинаций тем базиса.

Полученный базис тем использовался для инициализации модели ARTM. В статье был проведен сравнительный анализ инициализации модели ARTM полным набором и случайно по критериям, связанным с исследуемыми проблемами. Инициализация модели полным набором повышает перплексию на 13% и увеличивает скорость сходимости модели. Касательно устойчивости видно, что при фиксированной матрице Ф мо-

дель устойчива как по матрице Φ, так и по матрице Θ. Также способствует повышению устойчивости сильная разреженность матриц Φ и Θ. Однако полный набор позволяет строить решение на порядок более устойчивое, чем случайное начальное приближение.

Эксперимент показал, что около 40% тем модели ARTM похожи на темы базиса, остальные являются их комбинациями. Часто смешиваются разные тематики, что приводит к несогласованности тем.

Полный набор тем позволяет строить не только более устойчивые модели, но и повышает в разы интерпретируемость (когерентность) тем. Темы полного набора более крупные и общие, чем при случайной инициализации. Более того, содержание тем полного набора не изменилось при моделировании.

Таким образом, можно рекомендовать полный набор в качестве хорошего начального приближения матрицы Φ тематической модели. В дальнейшем планируется разработать модель, которая будет строить полный набор тем за меньшее число итераций.

Список литературы

[1] Hofmann T., "Probabilistic latent semantic indexing", *ACM SIGIR Forum*, **51**:2 (2017), 211–218.

[2] Blei D.M., Ng A.Y., Jordan M.I., "Latent dirichlet allocation", *Journal of machine Learning research*, **3**:Jan (2003), 993–1022.

[3] Steyvers M., Griffiths T., "Probabilistic topic models", *Handbook of latent semantic analysis*, **427**:7 (2007), 424–440.

[4] De Waal A., Barnard E., "Evaluating topic models with stability", 2008.

[5] Koltcov S., Koltsova O., Nikolenko S., "Latent dirichlet allocation: stability and applications to studies of user-generated content", *Proceedings of the 2014 ACM conference on Web science*, "ACM", 2014, 161–165.

[6] Greene D., O'Callaghan D., Cunningham P., "How many topics? stability analysis for topic models", *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, "Springer", Berlin, Heidelberg, 2014, 498–513.

[7] Balagopalan A., "Improving topic reproducibility in topic models.", 2012.

[8] Koltcov S. et al., "Stable topic modeling with local density regularization", *International Conference on Internet Science*, "Springer", 2016, 176–188.

[9] Vorontsov K., Potapenko A., "Additive regularization of topic models", *Machine Learning*, **101**:1–3 (2015), 303–323.

[10] Lancichinetti A. et al., "High-reproducibility and high-accuracy method for automated topic classification", *Physical Review X.*, **5**:1 (2015), 011007.

[11] Kuhn, Harold W., "The Hungarian method for the assignment problem", *Naval research logistics quarterly*, **2**:1-2 (1955), 83–97.

[12] Arora S. et al., “A practical algorithm for topic modeling with provable guarantees”, *International Conference on Machine Learning*, 2013, 280–288.

[13] Newman D. et al., “Automatic evaluation of topic coherence”, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, "Association for Computational Linguistics", 2010, 100–108.

[14] Mimno D. et al., “Optimizing semantic coherence in topic models”, *Proceedings of the conference on empirical methods in natural language processing*, "Association for Computational Linguistics", 2011, 262–272.

[15] Сухарева А.В., Воронцов К.В., “Построение полного набора тем вероятностных тематических моделей”, *Интеллектуальные системы. Теория и приложения*, **24**:4 (2019).

[16] Kraft D., “A software package for sequential quadratic programming”, *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

Completeness, stability and interpretability of probabilistic topic models

Sukhareva A.V.

Interpretability of the solution, the possibility of unsupervised learning, scalability made topic modeling one of the most popular tools for statistical text analysis. Topic models make it possible to reduce the dimension of the data space, since they describe each document as a probabilistic mixture of abstract topics, each topic as distribution over the vocabulary words of a collection. The transition from the space of words into the space of topics leads to a natural solution of the problems of synonymy and polysemy of terms. However, there are a number of disadvantages caused by the dependence of the solution on the initialization. The instability of topic models is a well-known fact, but the problem of completeness related to it is still not studied in the literature. To solve this problem, the article explores a new algorithm for finding a complete set of topics based on the building of the convex hull. Experimentally confirmed the effectiveness of this algorithm. In practice, a complete set of topics was used as the initialization of the ARTM (additive regularization for topic modeling) model. Compared with the randomized initial approximation, the basis topics allows to increase stability, perplexity by more than 10%, coherence by several times.

Keywords: probabilistic topic modeling, stability of topic models, complete set of topics of topic models, latent Dirichlet allocation, LDA, regularization, ARTM, BigARTM.