Information Extraction for Scholarly Digital Libraries

Kyle Williams[‡], Jian Wu[‡], Zhaohui Wu[†], C. Lee Giles[†‡]
[‡]Information Sciences and Technology, [†]Computer Science and Engineering
Pennsylvania State University, University Park, PA 16802, USA
kwilliams@psu.edu, jxw394@psu.edu, zzw109@psu.edu, giles@ist.psu.edu

ABSTRACT

Scholarly documents contain many data entities, such as titles, authors, affiliations, figures, and tables. These entities can be used to enhance digital library services through enhanced metadata and enable the development of new services and tools for interacting with and exploring scholarly data. However, in a world of *scholarly big data*, extracting these entities in a scalable, efficient and accurate manner can be challenging. In this tutorial, we introduce the broad field of information extraction for scholarly digital libraries. Drawing on our experience in running the CiteSeerX digital library, which has performed information extraction on over 7 million academic documents, we argue for the need for automatic information extraction, describe different approaches for performing information extraction, present tools and datasets that are readily available, and describe best practices and areas of research interest.

CCS Concepts

•Information systems → Extraction, transformation and loading; Information extraction; •Applied computing → Digital libraries and archives;

Keywords

Information extraction, scholarly big data, digital libraries

1. SCHOLARLY INFORMATION EXTRACTION

In recent years, there has been an unprecedented increase in the number of scholarly documents produced. In 2014 it was estimated that at least 114 million English scholarly documents were accessible on the Web with 27% being freely available [1]. Given this ever increasing corpus of documents, it has become increasingly necessary to organize and manage the documents; allow for the exploration and discovery of new documents; explore the relationships between entities in documents; and to explore the *science of science* [2]. Digital libraries facilitate this type of exploration and discovery through interfaces based on the metadata associated with scholarly documents and information extraction refers to the identification and labeling of this metadata. However, given the rate at which scholarly documents are produced and the heterogeneous nature of the data they contain, such as tables, figures and citations, it has become increasingly difficult to perform manual information extraction at scale, thus motivating the need for automatic methods.

Scholarly information extraction refers to the process by which metadata and entities are extracted from scholarly documents using automated algorithms and systems. These systems and algorithms need to be able to deal with the heterogeneity of the data, both in terms of the format of the documents and in terms of the data contained within the documents themselves. Furthermore, the approaches must be scalable in order to deal with the millions of documents that exist and that continue to be produced.

This half-day tutorial seeks to introduce the audience to the vast area of information extraction from scholarly documents. The tutorial will explore both the practical aspects of information extraction for scholarly digital libraries as well as the research opportunities that exist. The focus will be on information extraction in a world of *scholarly big data.*

2. ABOUT THE TUTORIAL

2.1 Scope

The tutorial will focus on digital libraries of scholarly documents, such as articles, slides, academic books, and technical reports. The benefit of focusing on the scholarly domain is that it is diverse and heterogeneous, while still being well defined and understood.

2.2 Learning Objectives

Attendees should leave the tutorial understanding:

- What is scholarly information extraction?
- What is the motivation for scholarly information extraction and what are the challenges?
- What approaches are there to scholarly information extraction and what readily available tools exist?
- How does information extraction fit into the larger digital library ingestion workflow?
- What research opportunities exist in information extraction and what are best practices?

JCDL '16 June 19-23, 2016, Newark, NJ, USA

ACM ISBN 978-1-4503-4229-2/16/06.
DOI: http://dx.doi.org/10.1145/2910896.2925430

2.3 Tutorial History

While this is the first time this tutorial is being presented, it is based on several conference presentations and publications [3, 4] as well as extensive experience in scholarly information extraction and digital libraries [5, 4].

2.4 Target Audience

The target audience is technical practitioners at a beginner or intermediate level who are interested in understanding how information extraction works, either for use in their own digital libraries or as an introduction to the research area.

2.5 Presenters

This tutorial will be prepared and presented by the following members of the CiteSeerX research group.

Kyle Williams A Ph.D. candidate in Information Sciences and Technology at Penn State University who has given several presentations on information extraction for scholarly digital libraries and integrated information extraction tools into document workflows.

