ИСПОЛЬЗОВАНИЕ МЕТОДОВ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ В ОБЛАСТИ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Иргизова Ксения Вячеславовна
аспирант кафедры английской филологии
ФГБОУ ВО «МГУ им. Н.П. Огарёва»
430005, Россия, Саранск, ул. Большевистская 68
irgizova1610@yandex.ru

Аннотация: *Данная статья предлагает рассуждения автора об использовании тематического моделирования для целей лингвистики и смежных областей. В статье описаны основные принципы работы тематических моделей, описана иерархия методов тематического моделирования, их плюсы и минусы. В конце статьи автором приведены современные проблемы тематического моделирования и пути развития данного научного направления.*

Ключевые слова: *тематическое моделирование, текстовые модели, семантический анализ, кластеризация текстовых данных, обработка естественного языка.*

TOPIC MODELING IN NATURAL LANGUAGE PROCESSING

Irgizova Kseniia Viacheslavovna
department of english philology, PhD student
National Research Mordovia State University
68, Bolshevistskaia st., Saransk, Russia, 430005
irgizova1610@yandex.ru

Abstract: *The article gives the author's reflections on the use of topic modeling for linguistics and related sciences. The author describes the basic principles of topic models, the hierarchy of topic modeling methods, their pros and cons. In the end the author presents modern problems in topic modeling and ways of its development.*

Keywords: *topic modeling, text models, semantic analysis, text data clustering, natural language processing.*

Процесс глобализации и увеличение объемов окружающей информации обуславливают появление различных методов ее правильной обработки и распределения. Одним из популярных методов сортировки данных выступает тематическое моделирование (ТМ), представляющее собой способ построения модели корпуса текстов, при этом основной целью является выявление набора тем, характеризующих данный корпус текстов. Иными словами, с помощью данного метода представляется возможным выделять из коллекции текстов тематические блоки, связанные с определенным множеством слов, и затем определять вероятность соотнесения текстов с данными темами.

Тематическое моделирование является сравнительно новым научным направлением, берущим начало в 1998 году после публикации трудов греческого и американского информатика Христоса Пападимитриу «Латентное семантическое индексирование: вероятностный анализ» представленных на Международном симпозиуме, посвященном принципам работы баз данных.

Затем феномен тематического моделирования обрел своих последователей по всему миру из числа представителей международного (Т. Хофман, М. Джордан, Д. Блэй) и отечественного (К.В. Воронцов, А.В. Коршунов, А.А. Потапенко) научных сообществ [4, с. 4].

Итак, тематическое моделирование используется для того чтобы извлекать обсуждаемые тематики из корпуса текстов. По факту, ТМ представляет собой способ кластеризации слов из словаря коллекции и документов из этой коллекции (би-кластеризация). У тематических моделей есть разные приложения. Они могут быть простыми, когда мы используем получаемые матрицы в качестве некоторых признаковых описаний для следующих моделей, и более сложные, когда тематическая модель используется самостоятельно в качестве полного инструмента решения некоторой задачи. Результатом моделирования являются два вида векторов, представляющих собой вероятностные распределения: либо вероятность слова в теме, либо вероятность темы в документе.

В тематическом моделировании для обработки текстовых данных используется метод «мешка слов», при котором порядок слов внутри текста и их взаимное расположение не имеет смысла. Согласно О.А. Митрофановой «документ описывается как набор тем, порождаемых семейством распределений» [5, с. 221]. При этом любой текст представляет собой набор слов, входящих в базовый вокабуляр нескольких тем, таким образом представляется возможным определить количество тем в тексте и их распределение. Так, например, в коллекции текстов про спорт с большой долей вероятности мы встретим слова «мяч», «матч», «вратарь», а наличие в текстах слов «кошка» и «лошадь» даст нам право отнести эти тексты к теме «Животные». Кроме того, стоит помнить, что каждый текст, как правило, включает в себя черты нескольких тематических блоков. Его общую тематическую направленность можно рассчитать исходя из процента присутствия лексики конкретной тематики. Если, например, в отдельно взятом тексте обнаруживается 20% слов на тему «Животные» и 80% слов на тему «Спорт», то определение темы станет очевидным. Также нельзя не упомянуть набор слов так называемой общей тематики, используемых в текстах с целью отображения связности/противоречивости идей (предлоги, союзы, отрицательные частицы), выражения количества (числительные), определения частотности совершения того или иного действия (наречия). Такие слова в той или иной степени встречаются в тексте любой тематики, поэтому нужно помнить, что их наличие может сказаться на точности определения количества и распределения тем в документе.

