Улучшение качества поиска источников при помощи тематического моделирования

Диев И. Н.[1], Огнева М. В.[2]
[1]*diev20503@gmail.com*, [2]*ognevamv@mail.ru*
Саратовский государственный университет имени Н. Г. Чернышевского

Аннотация. Традиционные методы для обработки и кластеризации текстовых данных могут не отражать семантическую связь между документами из-за сложной и разреженной структуры данных. В данной работе будет проведён сравнительный анализ различных алгоритмов для категоризации текстовых данных, таких как латентное размещение Дирихле и латентно-семантический анализ и их применение в задачи кластеризации для поиска научных статей. Использование этих моделей позволяет значительно улучшить качество анализа текстовых данных и повысить точность кластеризации, что позволит применять их в автоматизированных системах поиска документов.

Ключевые слова: обработка текстовых данных, машинное обучение, кластеризация, латентное размещение Дирихле, латентно-семантический анализ, вероятностные модели.

В условиях стремительного роста информационного объема найти релевантные статьи по заданной теме становится всё сложнее. При использовании традиционных поисковых систем учащиеся сталкиваются с проблемой избыточности и нерелевантности найденных источников и материалов, что затрудняет процесс подбора учебных и научных текстов для выполнения работ. Стандартные алгоритмы поиска часто опираются только на совпадение ключевых слов, игнорируя смысловую близость текстов и их контекст, что приводит к существенным временным затратам при поиске подходящих материалов для образовательных целей. Для преодоления этих ограничений необходимы более продвинутые автоматизированные системы, которые способны учитывать не только содержание текста, но и выявлять скрытые тематические структуры, объединяющие документы по общим концептам. Одним из наиболее перспективных подходов для решения этой задачи является тематическое моделирование, которое может стать важным инструментом в образовательном процессе.

1 Кластеризация и тематическое моделирование

Цель алгоритмов кластеризации – выявить скрытую структуру немаркированных данных с использованием признаков для организации экземпляров в различающиеся группы. При работе с текстовыми данными под экземпляром понимается единственный документ или высказывание, а под признаками – лексемы, словарь, слова и словосочетания.

Традиционно используемые алгоритмы кластеризации, такие как k-средних или иерархическая кластеризация, хорошо зарекомендовавшие себя в задачах анализа структурированных данных, имеют ограничения при работе с текстом. В отличие от числовых данных, текстовые документы имеют более сложную структуру, характеризующуюся высокой размерностью и разреженностью. Тогда алгоритмы могут терять свою эффективность и не отражать семантические связи между документами.

Тематическое моделирование – метод машинного обучения для определения тем коллекций документов, выделения основных категорий из

набора текстов. Тематические модели способны учитывать контекст в данных и лучше приспособлены для анализа естественного языка.

2 Латентно-семантический анализ

Латентно-семантический анализ (LSA) – метод обработки информации на естественном языке, анализирующий взаимосвязь между набором документов и терминами, в них встречающимися, и выявляющий характерные факторы, присущие текстовым данным с целью повышения эффективности работы информационно-поисковых систем.

Основная идея метода состоит в оценивании корреляции терминов путем анализа их совместной встречаемости в документах. Предположим, что в наборе из 100 документов встречаются термины «информатика» и «математика». Из них 95 документов содержат оба этих термина одновременно. Это значит, что если в каком-то документе есть только слово «информатика», но нет «математика», то можно предположить, что термин «математика» должен там тоже присутствовать. Поэтому такой документ можно считать релевантным запросу, содержащему оба слова. Подобные выводы можно делать не только из простой попарной корреляции терминов. С другой стороны, анализируя корреляцию терминов в запросе, можно более точно определять интересующий пользователя смысл основного термина и повышать позиции документов, соответствующих этому смыслу, в результатах поиска.

Таким образом, при латентно-семантическом анализе документов задача состоит в том, чтобы спроецировать часто встречающиеся вместе термины в одно и то же измерение семантического пространства, которое имеет пониженную размерность по сравнению с оригинальной терм-документной матрицей, которая обычно довольно разрежена.

