ОПУБЛИКОВАТЬ СТАТЬЮ

в изданиях НИЦ "Социосфера"

СОЦИОСФЕРА

- *Российский научный журнал*
- *ISSN 2078-7081*
- *РИНЦ*
- *Публикуются статьи по социально-гуманитарным наукам*

ПОДРОБНЕЕ

PARADIGMATA POZNÁNÍ

- *Чешский научный журнал*
- *ISSN 2336-2642*
- *Публикуются статьи по социально-гуманитарным, техническим и естественно-научным дисциплинам*

ПОДРОБНЕЕ

СБОРНИКИ КОНФЕРЕНЦИЙ

- *Широкий спектр тем международных конференций*
- *Издание сборника в Праге*
- *Публикуются материалы по информатике, истории, культурологии, медицине, педагогике, политологии, праву, психологии, религиоведению, социологии, технике, филологии, философии, экологии, экономике*

ПОДРОБНЕЕ

УДК 930.2

ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ В ИСТОРИЧЕСКОЙ НАУКЕ

А. В. Кузнецов

Кандидат исторических наук,
ORCID 0000-0003-4755-250X,
e-mail: historyras@gmail.com,
Институт всеобщей истории РАН,
г. Москва, Россия

TOPIC MODELING IN HISTORICAL SCIENCES

A. V. Kuznetsov

Candidate of Historical Sciences,
ORCID 0000-0003-4755-250X,
e-mail: historyras@gmail.com,
Institute of World History of RAS,
Moscow, Russia

Abstract. Topic modeling is a widely used technique in the digital humanities and digital history to explore the thematic structure of collection of documents. The most popular algorithm for topic modeling is Latent Dirichlet Allocation (LDA), which was proposed by David Blei, Andrew Ng and Michael Jordan in 2003. LDA is an unsupervised machine learning algorithm in which topics are represented as set of words, and documents are represented as a collection of topics. This article provides an overview of the technical foundations and assumptions underlying the LDA procedure, and on this basis discusses the possibilities and limitations related to historical research.

Keywords: digital history; topic modeling; computational text analysis; Latent Dirichlet Allocation.

Стремительное развитие информационных технологий в XXI веке коренным образом изменило характер исследований в гуманитарных науках в целом и в частности в исторической науке. Исследователям доступно всё большее число цифровых источников, вместе с этим существенно расширился набор инструментов и методов автоматического анализа текстов. Одним из наиболее популярных стал метод тематического моделирования (англ. topic modeling). Под тематическим моделированием понимается метод машинного обучения, который определяет, к каким темам относится каждый документ текстовой коллекции и какие слова (термины) образуют каждую тему. «Тема» в тематическом моделировании – это «повторяющийся набор совместно встречающихся слов» [12, с. 12]. В настоящее время разработано множество конкретных вариантов построения тематических моделей [4, с. 63; 9, с. 410], но наибольшую популярность в гуманитарных науках нашел алгоритм латентного размещения Дирихле, предложенный в 2003 году [7]. Его успех можно объяснить наличием большого количества готовых к использованию и хорошо документированных инструментов.

В данной статье предполагается на основе особенностей работы алгоритма латентного размещения Дирихле проанализировать потенциал и ограничения его применения в исторических исследованиях.

Тематическое моделирование оказалось полезным инструментом в самых разных областях знания [10]. Его очевидным преимуществом является возможность проанализировать тематическую структуру большого объема текстов. Второе преимущество заключается в том, что

тематическое моделирование легко комбинируется с другими методами, а полученные темы становятся признаками в реализации иных алгоритмов [11, с. 227]. Первым историческим исследованием, использующим это метод, стала статья 2006 года «Вероятностная тематическая декомпозиция американской газеты восемнадцатого века» Д. Ньюмана и Ш. Блок, посвященная анализу тематики газеты Pennsylvania Gazette в период с 1728 по 1800 год [19]. С 2010 года наблюдается рост интереса к тематическому моделированию со стороны гуманитариев, в том числе историков [22, с. 2]. Чаще всего тематическое моделирование применяется для анализа дневников [8], писем [18], периодических изданий [19; 21], сборников статей [14], хроник [3], записей парламентских дебатов [16], судебных решений [15]. Тематическое моделирование в гуманитарных дисциплинах рассматривают как пример «дальнего чтения» (англ. distant reading) – подхода литературоведа Ф. Моретти [5], основанного на анализе больших текстовых корпусов, подсчете статистики, построении графов и противопоставляемого привычному «пристальному чтению» (англ. close reading) [4].

