Реализация мультимодальных регуляризованных тематических моделей в библиотеке с открытым кодом BigARTM

Апишев Мурат Азаматович

Студент

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

E-mail: great-mel@yandex.ru

Тематическое моделирование является мощным инструментом статистического анализа текстов и используется для информационного поиска, выявления трендов в новостных потоках, создания рекомендательных систем. Оно основано на приближённом представлении матрицы частот слов в документах в виде произведения двух матриц: матрицы Φ вероятностей слов в темах и матрицы Θ вероятностей тем в документах. Темы являются скрытыми переменными, которые оцениваются в процессе обучения модели.

Во многих приложениях тексты, помимо слов, содержат метаинформацию и дополнительные модальности: имена авторов, ссылки, теги, категории, отметки времени, метки классов и т. п. Каждой модальности m соответствует своя матрица Φ_m.

Аддитивная регуляризация тематических моделей (APTM) [1, 2] позволяет вводить любое число дополнительных требований к тематической модели, комбинируя их с помощью взвешенной суммы регуляризаторов. В частности, каждой модальности соответствует отдельный регуляризатор, а их линейная комбинация позволяет строить темы, согласованные со всеми модальностями одновременно.

В данной работе рассматриваются особенности параллельной реализации алгоритмов обучения мультимодальных регуляризованных тематических моделей в библиотеке с открытым кодом BigARTM [3]. Параллельная обработка в ней производится с помощью онлайнового EM-алгоритма и основана на многопоточности. Коллекция документов, разделённая на пакеты, хранится на жёстком диске, и библиотека по мере необходимости подгружает их в очередь заданий. Создаётся множество потоков–обработчиков и один поток слияния. Каждый обработчик извлекает пакет документов из очереди, выводит для него соответствующую ему часть матрицы Θ и асинхронно отправляет результаты в очередь слияния, после чего обрабатывает новый пакет. Поток слияния извлекает из очереди эти результаты и обновляет глобальную матрицу Φ.

Механизм регуляризации реализован как для матрицы Θ, так и

для Φ. В первом случае регуляризирующие поправки вычисляются для каждого документа (т. е. для вектор-столбцов Θ) во время его обработки. Регуляризация Φ производится каждый раз при её обновлении. В текущей версии библиотеки реализованы регуляризаторы, описанные в [2]: сглаживание и разреживание матриц Θ и Φ, частичное обучение, декоррелирование тем в матрице Φ.

Реализация мультимодальных моделей потребовала обновления кода обработчиков. Кроме того, была необходимость определения способа хранения матриц Φ_m. Для этого каждое слово в словаре получило идентификатор модальности, к которой оно принадлежит. Это позволило использовать для хранения существующую структуру матрицы Φ с минимальными изменениями кода. Таким образом, хранящиеся физически вместе слова разных модальностей определяют логически разделённые матрицы Φ_m. Такой подход позволил легко обобщить регуляризацию на все матрицы Φ_m.

Был проведён ряд экспериментов по построению регуляризованных тематических моделей на коллекциях Wikipedia и Pubmed, а также мультимодальных моделей для задачи классификации на коллекции EUR-Lex. Аналогичные эксперименты проводились в [2] и [4]. BigARTM с регуляризованными мультимодальными моделями даёт результаты, сопоставимые с другими реализациями по качеству, и при этом почти линейно масштабируется по числу ядер.

Литература

1. Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014. Т. 455, № 3. С.268-271.
2. Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // AIST'2014, Analysis of Images, Social networks and Texts. — Springer International Publishing Switzerland, 2014. Communications in Computer and Information Science (CCIS). Vol. 436. pp. 29–46.
3. Страница библиотеки BigARTM: `http://bigartm.org`
4. T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers. Statistical topic models for multi-label document classification // Machine Learning, vol. 88, no. 1-2, pp. 157-208, 2012.