Text Mining Scholarly Publications using APIs

AUTHORS SECTION

Sarraf, Ishita Grinnell College & University of Illinois Urbana-Champaign, USA | sarraf.ishita@gmail.com
Fu, Yuanxi University of Illinois Urbana-Champaign, USA | fu5@illinois.edu
Schneider, Jodi University of Illinois Urbana-Champaign, USA | jodi@illinois.edu

KEYWORDS

Student submission; text mining; scholarly publications; full text, requirements analysis; scholarly data mining; XML

INTRODUCTION

New research is hardly possible without studying previous research works. But with the huge volume of digital publications at our disposal, researchers often create custom datasets for their work instead of using whole corpora of scholarly publications. Some researchers might want to study entire venues; for instance, Bolaños (2022) constructed and analyzed a corpus of 227 papers published in the Bibliometric-enhanced Information Retrieval workshop series. Other researchers create new datasets from large corpora, such as the InTeReC dataset that contains sets of sentences with in-text references to understand citation contexts in scientific papers (Bertin & Atanassova, 2018). The creation of such datasets is a complex task and requires important skills (Bertin & Atanassova, 2018) not only for acquiring the full text but also for dealing with variations in copyright licenses.

In this extended abstract, I describe my work constructing a pipeline that will make the creation of these custom datasets easy. My pipeline will be reusable such that given any Digital Object Identifier (DOI) of scholarly papers it can extract the full texts, if available, and researchers can create their own datasets to analyze the papers. The pipeline will extract the papers' PDF and XML full texts and store them in a database. I will also identify scientific analysis tasks that can be done after the extraction such as Citation Context Analysis (Anderson & Lemken, 2023) and Funder Information Extraction (Alexander & Vries, 2021). The data extraction pipeline is important because it allows different datasets to be created using a single pipeline making it easier for researchers to construct their custom datasets.

BACKGROUND

The switch from traditional paper journals to digital scholarly publications (Dai et al., 2010) resulted in the creation of corpora such as PubMed Central, PLOS Articles, the Association of Computational Linguistics proceedings, and more that contain huge volumes of open access scholarly publications that can be text mined. Text Mining, defined as the process of finding and analyzing patterns in digital text (Lammey, 2014) has become an important tool for researchers to navigate these large datasets (Dai et al., 2010). However, digital publications come with copyright licenses, and many are not open access, creating obstacles for researchers. To collect license information, I use Crossref, an association that has its own Text and Data Mining (TDM) Application Programming Interface (API) (Polischuk, 2020), that provides metadata for scholarly publications from their DOIs. However, if the publications are licensed by Elsevier or Wiley, Crossref cannot provide the full texts (Vikery, 2021), therefore, I need to use their respective TDM APIs to download the full texts.

METHODS

Data Extraction Pipeline

My full text extraction pipeline gets the DOI of the scholarly paper from the user and supplies it to the Crossref TDM API which returns the metadata of the paper. The pipeline then searches for the license URL and full-text URL. Then, if available, it downloads the PDF and XML versions of the full text and stores the information in a database as shown in Figure 1.

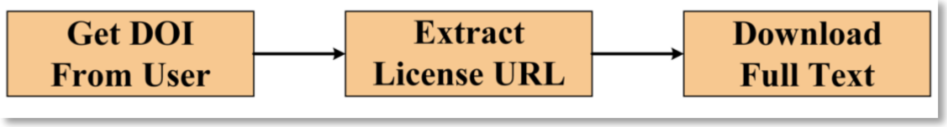


Figure 1: Data Extraction Pipeline

An important consideration is that as Wiley and Elsevier have loaded licenses to Crossref’s TDM API, users cannot access full texts from these publishers and have to use the publisher’s own APIs to access the full text. Therefore, my pipeline checks whether the publisher is Wiley or Elsevier and if so, accesses the full text from their TDM APIs.

86th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – 31, 2023 |London, United Kingdom. Author(s) retain copyright, but ASIS&T receives an exclusive publication license.

