Archive Scholar Citation Dataset

Martin Czygan

Internet Archive
San Francisco, California, USA
martin@archive.org

Bryan Newbold

Internet Archive
San Francisco, California, USA
bnewbold@archive.org

Abstract

As part of its scholarly data efforts, the Internet Archive releases a citation graph dataset derived from scholarly publications and additional data sources. It is composed of data gathered by the fatcat cataloging project and related web-scale crawls targeting primary and secondary scholarly outputs. In addition, relations are worked out between scholarly publications, web pages and their archived copies, books from the Open Library project as well as Wikipedia articles. This first version of the graph consists of over X nodes and over Y edges. We release this dataset under a Z open license under the collection at https://archive.org/details/TODO-citation_graph, as well as all code used for derivation under an MIT license.

Index terms— Citation Graph, Web Archiving

1 Introduction

The Internet Archive releases a first version of a citation graph dataset derived from a raw corpus of about 2.5B references gathered from metadata and from data obtained by PDF extraction tools such as GROBID[5]. The goal of this report is to describe briefly the current contents and the derivation of the Archive Scholar Citations Dataset (ASC). We expect this dataset to be iterated upon, with changes both in content and processing.

Modern citation indexes can be traced back to the early computing age, when projects like the Science Citation Index (1955)[3] were first devised, living on in existing commercial knowledge bases today. Open alternatives were started such as the Open Citations Corpus (OCC) in 2010 - the first version of which contained 6,325,178 individual references[6]. Other notable sources from that time include CiteSeerX[9] and CitEc[1]. The last decade has seen an increase of more openly available reference dataset and citation projects, like Microsoft Academic[8] and Initiative for Open Citations[2][7]. In 2021, according to [4] over 1B citations are publicly available, marking a tipping point for open citations.

2 Related Work

3 Citation Dataset

4 System Design

The constraints for the systems design are informed by the volume and the variety of the data. In total, the raw inputs amount to a few TB of textual content, mostly newline delimited JSON. More importantly, while the number of data fields is low, certain schemas are very partial with hundreds of different combinations of available field values found in the raw reference data. This is most likely caused by aggregators passing on reference data coming from hundreds of sources, each of which not necessarily agreeing on a common granularity for citation data and from artifacts of machine learning based structured data extraction tools.

Each combination of fields may require a slightly different processing path. For example, references with an Arxiv identifier can be processed differently from references with only a title. Over 50% of the raw reference data comes from a set of eight field manifestations, as listed in Table 1.

Overall, a map-reduce style approach is followed, which allows for some uniformity in the overall processing. We extract (key, document) tuples (as TSV) from the raw JSON data and sort by key. Then we group documents with the same key into groups and apply a function on each group in order to generate our target schema (currently named biblioref, or bref for short) or perform addition operations (such as deduplication).

The key derivation can be exact (like an identifier like DOI, PMID, etc) or based on a normalization procedure, like a slugified title string. For identifier based matches we

Fields	Share
CN CRN—P—T— U— V— Y	14%
DOI	14%
CN—CRN—IS—P—T—U—V—Y	5%
CN—CRN—DOI—U—V—Y	4%
PMID—U	4%
CN—CRN—DOI—T—V—Y	4%
CN—CRN—Y	4%
CN—CRN—DOI—V—Y	4%

Table 1. Top 8 combinations of available fields in raw reference data accounting for about 53% of the total data (CN = container name, CRN = contrib raw name, P = pages, T = title, U = unstructured, V = volume, IS = issue, Y = year, DOI = doi, PMID = pmid). Unstructured fields may contain any value.

can generate the target biblioref schema directly. For fuzzy matching candidates, we pass possible match pairs through a verification procedure, which is implemented for release entity schema pairs. The current verification procedure is a domain dependent rule based verification, able to identify different versions of a publication, preprint-published pairs or or other kind of similar documents by calculating similarity metrics across title and authors. The fuzzy matching approach is applied on all reference documents, which only have a title, but no identifier.

With a few schema conversions, fuzzy matching can be applied to Wikipedia articles and Open Library (edition) records as well. The aspect of precision and recall are represented by the two stages: we are generous in the match candidate generation phase in order to improve recall, but we are strict during verification, in order to control precision.

5 Fuzzy Matching Approach

6 Quality Assurance

In general a short summarizing paragraph will do, and under no circumstances should the paragraph simply repeat material from the Abstract or Introduction. In some cases it's possible to now make the original claims more concrete, e.g., by referring to quantitative performance results.

7 Future Work

This material is important – part of the value of a paper is showing how the work sets new research directions. I like bullet lists here. A couple of things to keep in mind:

- If you're actively engaged in follow-up work, say so. E.g.: "We are currently extending the algorithm to... blah blah, and preliminary results are encouraging." This statement serves to mark your territory.
- Conversely, be aware that some researchers look to Future Work sections for research topics. My opinion is that there's nothing wrong with that – consider it a compliment.

8 Acknowledgements

Don't forget them or you'll have people with hurt feelings. Acknowledge anyone who contributed in any way: through discussions, feedback on drafts, implementation, etc. If in doubt about whether to include someone, include them.

9 Citations

10 Appendix A

References

[1] Citations in economics. `https://citec.repec.org/`. Accessed: 2021-07-30.

[2] Initiative for open citations. `https://i4oc.org/`. Accessed: 2021-07-30.

[3] E. Garfield. The evolution of the science citation index. *International microbiology*, 10(1):65, 2007.

[4] B. I. Hutchins. A tipping point for open citation data. *Quantitative Science Studies*, pages 1–5, 2021.

[5] P. Lopez. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *International conference on theory and practice of digital libraries*, pages 473–474. Springer, 2009.

[6] D. Shotton. Publishing: open citations. *Nature News*, 502(7471):295, 2013.

[7] D. Shotton. Funders should mandate open citations. *Nature*, 553(7686):129–130, 2018.

[8] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. Hsu, and K. Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246, 2015.

[9] J. Wu, K. Kim, and C. L. Giles. Citeseerx: 20 years of service to scholarly big data. In *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse*, pages 1–4, 2019.