Алгоритм латентного размещения Дирихле относится к методам машинного обучения без учителя. Как и у большинства методов машинного обучения, результаты его выполнения зависят от ряда шагов, которые должны быть выполнены исследователем в процессе моделирования. Во-первых, корпус текстов должен быть предварительно обработан, чтобы преобразовать неструктурированные тексты в структурированный набор данных пригодный для компьютерного анализа [13]. Во-вторых, исследователь должен установить параметры моделирования: количество создаваемых тем, а также параметры α-альфа, управляющий вероятностным распределением тем по документам, и параметр β-бета, определяющий распределение слов по темам [6].

В настоящее время не выработано однозначных рекомендаций по предобработке текста и подбору параметров моделирования. Выбор необходимого количества тем – одна из самых больших проблем тематического моделирования. Их оптимальное количество зависит от объема данных, поставленных задач и содержания текста. При большом корпусе разнообразных текстов следует ожидать большого количества тем. Если набор данных небольшой, то и количество тем будет сравнительно небольшим [17]. Алгоритм латентного размещения Дирихле весьма чувствителен также и к настройкам параметров альфа и бета. При высоком значении альфа документ соотносится с несколькими темами, больше документов классифицируются как похожие друг на друга, и наоборот. Высокое значение бета приводит к тому, что слова часто повторяются в разных темах, и темы будут излишне похожими.

Результатом тематического моделирования становится построение двух вероятностных распределений: распределение слов по темам и распределение тем по документам. Распределение слов по темам представляет собой списки слов, наиболее вероятных для каждой темы, ранжированные в зависимости от их значимости. Распределение тем по документам представляет собой перечень документов с указанием присутствующих в них тем в процентном отношении. Необходимо подчеркнуть, что тематическое моделирование является вероятностным методом. Результаты сильно различаются в зависимости от предварительной обработки данных и выбранных параметров. Более того, тематическая модель, построенная на одном и том же наборе данных и с одинаковыми настройками несколько раз, может дать разные результаты [20, с. 437]. В такой неустойчивости результатов моделирования заключается основное ограничение метода тематического моделирования.

Полученные распределения слов и тем не могут быть итогом исследования, но лишь его промежуточным результатом, основой последующего анализа. Построив тематическую модель, исследователю необходимо оценить её качество, а полученные темы интерпретировать. В настоящее время предложено множество критериев оценки качества моделей [2, с. 245–249], но основным критерием стала именно интерпретируемость тем [9, с. 410]. Тема считается интерпретируемой, если по списку наиболее частотных слов и документов темы, специалист в данной предметной области сможет понять, о чём эта тема и дать ей адекватное название [1, с. 709].

Сложность интерпретации тематических моделей исторических текстов хорошо иллюстрирует пример модели дневника акушерки Марты Баллард (1734/1735–1812), [8] Так в списке наиболее частотных слов для темы «Смерть» семантически значимым становится только шестое по счету слово – «death»: day, yesterday, informd, morn, years, death, ye, hear, expired, expird, weak, dead, las, past, heard, days, drowned, departed, evinn. Как заметил К. Блевинс, более значимым для этой темы были слова «сообщил» (informed) или «слышала» (hear), поскольку Марта Баллард в дневнике в основном пишет о смерти в контексте новостей, распространяемых посредством личного общения. В качестве ещё одного примера укажем на тематическую модель небольшого корпуса текстов, состоящего из газетных публикаций, посвященных визиту в СССР Ива Монтана [17]. Исследователи выделили 10 тем. Первая тема включала следующие слова: Монтан, Ив, Москва, Синьоре, Симона, СССР, Франция, Париж, певец, вестибюль, декабрь, связи, вечер, пресса, газета, зал, смена, Шурупова. Только внимательное прочтение текстов позволило интерпретировать данное сочетание слов как тему «Прием И. Монтана» в СССР в декабре 1956 года. Советские газеты писали, что люди с нетерпением ожидали возможности купить билеты на его концерты. Слова «вестибюль», «зал», «смена» и «Шурупова» относятся к конкретной статье, в которой рассказывалось об очередях за билетами [20, p. 434].

