Автоматическая настройка параметров BigARTM под широкий класс задач.*

Гришанов А. В.[1]***, Булатов В. Г.***[1]***, Воронцов К. В.***[1]
grishanov.av@phystech.ru, bt.uytya@gmail.com, vokov@forecsys.ru
[1] Московский физико-технический институт

Открытая библиотека BigARTM позволяет строить тематические модели, используя широкий класс возможных регуляризаторов. При этом задача настройки коэффициентов оказывается весьма сложной. В данной статье мы ищем набор параметров, дающий «достаточно хорошие» результаты на широком классе задач, используя механизм относительных коэффициентов регуляризации и автоматический выбор N-грамм. Для эксперимента использовались наборы данных Victorian Era Authorship Attribution, 20 Newsgroups, МКБ-10. Модель с подобранными коэффициентами имеет качество не более чем на <X>% хуже «локально лучших моделей»

Ключевые слова: *тематическое моделирование, аддитивная регуляризация тематических моделей, PLSA, BigARTM.*

1 Введение

Тематические модели широко используются на практике для решения задач классификации и ранжирования документов, а также для разведочного поиска [1]. Одним из распространённых инструментов в тематическом моделировании является библиотека bigARTM [2]. Она предоставляет широкий выбор для настройки модели, используя обширный класс регуляризаторов. При этом возникает потребность в аккуратном подборе параметров.

ARTM имеет 5 типов параметров: α_1, β_1 — коэффициенты сглаживания для распределений тем в документах и для распределений терминов в темах, α_2, β_2 — коэффициенты разреживания распределений тем в документах и распределений терминов в темах, γ — коэффициент декоррелирования. Текущий подход к стратегии тюнинга параметров описан в работе [1]. Сначала подбирается один из регуляризатооров (например декоррелирования). Его значение находится приблизительно, затем в зависимости от цели исследуется следующий параметр или продолжает улучшаться текущий. Существенно то, что имеются различные регуляризаторы для различных итераций. Поэтому процесс перебора последовательный: добавляется один регуляризатор, оптимизируется, затем добавляется следующий. Из-за этого для оптимизации трудно использовать продвинутые методы, такие как байесовская оптимизация. Остаётся жадный или рандомизированный поиск.

Существующий процесс перебора не лишён недостатков. Используемые коэффициенты имеют характер абсолютных, а не относительных. Из-за этого у них появляются следующие негативные свойства. Во-первых абсолютные коэффициенты привязаны к одной коллекции и трудно переносятся на другие. Во-вторых, для них нет интуитивной интерпретации и советов по подбору. Стоит также отметить, что не существует единого критерия качества для тематической модели, это дополнительно усложняет задачу.

Цель данной работы — исследовать процесс подбора параметров и найти критерии, по которым можно выбирать начальные параметры для широкого класса задач. Предлагается использовать относительные коэффициенты регуляризации и автоматический подбор N-граммм.

*Задачу поставил: Булатов В. Г. Консультант: Воронцов К. В.

2 Название раздела

TODO

3 Заключение

TODO

Литература

[1] A.O. Ianina and K.V. Vorontsov. Multimodal topic modeling for exploratory search in collective blog. *Machine Learning and Data Analysis*, 2(2):173–186, 2016.

[2] 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.