Сетевой кластерный подход к анализу естественного языка и его применение

А. Д. Богомолова[1]

Кластерный подход к анализу естественного языка решает проблему трудоемкости и больших временных затрат при анализе подобного типа данных. Он позволяет увидеть общую объективную картину научных исследований как в динамической форме, так и в моменте времени для анализа текущей ситуации. Метод совмещает в себе возможности машинного обучения и когнитивных способностей человека, являясь эффективным способом для анализа больших объемов данных. Метод подходит для разнообразных целей и здесь представлены примеры использования в трех различных проектах.

Ключевые слова: Кластерный подход, анализ естественного языка, патентный анализ, токенизация, сетевой граф

Проблема анализа текстовой информации является сложной и актуальной задачей. Текст имеет более сложную структуру, чем числовые или категориальные данные, поэтому его труднее анализировать методами машинного обучения. Тем не менее, в век технологического господства необходимо использовать всю доступную информацию, в том числе тексты, содержащие огромное количество знаний [1].

Существует множество методов анализа текста [2]. Прежде всего, это метод Word2Vec, затем идет длинная цепь элементов краткосрочной памяти (long short-term memory; LSTM), модель классификации K-ближайшего соседа (K-Nearest Neighbor classification model), свёрточные нейронные сети (convolutional neural network, CNN) и другие модели, основанные на них. Однако анализ полных текстов требует больших вычислительных ресурсов и трудоемкой очистки данных. Для большинства методов необходима обучающая выборка. Также стоит задача составить обширный словарь устойчивых словосочетаний и «ненужных сло», таких как вводные слова, союзы, артикли и т.д. Семантический словарь может занять годы, а обучение модели может оказаться непростой задачей. Когда речь идет об анализе научных трудов, таких как патенты и публикации, задачу еще больше усложняет обилие научных терминов, формул и специальных обозначений.

Тем не менее, анализировать научные труды – это хороший способ найти актуальные технические тренды, потенциальные точки роста в

[1] *Богомолова Арина Дмитриевна* — аспирант МГУ им. М. В. Ломоносова; e-mail: arina.bog@gmail.com

Bogomolova Arina Dmitrievna — graduate student, Lomonosov Moscow State University.

междисциплинарных темах и даже предугадать инновации в отдельных компаниях. С точки зрения университета, анализ развития узких научных направлений является полезной возможностью для предвосхищения потребности в определенных компетенциях и, соответственно, обогащения учебных курсов для подготовки востребованных специалистов. Очевидно, что для эффективного использования результатов анализа технологических трендов, исследование не может занимать долгое время, так как ситуация может измениться буквально за месяцы, так как каждый день появляются десятки и даже сотни статей и новых патентов. Обработка таких объемов научной информации уже превышают возможности человека, но машинное обучение также займет немало времени и усилий, если использовать распространенные методы.

Таким образом появляется необходимость разработать метод анализа научных текстов, таких, как патенты и публикации, соединяющий когнитивные способности человека и вычислительные мощности машины для получения актуальных результатов.

Суть предлагаемого нами метода состоит в следующем алгоритме действий:

- **Подготовка данных**:
  - Основным ограничением может быть узкая научная тематика, одна или несколько организаций, или даже некоторая группа ученых.
  - Тип данных может состоять как из полных текстов патентов или публикаций, так и только из названий, ключевых слов или аннотаций. От выбора типа данных зависит трудоемкость и точность результатов.
  - Выбор временного промежутка также влияет на результат. Хотим ли мы проследить зарождение и развитие технологических инноваций в динамике (тогда мы выбираем набор временных промежутков) или посмотреть в моменте тематическую близость и междисциплинарный потенциал в рамках основного ограничения.
- **Сбор данных и оценка их качества:**: на данном этапе мы проверяем достоверность данных и их пригодность для анализа. Также на этом этапе мы можем оценить полноту выборки.
- **Первичная очистка данных** включает в себя удаление знаков препинания и «ненужных слов», а также лемматизация текстов.