Эти примеры ярко демонстрируют необходимость осведомленности исследователя в контексте создания изучаемых источников, невозможность оценить точность составленной тематической модели и интерпретировать полученные результаты без традиционного «пристального чтения» текстов. Существенным подспорьем в оценке и понимании тематических моделей служат современные, в том числе интерактивные визуализации тем [4, с. 65; 10, с. 185–190].

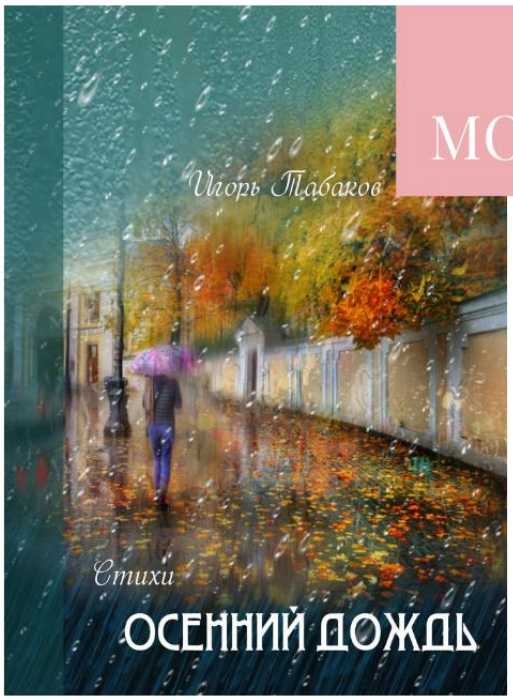
Тематическое моделирование, таким образом, является полезным методом для анализа больших массивов текстовых данных и помогает обнаруживать закономерности, которые не являются очевидными для читателя-человека. Результаты тематического моделирования сильно различаются в зависимости от предварительной обработки корпуса текстов и выбранных параметров моделирования. Построение интерпретируемой тематической модели зависит как от наличия у исследователя знаний о технических аспектах работы алгоритма, так и понимания контекста создания и природы анализируемых исторических источников. В настоящее время в исторической науке тематическое моделирование не может рассматриваться как самостоятельный метод исследования, но лишь в сочетании с традиционным «пристальным чтением» становится ценным способом изучения текстовых источников. При таком подходе количественный метод дополняет качественное исследование, выводя его на новый уровень анализа.

