Тематическое моделирование и кластеризация текстов на арабском языке

Н. А. Кижаева, аспирант[1]
Санкт-Петербургский государственный университет
"Лаборатория анализа и моделирования социальных процессов: политический ислам / исламизм: теория и практика в сравнительной и исторической перспективе"
natalia.kizhaeva@gmail.com

В связи с возрастающим количеством электронных книг, журналов и газет возникает необходимость в эффективных методах их классификации, индексирования и суммаризации. Полученные структурированные данные могут быть использованы в дальнейших исследованиях в области гуманитарных наук. В работе проведен обзор возможных подходов к решению задачи, включая тематическое моделирование и кластеризацию на основе "опознания по сжатию".

Ключевые слова: тематическое моделирование, арабский язык, машинное обучение, кластеризация.

1. Введение

Использование вычислительных методов в исследованиях в области социальных и гуманитарных наук становится все более распространенным, так как возрастает объем данных, охватывающий сферы человеческого общения (включая тексты, аудио, видео и т. п.). Автоматизированный анализ позволяет исследователям не только собирать и изучать количество материала, анализ которого вручную невозможен, но и выявлять закономерности, незаметные при простом прочтении.

Представляет интерес семантический анализ текстов на арабском языке в широком временном диапазоне. Ставится цель определить развитие и преобразование важнейших концептов, для достижения которой возможно применение различных подходов.

Тематическое моделирование (topic modeling) — область машинного обучения, находящая применение в анализе текстов. Тематическая модель коллекции текстовых документов определяет, к каким темам относится каждый документ в тексте и какие слова (термины) образуют каждую тему [1].

2. Математические модели

В основе большинства методов информационного поиска лежит векторное представление текстов [2], в котором координаты вектора соответствуют словам, а значения — некоторым статистическим характеристикам. Таким образом, каждый документ в коллекции является вектором "терм-частота". Классический метод информационного поиска *tf-idf* (от англ. *tf* — term frequency, *idf* — inverse document frequency) использует векторное представление документов, сопоставляя каждому слову в векторе некоторый вес:

$$\textit{tf-idf} = tf(t,d) \times idf(t,D), \tag{1}$$

где tf — нормализованная частота слова в тексте:

$$tf = \frac{freq(t,d)}{max_{w \in D} freq(w,d)}, \tag{2}$$

idf – обратная частота документов

$$idf(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}. \tag{3}$$

В формуле (1) $freq(t,d)$ — число вхождений слова t в документе d, а в (3) числитель — общее число документов в коллекции, знаменатель – количество документов, в которых встречается слово t.

В тематическом моделировании документ в коллекции представляется вектором тем. Происходит уменьшение размерности данных, т. к. количество тем меньше количества различных слов в коллекции документов.

Пусть задана коллекция документов D. Каждый документ d представляет собой последовательность слов $\mathbf{w} = (w_1, \ldots, w_{n_d})$ из словаря $\mathbb{W}$, где n_d — длина документа d. В основе вероятностных тематических моделей лежат следующие предположения:

1. Предполагается, что существует конечное множество тем T, и коллекция порождается дискретным распределением $p(d,w,t)$. Переменные d и w являются наблюдаемыми, t — скрытой.

2. Предполагается, что условное распределение вероятностей терминов $p(w|t,d)$ зависит только от тем, но не от документа d (гипотеза условной независимости): $p(w|d,t) = p(w|t)$.

3. Порядок слов в документе и порядок документов в коллекции не важны для выявления тематики (*гипотеза "bag or words"*).

Построить тематическую модель текста — значит найти распределение терминов в темах $\phi_{wd} \equiv p(w|t)$ и распределение тем в документах $\theta_{td} \equiv p(t|d)$.

Поиск документа по короткому запросу заключается в нахождении такого вектора, в котором часто встречаются слова из запроса. Основными проблемами такого подхода являются проблемы синонимии и полисемии. Для решения этих проблем вводится понятие терм-документной матрицы — матричное представление коллекции документов, совокупность векторов терм-частота.

В 1988 Dumais *et al.* [3] предложили метод *латентно-семантического анализа (Latent Semantic Analysis, LSA)*, который отходит от концепции простого поиска совпадений слов. Его основная идея заключается в оценивании корреляции терминов путем анализа их совместного появления в документах. Такое представление связи терм-документ (так называемой семантической структуры) становится ближе к понятию *темы*. Например, LSA позволяет вычислить похожи ли тематически два документа, даже если в них нет общих слов. Повышение эффективности информационного поиска достигается за счет понижения размерности: документы и термины проецируются в некоторое пространство, представляющее семантическую структуру коллекции документов и имеющее меньшую размерность в сравнении с исходной терм-документной матрицей. В основе распространенных реализаций LSA лежит *разложение матрицы по сингулярным значениям (Singular Value Decomposition, SVD)* или *неотрицательная матричная факторизация (non-negative matrix factorization, NMF)* [4].

