С.В. Ким, О.В. Желудкова

Тюменский государственный университет, г. Тюмень

УДК 004.67

ПОСТРОЕНИЕ РЕГУЛЯРИЗОВАННЫХ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ В BIGARTM

Аннотация. В статье рассмотрены принципы работы тематического моделирования и аддитивной регуляризации. Также изучена библиотека для построения тематических моделей BigARTM. Описан эксперимент работы с этой библиотекой. Рассмотрены сферы применения.

Ключевые слова: тематическое моделирование, аддитивная регуляризация тематического моделирования, ARTM, APTM, BigARTM, VowpalWabbit, batch, батч.

Тематическое моделирование

С приходом компьютеров и сети Интернет информация стала более доступной для всех. Существует множество поисковых систем (например, Google и Яндекс) и архивов научных статей. Для поиска информации необходимо четко формулировать запрос, знать ключевые слова и термины. Однако не всегда это становится возможным. Например, при исследовании новой предметной области. Если дана только тема, то нахождение информации становится проблематичным. Для решения таких задач существует тематическое моделирование.

Тематическое моделирование (topic modeling) – статистический анализ текстов для выявления тематики в коллекциях документов.

Тематическая модель представляет собой условное распределение $p(w|d)$ слов w в документах d коллекции D.

$p(w|d) = \sum_{t \in T} p(w|t)p(t|d)$, где T – множество тем;

$\phi_{wt} = p(w|t)$ – распределение слов w в темах t;

$\theta_{td} = p(t|d)$ – распределение тем t в документах d.

Параметры $\Phi = (\phi_{wt})$ и $\Theta = (\theta_{td})$ находятся путем решения задачи максимального правдоподобия $\sum_{d\in D}\sum_{w\in d} n_{dw} \log \sum_{t\in T} \phi_{wt}\theta_{td} \to max_{\Phi\Theta}$, при $\sum_w \phi_{wt} = 1$, $\sum_t \theta_{td} = 1$, $\phi_{wt} \geq 0$, $\theta_{td} \geq 0$ [2].

Данная задача поставлена некорректно и имеет в общем случае бесконечное множество решений вида $(\Phi S)(S^{-1}\Theta) = \Phi\Theta$.

Для решения проблемы неединственности используется регуляризация. На искомое решение накладываются дополнительные ограничения.

Библиотека BigARTM

Подход аддитивной регуляризации тематических моделей основан на идее многокритериальной регуляризации. Он позволяет строить модели, удовлетворяющие многим ограничениям одновременно. Каждое ограничение формализуется в виде регуляризатора $R_i(\Phi, \Theta) \to max$, зависящего от параметров модели. Взвешенная сумма $R(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta)$ максимизируется совместно с основным критерием правдоподобия $\sum_{d\in D}\sum_{w\in d} n_{dw} \log \sum_{t\in T} \phi_{wt}\theta_{td} + R(\Phi, \Theta) \to max_{\Phi\Theta}$ при тех же ограничениях нормировки и неотрицательности [2].

BigARTM (additive regularization topic modeling, ARTM) – открытая библиотека для тематического моделирования больших коллекций текстовых документов на основе аддитивной регуляризации.

Разработчики: Александр Фрей, Константин Воронцов, Мурат Алишев.

Существует два алгоритма обучения модели:

1) *Online-алгоритм*. Выполняет один проход по коллекции и несколько проходов по документу. Применяется для больших коллекций в потоковом режиме.

2) *Offline-алгоритм*. Выполняет несколько проходов по коллекции и один по документу. Применяется для маленьких коллекций

Основные регуляризаторы, применяемые в BigARTM:

- *SmoothSparsePhiRegularizer* – сглаживание матрицы Φ.
- *SmoothThetaRegularizer* – сглаживание матрицы Θ

- *DecorrelatorPhiRegularizer* – декоррелирование тем в матрице Φ

Метрики качества, используемые в BigARTM

- *PerplexitySore* - перплексия
- *SparsityPhiSore* - разреженность Φ
- *SparsityThetaSore* - разреженность Θ
- *TopTokensSore* - топ наиболее вероятных слов в темах

Полный список регуляризаторов и метрик с подробным описанием можно найти в онлайн-документации [1].

