ОБЗОР ИСПОЛЬЗОВАНИЯ ЛАТЕНТНОГО РАЗМЕЩЕНИЯ ДИРИХЛЕ В КОНТЕКСТЕ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ ТЕКСТОВ НА РУССКОМ И АНГЛИЙСКОМ ЯЗЫКЕ

А.В. Семенюта, студент гр. 8ПМ11,
А.Ю. Кайда, ассистент ОИТ ИШИТР
Томский политехнический университет
E-mail: avs183@tpu.ru

Введение

Важная задача интеллектуального анализа текста состоит в том, чтобы найти в большой коллекции текстов документы, относящиеся к определенным темам, а затем определить дальнейшую структуру среди найденных текстов. Зачастую для решения этой задачи применяется латентное размещение Дирихле (LDA), модель, придающая нечёткость определяемым темам, что полезно для совладания с гибкостью языка.

Целью данной работы является исследование использования латентного размещения Дирихле в контексте тематического моделирования в работах зарубежных и российских авторов. В ходе работы была исследована 21 статья, касающаяся LDA-модели и тематического моделирования.

Обзор работ по тематическому моделированию текстов на английском языке

За время использования и развития модели латентного размещения Дирихле было создано множество модификаций данной модели. В частности, модификации касаются следующих пунктов: адаптация внутренней структуры модели; модификация алгоритма обучения модели; разработка новых алгоритмов по мотивам LDA и их сравнение с предыдущими.

Для создания временных рядов тем создан алгоритм RollingLDA, позволяющий прослеживать в режиме реального времени изменения и нарушения в структуре наблюдаемых текстов [1].

Для назначение темы не только словам, но и сгруппированным объектам, будь то твиты в Twitter или видео на Youtube, была разработана модель логистического LDA.

Была придумана модель, адаптированная для работы в децентрализованных сетях и оптимизированная таким образом, чтобы сохранять производительность при использовании стохастических методов.

Для эффективного параллельного обучения модели на больших корпусах разработан дважды разреженный массивно параллельный сэмплер, в противоположность сэмплирования по Гиббсу.

Ещё одной модификацией алгоритма обучения модели является дискриминационный подход к обучению для модели контролируемого латентного размещения Дирихле (LDA) с использованием обратного распространения (BP-SLDA), который максимизирует апостериорную вероятность переменной прогнозирования с учетом входного документа [2].

Помимо этого, был разработан быстрый и точный пакетный алгоритм активного распространения убеждений (ABP) для обучения LDA. Для обработки массивных корпусов с большим количеством тем обучающая итерация пакетных алгоритмов LDA часто неэффективна и отнимает много времени. Чтобы ускорить скорость обучения, ABP активно сканирует подмножество корпуса и выполняет поиск в подмножестве тематического пространства для тематического моделирования.

Одной из проблем при обучении являются “эффекты порядка”, то есть при перетасовке обучающих данных генерируются разные темы. LDADE - метод на основе дифференциальной эволюции, благодаря которому распределения, генерируемые LDA, становятся более стабильными.

Для более эффективного обучения модели на больших корпусах интерпретируемых тем разработана встроенная тематическая модель (ETM), которая эффективнее LDA обнаруживает поддающиеся интерпретации темы даже с большим словарным запасом, включающим редкие слова и стоп-слова.

Были проведены множественные сравнения алгоритма LDA с аналогами на разных задачах. К примеру, при сравнении работы динамического латентного размещения Дирихле (D-LDA) с динамической встроенной тематической моделью (D-ETM) на документах из области права и науки более эффективным оказался последний [3].

Обзор работ по тематическому моделированию текстов на русском языке

В данной секции работы, как правило, более прикладные. В большинстве случаев представляются не модификации алгоритма как такового, а анализ, сделанный с помощью тематической модели, и системы на основе рассматриваемой модели.

С помощью алгоритма LDAMultiCore (модификации LDA для работы на множестве ядер) было проанализировано развитие поэтических традиций избранных тем и сравнено с траекториями традиций других языков.

С помощью LDA с определенными модификациями были проанализированы изменения в тематической структуре Живого Журнала после выборов 2011 года. В работе применен метод оценки диахронических изменений, основанный на результатах применения стандартного LDA, позволяющий отслеживать временные изменения в коллекции текстов [4].

Группой ученых был разработан фрактальный подход для определения оптимального количества тем в области тематического моделирования. Численные результаты были представлены для трех моделей: PLSA, ARTM, and LDA.

В одной из работ был представлен обобщенный алгоритм обучения для вероятностных тематических моделей (PTM). Многие известные и новые алгоритмы для моделей PLSA, LDA и SWB могут быть получены в качестве его особых случаев, выбрав подмножество следующих “опций”. Была создана устойчивая тематическая модель, которой не нужна регуляризация Дирихле.

Был собран текстовый корпус среднего уровня конфликтной дискуссии в Twitter, и проанализирована интерпретируемость тем с помощью моделей LDA, WNTM и BTM. Их качество оценивалось как с помощью автоматизированных средств, так и с помощью ручных.

В одной из работ обозреваются различные конвейеры предварительной обработки для тематического моделирования, и выделяется модель LDA-Mallet, которая демонстрирует лучшая производительность. Корпус был составлен из комментариев к различным курсам на Coursera и подборки постов из Twitter [5].

Как правило, для выделения удачности получившихся с помощью LDA тем использовалось ручное выделение. В одной из работ. рассматривается первый автоматизированный неконтролируемый анализ моделей LDA для определения нежелательных тем от законных и ранжирования значимости темы.

Заключение

В результате исследования работ, связанных с LDA и тематическим моделированием, можно выделить следующие пункты:

1. LDA является одной из популярных тематических моделей.
2. LDA имеет большое количество модификаций, решающих различные проблемы.
3. Большое количество работ, связанных с анализом тем исследуемых документов, используют в своей основе модель LDA.
4. Существует различие между тематикой работ по тематическому моделированию текстов на разных языках: так, работы по моделированию англоязычных текстов обычно рассматривают модификации алгоритма, в то время как работы по моделированию русскоязычных текстов акцентируют внимание на использование алгоритма в анализе и построении прикладных систем.

Список использованных источников

1. J. Rieger, C. Jentsch, J. Rahnenfuhrer. RollingLDA: An Update Algorithm of Latent DIrichlet Allocation to Construct Consistent Time Series from Textual Data. – Findings (EMNLP), 2021.
2. J. Chen, J. He, Y. Shen, L. Xiao, X. He, J. Gao, X. Song, L. Deng. End-to-end Learning of LDA by Mirror-Descent Back Propagation over a Deep Architecture. – NeurlPS, 2015.
3. A. Dieng, F. Ruiz, D. Blei. The Dynamic Embedded Topic Model. – arXiv preprint arXiv 1907.05545, 2019.
4. K. Maslinsky, S. Koltcov, O. Koltsova. Changes in the Topical Structure of Russian-Language LiveJournal: The Impact of Elections 2011. – Higher School of Economics Research Paper No. WP BPR 14/SOC/2013, 2013.
5. D. Bogoradnikova, O. Makhnytkina, A. Matveev, A. Zakharova, A. Akulov. Multilingual Sentiment Analysis and Toxicity Detection for Text Messages in Russian. – 29th Conference of Open Innovations Association (FRUCT), 2021.