Requirements Analysis

To identify scientific analysis tasks that could build on my pipeline, I interviewed researchers in the Information Quality Lab. After identifying use cases for full text datasets, in the future I will implement their work in my pipeline.

RESULTS

As shown in Figure 2, I am already working on two possible applications for XML full text: Funder Information Extraction and Citation Context Analysis, to support ongoing projects conducted by other lab members.

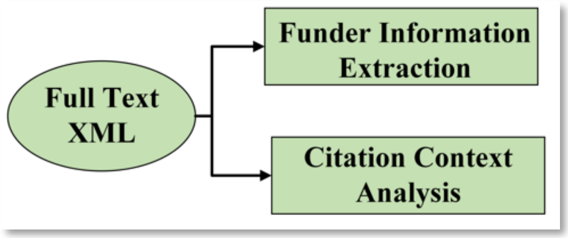
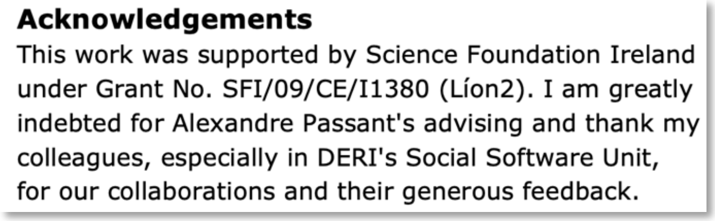


Figure 2: Possible Applications of Full Text XML

Funder Information Extraction

I will parse the XML full text to extract the Acknowledgment section as shown in Figure 3 and along with that, the authors' information such as full name, ORCID ID, and affiliation, DOI, year published, and type of paper. This will create a dataset with funding information for many papers using their DOIs. This information is crucial to analyze the funder's relationship with the research being conducted to detect funding bias. Various scholarly bibliometric databases do have funding information that can be accessed using REST APIs. However, to our knowledge the only openly licensed funding information comes from the Crossref Funder Registry and OpenAlex; while bother are CC0-licensed, neither is comprehensive. The Open Funder Registry currently has only 35,414 funders (Crossref, 2023); contributions are reviewed by Crossref with updates shared monthly (Michaud, 2021). OpenAlex incorporates the Open Funder Registry, enriching it with Wikidata and ROR (OpenAlex, 2023); designed to replace the now-defunct Microsoft Academic Graph (Priem et. al., 2022), OpenAlex is still rapidly evolving. This openly available funder information is not comprehensive, for instance when authors omit funding institutions or institutions are not covered. Therefore, it is important to have the option of extracting funders from full text so that the Acknowledgments section can be checked for all funding institutions from any scholarly publication.



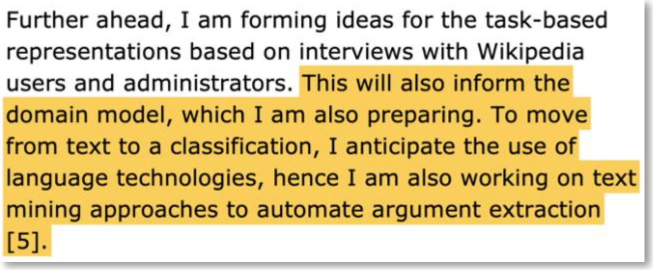
Acknowledgements

This work was supported by Science Foundation Ireland under Grant No. SFI/09/CE/I1380 (Líon2). I am greatly indebted for Alexandre Passant's advising and thank my colleagues, especially in DERI's Social Software Unit, for our collaborations and their generous feedback.

Figure 3: Acknowledgment Section of a Scholarly Publication

Citation Context Analysis

I seek to support two related citation context analysis projects. The first is to extract sentences that occur before and after an in-text citation as shown in Figure 4 along with the reference information to analyze the citation context. By doing so, I will create a dataset that will help check the citation contexts of many scholarly papers with their references.



Further ahead, I am forming ideas for the task-based representations based on interviews with Wikipedia users and administrators. This will also inform the domain model, which I am also preparing. To move from text to a classification, I anticipate the use of language technologies, hence I am also working on text mining approaches to automate argument extraction [5].