- **Токенизация** представляет собой превращение каждого отдельного слова в токен (или точку). Вес или величина точки зависит от частоты присутствия слова в исследуемой выборке.

- **Построение сетевого графа,** вершинами которого являются токены. Ребро между двумя токенами строится в том случае, если эти два токена встречаются в одном текстовом объекте. Толщина ребра пропорциональна частоте употребления двух токенов вместе.

- **Кластеризация** точек сетевого графа одним из классических методов кластеризации. В своем исследовании [1] мы выяснили, что наиболее подходящим для наших целей являются методы Лувейна и Лейдена [3] при условии использования показателя модулярности [4] как функции качества (quality function).

- **Интерпретация результатов** является последним шагом данного алгоритма, и здесь необходимы когнитивные способности человека, имеющего хотя бы начальные познания в исследуемой теме. Именно поэтому данный метод наиболее полезен для научных сотрудников в университетах, так как они обладают необходимыми знаниями в своих темах, а представленный кластерный подход к анализу естественного языка позволяет им увидеть картину в целом и проанализировать сотни и тысячи статей без непосредственного чтения полных текстов.

Данный метод используется нами для различных аналитических проектов. Далее приводятся примеры практического использования.

Во-первых, данный метод подходит для анализа узких научных направлений. В нашем случае мы проводили анализ патентов по теме возобновляемой энергетики [5]. В этом исследовании мы рассмотрели патенты, связанные с энергоэффективностью, выданные с 1973 по 2019 год. В ходе работы мы проанализировали более 108 000 патентов, выданных за этот период по всему миру. Чтобы продемонстрировать тенденции в данной научной области, мы разделили патенты на семь групп, первая группа охватывает вопросы патентов с 1973 по 1989 год, а остальные охватывают оставшийся период времени с шагом в пять лет. Стоит отметить, что количество патентов в этой области стремительно растет, особенно в последние годы.

В результате мы ясно видим, что исследования, ведущие к повышению энергоэффективности и, следовательно, достижению целей в области устойчивого развития, быстро расширяются за последние три десятилетия. В последние несколько лет произошел значительный сдвиг в темах исследований с эффективности сжигания природного газа в 1970-х

и 1980-х годах на исследования топливных элементов и солнечной энергии. Можно с уверенностью сказать, что технологии, связанные с возобновляемой энергией, являются основным направлением патентования в области, связанной с энергоэффективностью. Далее можно подробно рассмотреть трансформацию каждого кластера тем и выделить более конкретные технологические решения, которые в будущем могут стать хорошим конкурентным преимуществом для компаний.

Во-вторых, описанный нами метод можно использовать для анализа деятельности конкретных компаний. Для того, чтобы продемонстрировать работу алгоритма мы взяли компанию Apple в качестве примера [1]. Таким образом мы анализировали патенты, принадлежащие Apple Inc., опубликованные с января 2019 года по сентябрь 2020 года. В выборку вошли 13 500 патентов по различным темам.

Согласно полученным результатам, можно было предположить, что компания провела большинство исследований за эти полтора года в следующих областях: коммуникации (то есть налаживание технологий связи и распространения сигнала), физический интерфейс и электроника, пользовательский интерфейс и виртуальный помощник. Также можно обнаружить задел на разработку более совершенной системы биометрической идентификации и ряд патентов, посвященных теме снижения энергопотребления экраном. Большинство из этих предположений оправдались и нашли свое отражение в вышедшем уже после публикации статьи iPhone 13.

И наконец третий пример использования данного алгоритма обработки информации в форме естественного языка – это анализ компетенций и междисциплинарного взаимодействия внутри одного из крупных технических региональных университетов, впоследствии вошедшем в программу «Приоритет 2030». За основу были взяты публикации сотрудников университета в рейтинговых журналах.