Дальнейшим развитием латентно-семантического анализа является вероятностный латентно-семантический анализ *(probabilictic Latent Semantic Analysis, pLSA)*, впервые предложенный Томасом Хофманном в 1999 году [5]. Основное отличие метода заключается в моделировании вероятности тем в документах и слов в теме, что позволяет относить тот или иной документ к нескольким темам (с некоторой вероятностью). Вероятностный LSA имеет явные преимущества перед ранними подходами, но тем не менее, можно выделить следующие недостатки: модель склонна к переобучению (в следствие того, что число параметров растет в линейной зависимости от числа документов в коллекции), невозможно вычислить

вероятность "нового" документа, отсутствует закономерность в генеративном процессе (генерации документа из сочетания полученных тем).

Перечисленные недостатки были устранены в модели *скрытого размещения Дирихле (Latent Dirichlet Allocation)*, считающейся основной для вероятностного тематического моделирования [6].

2.1. Кластеризация и "опознание по сжатию"

Как уже упоминалось, традиционным является представление документа в виде вектора терм-частота. Одним из упрощающих предположений алгоритмов обработки естественного языка является гипотеза "мешка со словами" ("bag of words"), означающая, что порядок слов в документе не имеет значения. Детальное описание алгоритма формирования векторов и назначения весов дано в соответствующей главе книги [7].

Результирующее представление документа — разреженный вектор высокой размерности. Это может препятствовать построению хорошего классификатора из-за высокой вычислительной нагрузки. Для уменьшения размерности можно применять различные методы. В последнее время при обработке многоразмерных сигналов, допускающих разреженное представление в некотором базисе, начинает активно применяться новая парадигма обработки информации — "опознания по сжатию" [8]. При этом наиболее трудоемкой с вычислительной точки зрения является процедура восстановления сигнала. В задачах кластеризации, в которых для нас важна только группировка данных, эту трудоемкую процедуру можно исключить на этапе обучения, применив ее только в конце работы алгоритма к полученным "центрам" классов.

Рассмотрим задачу кластеризации векторов большой размерности $\mathbf{w} \in \mathbb{W} \subset \mathbb{R}^N$, имеющих разреженное представление в некотором базисе. Предположим, что эти вектора можно разделить на два класса $\mathbb{X}_1$ и $\mathbb{X}_2$. Это обозначение основывается на том факте, что проекции двух разреженных элементов $\mathbf{w}', \mathbf{w}'' \in \mathbb{W}$ с использованием случайной $m \times N$ матрицы $\mathbf{A}$ с высокой вероятностью сохраняет расстояние между $\mathbf{w}'$ и $\mathbf{w}''$. Линейная разделимость — это важное свойство классификации, которое сохраняется при сохранении расстояний. Следовательно, если существует гиперплоскость, разделяющая классы, то так же с высокой вероятностью будет су-

ществовать гиперплоскость, разделяющая их проекции $\mathbf{y}' = \mathbf{A}\mathbf{w}'$ b $\mathbf{y}'' = \mathbf{A}\mathbf{w}''$. Т. е. мы можем искать разделение на кластеры множества $\mathbb{Y} \subset \mathbb{R}^m$ вместо разделения на кластеры высокоразмерного множества $\mathbb{W} \subset \mathbb{R}^N$.

Еще раз следует упомянуть важное отличие такого подхода с использванием "опознания по сжатию": сжатый вектор *не требует реконструкции* и его обработка проходит сразу в сжатом виде.

Описанный способ кластеризации является одной из разновидностей метода "случайных проекций".

3. Применение в исследованиях

Метод "случайных проекций" успешно применялся для определения авторского стиля [9].

К. Темплтон сделал обзор работ по тематическому моделированию в гуманитарных науках, сгруппировав их по диахроническому (предмет анализа – временной интервал) и синхроническому (предметом анализа не является временным отрезком) подходам [10].

Методы тематического моделирования применялись в работах [11] и [12] при анализе американский газет XVIII-XIX веков для определения тем и их изменения во времени.

Результат тематического моделирования не всегда нагляден. В то время как некоторые темы кажутся очевидными, понять другие можно лишь зная контекст корпуса текстов. Поэтому даже при наличии вспомогательных средств для визуализации выходных данных, остается открытым вопрос интерпретации полученных результатов [13].

4. Подготовка базы данных

Для применения алгоритмов необходимо формирование тренировочной базы данных. В качестве ресурсов статей на арабском языке могут выступать официальные новостные порталы, сайты международных информационных агенств. В рамках проекта "Лаборатория анализа и моделирования социальных процессов: политический ислам / исламизм: теория и практика в сравнительной и исторической перспективе"ведется работа по оптическому распознаванию рукописных текстов на арабском языке [14]. Полученные в результате этого исследования оцифрованные варианты ма-

нускриптов представляют интерес для апробации алгоритмов тематического моделирования и кластеризации.

