Internet Archive Scholar Citation Graph Dataset

Technical Report

Martin Czygan
Internet Archive
San Francisco, CA 94118
martin@archive.org

Bryan Newbold
Internet Archive
San Francisco, CA 94118
bnewbold@archive.org

August 10, 2021

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.

Keywords Citation Graph Dataset · Scholarly Communications · 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[Lopez, 2009]. The goal of this report is to describe briefly the current contents and the derivation of the Internet Archive Scholar Citation Graph Dataset (IASCG). 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)[Garfield, 2007] 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[Shotton, 2013]. Other notable sources from that time include CiteSeerX[Wu et al., 2019] and CitEc[Cit]. The last decade has seen an increase of more openly available reference dataset and citation projects, like Microsoft Academic[Sinha et al., 2015] and Initiative for Open Citations[i4o][Shotton, 2018]. In 2021, according to [Hutchins, 2021] over 1B citations are publicly available, marking a tipping point for open citations.

2 Citation Graph Contents

3 System Design

TODO: describe limitations, single machine, prohibitive external data store lookups, and performance advantages of stream processing; “miniature map-reduce”, id based matching; fuzzy matching; funnel approach; data quality issues; live system design (es, pg, . . .)

The constraints for the system design are informed by the volume and the variety of the data. In total, the raw inputs amount to about X TB uncompressed textual data. More importantly, while the number of data fields is low, over Y different combinations of fields are found in the raw reference data. 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. We identify about X types of manifestations which in total amount for Y% of the reference documents.

Overall, a map-reduce style approach is followed, which e.g. allows for some uniformity in the overall processing. We extract key value tuples (as TSV) from the raw JSON data and sort by key. Finally we group pairs with the same key into groups and apply a function of the elements of the group in order to generate our target schema (biblioref, called bref, for short).

The key derivation can be exact (e.g. an id like doi, pmid, etc) or based on a normalization procedure, like a slugified title string. For id based matches we can generate the bref schema directly. For fuzzy matching candidates, we pass possible match pairs through a verification procedure, which is implemented for documents of one specific catalog record schema.

With a few schema conversions, fuzzy matching can be applied to Wikipedia articles and Open Library editions 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 ensure precision.

4 Fuzzy Matching Approach

5 Discussion

References

Patrice 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.

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

David Shotton. Publishing: open citations. *Nature News*, 502(7471):295, 2013.

Jian Wu, Kunho Kim, and C Lee 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.

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

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan 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.

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

David Shotton. Funders should mandate open citations. *Nature*, 553(7686):129–130, 2018.

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