Для практической реализации данного подхода был использован датасет с научными статьями, содержащий тексты статей на английском языке, каждая из которых может относиться к одной из четырех категорий: «компьютерные науки», «математика», «физика», «статистика», общим размером 15 тысяч документов [1]. Данные были очищены от стоп-слов, сформированы биграммы.

После обучения модели семантического анализа на 4 темах, с помощью облаков слов было выявлено, какие слова попали в каждую из категорий.

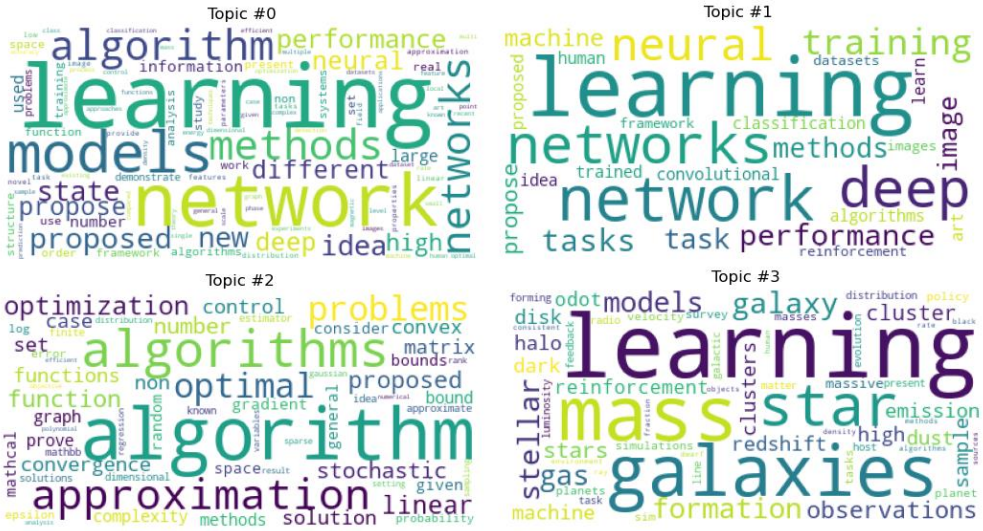

Рис. 1. Облака слов, выявленных латентно-семантической моделью

Таблица 1 — Результаты и точность обучения LSA.

Кластер	Точность	Полнота	F1-мера
0	0.43	0.97	0.60
1	0.25	0.00	0.00
2	0.47	0.08	0.14
3	0.99	0.10	0.18
Взвешенная мера			**0.33**

Можно заметить следующее: тема №0 предположительно относится к категории «Компьютерные науки», тема №1 больше похожа на «Статистику», тема №2 скорее всего является «Математикой», и тема №3 сочетает в себе слова из категории «Физика».

Можно осуществить переход в семантическое пространство – условную систему координат, где каждый текст представлен в виде точки. В семантическом пространстве можно наглядно увидеть, насколько один текст далеко расположен от другого.

