Иерархические вероятностные тематические модели

Чиркова Надежда Александровна

Студентка

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

E-mail: nadiinchi@gmail.com

Тематическое моделирование применяется для мягкой кластеризации коллекции текстовых документов по темам в задачах информационного поиска и анализа текстов. Документ при этом представляется в виде «мешка слов». Иерархические тематические модели позволяют представить структуру коллекции в виде ориентированного графа или дерева тем, на верхних уровнях которого находятся крупные темы, а в листьях — узко-специализированные темы [1, 2]. Каждой теме t соответствует дискретное распределение на множестве слов $p(w|t)$, а каждому документу d — дискретное распределение на множестве тем $p(t|d)$. В случае дерева для каждой темы t, имеющей множество дочерних тем S_t, согласно формуле полной вероятности, $p(t|d) = \sum\limits_{s \in S_t} p(s|d)$.

В литературе описаны десятки подходов к построению тематических иерархий, но признаётся, что проблема автоматического построения иерархии тем, которая была бы интерпретируемой для человека, остаётся открытой. Задача построения тематических моделей, и, тем более, тематических иерархий, является некорректно поставленной. Для получения её адекватного решения требуется введение не одного, а многих регуляризаторов, формализующих различные дополнительные требования к модели. В данной работе применяется аддитивная регуляризация тематических моделей (АРТМ), позволяющая комбинировать любое число регуляризаторов без существенного усложнения алгоритма обучения модели [3].

Иерархия строится по уровням сверху вниз. Для перехода от родительского уровня с множеством тем T к дочернему уровню с множеством тем S строится тематическая модель, описывающая появление слов w в документах d с помощью трёхматричного разложения $p(w|d) = \sum\limits_{s \in S} \sum\limits_{t \in T} p(w|s)p(s|t)p(t|d)$, где распределение $p(t|d)$ известно, так как родительский уровень к данному моменту уже построен.

АРТМ позволяет применить несколько регуляризаторов одновременно для улучшения интерпретируемости тем и выявления взамосвязей между темами и подтемами.

Сочетание регуляризаторов сглаживания фоновой темы с разреживанием и декоррелированием предметных тем [3] позволяет выделить для каждой родительской темы лексическое ядро — общую лексику данной темы и всех дочерних тем. Благодаря декоррелированию в каждой подтеме выделяется своё лексическое ядро, состоящее из слов, редко встречающихся в других подтемах той же темы. Таким образом, мы получаем дополнительную информацию о том, что есть общего и в чём отличия подтем каждой темы.

Регуляризатор отбора тем [4] позволяет отбрасывать дублирующие и линейно зависимые темы при построении каждого уровня.

Регуляризатор разреживания матрицы вероятностей подтем в темах $p(s|t)$ позволяет регулировать число связей каждой подтемы с темами родительского уровня. При максимальной разреженности каждая подтема имеет одного родителя, следовательно, иерархия представляется деревом.

Построена иерархическая тематическая модель русскоязычных научных конференций в области анализа данных ММРО и ИОИ с 2007 по 2013 гг. Модель представлена в виде тематического навигатора с веб-интерфейсом.

Литература

1. Wang C., Liu X., Song Y., Han J. Scalable and robust construction of topical hierarchies // arXiv:1403.3460v1 [cs.LG] 13 Mar 2014
2. Wang C., Danilevsky M., Liu L., Desai N., Ji H. Constructing topical hierarchies in heterogeneous information networks // Proc. 2013 IEEE Int. Conf.on Data Mining (ICDM'13), Dallas, TX, Dec. 2013
3. 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.
4. Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive regularization of topic models for topic selection and sparse factorization // Springer International Publishing Switzerland, A. Gammerman et al. (Eds.): SLDS 2015, LNAI 9047, pp. 193–202, 2015.