Автоматическая настройка параметров APTM под широкий класс задач

Е. В. Иванова[1], С. Н. Матвеева[2], Т. А. Голубева[3], А. В. Трусов[4], В. В. Черноног[5], М. В. Царицын[6]

Аннотация: В работе рассматривается задача тематического моделирования. Тематическое моделирование нашло своё применение в таких областях как машинное обучение и обработка естественного языка. В работе используется широко известная библиотека BigARTM, использование которой требует настройки большого числа параметров. В работе рассматривается возможность нахождения универсального набора значений параметров для широкого класса задач. Для нахождения этих значений, используется метод <...>. Для оценки качества используется критерий <...>. Полученная, с помощью подобранных коэффициентов, ошибка не больше чем на <X>% больше чем в "локально лучших моделях".

Ключевые слова: тематической моделирование, ARTM, BigARTM, глобальная оптимизация.

1 Введение

Тематическое моделирование — способ построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов[4]. Тематическая модель (англ. topic model) коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова образуют каждую тему. Тематические модели широко используются на практике для решения задач классификации и ранжирования документов, а также для разведочного поиска[6]. Открытая библиотека bigARTM[3] позволяет строить тематические модели, используя широкий класс возможных регуляризаторов[5]. Однако такая гибкость приводит к тому, что задача настройки коэффициентов оказывается очень сложной.

[1]Московский физико-технический институт, ivanova.ev@phystech.edu
[2]Московский физико-технический институт, matveeva.sn@phystech.edu
[3]Московский физико-технический институт, golubeva.ta@phystech.edu,
[4]Московский физико-технический институт, trusov.av@phystech.edu,
[5]Московский физико-технический институт, chernonog.vv@phystech.edu,
[6]Московский физико-технический институт, tsaritsyn.mv@phystech.edu,

Эту настройку можно значительно упростить, используя механизм относительных коэффициентов регуляризации и автоматический выбор N-грамм. В работе проверяется гипотеза о том, что существует универсальный набор относительных коэффициентов регуляризации, дающий "достаточно хорошие"результаты на широком классе задач. Дано несколько датасетов с внешним критерием качества (классификация документов по категориям и ранжирование). Находятся лучшие параметры для конкретного датасета, дающие "локально лучшую модель". Ищется алгоритм инициализации bigARTM, производящий тематические модели с качеством, сравнимым с "локально лучшей моделью"на её датасете. Критерий сравнимости по качеству: на данном датасете качество "универсальной модели"не более чем на 5% хуже, чем у "локально лучшей модели".Производится сравнение с другими более простыми моделями: PLSA(Probabilistic latent semantic analysis)[2] и LDA(Latent Dirichlet allocation)[1]

Список литературы

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

[3] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. Bigartm: open source library for regularized multimodal topic modeling of large collections. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 370–381. Springer, 2015.

[4] Hanna M Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM, 2006.

[5] КВ Воронцов. Вероятностное тематическое моделирование. *Москва*, 2013.

[6] Анастасия Олеговна Янина and Константин Вячеславович Воронцов. Мультимодальные тематические модели для разведочного поиска в коллективном блоге. *Машинное обучение и анализ данных*, 2(2):173–186, 2016.