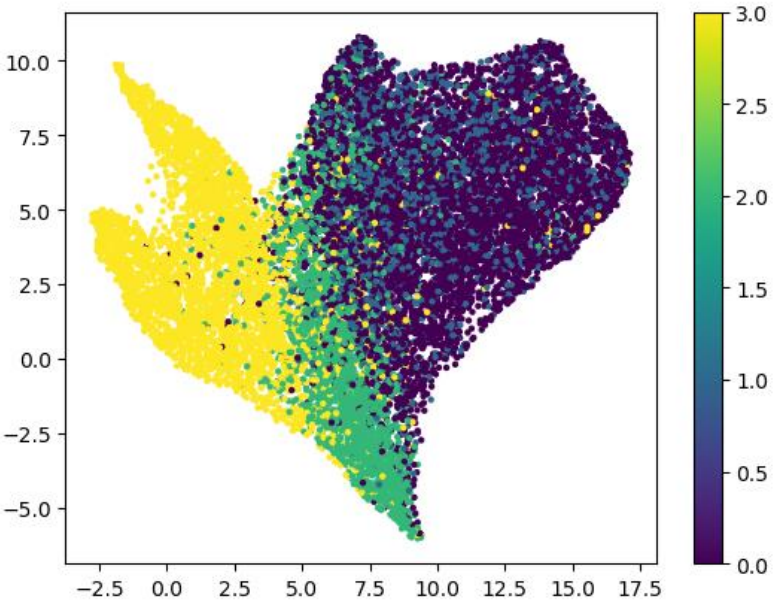

Рис. 2. Семантическое пространство текстов

Как можно заметить, тема №3 – «Физика», наиболее далека от темы №0 – «Компьютерные науки», а тема №1 – «Статистика», почти незаметна на фоне темы №0. Последний факт также является объяснением, почему в таблице точности модели тема №1 не была определена вовсе – 0% точности.

3 Латентное размещение Дирихле

Впервые предложенный Дэвидом Блеем, Эндрю Ыном и Майклом Джорданом в 2003 г., метод латентного размещения Дирихле (LDA) принадлежит семейству порождающих вероятностных моделей, в которых темы представлены вероятностями появления каждого слова из заданного набора. Документы, в свою очередь, могут быть представлены как сочетания этих тем. Уникальная особенность моделей LDA состоит в том, что темы не обязательно должны быть различными и слова могут встречаться в нескольких темах; это придает некоторую нечеткость определяемым темам. Распределение Дирихле из семейства непрерывных распределений позволяет по наблюдаемой лексеме определить вероятность принадлежности слова к теме, распределение слов в каждой теме и сочетание тем в документе [2].

Пусть имеется документ, полностью состоящий из слов p и q, при этом документ генерируется путем просчитывания вероятности двух событий – $P(p)$ и $P(q)$, где $P(q) = 1 - P(p)$. Если заранее знать эти вероятности, можно просчитать вероятность получения документа, содержащего n слов, среди которых k слов – p, с помощью формулы Бернулли: $P = C_n^k P(p)^k P(q)^{n-k}$. Однако вероятности $P(p)$ и $P(q)$ заранее неизвестны, но если рассматривать конкретный документ, состоящий, например, из семи слов p и трех слов q, можно было бы предположить, что $P(p) = 0.6$, а $P(q) = 0.4$. Рассмотрение документа, состоящего из большего числа слов, могло бы уточнить эти вероятности, если в документе из 1000 слов встретилось бы 600 раз слово p и 400 раз слово q. Именно распределение Дирихле дает возможность количественно оценить верность этих убеждений после получения дополнительных доказательств. Распределение принимает два параметра α и β, а затем создает распределение вероятностей. Параметры представляют, сколько предварительных (априорных) знаний имеется о вероятностях. Более низкие значения приводят к более широкому распределению и отражают неопределенность и отсутствие предварительных знаний. С другой стороны, большие значения параметров дают распределение с резким пиком около определенного значения [3].