Библиографический список

1. 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.). – 2014. – №. 13. – С. 707–719.
2. Воронцов К.В. Обзор вероятностных тематических моделей // Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие. – М.: Изд-во НИУ ВШЭ, 2017. – С. 195–268.
3. Кузнецов А. В. Компьютерный анализ текстов на латинском языке: тематическое моделирование «Истории готов, вандалов и свевов» Исидора Севильского // Электронный научно-образовательный журнал «История». 2020. Т. 11. Выпуск 3 (89) [Электронный ресурс]. URL: https://history.jes.su/s207987840009681-8-1 (дата обращения: 25.10.2022).
4. Милкова М. А. Тематические модели как инструмент «дальнего чтения» // Цифровая экономика. – 2019. – №. 1. – С. 57–70.
5. Моретти Ф. Дальнее чтение. – М., 2016. – 352 с.
6. Binkley D., Heinz D., Lawriey D., Overfelt J. Understanding LDA in source code analysis // Proceedings of the 22nd international conference on program comprehension. – ACM, 2014. Pp. 26–36.
7. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. Vol. 3. Pp. 993–1022.
8. Blevins C. Topic Modeling Martha Ballard's Diary, April 1, 2010 [Электронный ресурс]. URL: http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/ (дата обращения 25.10.2022).
9. Bodrunova S. S. Topic Modeling in Russia: Current Approaches and Issues in Methodology // The Palgrave Handbook of Digital Russia Studies. – 2021. – Pp. 409–426.
10. Boyd-Graber J., Hu Y., Mimno D. Applications of Topic Models // Foundations and Trends® in Information Retrieval. 2017. Vol. 11: №. 2/3. – Pp. 143–296.
11. Boyd-Graber J., Mimno D., Newman D. Care and feeding of topic models: Problems, diagnostics, and improvements // Handbook of mixed membership models and their applications. – 2014. – Pp. 225–255.
12. Brett M. R. Topic modeling: A basic introduction // Journal of digital humanities. – 2012. – Т. 2. – №. 1. – Pp. 12–16.
13. Denny M. J., Spirling A. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it // Political Analysis. – 2018. – Т. 26. – № 2. – Pp. 168–189.
14. Goldstone A., Underwood T. The quiet transformations of literary studies: What thirteen thousand scholars could tell us // New Literary History. – 2014. – Т. 45. – № 3. – Pp. 359–384.
15. Grajzl P., Murrell P. A machine-learning history of English caselaw and legal ideas prior to the Industrial Revolution I: generating and interpreting the estimates // Journal of Institutional Economics. – 2021. – Т. 17. – №. 1. – Pp. 1–19.
16. Guldi J. Parliament's debates about infrastructure: an exercise in using dynamic topic models to synthesize historical change // Technology and Culture. – 2019. – Т. 60. – №. 1. – Pp. 1–33.
17. Johnson B., Oiva M., Salmi H. Yves Montand in the USSR. Mixed Messages of Post-Stalinist/Western Cultural Encounters // Entangled East and West: Cultural Diplomacy and Artistic Interaction during the Cold War. – 2019. – Pp. 241–261.
18. McGillivray B., Buning B., Hengchen S. Topic Modelling: Hartlib's Correspondence before and after 1650 // Reassembling the Republic of Letters in the Digital Age. Göttingen, 2019. Pp. 426–428.
19. Newman D. J., Block Sh. Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper // Journal of the American Society for Information Science and Technology Volume 57, Issue 6. 2006. Pp. 753–767.
20. Oiva M. Topic Modeling Russian History // The Palgrave Handbook of Digital Russia Studies. – Palgrave Macmillan, Cham, 2021. – Pp. 427–442.
21. Wehrheim L. Economic History Goes Digital: Topic Modeling the Journal of Economic History // Cliometrica. January 2019. Vol. 13. Issue 1. Pp 83–125.
22. Weingart S. B., Meeks E. The Digital Humanities Contribution to Topic Modeling // The Journal of Digital Humanities. Vol. 2 (1). Winter 2012. Pp. 2–6.

© Кузнецов А. В., 2022

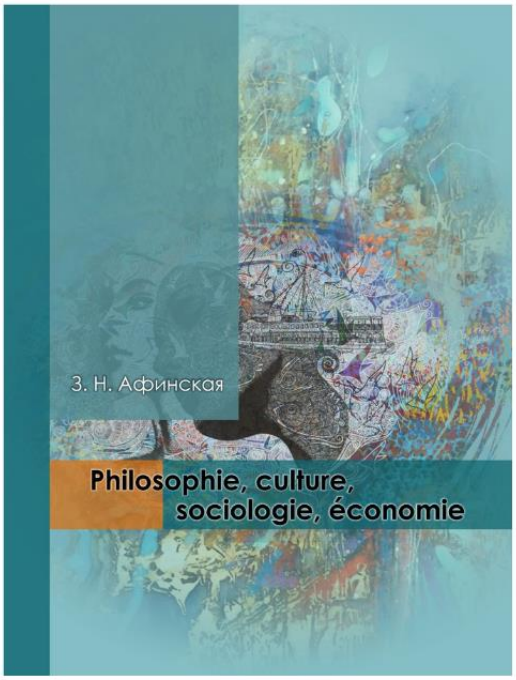


СРОЧНОЕ ИЗДАНИЕ МОНОГРАФИЙ И ДРУГИХ КНИГ

Два места издания Чехия или Россия.
В выходных данных издания
будет значиться
Прага: Vědecko vydavatelské centrum "Sociosféra-CZ"
или
Пенза: Научно-издательский центр "Социосфера"

РАССЧИТАТЬ СТОИМОСТЬ

- Корректура текста
- Изготовление оригинал-макета
- Дизайн обложки
- Присвоение ISBN



У НАС ДЕШЕВЛЕ

- Печать тиража в типографии
- Обязательная рассылка
- Отсулка тиража автору