Вышеперечисленные действия достаточно затратны по времени и скорости исполнения если лингвист работает с лексикой, выписывая и группируя слова из текста, как это было принято в доцифровую эпоху, или даже при использовании им простейших текстовых редакторов. Однако стоит упомянуть, что тематическое моделирование как особый вид статистической модели, берет свое начало в области машинного обучения. Оно позволяет выразить интуитивное распределение документов по темам в точной математической структуре, тем самым выступая более надежным и релевантным подходом к анализу данных. На настоящий момент проводится много исследований по улучшению автоматизированных методов понимания полного контекста коллекции документов. Принято подразделять их на алгебраические и вероятностные (рис. 1).

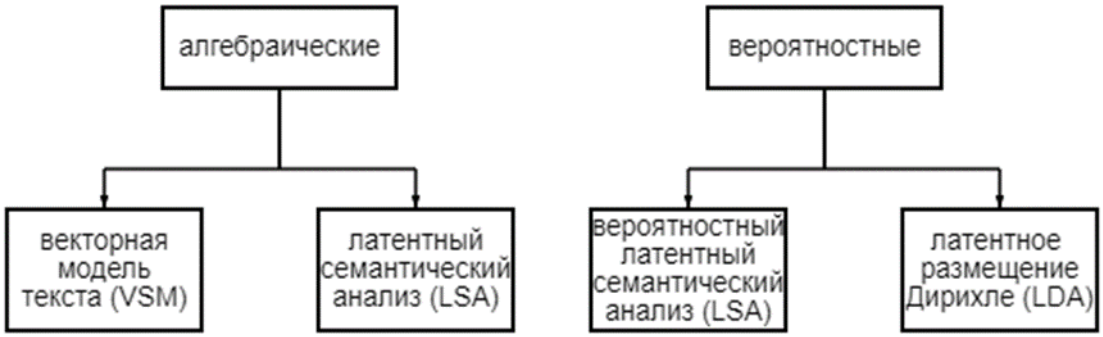

Рис. 1. Методы тематического моделирования

Векторная модель текстов (VSM) представляет собой способ представления текстов в виде векторов. Такая модель используется для решения задач ускоренного текстового анализа, поиска по документу, целей классификации и кластеризации, и что самое важное – выступает фундаментальной основой для других методов [3, с. 207]. При использовании модели VSM документ рассматривается в качестве неупорядоченного множества слов и элементов (иначе термов), включенных в текст, за исключением знаков препинания. Термы внутри документа представляются в виде матрицы «слово-документ», где под словом подразумевается каждое уникальное. В полученные ячейки матрицы вписывается вес слова в документе. Способы вычисления этого веса различны, применяются разные алгоритмы, однако наиболее популярным является вычисление величины TF-IDF (term frequency – inverse document frequency), популярность которого обусловлена не только простейшим вычислением частотности появления слова в тексте (как при методе «мешка слов»), а также выявлением того, насколько конкретное слово отличает конкретный документ от других документов в коллекции.

Продолжением метода VSM, который хоть и является популярным, однако имеет ограничения по объемам и количеству документов, является латентно-сематический анализ (LSA). Данный метод является одним из первых методов тематического моделирования, наиболее близких к современным модификациям. Его теоретической основой является предположение о том, что слова, близкие по своему словарному значению, будут встречаться в похожих контекстах. Метод предполагает создание матрицы слов из отдельно взятых частей (например, предложений) текстов корпуса (контекстов). Строки матрицы в данном случае представляют собой уникальные слова, присутствующие в каждом предложении, а в столбцах указываются непосредственно предложения. Рассмотрим следующий пример:

Таблица 1. Простейшая прямоугольная матрица

Document_A:	Alpine snow winter boots.		
Document_B:	Snow winter jacket.		
Document_C:	Alpine winter gloves.		
Words:	Document_A	Document_B	Document_C
alpine	1	0	1
snow	1	1	0
winter	1	1	1
boots	1	0	0
jacket	0	1	0
gloves	0	0	1

Согласно полученным данным, можно заключить, что если пара слов «snow-winter» встречается в контекстах чаще, то их семантическое значение выше, чем, например, у пары слов «alpine-jacket». Пара слов с наиболее высоким показателем семантического значения определяет тематическую принадлежность текстов, что легко прослеживается в предложенных примерах (snow, winter). Однако такой простейший способ работает лишь при условии ограниченного количества контекстов и слов. При увеличении данных переменных простая прямоугольная матрица может отличаться наличием шумов (нерелевантных данных), которые будут в прямом смысле мешать в выделении ключевых составляющих. В этой связи следующим шагом в рамках LSA выступает сингулярное разложение матрицы (SVD), то есть ее разложение на 3 составляющие – две ортогональные и одну диагональную матрицу – и ее дальнейшее транспонирование. Существует ряд программных реализаций для автоматического разложения матриц.