Jian Wu A postdoctoral fellow in Information Sciences and Technology at Penn State University and the technical director of the CiteSeerX digital library. Dr Wu has experience in designing, implementing and maintaining information extraction workflows as part of the CiteSeerX digital library.

Zhaohui Wu A Ph.D. candidate in the Computer Science and Engineering at Penn State University with experience in extracting entities from heterogeneous data types and using them to build novel digital libraries.

C. Lee Giles The director of the CiteSeerX digital library project, with extensive experience in complex systems, digital libraries and the Web.

3. TOPICAL OUTLINE

The tutorial will cover the following topics:

Motivation and Challenges

The tutorial will begin by answering questions such as: *Why perform information extraction in scholarly digital libraries*? *What are the challenges*? *What does automatic information extraction enable*? The focus will be on the heterogeneous data entities that exist in scholarly data (or scholarly entities) and on the challenges and opportunities for their extraction in the big data setting. We will cover the definitions of various types of scholarly entities and the semantic relationships among them that could form a heterogeneous scholarly knowledge base. We will show how extracting these entities can enable enhanced digital library services and tools for exploring scholarly data. This will include a description of various specialized digital libraries that we have created based on specialized information extractors.

Approaches

Information extraction is an important task with increasing research attention not only in digital libraries, but also in NLP, IR, Web/Semantic Web, data mining, and knowledge engineering. Various approaches have be used for scholarly information extraction, by taking advantage of document templates, heuristics, crowdsourcing, knowledge bases, machine learning, and the Web [6]. We will describe various approaches that exist, identify their pros and cons, and describe situations in which each method is best suited.

Tools and Data

Various tools are readily available for extracting information from scholarly documents. We will cover a variety of existing open source tools, such as SVMHeaderParse, Grobid and ParsCit, as well as tools we have developed for CiteSeerX in order to give the audience a broad overview of tools available. These will include tools for well known extraction use cases, such as text, keyphrase and citation extraction, as well as tools for richer information, such as tables, figures and data. Where appropriate, we will provide empirical comparisons of different extractors [7]. We will also describe datasets and methods that exist for evaluating information extractors, including data from CiteSeerX repositories.

Information Integration

Drawing on our experience in running the CiteSeerX digital library, we will show how various tools for information extraction can be integrated into the digital library workflow as part of the document ingestion process. We will discuss issues related to pipelining, scalability, databases and indexing and describe APIs that can be integrated into existing document workflows.

Best Practices and Going Forward

We will describe best practices in information extraction for scholarly digital libraries and highlight ongoing challenges and research problems that may be of interest to the digital libraries, information retrieval, data mining, Web, and NLP research communities.

Acknowledgments

We gratefully acknowledge partial support by the National Science Foundation.

4. REFERENCES

[1] M. Khabsa and C. L. Giles. The number of scholarly documents on the public web. *PloS one*, 9(5):e93949, 2014.

[2] *The Science of Science Policy: A Federal Research Roadmap. Report on the Science of Science Policy.* National Science and Technology Council, 2008.

[3] C. L. Giles. Scholarly big data: information extraction and data mining. In *Proceedings of CIKM*, pages 1–2, 2013.

[4] K. Williams, J. Wu, S. R. Choudhury, M. Khabsa, and C. L. Giles. Scholarly Big Data Information Extraction and Integration in the CiteSeerX Digital Library. In *Proceedings of IIWeb*, pages 68–73, 2014.

[5] H. Li, I. Councill, W.-C. Lee, and C. L. Giles. CiteSeerX: An Architecture and Webservice Design for an Academic Document Search Engine. In *Proceedings of WWW*, pages 883–884, 2006.

[6] Z. Wu, J. Wu, M. Khabsa, K. Williams, H. H. Chen, W. Huang, S. Tuarob, S. R. Choudhury, A. Ororbia, P. Mitra, C. L. Giles. Towards building a scholarly big data platform: Challenges, lessons and opportunities. In *Proceedings of JCDL*, pages 117–126, 2014.

[7] M. Lipinski, K. Yao, C. Breitinger, J. Beel, and B. Gipp. Evaluation of header metadata extraction approaches and tools for scientific PDF documents. In *Proceedings of JCDL*, pages 385–386, 2013.