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

Helge Holzmann
Internet Archive
San Francisco, CA 94118
helge@archive.org

Jefferson Bailey
Internet Archive
San Francisco, CA 94118
jefferson@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.

As of version "20210810", 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/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].

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.