ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 508, с. 106–108

ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА И МАШИННОГО ОБУЧЕНИЯ

УДК 004.8

ИНКРЕМЕНТАЛЬНОЕ ОБУЧЕНИЕ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ ДЛЯ ПОИСКА ТРЕНДОВЫХ ТЕМ В НАУЧНЫХ ПУБЛИКАЦИЯХ

© 2022 г. Н. А. Герасименко[1,*], А. С. Чернявский[1], М. А. Никифорова[1], М. Д. Никитин[1], К. В. Воронцов[2]

Представлено академиком РАН В.Б. Бетелиным
Поступило 28.10.2022 г.
После доработки 28.10.2022 г.
Принято к публикации 01.11.2022 г.

Стремительный рост числа научных публикаций, интенсивное появление новых направлений и подходов ставят перед научным сообществом задачу своевременного выявления трендов. Под трендом мы понимаем семантически однородную тему, которая характеризуется устойчивым во времени лексическим ядром и резким, зачастую экспоненциальным, ростом числа публикаций [1]. Примерами трендов в машинном обучении являются “LSTM”, “deep learning”, “word2vec”, “BERT”, “fake news detection”. Для выделения трендовых тем в потоке научных публикаций в реальном времени мы используем инкрементальные методы вероятностного тематического моделирования. При помощи подхода, основанного на ARTM, мы превзошли результаты популярных классических и нейросетевых подходов к задаче ранней детекции трендов. Для оценки качества мы вручную сформировали и сделали общедоступным датасет из 91 тренда.

Ключевые слова: инкрементальное тематическое моделирование, детектирование научных трендов, ARTM

DOI: 10.31857/S2686954322070086

Мы рассматриваем задачу ранней детекции трендовых тем. Эксперименты по выделению трендов производились на коллекции из 73 959 статей, опубликованных с 2000 по 2021 г. на конференциях по машинному обучению с h-индексом, превышающим 100. Валидационный датасет охватывает тренды в области машинного обучения и искусственного интеллекта 2009–2021 гг., каждый из которых характеризуется набором из не менее, чем 10 ключевых статей и 5 ключевых терминов. Обучение моделей производилось без учителя, а валидационная разметка использовалась только для финальной оценки качества.

Чтобы отслеживать появление новых тем, мы обучали отдельные модели для каждого временного шага. При поступлении новой порции документов D' словарь пополняется новыми терминами W' и могут образоваться новые темы T'. Предполагается, что новая лексика, появившаяся в новых документах, относится преимущественно к новым темам (рис. 1). Дополнительные ограничения на тематическую модель накладываются в рамках подхода аддитивной регуляризации ARTM с использованием библиотеки BigARTM [2]. В частности, для повышения различности тем используется регуляризатор декоррелирования.

Для определения количества новых тем для каждого временного шага использовалась метрика на основе относительного изменения количества токенов в словаре на текущем временном шаге, регулируемая гиперпараметром β. Это было сделано из предположения о том, что вместе с новыми темами так же появляются новые слова или начинают увеличиваться в употреблении уже известные.

На выходе модели каждой теме соответствуют ранжированные списки документов D_{topic} и ключевых слов W_{topic}. Валидационный датасет, предложенный в этой работе, также состоит из множества трендов, которым соответствуют ранжированные списки D_{trend} и ключевых слов W_{trend}, а так же названия трендов S_{trend}, для полученных моделью тем мы используем ключевые слова как возможные варианты названия тренда $S_{topic} := W_{topic}$. Чтобы сопоставить результаты реальным трендам, мы считаем три метрики Recall@k:

[1] *ПАО “Сбербанк”, Москва, Россия*
[2] *Федеральный исследовательский центр “Информатика и управление” Российской академии наук, Москва, Россия*
*****E-mail: nikgerasimenko@gmail.com*

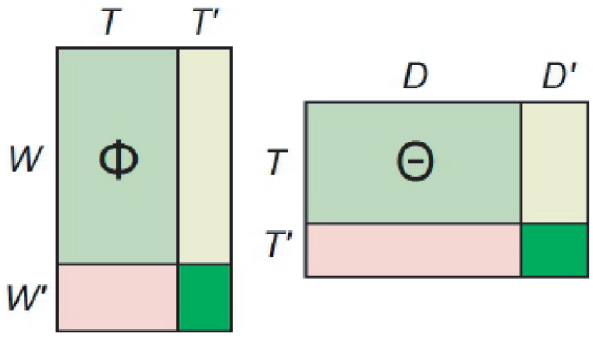

Рис. 1. Инкрементная тематическая модель. Нулевые блоки выделены красным цветом, а сильно разреженные – светло-зеленым.

$$XRecall@k = \frac{\left|X_{topic}[:k] \cap X_{trend}\right|}{K},$$

где X[:k] – первые k элементов списка X, а X заменяется на W, D, S. Для подсчета метрики для документов, слов и названий используются разные значения k, которые обозначаются как k_D, k_W и $k_S \le k_W$ соответственно.

