Contribution ID: **296** Type: **not specified**

Алгоритм поиска научных публикаций на основе информации о внешнем цитировании с применением нейросетевых моделей

Thursday 6 July 2023 15:00 (15 minutes)

Алгоритм поиска научных публикаций на основе информации о внешнем цитировании с применением нейросетевых моделей

Базы данных научных публикаций в настоящее время насчитывают миллионы статей, методы поиска в них непрерывно развиваются: от традиционного текстового поиска к системам, которые учитывают дополнительную библиометрическую информацию (индексы цитирований), семантический поиск, нейросетевые модели и другие. В частности, популярными поисковыми системами по научной литературе являются Google Scholar и Scopus, алгоритмы ранжирования которых не только выполняют полнотекстовый поиск, но и учитывают данные о цитированиях одних статей другими [1, 2, 3]. Также существуют другие системы с возможностями анализа частоты совместного цитирования и представления результатов в виде графа близких по смыслу статей (CoCites, Connected Papers) [4, 5].
Для увеличения точности поиска по научной литературе важно использовать дополнительные факторы, которые отражают ключевую суть искомой публикации.
В рамках исследования новых методов поиска по научной литературе была разработана система семантического поиска научных публикаций на основе информации о внешнем цитировании с использованием нейросетевых моделей по большим базам научных публикаций.
В качестве источника данных был выбран полнотекстовый архив научных публикаций по биомедицине PubMed Central (PMC) объемом 7,6 миллиона статей (9,1 Тб) [6].
Для увеличения точности поиска по научной литературе были совмещены два подхода:

- использовалась информация о цитировании одних статей другими, а именно текст авторского упоминания ключевых результатов другой работы и ссылка на нее
- были применены современные нейросетевые модели на основе алгоритмов трансформеров для поиска по смыслу

В текстах статей отбирались предложения, содержащие краткие описания основных результатов других статей и ссылки на них. Такие «сутевые упоминания» были собраны в единый набор данных для последующего поиска. В результате было обработано 350000 статей банка данных PMC open access в виде файлов формата XML с использованием библиотеки Python lxml. Таким образом, было собрано более 550000 упоминаний работ в единый набор данных. Также был собран дополнительный набор данных с метаинформацией статей (идентификатор, название, авторы, аннотация).
В рамках работы над поисковой системой было проведено дообучение нейросетевой модели BERT [7] для задачи мультиклассовой классификации на наборе из 10000 цитат с использованием библиотек transformers, torch, scikit-learn, pandas [8-11]. При дообучении модели ставилась задача сделать векторные представления разных упоминаний одной и той же работы ближе друг к другу в векторном пространстве [12].
В результате работы был реализован поисковый сервис на основе библиотеки flask python [13]. База данных с информацией об упоминаниях статей была токенизирована и подана на вход дообученной нейросетевой модели BERT, после чего построено дерево числовых векторов упоминаний с помощью библиотеки scikit-learn. Сервис принимает запросы, содержащие ключевые слова пользователя, и

выполняет поиск ближайших соседей в построенном дереве. По найденным упоминаниям соответствующие им статьи выдаются в качестве результатов. Для отображения найденных статей был реализован веб-интерфейс на основе библиотеки React [14].

Литература

1. Suzuki H. Google Scholar metrics for publications //Google Scholar Blog. –2012.
2. Burnham J. F. Scopus database: a review //Biomedical digital libraries. –2006. –Т. 3. –№. 1. –С. 1-8.
3. Falagas M. E. et al. Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses //The FASEB journal. –2008. –Т. 22. –№. 2. –С. 338-342.
4. Small H. Co-citation in the scientific literature: A new measure of the relationship between two documents //Journal of the American Society for information Science. –1973. –Т. 24. –№. 4. –С. 265-269.
5. Eitan A. et al. Connected Papers: Find and explore academic papers. –2020.
6. Roberts R. J. PubMed Central: The GenBank of the published literature //Proceedings of the National Academy of Sciences. –2001. –Т. 98. –№. 2. –С. 381-382.
7. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. –2018.
8. Wolf T. et al. Huggingface's transformers: State-of-the-art natural language processing //arXiv preprint arXiv:1910.03771. –2019.
9. Collobert R., Bengio S., Mariéthoz J. Torch: a modular machine learning software library. –Idiap, 2002. –№. REP_WORK.
10. Kramer O., Kramer O. Scikit-learn //Machine learning for evolution strategies. –2016. –С. 45-53.
11. McKinney W. et al. pandas: a foundational Python library for data analysis and statistics //Python for high performance and scientific computing. –2011. –Т. 14. –№. 9. –С. 1-9.
12. Sun C. et al. How to fine-tune bert for text classification? //Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18. – Springer International Publishing, 2019. –С. 194-206.
13. Grinberg M. Flask web development: developing web applications with python. –" O'Reilly Media, Inc.", 2018.
14. Fedosejev A. React. js essentials. –Packt Publishing Ltd, 2015.

Summary

Authors: ДОРОВСКИХ, Дарья (МФТИ, НИЦ "Курчатовский Институт"); ТЕСЛЮК, Антон (МФТИ); БОБКОВ, Сергей (НИЦ "Курчатовский институт")

Presenter: ДОРОВСКИХ, Дарья (МФТИ, НИЦ "Курчатовский Институт")

Session Classification: Big Data, Machine Learning and Artificial Intelligence

Track Classification: Big Data, Machine Learning and Artificial Intelligence