Библиотека работает с данными во внутреннем бинарном представлении, называемыми батчами. Получить их можно с помощью встроенного метода, получающего на вход текст разного рода форматов (основной – VowpalWabbit).

Батч (batch) – текстовый файл, в котором каждая строка – это один документ. Формат строки: *[<название документа>] {[@модальность] {слово: счетчик}}*

Применение библиотеки BigARTM

Библиотека реализована на языке программирования C++ и имеет интерфейсы на C++, Python.

Входные данные: 800 статей на тему информационной безопасности. Источники [3][4]

Выходные данные: матрица Θ – распределение тем в документах.

Для получения батчей была проведена предварительная обработка документов. Выполнена лемматизация (постановка всех слов в начальную форму), отброшены стоп-слова и редко встречающиеся слова. В результате получается текст в формате VowpalWabbit (рис.1)

1|text концепция:5 рассматриваться:2 точка:3 лгать:3 стандарт:6 особенно:2 либо:2 длительный:2 защита:4 настоящий:4 перспективный:2 система:7 конкурентный:2 личный:2 друг:6 случай:5 получать:3 технический:3 качество:5 среда:4 связывать:4 несколько:2 предоставлять:3 приложение:11 компьютерный:2 аппаратный:2 появление:2 использовать:4 интегрировать:4 интерес:2 комплексный:4 возникать:2 первый:2 например:2 развитие:3 веб:5 условие:9 приобретать:2 иметь:4 момент:4 форма:2 вычисление:7 рассматривать:2

Рис. 1. Пример документа в фрмате VowpalWabbit

С помощью библиотеки ARTM для языка программирования Python 2.7 создан объект модели. Для ее обучения используется offline-алгоритм со следующими параметрами: количество тем – 12, количество проходов по документу – 1, количество проходов по коллекции – 40. Для изучения сходимости добавлены метрика *PerplexitySore*, метрики разреженности матриц *SparsityPhiSore* и *SparsityThetaSore* и метрика *TopTokensSore*. По графику (рис.2) видно, что модель сошлась уже на 15 проходе.

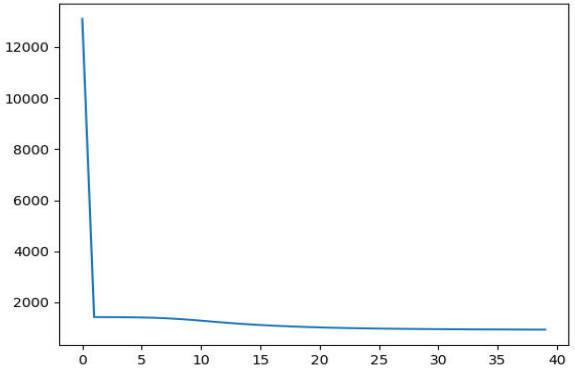

Рис. 2. График изменения перплексии на каждом проходе

Добавление регуляризаторов *SmoothSparsePhiRegularizer* ($\tau = -10$), *SmoothThetaRegularizer* ($\tau = -100$), *DecorrelatorPhiRegularizer* ($\tau = -2 * 1e6$) дает следующий результат (рис.3, рис.4). Регуляризаторы добавлялись последовательно, параметры подбирались в зависимости от топов слов и метрик разреженности.

Разреженность Phi = 0.955234467983
Разреженность Theta = 0.693575322628

Рис. 3. Метрики разреженности матриц

topic_0: этап сибс показатель характеристика целевой рациональный замечание условный обоснование изготовление
topic_1: автомат клеточный ячейка лавинный класть криптография блочный гомоморфный кеб заполнение
topic_2: уровень социальный человек общество это свой проблема развитие личность современный
topic_3: процесс оценка риск угроза управление иб вероятность объект требование критерий
topic_4: российский государственный государство россия сфера национальный страна федерация правовой право
topic_5: доверенность изделие веб мобильный эксперт закладка составной разработчик значимость
topic_6: сеть доступ пользователь атака связь узел сетевой сервер ресурс устройство
topic_7: алгоритм вектор функция код параметр ключ число сигнал схема случай
topic_8: подготовка образовательный обучение образование специалист профессиональный учебный студент компетенция дисциплина
topic_9: информация защита данные являться средство обеспечение мочь использование технология данный
topic_10: элемент вершина индекс дерево сжатый вскрытие орграф порядок подавление
topic_11: задача решение метод применение модель анализ время множество мочь данный

