Усовершенствование классификации научных событий с помощью кластеризации смежных областей исследований

А.Г. Морковкин
Национальный исследовательский университет Высшая школа экономики
Москва, Россия
ag.morkovkin@gmail.com

Д. А. Ильвовский
Национальный исследовательский университет Высшая школа экономики
Москва, Россия
dilvovsky@hse.ru

***Аннотация* — Классификация научных событий – это сложная задача, требующая точного определения и назначения тематических категорий. Задача усложняется, когда встречаются общие или неточные категории, не отражающие суть конкретной научной дисциплины. В нашем исследовании представлена методология, базирующаяся на принципах кластеризации областей исследований (FOS). Этот подход позволил значительно улучшить процесс классификации научных мероприятий, обеспечивая более точное и полное представление каждого события. Таким образом, данная методология предлагает ценные возможности для ученых, научных сообществ и организаторов мероприятий, способствуя более эффективному управлению и анализу научной информации.**

Ключевые слова — обработка естественного языка, классификация научных событий, open academic graph

I. Введение

Актуальность задачи классификации научных событий в современном академическом мире неоспорима, учитывая необходимость систематизации обширных массивов информации. В эпоху быстрого научного прогресса и возрастания междисциплинарных исследований, точная классификация и анализ научных данных становятся критически важными для развития академического сообщества и науки в целом. В данной работе мы сфокусировались на этой задаче, используя данные, полученные через веб-скрейпинг календарей ведущих университетов по всему миру. Основной задачей было не просто собрать обширный массив данных, но и создать эффективные методы для их анализа. Для этого мы использовали модель SciBERT, которая помогла нам более глубоко понять собранную научную информацию. Также мы включили в работу базу данных Open Academic Graph (OAG), что позволило нам устранить неточности и несоответствия в категоризации данных. Этот комплексный подход значительно улучшил процесс категоризации научных событий, обогатив наш анализ и понимание представленных данных.

II. Описание данных

В данном исследовании мы использовали обширный набор данных, собранных с научных мероприятий, организованных крупными университетами по всему миру. Для сбора данных был проведен веб-скрейпинг, анализируя календари и каталоги университетов. Эти академические ресурсы, как правило, хорошо структурированы и соответствуют формату iCalendar, который является универсально принятым стандартом для обмена календарными данными в интернете [1].

Структура нашего набора данных спроектирована для предоставления подробной информации о каждом событии. Каждое событие описывается пятью ключевыми атрибутами: «dtstart», «url», «summary», «description» и «categories». Атрибут «dtstart» указывает время начала события, «url» ведет на страницу события, «summary» дает краткое описание темы или цели события, «description» предоставляет детальную информацию о событии, а «categories» обозначает его тип или характер.

При анализе набора данных было обнаружено около 207,000 уникальных календарей после удаления повторов. Особо примечательным было обнаружение 4,510 различных меток в атрибуте «categories». Следует отметить, что большая часть (более 65% или 2,960) этих категорий встречается редко, менее десяти раз. Самой частой оказалась категория «Meeting», но ее общий характер не указывает на конкретную научную дисциплину, связанную с событием.

Стоит подчеркнуть, что наш набор данных, хотя и обширен, не идеален. Были выявлены случаи неполных или несогласованных данных в этих атрибутах. В частности, атрибут «categories», предназначенный для классификации событий, иногда содержал несколько, одиночные или отсутствующие метки категорий. Отсутствие стандартизированного списка категорий усложняет ситуацию, приводя к созданию уникальных категорий, несоответствий в данных и другим аномалиям. Это подчеркивает необходимость тщательной очистки и стандартизации данных, особенно в отношении атрибута «categories».

III. Классификация событий

Оригинальная модель BERT, предложенная Devlin и другими в 2019 году [2], привнесла значительные инновации в сферу обработки естественного языка (NLP). Основываясь на архитектуре трансформера [3], BERT отличается своей способностью двунаправленного контекстуального кодирования, что позволяет ей создавать более глубокие и точные векторные представления текста, улучшая тем самым результаты в различных задачах классификации.

Несмотря на то, что BERT установила новые стандарты в NLP, дальнейшие улучшения стали возможными за счет ее адаптации под конкретные области применения. Это привело к разработке модели SciBERT, варианта BERT, оптимизированной для работы с научными текстами [4]. SciBERT была обучена на большом массиве научных статей, что позволило ей эффективно распознавать сложную научную терминологию, что делает ее идеальным инструментом

для задач, связанных с глубоким анализом и классификацией научных текстов.

В рамках нашего исследования мы сначала провели предварительную обработку текстов событий. Это включало объединение атрибутов «summary» и «description», удаление HTML-тегов и ссылок и приведение текста к нижнему регистру для унификации.

