РОБАСТНЫЕ АЛГОРИТМЫ С РАЗРЕЖИВАНИЕМ ДЛЯ ВЕРОЯТНОСТНОГО ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

Изгачёв И.Ю., Аниховский М.А.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Стержанов М.В. – канд. техн. наук, доцент

В данной работе рассмотрены обобщенный EM-алгоритм и робастный алгоритм с разреживанием для решения задач тематического моделирования, рассмотрены тематическое моделирование и его стандартный критерий качества – перплексия.

Тематическое моделирование (topic modeling) – способ семантического анализа коллекции текстовых документов, первое описание которого датируется 1998 годом. С помощью тематического моделирования можно определить, какие темы содержатся в большой коллекции текстовых документов, какие слова характеризуют определенную тему, к каким темам относится каждый документ. Согласно вероятностной тематической модели, коллекция текстовых документов представляет собой распределение на множестве терминов, а отдельный документ – распределение на множестве тем. Тематическое моделирование можно использовать, например, для определения авторства документов.

Предположение о том, что лишь определенные слова в текстах относятся к каким-либо темам, можно формализовать как робастную тематическую модель. Такая модель может быть представлена как вероятностное сочетание тематической, шумовой и фоновой компоненты.

Тематическая компонента – термины, которые характеризуют текст как принадлежащий к определенной теме. Шумовая компонента – это термины, которые относятся к темам, слабо представленным в текущей коллекции. Также к шумовой компоненте относят слова, характерные только для конкретного документа. Фоновая компонента – часто употребляемые слова, а также стоп-слова, не отброшенные на предварительной стадии получения ключевых терминов. Стоп-слова – это слова, связывающие части речи в предложении. Они должны пропускаться автоматически, т. к. не несут в себе смысловой нагрузки и не относятся к каким-либо темам.

Перплексия (perplexity) – это стандартный критерий качества тематических моделей. Это мера несоответствия терминов, наблюдаемых в документах коллекции, текущей модели. Чем меньше перплексия, тем лучше модель подходит для предсказания появления терминов.

Рассмотрим обобщенный EM-алгоритм с эвристиками сглаживания, сэмплирования, робастности и разреживания, который позволяет при различных сочетаниях этих эвристик получать известные тематические модели PLSA (probabilistic latent semantic analysis), LDA (latent Dirichlet allocation), SWB (special words with background), а также новые.

Возьмем упрощенный робастный алгоритм, который хорошо сочетается с эвристикой разреживания. Указанный алгоритм не требует ни дополнительных вычислений, ни хранения матрицы параметров шума. Такая робастная модель без сглаживания позволяет разреживать искомые распределения на 99 % без ухудшения качества перплексии модели.

Робастные алгоритмы с разреживанием являются лучшими по критерию контрольной перплексии. Указанный алгоритм не требует введения априорных распределений Дирихле. Расчет

сглаживания для таких алгоритмов не является необходимым, т. к. не приводит к значительному уменьшению показателя перплексии.

Список использованных источников:

1. К. В. Воронцов. Вероятностное тематическое моделирование [Электронный ресурс]. – 2013. – Режим доступа : http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf.

2. Chemudugunta, Ch. Text Mining with Probabilistic Topic Models: Applications in Information Retrieval and Concept Modeling / Ch. Chemudugunta. – LAP LAMBERT Academic Publishing, 2010. – 140 p.