В частности, этот алгоритм встроен в математические пакеты систем MATLAB и GNU Octave. Алгоритм вызывается командой [U, S, V] = svd(M), где под M подразумевается изначальная прямоугольная матрица (таблица 1). Метод латентного семантического анализа успешно используется при работе с большими корпусами и позволяет частично разрешать контекстную синонимию, однако совсем не справляется с явлением полисемии. Кроме того, используя данный метод очень сложно задать ожидаемое число тем заранее.

Метод вероятностного латентного семантического анализа (PLSA) представляет собой модификацию вышеописанного метода с разницей в наличии вероятностных элементов, а именно вероятностей появления слов и текстовых документов в темах. Расчет данных вероятностей производится в соответствии с методом Байеса, с помощью которого можно «узнать вероятность определенного события при условии, что произошло какое-то другое, статистически взаимосвязанное с ним» [6]. Данный метод отчасти разрешает контекстную полисемию, то есть учитывает тот факт, что полисемантичные слова в своих разных значениях могут относиться к разным темам. Основными недостатками метода признаются сложность добавления в структуру модели нового документа, а также зависимость количества параметров от количества текстовых документов.

Латентное размещение Дирихле (LDA) выступает еще более надежным методом тематического моделирования. Данная модель является улучшенной модификацией PLSA и позволяет корректировать ее недостатки путем использования распределения Дирихле в качестве априори распределения, тем самым позволяя получить более четкий и конкретный набор тем [2, с. 103].

На сегодняшний день тематическое моделирование активно применяется в различных сферах. Так, например, описанные выше тематические модели полезны:

при осуществлении семантического поиска документов, близких друг к другу по смыслу;

в трендовой аналитике и анализе новостных потоков, когда в качестве тем выступают некоторые обсуждаемые события в блогосфере или в социальных сетях и для отслеживания их дальнейшего развития во времени;

для целей классификации и категоризации текстов;

для суммаризации текстов, поскольку некоторые тематические модели позволяют делать выжимки из текстов с основной, самой важной информацией.

Тематическое моделирование, будучи новым научным направлением, неизбежно сталкивается с некоторыми вызовами относительно своего будущего развития. Наиболее популярными вопросами в современной науке о тематическом моделировании принято считать:

совершенствование способов визуализации тематического пространства документа (графы, облако слов и т.д.);

интерпретация полученного набора тем человеком;

создание нового программного обеспечения для целей тематического моделирования с повышенной скоростью работы и увеличением операционной памяти для работы с большими текстовыми коллекциями;

выбор оптимального количества тем и способа определения данного количества;

созависимость между предметной областью и выбором метода тематического моделирования для работы с ней;

оценка качества работы тематической модели [7, с. 12].

Заключая, следует сказать, что тематическое моделирование помогает разрешать задачи обработки текстовых данных во многих областях, позволяя работать с информацией более быстро и продуктивно за счет качественной систематизации. Тем не менее, приведенный выше список проблем современного тематического моделирования может говорить только о том, что данной науке предстоит активное развитие в ближайшие годы. Необходимо четко понимать, как устроены методы тематического моделирования, чтобы выбрать для своего исследования самый необходимый, с учетом всех плюсов и минусов. Кроме того, процесс построения тематической модели требует от исследователя особых навыков в области статистики, математики, а также программирования, если целью является автоматическая категоризация данных.

Библиографический список:

1. Глушков Н.А. Анализ методов тематического моделирования текстов на естественном языке // Молодой ученый, 2018. – № 19 (205). – С. 101–103.

2. Едунов С. Латентно-семантический анализ. Режим доступа: https://habr.com/ru/post/110078/ (дата обращения 11.03.2023).

3. Иргизова К.В. Автоматическая категоризация текстов методами машинного обучения // XXXIII Международная научно-практическая конференция «Вопросы иноязычной филологии в свете современных исследований». – Чебоксары: Издательство Чувашского государственного педагогического университета им. И.Я. Яковлева, 2022. – С. 206–211.

4. Карпович С.Н. Математическое и программное обеспечение вероятностного тематического моделирования потока текстовых документов. Дисс. канд. техн. наук. Санкт-Петербург, 2017. – 153 с.

5. Митрофанова О.А. Моделирование тематики специальных текстов на основе алгоритма DLA. XLII Международная филологическая конференция: избранные труды. – СПб: Издательство Санкт-Петербургского государственного университета. 2013. С. 220–233.

6. Скворцова О. Скажи Байесу «да!». Забудь про интуицию – просто думай, как Байес завещал. ТАСС Наука. Режим доступа: https://nauka.tass.ru/sci/6815287 (дата обращения 11.03.2023).

7. Kherwa, P. & Bansal, P. (2019). Topic Modeling: A Comprehensive Review. EAI Endorsed Transactions on Scalable Information Systems. Volume 7. Issue 24. Pp. 1–16.