Использование методов кластеризации текстов на естественном языке в рекомендательных системах

Е.Д. Пуговкина
Самарский национальный исследовательский университет им. академика С.П. Королева
Самара, Россия
pugovkinaed@gmail.com

А.А. Белоусов
Самарский национальный исследовательский университет им. академика С.П. Королева
Самара, Россия
adark@narod.ru

***Аннотация*—Рекомендательные системы достаточно новое направление в области искусственного интеллекта. Изучение подходов, позволяющих находить предпочтения пользователей, является важной задачей, в частоте, наибольший интерес представляют текстовые данные, генерируемые пользователем или системой. Использование алгоритмов тематического моделирования текстов для поиска схожих пользователей или элементов представляет интерес как с точки зрения обработки естественного языка, так и соо стороны построения рекомендательных алгоритмов.**

Ключевые слова— рекомендательные алгоритмы, рекомендательные системы, NLP, тематическое моделирование.

1. Введение

Изучение рекомендательных систем достаточно ново относительно исследования других классических инструментов и методов информационных систем (например, баз данных или поисковых систем). В последние годы интерес к рекомендательным системам возрастает, так как объем данных, генерируемых пользователями, увеличивается, а следовательно, усложняется задача персонализации рекомендательных алгоритмов. Одним из самых популярных генерируемых пользователями контентом является текст, поэтому рассмотрение способов кластеризации текстов на естественном языке как методов поиска схожих пользователей, а следовательно, определение похожих пользователей, является важной задачей. Особенный интерес представляют способы кластеризации коротких текстов.

В рамках текущей работы будут рассмотрены способы построения тематической модели корпуса текстов и будет проанализировано их использование в рекомендательных алгоритмах. В статье приводятся как классические модели тематического моделирования текстов, так и более современные подходы, основанные на нейросетевых моделях.

2. Алгоритмы тематического моделирования текста

Тематическое моделирование - это метод извлечения скрытых тем из больших объемов текста. На данный момент самым популярным алгоритмом является LDA (Latent Dirichlet Allocation), но существует множество других подходов для разных целей. Например, тематическое моделирование короткого текста - очень сложная задача со специальными алгоритмами. В рамках текущей работы будут рассмотренные следующие алгоритмы тематического моделирования текста: GSDMM, LDA, LSI, контекстно-нейронные тематические модели.

Кластеризация короткого текста представляет собой сложную проблему из-за его разреженных, многомерных и объемных характеристик. Алгоритм для моделирования коротких текстов, GSDMM, имеет несколько важных свойств: во-первых, GSDMM может автоматически определять количество кластеров, во-вторых, GSDMM предлагает четкий способ сбалансировать полноту и однородность результатов кластеризации, в-третьих, GSDMM быстро сходится, в-четвертых, в отличие от подходов, основанных на модели векторного пространства, GSDMM может справиться с разреженными и многомерными проблемами коротких текстов, в-пятых, подобно тематическим моделям (например, LDA), GSDMM также может получать репрезентативные слова для каждого кластера. Приведённые выше тезисы делают GSDMM хорошим выбором для моделирования тем коротких текстов.

Следующий алгоритм – LDA. Идея латентного размещения Дирихле (LDA) основывается на двух предположениях: человек, который пишет документ, закладывает в текст определенные темы и выбор этой̆ темы, означает целенаправленный подбор слова с определённой вероятностью из некоторого набора слов, относящихся к рассматриваемой̆ теме. В этом случае документ представляется как смесь различных тем. В более общем смысле, LDA-модель помогает объяснить сходство данных посредством группировки свойств этих данных в ненаблюдаемые наборы [1,2].

Третий̆ алгоритм – LSI. LSI является методом уменьшения размерности, который̆ проецирует документы в семантическое пространство более низкой размерности и при этом заставляет документы с аналогичным тематическим содержанием располагаться близко друг к другу в результирующем пространстве. Скрытое пространство создается автоматически на основе совпадения слов в коллекции документов, поэтому степень семантической взаимосвязи между документами в скрытом пространстве будет зависеть от других документов в коллекции [2].

