Использование лингвистической информации в тематической модели PLSA

Нокель Михаил Алексеевич

Аспирант

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: mnokel@yandex.ru

Тематическое моделирование — одно из современных приложений машинного обучения к обработке текстов, успешно применяющееся в задачах информационного поиска, анализа изображений, аудио- и видеосигналов. *Тематические модели* определяют, к каким темам относится каждый документ в коллекции и какие слова образуют каждую тему [1]. В данной работе рассматривается одна из наиболее известных моделей – метод вероятностного латентного семантического анализа (PLSA) [2].

В тематических моделях используется модель мешка слов, основывающаяся на предположении о независимости слов друг от друга. Однако в документах, как правило, присутствует очень много слов, связанных между собой по смыслу — в частности, однокоренные слова, например: *банк–банковский–банкир*, *кредит–кредитный–кредитовать–кредитование*. В данной работе выдвигается гипотеза, что учет таких слов может улучшить качество работы тематических моделей в обработке текстов.

Как известно [1], для выделения тем тематические модели используют совместную встречаемость слов в документе. В предлагаемом подходе для учета однокоренных слов рассматриваются частоты совместной встречаемости всей совокупности похожих слов в документах. Для этого на этапе предобработки текстов определяются множества похожих слов, т.е. слов, начинающихся на один и тот же набор букв. Затем для каждого документа строятся пересечения множества его слов с найденными множествами похожих слов, и частота каждого слова в каждом пересечении увеличивается на сумму частот остальных слов в пересечении. Тем самым осуществляется репликация похожих слов в рамках каждого документа и повышается «вес» этих слов для последующей работы тематических моделей. В экспериментах в качестве подобных слов рассматривались только существительные и прилагательные, поскольку темы, как правило, задаются именными группами.

Для проверки данной гипотезы была использована коллекция банковских русскоязычных текстов (10422 документа, примерно 15.5

млн слов), взятых из различных электронных банковских журналов. В качестве метрик качества была выбрана перплексия, являющаяся стандартным критерием качества тематических моделей, вычисляемая по контрольной выборке [1]. Кроме того, использовалась и предложенная в работе [3] мера согласованности топиков *TC-PMI*, вычисляемая по первым 10 словам каждой выделенной темы.

В таблице [1] приведены результаты работы метода PLSA в зависимости от числа одинаковых букв в начале слов, по которым слова считаются похожими.

Метрики	*0 букв*	*2 буквы*	*3 буквы*	*4 буквы*	*5 букв*	*6 букв*
Перплексия	1694	1852	1565	**1434**	1620	1610
TC-PMI	2	4	10	**54**	54	29

Таблица 1: Зависимости метрик качества от числа одинаковых букв в начале похожих слов для метода PLSA

Как видно, наилучший результат показывает модель, рассматривающая в качестве похожих слова, начинающиеся с 4 одинаковых букв. Однако в русском языке есть много приставок длины 4 буквы и больше. Учитывая это, был составлен список наиболее широко использующихся таких приставок и введен дополнительный критерий: если слова начинаются на одну и ту же приставку, то они считаются похожими, если следующая буква после приставки также одна и та же. Данный критерий позволил еще снизить перплексию до 1376 и оставить согласованность топиков примерно на лучшем уровне — 50.

Таким образом, в данной работе предложен метод предобработки коллекции текстов на русском языке, улучшающий качество работы тематических моделей.

Литература

1. Воронцов К. В., Потапенко А. А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных, 2013. Т. 1, № 6, С. 657–686.

2. Hoffman T. Probabilistic Latent Semantic Indexing // In Proceedings of the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval, ACM New York, USA, 1999, P. 50–57.

3. Lau J. H., Baldwin T., Newman D. On Collocations and Topic Models // In ACM Transactions on Speech and Language Processing (TSLP), New York, USA, 2013, V. 10, N. 3, P. 4–13.