Мы провели серию экспериментов, где рассмотрели вероятностные тематические модели, такие как PLSA, LDA и ARTM с декоррелирующим регуляризатором матрицы Φ, и нейронные сети, в частности BERTopic. Несмотря на то что BERTopic поддерживает динамическое тематическое моделирование, модель не соответствует нашим целям и критериям. Сначала BERTopic создает общую тематическую модель, как если бы в документах не было временного аспекта. Затем для каждой темы и временного шага он вычисляет представление c-TF-IDF, что приводит к различным формулировкам одних и тех же тем на разных временных шагах.

Мы сравнили наше решение с вышеперечисленными на основе трех конфигураций, включающих в себя следующие параметры:

1) Config1: D$Recall@k > 0.1$

2) Config2: W$Recall@k > 0.3$ and S$Recall@k > 0$

3) Config3: D$Recall@k > 0.1$, W$Recall@k > 0.3$ and S$Recall@k > 0$

Здесь Config1 соответствует сопоставлению извлеченных тем и трендов по документам, Config2 – только по ключевым словам, и Config3 объединяет в себе две предыдущие опции.

Модель BERTopic достигла наилучших результатов для Config1, и выявила почти все тренды: 90 из 91. Это связано с тем, что модель имеет большее количество тем и способна успешно различать документы среди них. Однако модель показывает результаты хуже для извлечения ключевых слов в других конфигурациях, так как это не является ее основной задачей.

ARTM детектирует правильные темы достаточно быстро даже в наиболее сложной конфигурации Config3, хотя в некоторых случаях может извлекать суммарно меньше трендов. В конфигурации Config1, основной целью которой является правильное разделение документов по темам, ARTM извлекает почти половину трендов за первые два месяца. Таким образом, модель подходит для качественного выявления трендов в задаче ранней детекции. Также стоит отметить, что модели BERTopic и ARTM способны извлекать тренд прямо в момент его возникновения для Config1. Это связано с тем, что некоторые из трендов относятся к типу “задача” и не имеют конкретного первого документа.

В нашем подходе есть несколько вариантов выбора набора данных для переобучения на каждом этапе. Эксперименты проводились для двух возможных вариантов: обучение по всей истории документов и обучение только по новым документам (инкрементально). Мы обозначили подход с инкрементальным обучением ARTMi. По результатам экспериментов ARTMi извлекает в общей сложности больше трендов, чем ARTM во всех рассматриваемых конфигурациях.

В своей работе мы исследовали задачу ранней детекции научных трендов. Мы адаптировали стандартный подход, основанный на ARTM, и предложили инкрементальное обучение, состоящее из инкрементальной инициализации, инкрементального набора данных и обновления количества тем на основе текущего словаря трендовых словосочетаний. Кроме того, мы включили дополнительную регуляризацию разреженности в наш подход для достижения наилучших результатов. Наш подход универсален и не зависит от конкретной модели.

Эксперименты показали, что базовая модель ARTM получила один из лучших результатов по

сравнению с другими базовыми моделями во всех рассмотренных конфигурациях подсчета качества. Более того, методы инкрементального обучения и дополнительная регуляризация позволили значительно улучшить качество. Итоговый подход, основанный на ARTM, способен выделить наибольшее количество трендов на ранних стадиях их развития и может работать в режиме реального времени.

СПИСОК ЛИТЕРАТУРЫ

1. *Kontostathis A., Galitsky M.L., Pottenger M.W., Roy S., Phelps J.D.* A Survey of emerging trend detection in textual data mining // Springer New York, 2004. p. 185–224.

2. *Vorontsov K., Frei O., Apishev M., Romov P., and Dudarenko M.* BigARTM: Open source library for regularized multimodal topic modeling of large collections // AIST, Springer International Publishing, 2015. p. 370–381.