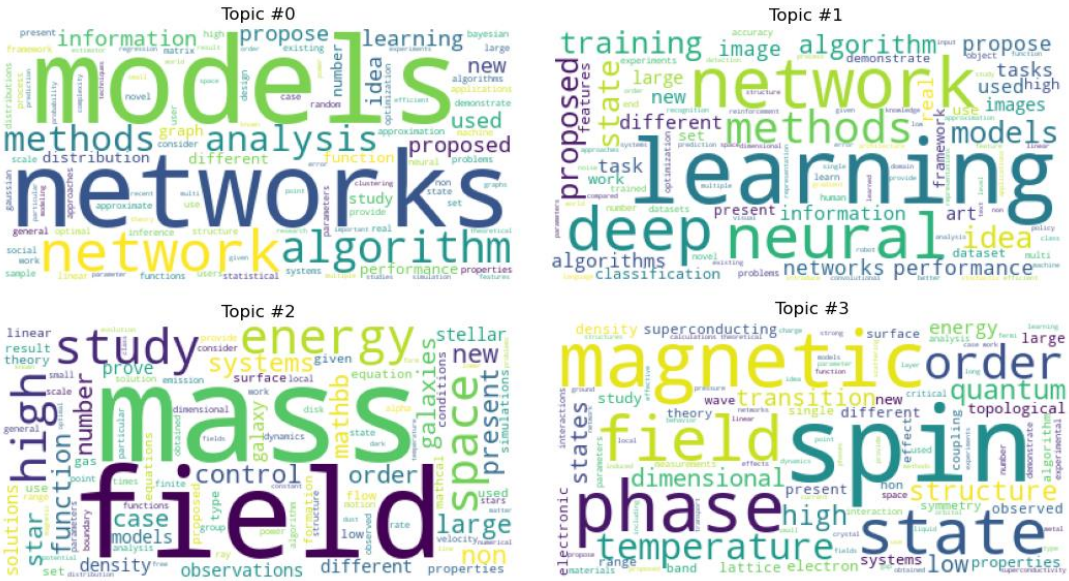

Рис. 3. Облака слов, выявленных латентным размещением Дирихле

После обучения модели LDA и выявления тем, можно заметить, что тема №0 может определять категорию «Статистика», тема №1 скорее относится к категории «Компьютерные науки», тема №2 может определять как «Математику», так и «Физику», а в теме №3 четко прослеживается категория «Физика».

Таблица 2 — Результаты и точность обучения LDA.

Кластер	**Точность**	**Полнота**	**F1-мера**
0	0.32	0.61	0.42
1	0.80	0.59	0.68
2	0.45	0.77	0.57
3	0.97	0.46	0.63
Взвешенная мера			**0.61**

4 Параметры моделей

Важный параметр для тематической модели – число тематик. Неверный выбор количества тем приведет к меньшей точности модели. Существуют механизмы, позволяющие оценить наиболее подходящее число тем, например, модель CoherenceModel, которая обучает несколько моделей на разном исходном количестве тем и выдает соответствующие оценки когерентности. Когерентность, или показатель согласованности используется в тематическом моделировании, чтобы измерить, насколько темы согласованы, то есть насколько вероятно встретить слова в одной теме [4]. Так как изначально количество тем в документах известно, для всех моделей в работе было выбрано число 4.

Заключение

В заключение можно отметить, что использование тематических моделей эффективно для работы с текстовыми данными, поскольку они учитывают семантические связи между словами и контекстуальные особенности документов. В условиях постоянно растущего объема информации

тематическое моделирование предоставляет более точный и осмысленный подход для анализа и поиска текстовых данных. Методы тематического моделирования позволяют строить более глубокие семантические представления текстов, выявляя скрытые темы и улучшая качество поиска документов. Это может значительно облегчить процесс подбора релевантных источников и литературы для написания рефератов, курсовых, магистерских и других учебных работ, а также повысить общую эффективность учебного процесса.

Практическая оценка этих методов показала, что LDA продемонстрировал наилучшие результаты с точки зрения точности и когерентности тем, в то время как LSA продемонстрировал нестабильные результаты и ограниченную применимость для задач тематического анализа.

Список литературы

[1]. Topic modeling for research articles [Электронный ресурс]. — URL: https://www.kaggle.com/datasets/abisheksudarshan/topic-modeling-for-research-articles (дата обращения 07.10.2024).

[2]. Latent Dirichlet Allocation [Электронный ресурс]. — URL: https://medium.com/analytics-vidhya/latent-dirichelt-allocation-1ec8729589d4 (дата обращения 07.10.2024).

[3]. Тематическое моделирование с использованием LDA [Электронный ресурс]. — URL: https://dev-gang.ru/article/tematiczeskoe-modelirovanie-s-ispolzovaniem-lda-map81iwksb/ (дата обращения 07.10.2024).

[4]. Оценка оптимального количества тематик в тематической модели: подход на основе количества кластеров [Электронный ресурс]. — URL: https://cyberleninka.ru/article/n/otsenka-optimalnogo-kolichestva-tematik-v-tematicheskoy-modeli-podhod-na-osnove-kachestva-klasterov/ (дата обращения 07.10.2024).