Для упорядочивания данных мы ограничились событиями с одной определенной категорией в атрибуте «categories» и исключили категории с менее чем 10 связанными событиями. Анализ длины обработанных текстов показал, что в среднем они содержат 80 слов, с медианной длиной в 28 слов. В результате этих процедур мы сформировали корпус, включающий около 46,000 текстов.

Тщательное изучение нашего набора данных выявило значительный дисбаланс в распределении категорий. Наиболее частые категории – «Meetings», «Exhibition», «Seminar», «Education» и «Social» – не давали четкого представления о специфических научных областях. Для улучшения ситуации мы использовали SciBERT для классификации данных, что позволило нам разделить их на 255 различных категорий. Этот подход оказался успешным, достигнув точности в 81% и F1-оценки 0,51.

IV. Улучшение классификации с помощью Open Acadimic Graph

В области анализа научных данных часто возникает проблема несоответствия и неточности в аннотациях категорий. Распространенные трудности включают перекрывающиеся категории, такие как «Mathematics» и «Math», которые, хотя и относятся к одной области, иногда рассматриваются как разные. Такие несоответствия могут уменьшать точность анализа и затруднять его интерпретацию. В ответ на эту проблему мы предложили методологию, основанную на надежности базы данных Open Academic Graph (OAG), представленной Zhang и соавторами в 2019 году [5]. OAG представляет собой объединение двух значительных академических графов: Microsoft Academic Graph (MAG) [6] и AMiner [7]. Основная цель этого слияния – повышение точности аннотаций категорий.

База данных OAG известна своим огромным массивом научных статей, снабженных подробными метаданными. Многие статьи в OAG помечены тегами FOS (fields of studies), указывающими на их академические области. Мы используем эти теги для кластеризации абстрактов статей. Этот метод не только предоставляет ценные сведения о различных областях, но и создает прочную основу для последующих процессов классификации.

Первый этап нашего подхода заключается в обработке анотаций из базы данных OAG с помощью модели SciBERT. С применением векторных представлений, созданных SciBERT, мы используем алгоритм кластеризации k-средних. Этот алгоритм упорядочивает статьи в кластеры на основе близости их векторных представлений, обеспечивая, чтобы статьи в каждом кластере были более похожи друг на друга, чем на статьи в других кластерах. Это позволяет эффективно объединить пересекающиеся или почти идентичные FOS в единые кластеры.

Полученные кластеры затем используются в качестве меток для обучения классификатора. Этот важный шаг гарантирует, что классификатор учитывает уникальные характеристики каждого кластера, подготавливая его к точному определению наиболее подходящего кластера для новых данных. После обучения классификатор интегрируется с календарными данными, что позволяет определять наиболее подходящий кластер для каждой записи на основе ее содержания.

Применение нашей комплексной методологии привело к значительному уменьшению перекрывающихся и избыточных категорий. Благодаря этому категории стали более согласованными, что позволило глубже понять содержание. Сочетание кластеризации k-средних и векторных представлений SciBERT оказалось эффективным в объединении похожих FOS в координированные кластеры. Кроме того, классификатор при работе с календарными данными продемонстрировал высокие показатели точности, обеспечивая актуальные и точные метки. В частности, применение нашего подхода позволило повысить точность классификации до 87%.

V. Заключение

В заключение, наше исследование открывает новые перспективы в сборе и анализе глобальных академических событий. Применение модели SciBERT и базы данных Open Academic Graph позволило нам эффективно решить проблемы несогласованности и перекрытия категорий, достигая более точной категоризации. Наши результаты демонстрируют сокращение избыточности и повышение точности. В дальнейших планах – использование более обширных наборов данных и применение иерархической классификации для повышения точности.

Литература

[1] Desruisseaux, B. Internet calendaring and scheduling core object specification (icalendar) [Electronic resource]. – Access mode: https://www.rfc-editor.org/rfc/rfc5545.

[2] Devlin, Jacob BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / Jacob Devlin et al. // Proceedings of NAACL-HLT 2019. – 2019. – P. 4171–4186

[3] Vaswani, A. Attention is all you need / A. Vaswani et al. // Advances in neural information processing systems. – 2017. – Vol. 30.

[4] Beltagy, Iz SciBERT: A Pretrained Language Model for Scientific Text / Beltagy, Iz et al. //Conference on Empirical Methods in Natural Language Processing (2019). – 2019. – P. 3615-3620.

[5] Zhang, F. Oag: Toward linking large-scale heterogeneous entity graphs / Zhang, F. et al. //Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. – 2019. – P. 2585-2595.

[6] Sinha, A. et al. An overview of microsoft academic service (mas) and applications / Sinha, A. et al. //Proceedings of the 24th international conference on world wide web. – 2015. – P. 243-246.

Tang, J. Arnetminer: extraction and mining of academic social networks / Tang, J. et al. //Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. – 2008. – P. 990-998.