В ходе проекта были выделены 5 сильных научных школ на базе университета, а также интенсивность междисциплинарного взаимодействия между этими школами. Самый основательный и емкий химический научный кластер в тоже время является наиболее закрытым и обособленным. В нем также не задействовано новых перспективных технологий, таких как искусственный интеллект и машинное обучение. Поэтому можно сделать вывод о том, что для развития этой научной школы на уже существующем крепком фундаменте исследований необходимо выходить за рамки привычных тем и вести больше совместных проектов с другими школами для обогащения и актуализации тем и методов. Наука о материалах, которая чаще всего является связующим звеном между физикой и химией в данном случае очень тесно интегрирована с физикой, но не с химией. Математика также в большинстве случаев встречается в статьях

с физической тематикой. Биология и геология являются обособленными и слаборазвитыми школами. Таким образом физическая научная школа является основной двигательной силой университета, применяет современные методы и взаимодействует с коллегами из других сфер науки. Далее полученные выводы на основе такой общей картины научной деятельности были использованы для формирования стратегии развития университета и эффективных мер по стимуляции исследовательской деятельности.

Таким образом, в данной работе был рассмотрен метод кластерного анализа естественного языка и возможности его применения на практике. В наших проектах мы использовали его чаще всего для анализа названий и аннотаций патентов и публикаций, но также существует потенциал для использования этого метода на других формах данных в форме естественного языка. Главное преимущество этого метода анализа - его скорость, простота и универсальность. Таким образом, этот метод - лучший выбор для быстрой оценки ситуации в определенной компании или сфере деятельности. Дальнейшие исследования будут связаны с улучшением методов очистки данных и упрощением анализа более длинных текстов с использованием того же метода кластеризации. Мы также заинтересованы в сокращении экспертного элемента в алгоритме принятия решений, но при этом с сохранением преимущества небольших временных затрат на реализацию.

Список литературы

[1] Bogomolova A., Ryazanova M. Balk I., *Cluster approach to analysis of publication titles*, Journal of Physics: Conference Series 1727, 2021, 012016 pp.

[2] Kim, JM., Kim, NK., Jung, Y. et al., "Patent data analysis using functional count data model", *Soft Comput*, **23** (2019), 8815–8826

[3] Traag V.A., Waltman L., van Eck N.J., "From Louvain to Leiden: guaranteeing well-connected communities", *Scientific Reports*, **9** (2019).

[4] Blondel V D, Guillaume J.L., Lam biotte R., Lefebvre E., "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics: Theory and Experiment*, 2008, 10008 pp.

[5] Bogomolova A., Balk I., Semenov E., Ivaschenko N., "Network Analysis of Patenting Trends in Energy Efficiency", *IOP Conference Series Earth and Environmental Science*, **317**:1 (2019), 012005 pp.

Cluster network approach to natural language analysis and its applications

Bogomolova A.D.

The cluster approach to natural language analysis solves the problem of labor and time consuming in this data form analysis. It gives

opportunity to see the overall objective picture of scientific research in both dynamic form and in static to analyze the current situation. The method combines the capabilities of machine learning and human cognitive abilities which is an efficient way to analyze large amounts of data. The method is suitable for a variety of purposes and in this article there are some examples of its applications.

Keywords: Cluster approach, natural language analysis, patent analysis, tokenization, network

References

[1] Bogomolova A., Ryazanova M. Balk I., *Cluster approach to analysis of publication titles*, Journal of Physics: Conference Series 1727, 2021, 012016 pp.

[2] Kim, JM., Kim, NK., Jung, Y. et al., "Patent data analysis using functional count data model", *Soft Comput*, **23** (2019), 8815–8826

[3] Traag V.A., Waltman L., van Eck N.J., "From Louvain to Leiden: guaranteeing well-connected communities", *Scientific Reports*, **9** (2019).

[4] Blondel V D, Guillaume J.L., Lambiotte R., Lefebvre E., "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics: Theory and Experiment*, 2008, 10008 pp.

[5] Bogomolova A., Balk I., Semenov E., Ivaschenko N., "Network Analysis of Patenting Trends in Energy Efficiency", *IOP Conference Series Earth and Environmental Science*, **317**:1 (2019), 012005 pp.