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 · 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 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)[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

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.

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.

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 2.

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 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.

4 Fuzzy Matching Approach

The fuzzy matching approach currently implemented works in two phases: match candidate generation and verification. For candidate generation, we map each document to a key. We implemented a number of algorithms to form these clusters, e.g. title normalizations (including lowercasing, whitespace removal, unicode normalization and other measures) or transformations like NYSIIS[Silbert, 1970].

The verification approach is based on a set of rules, which are tested sequentially, yielding a match signal from weak to exact. We use a suite of over 300 manually curated match examples[1] as part of a unit test suite to allow for a controlled, continuous adjustement to the verification procedure. If the verification yields either an exact or strong signal, we include consider it a match.

We try to keep the processing steps performant to keep the overall derivation time limited. Map and reduce operations are parallelized and certain processing steps can process 100K documents per second or even more on commodity hardware with spinning disks.

5 Quality Assurance

Understanding data quality plays a role, as the data is coming from a myriad of sources, each with possible idiosyncratic features or missing values. We employ a few QA measures during the process. First, we try to pass each data item through only one processing pipeline (e.g. items matched by any identifier should not even be considered for fuzzy

[1]The table can be found here: https://gitlab.com/internetarchive/fuzzycat/-/blob/master/tests/data/verify.csv

matching). If duplicate links appear in the final dataset nonetheless, we remove them, prefering exact over fuzzy matches.

We employ a couple of data cleaning techniques, e.g. to find and verify identifiers like ISBN or to sanitize URLs found in the data. Many of these artifacts stem from the fact that large chunks of the raw data come from heuristic data extraction from PDF documents.

6 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.

Jeffrey M Silbert. The world's first computerized criminal-justice information-sharing system-the new york state identification and intelligence system (nysiis). *Criminology*, 8:107, 1970.

7 Appendix

Number of matches	Citation Provenance	Match Status	Match Reason
934932865	crossref	exact	doi
151366108	fatcat-datacite	exact	doi
65345275	fatcat-pubmed	exact	pmid
48778607	fuzzy	strong	jaccardauthors
42465250	grobid	exact	doi
29197902	fatcat-pubmed	exact	doi
19996327	fatcat-crossref	exact	doi
11996694	fuzzy	strong	slugtitleauthormatch
9157498	fuzzy	strong	tokenizedauthors
3547594	grobid	exact	arxiv
2310025	fuzzy	exact	titleauthormatch
1496515	grobid	exact	pmid
680722	crossref	strong	jaccardauthors
476331	fuzzy	strong	versioneddoi
449271	grobid	exact	isbn
230645	fatcat-crossref	strong	jaccardauthors
190578	grobid	strong	jaccardauthors
156657	crossref	exact	isbn
123681	fatcat-pubmed	strong	jaccardauthors
79328	crossref	exact	arxiv
57414	crossref	strong	tokenizedauthors
53480	fuzzy	strong	pmiddoipair
52453	fuzzy	strong	dataciterelatedid
47119	grobid	strong	slugtitleauthormatch
36774	fuzzy	strong	arxivversion
35311	fuzzy	strong	customieeearxiv
33863	grobid	exact	pmcid
23504	crossref	strong	slugtitleauthormatch
22753	fatcat-crossref	strong	tokenizedauthors
17720	grobid	exact	titleauthormatch
14656	crossref	exact	titleauthormatch
14438	grobid	strong	tokenizedauthors
7682	fatcat-crossref	exact	arxiv
5972	fatcat-crossref	exact	isbn
5525	fatcat-pubmed	exact	arxiv
4290	fatcat-pubmed	strong	tokenizedauthors
2745	fatcat-pubmed	exact	isbn
2342	fatcat-pubmed	strong	slugtitleauthormatch
2273	fatcat-crossref	strong	slugtitleauthormatch
1960	fuzzy	exact	workid
1150	fatcat-crossref	exact	titleauthormatch
1041	fatcat-pubmed	exact	titleauthormatch
895	fuzzy	strong	figshareversion
317	fuzzy	strong	titleartifact
82	grobid	strong	titleartifact
33	crossref	strong	titleartifact
5	fuzzy	strong	custombsiundated
1	fuzzy	strong	custombsisubdoc
1	fatcat	exact	doi

Table 2: Table of match counts, reference provenance, match status and match reason. The match reason identifier encode a specific rule in the domain dependent verification process and are included for completeness - we do not include the details of each rule in this report.