Иное семейство тематических моделей – контекстно-нейронные. В статье [3] авторы объединили контекстуализированные представления с нейронными тематическими моделями. Было обнаружено, что такой подход создает более значимые и последовательные темы, чем традиционные тематические модели набора слов и недавние нейронные модели. Авторы вводят

комбинированную тематическую модель (CombinedTM), чтобы исследовать включение контекстуализированных представлений в тематические модели. Модель строится вокруг двух основных компонентов: нейронной тематической модели ProdLDA [24] и встроенного представления SBERT [4].

В статье [5] авторы предлагают похожую тематическую модель - Zero-Shot Topic Model (ZeroShotTM). Она обучается с использованием представлений входных документов, которые учитывают порядок слов и контекстную информацию, преодолевая одно из основных ограничений моделей BoW. Более того, использование независимого от языка представления документов позволяет выполнять моделирование тем с нулевым выстрелом для невидимых языков. Это свойство важно в условиях нехватки ресурсов, когда мало данных для новых языков.

3. Тематическое моделирование текста в рекомендательных системах

Алгоритмы тематического моделирования текста могут быть использованы при построении как контентных, так и коллаборативных рекомендательных моделей. В зависимости от типа системы персональных рекомендаций появление текстовых данных в системе будет носить различный характер. Построение тематической модели корпуса текстов дает возможность осуществить мягкую кластеризацию пользователей и элементов в системе, а далее, применяя различные метрики текстового сходства, такие как меры семантического сходства с использованием тезаурусов или модели с использованием встраивания слов, можно получать матрицы сходства между пользователями и элементами системы. Также одним из вариантов применения алгоритмов кластеризации текста является ускорение вычислений в системе персональных рекомендаций, так как появляется возможность разбивать пользователей на группы и осуществлять подбор рекомендаций по иным параметрам.

Таблица I. Значение когерентности

Сравнение алгоритмов	**Лучший результат**		
	Название модели	***Количество тематических топиков***	***Когерентность***
	LSI	14	0,412
	LDA	11	0,406
	GSDMM	30	0,491
	CTM	45	0,663
	ZSTM	30	0,674

4. Вычислительный эксперимент

Для проведения вычислительного эксперимента были использованы все модели, рассмотренные в разделе 3 текущей работы. В качестве метрики оценки качества была использована когерентность. Так как выбор оптимального количества тематических топиков во многом отражает качество построенной тематической модели текстов на естественном языке, то для каждого тематического топика была построена модель и оценено ее значение когерентности. Перебор осуществлялся от 10 до 35 тематических топиков. Результаты представлены в таблице 1.

5. Заключение

Вопросы, связанные с кластеризацией текста на естественном языке, являются актуальными в связи с колоссальным объемом текстовых данных, генерируемых пользователями в социальных сетях. Подходы и методы, рассмотренные в статье, планируются к апробации над текстовыми данными, полученными из вопросно-ответной системы. В рамках работы произведено сравнение тематических моделей на корпусе текстов и использована мера семантического сходства, представленная языковой моделью RoBERTA, для построения рекомендательной системы с использованием методов тематического моделирования текста на естественном языке.

Литература

[1] Вероятностный латентно-семантический анализ // Википедия: свободная энцикл. [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Вероятностный_латентно-семантический_анализ (19.11.2020).

[2] Воронцов, К.В. Вероятностные тематические модели / К.В. Воронцов // Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных [Электронный ресурс]. — Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=Вероятностные_тематические_модели (21.11.2020).

[3] Bianchi, F. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence / F. Bianchi, S. Terragni, D. Hovy // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). — 2021. — P. 759-766.

[4] Reimers, N. SentenceBERT: Sentence embeddings using Siamese BERTnetworks / N. Reimers, I. Gurevych // Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). – 2019. – P. 3982-3992.

[5] Bianchi, F. Cross-lingual Contextualized Topic Models with Zero-shot Learning / F. Bianchi, S. Terragni, D. Hovy, D. Nozza, E. Fersini // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. – 2021. – P. 1676-1683.

[6] Kadomcev, B.B. Dynamics and the Information / B.B. Kadomcev // Izbrannye trudy: in 6 volumes. – Moscow: “Fizmatlit” Publisher, 2003. – Vol. 2. – P. 508-515.