УДК 004.822

“Подходы для усовершенствования метода латентного размещения Дирихле при категоризации текстов”

Автор: Чепрасова Е. В., Университет ИТМО, Санкт-Петербург
Научный руководитель: Николаев В.В., Университет ИТМО, Санкт-Петербург

Объем данных в сети Интернет увеличивается с каждой секундой: появляются новости, публикуются картинки, создаются записи в сообществах в социальных сетях. Согласно данным журнала Форбс, 90% данных в Интернете было создано за последние два года.

В данной ситуации наиболее резко встает вопрос получения пользователем контента, наиболее полно и точно удовлетворяющего смысловому значению запроса. В то время как человек с легкостью сопоставляет определенное слово с его смысловым значением на основании содержания остального текста, для вычислительной машины требуются специальные алгоритмы для определения принадлежности слова к какой-либо тематике. Сложность определения принадлежности конкретного слова к той или иной теме заключается в проблемах синонимии (один и тот же смысл может быть передан посредством различных слов), полисемии (один и тот же термин может использоваться в различных значениях) и омонимии (одинаковые по написанию слова с разными значениями морфемы).

Для решения задач категоризации существует тематическое моделирование (topic modeling) - подход построения модели для определения тематики документов. Документ рассматривается как совокупность тем, а каждая тема, в свою очередь, определяется набором слов. Тематическая модель служит для обобщения, систематизации, категоризации большого количества текстов, а также выявления неявных (латентных) связей данных. Рассматриваемый подход находит применение при решении таких задач как: классификация и категоризация документов (научных статей, книг, рефератов); тематический поиск документов; фильтрация спама; анализ данных социальных сетей, новостных источников; выявление похожих медиаматериалов.

Отношение документа к той или иной теме становится понятным благодаря используемым в нем терминам. Так, например, если в тексте часто встречаются такие слова как “молоко”, “хлеб”, “фрукты” можно сделать вывод о его принадлежности к тематике пищи. Обычно в тексте можно выделить не одну, а несколько идей (например, “война” и “политика”). По частоте вхождения определенных слов в текст можно судить о его схожести с другими документами и о его содержании (темах). Таким образом, можно полагать, что документ представляет собой набор тем, каждая из которых встречается в нем с вероятностью $p(t|d)$, а каждая тема, в свою очередь, описывается вероятностью вхождения в нее определенных слов $p(w|t)$.

Модель латентного размещения Дирихле (Latent Dirichlet allocation, LDA) широко применяется при классификации текстов и создании рекомендательных систем. Непосредственным достоинством LDA в сравнении с другими популярными алгоритмами является тот факт, что слова и документы могут принадлежать сразу нескольким темам. В популярных пакетах машинного обучения (Mallet, Yahoo! LDA, Apache Mahout и пр.) при построении модели принимается допущение, что количество тем является заранее заданным параметром. Число тем играет важную роль при построении модели, так как данный параметр непосредственно влияет на качество модели: при малом значении параметра темы становятся слишком похожими и теряют свою идентичность, в то время как при слишком большом значении параметра темы утратят свою описательную концентрацию, то есть документы

будут распределены по темам, которые практически отображают наиболее часто встречающиеся в них слова, без учета связности слов в концепциях.

Одним из возможных вариантов является иерархический процесс Дирихле (hierarchical Dirichlet process, HDP), однако при использовании данного алгоритма при каждой итерации генерируется различное количество тем, что не позволяет переиспользовать модель в дальнейшем (например, тем с их весами в качестве входных данных для других алгоритмов).

В работе предложен способ нахождения входного параметра, при котором производится параллельное обучение нескольких моделей с разным входным параметром и их оценка. Используется модель распределенных вычислений MapReduce. На этапе map производится параллелизация тренируемых моделей с их последующей агрегацией на этапе reduce.