Подготовка данных может включать в себя: проверку правописания, выделение именованных сущностей (named entity recognition), тэгирование частей речи (POS tagging), стемминг (stemming), разделение текста на части (text chunking). Идея комбинирования этих методов для улучшения результатов тематического моделирования приведена в работе [15].

Следует отметить, что арабский язык обладает сложной морфологической структурой, поэтому препроцессинг текстов является важным и трудоемким этапом исследования [16].

Сравнительный анализ существующих инструментов для морфологического анализа текстов дан в статье [17], в частности, некоторые программные решения, разработанные специально для тематического моделирования, описаны в работе [18].

5. Заключение и перспективы дальнейших исследований

Описанные алгоритмы могут использоваться для решения ряда практических задач. Например, систематизирование неструктурированной коллекции документов, осуществление тематического поиска по коллекции [19]. Также можно будет обнаружить динамику тем в определенный период времени.

В дальнейшем исследовании предлагается выявить ограничения, которые накладывает структура арабского языка на использование алгоритмов, и предложить способы их преодоления. Является перспективным рассмотрение возможности рандомизации алгоритмов с целью повысить их качество и стабильность. Недавно полученные результаты в решении задач кластеризации [20,21] могли бы быть апробированы на корпусе арабских текстов.

Список литературы

[1] *Коршунов А., Гомзин А.* Тематическое моделирование текстов на естественном языке // Труды ИСП РАН. 2012. С. 215–244.

[2] *Salton G., McGill M.J.* Introduction to Modern Information Retrieval. 1983.

[3] *Deerwester!S., Dumais G.W., Furnas S.T., Landauer T.K., Harshman R.* Indexing by latent semantic analysis // Journal of the American Society of Information Science. Vol. 41. 1990. P. 391–407.

[4] *Arora S., Ge R., Moitra A.* Learning topic models–Going beyond SVD // IEEE 53rd Annual Symposium on Foundations of Computer Science (FOCS). 2012. P. 1–10.

[5] *Hofmann T.* Probabilistic latent semantic indexing // Proceedings of he Twenty-Second Annual International SIGIR Conference. 1999.

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

[7] *Dhillon I., Kogan J., Nicholas C.* In: A Comprehensive Survey of Text Mining. Springer, Berlin Heildelberg New York. 2003. P.73–100.

[8] *Граничин О.Н., Павленко Д.В.* Рандомизация получения данных и l_1-оптимизация (опознание со сжатием) (обзор) // Автоматика и телемеханика. 2010. №11. С. 3–28.

[9] *Efimov K., Titievsky A., Rave E., Volkovich Z.* Style Determination via Random Projection. 2013.

[10] *Templeton C.* Topic Modeling in the Humanities: An Overview. Maryland Institute for Technology in the Humanities, http://mith.umd.edu/topic-modeling-in-the-humanities-an-overview/

[11] *Nelson R. K.* Mining the Dispatch. 2010. http://dsl.richmond.edu/dispatch/

[12] *Newman D. J., Block S.* Probabilistic topic decomposition of an eighteenth-century American newspaper // Journal of the American Society for Information Science and Technology. Vol. 57. No. 6. 2006. P. 753–767.

[13] *Chang J., Boyd-Graber J., Wang C., Gerrish S., Blei D.M.* Reading tea leaves: how humans interpret topic models // Neural Information Processing Systems. 2009.

[14] *Бернikова О.А., Бояров А.А., Редькин О.И., Сенов А.А.* Методы оптического распознавания текста на арабском языке // Стохастическая оптимизация в информатике. Т. 9 (2). 2013.

[15] *Zhu X. J., Blei D.M., Lafferty J.* TagLDA: Bringing Document Structure Knowledge Into Topic Models. 2006.

[16] *Redkin O.I., Bernikova O.A.* Problems of the Arabic OCR: new attitudes // Proceedings of the 2013 International Conference on Artificial Intelligence. Las Vegas, 2013. P. 777–782.

[17] *Said D. A. et al.* A study of text preprocessing tools for Arabic text categorization // The Second International Conference on Arabic Language. 2009. P. 230–236.

[18] *Brahmi A., Ech-Cherif A., Benyettou A.* An arabic lemma-based stemmer for latent topic modeling // Int. Arab J. Inf. Technol. Vol. 10. No. 2. 2013. P. 160–168.

[19] *Griffits T., Steyvers M.* Finding Scientific topics // Proceedings of the National Academy of Sciences. 101. 2004. P. 5228–5235.

[20] *Граничин О.Н., Измакова О.А.* Рандомизированный алгоритм стохастической аппроксимации в задаче самообучения // Автоматика и телемеханика. 2005. №8. С. 52–63.

[21] *Граничин О.Н., Шалымов Д.С., Аврос Р., Волкович З.* Рандомизированный алгоритм нахождения количества кластеров // Автоматика и телемеханика. 2011. №4. С. 86–98.