Figure 4: Sentences Around In-Text Citation

The second is to extend the data extraction pipeline to identify documents using their PubMed IDs and then use the full text to check for citation errors. This application will also involve extracting sentences around in-text citations as shown in Figure 4 and then checking their context in relation to the reference to verify whether the meaning of the citation has been retained in the paper citing it.

FUTURE WORK

I will continue to work on developing my pipeline to implement these applications. With further requirements analysis, I will investigate and incorporate additional possible applications for analyzing full texts of various scholarly publications.

CONCLUSION

I have constructed a pipeline that will make it easy for researchers to create their custom collection of data for research using DOIs and Crossref, Elsevier, and Wiley's TDM APIs. My pipeline will help navigate the license problems and other access issues related to full-text extraction and allow researchers to focus on their analysis work. Various applications will be able to analyze the full texts already acquired from the pipeline.

ACKNOWLEDGMENTS

The work of Ishita Sarraf was supported in part by the Distributed Research Experiences for Undergraduates (DREU) program, a joint project of the CRA Committee on the Status of Women in Computing Research (CRA-W) and the Coalition to Diversify Computing (CDC), which is funded in part by the NSF Broadening Participation in Computing program (NSF BPC-A #1246649). Yuanxi Fu and Jodi Schneider were funded by NSF 2046454 CAREER: Using network analysis to assess confidence in research synthesis.

REFERENCES

Alexander, D., & Vries, A. P. de. (2021). "This research is funded by...": named entity recognition of financial information in research papers. In I. Frommholz, P. Mayr, G. Cabanac, & S. Verberne (Eds.), *Proceedings of the 11th International Workshop on Bibliometric-enhanced Information Retrieval* (Vol. 2847, pp. 102–110). CEUR. https://ceur-ws.org/Vol-2847/paper-10.pdf

Anderson, M. H., & Lemken R. K. (2023). Citation context analysis as a method for conducting rigorous and impactful literature reviews. *Organizational Research Methods*, *26*(1), 77–106. https://doi.org/10.1177/1094428120969905

Bertin, M., & Atanassova, I. (2018). InTeReC: In-text Reference Corpus for applying natural language processing to bibliometrics. In P. Mayr, I. Frommholz, & G. Cabanac (Eds.), *Proceedings of the 7th International Workshop on Bibliometric-enhanced Information Retrieval* (Vol. 2080, pp. 54–62). CEUR. https://ceur-ws.org/Vol-2080/paper6.pdf

Bolaños, F. (2022). Mapping the trending topics of bibliometric-enhanced information retrieval. In I. Frommholz, P. Mayr, G. Cabanac, & S. Verberne (Eds.), *Proceedings of the 12th International Workshop on Bibliometric-enhanced Information Retrieval* (Vol. 3230, pp. 61–70). CEUR. https://ceur-ws.org/Vol-3230/paper-08.pdf

Crossref. (2023, August). Crossref Funder Registry. v1.52. https://gitlab.com/crossref/open_funder_registry

Dai, HJ., Chang, YC., Tsai, R. TH., Hsu, WL. (2010). New challenges for biological text-mining in the next decade. *Journal of Computer Science and Technology*, *25*(1), 169–179. https://doi.org/10.1007/s11390-010-9313-5

Lammey R. (2014). CrossRef's text and data mining services. Learned Publishing, 27(4), 245–250. https://doi.org/10.1087/20140402

Michaud, F. (2021, April 9). Accessing the open funder registry. Crossref. https://www.crossref.org/documentation/funder-registry/accessing-the-funder-registry/

OpenAlex API entities: Funders (2023, July 31). OpenAlex API documentation. https://docs.openalex.org/api-entities/funders

Polischuk, P. (2020, April 8). Text and data mining. Crossref. Retrieved July 5, 2023 from https://www.crossref.org/documentation/retrieve-metadata/rest-api/text-and-data-mining/

Priem, J., Piwowar, H., Orr R. (2022). OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. In 26th International Conference on Science and Technology Indicator, STI 2022. https://doi.org/10.5281/zenodo.6936227