Рис. 4. Топ слов по каждой теме

Интерпретируем темы:

- *Тема 1*: криптография, шифрование
- *Тема 2*: информационная безопасность в социальной сфере
- *Тема 3*: управление рисками
- *Тема 4*: информационная безопасность в России
- *Тема 6*: сетевая защита
- *Тема 7*: алгоритмы обработки
- *Тема 8*: образование
- *Тема 9*: технологии защиты данных
- *Тема 10*: графы и деревья

	0	1	2	3	4	5	6 \
topic_0	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000
topic_1	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000
topic_2	0.092683	0.647980	0.0	0.167856	0.365248	0.931653	0.065254
topic_3	0.000000	0.052605	0.0	0.051439	0.000000	0.000000	0.577616
topic_4	0.043590	0.143212	0.0	0.029909	0.634752	0.043249	0.020996
topic_5	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000
topic_6	0.123861	0.000000	0.0	0.080871	0.000000	0.000000	0.000000
topic_7	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000
topic_8	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000
topic_9	0.735812	0.156202	0.0	0.582315	0.000000	0.025098	0.330185
topic_10	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000
topic_11	0.004054	0.000000	0.0	0.087610	0.000000	0.000000	0.005950

Рис. 5. Матрица Theta

Чтобы использовать обученную модель, необходимо извлечь матрицу Θ, по которой можно определять тематику документов (рис.5). Например, документ «1» с вероятностью 0.64 принадлежит «topic_2». Для пополнения коллекции необходимо новые документы преобразовать в батчи и вызвать метод трансформирования модели.

Проблемы, возникшие в ходе эксперимента: не все темы можно однозначно интерпретировать

Следующие шаги помогут решить возникшие проблемы.

Способы улучшения алгоритма:

1) Расширить коллекцию, что увеличит количество уникальных слов (тестовая выборка составляла всего 800 документов)

2) Добавление n-грамм, что увеличит точность распознания тем, так как некоторые слова в совокупности имеют больше смысла, чем по отдельности (при предварительной обработке выделялись только отдельные слова)

3) Разделение модальностей, что позволит придавать ключевым словам больший вес по отношению к остальным модальностям (авторы, ключевые слова и остальной текст были объединены в одну модальность)

Заключение

Развивая полученную модель, можно получить хороший инструмент для разведочного поиска, который значительно облегчит выполнение научных работ на тему информационной безопасности.

Однако это не единственное применение тематического моделирования. Тематические модели можно использовать для кластеризации, сегментации и классификации большого объема данных, для анализа новостных лент и социальных сетей. Например, анализируя новости в экономической сфере, можно следить за развитием какой-либо компании, да и в целом за состоянием рынка. Исследование социальных сетей может помочь выявлять публикации террористического и экстремистского характера, а также национальные и религиозные конфликты.

Тематическое моделирование – перспективный метод статистического анализа, который в уже имеет применение в самых разных областях деятельности.

СПИСОК ЛИТЕРАТУРЫ

1. Документация BigARTM URL: https://bigartm.readthedocs.io/en/stable/index.html (дата обращения 1.04.2018)

2. Аддитивная регуляризация тематических моделей URL: http://machinelearning.ru/wiki/index.php?title=Аддитивная_регуляризация_тематических_моделей (дата обращения 13.04.2018)

3. Научная электронная библиотека «киберленинка» URL: https://cyberleninka.ru/search?q=информационная+безопасность&page=3 (дата обращения 26.04.2018)

4. Журнал «Вопросы кибербезопасности» URL: http://cyberrus.com/ (дата обращения 26.04.2018)