Scholarly Big Data Information Extraction and Integration in the Digital Library

V.Anbarasu, B.Elakkia, V.Bhuvaneshwari

***Abstract*— A digital library search engine provides access to million numbers of documents. It stores and indexes research articles in Computer Science and other related articles. Its main purpose is to make search easier for users to search the information. It makes document available via public Website, the data is also used to facilitate research in areas like citation analysis, co-author network analysis, and scalability evaluation and information extraction. The papers are gathered from Web by means of continuous automatic focused crawling and go through a series of automatic processing steps as part of the ingestion process. In this paper, we describe how we aggregate data from multiple sources on the Web, store and manage data, process data as part of an automatic ingestion pipeline that includes automatic metadata and information extraction, perform document and citation clustering, perform entity linking and name disambiguation and make our data and source code available to enable research and collaboration. We also provide how it integrates data across the Web.**

***Index Terms*— automatic crawling, clustering, metadata.**

I. INTRODUCTION

Generally Scholarly Big Data composed of a large quantity of data related to scholarly undertaking such as journal articles, conference proceedings, theses, books, presentation slides and experimental data. There are a large number of research articles are emerging as science advances. A very large proportion of data is freely available on the Web with many articles. It is to deal with scale of the data that they collect, integrating information from multiple sources and extracting information from the data. This is become a significant as it is used for decision making in funding, education and government and also by scientists and business. The benefit of this is that the automated process which is for better scalability to collect and process the scholarly big data. It is relational such as citations, co-authorship results, research projects and results. It provides better library as it overcomes some of the challenges such as entity linking, name disambiguation and sharing data. In this we propose how documents are integrated automatically from the Web, how they extract information and how the challenges such as mentioned above are faced.

Manuscript received March 11, 2015

V.Anbarasu, Department of Information Technology, Jeppiaar Engineering College, Chennai, India

B.Elakkia, Department of Information Technology, Jeppiaar Engineering College, Chennai, India

V.Bhuvaneshwari, Department of Information Technology, Jeppiaar Engineering College, Chennai, India

II. LITERATURE SURVEY

To extract and integrate data information from the library has been widely studied by some researchers. Choudury[1] deals with search on figures in academic documents. This search engine is used to search on figures in chemistry journals articles. This system indexes figure caption and mentions extracted from the PDF using custom built extractor from the documents. It provides the framework for the extraction algorithm, architecture and ranking function. But in this system, there is no discussion of the process of figure and metadata extraction. Carman[2] deals with the search for equations. It proposes how equations in a document are extracted. It also proposes a toolkit to process and to enable users to browse and search equations in a document. But its drawback is that it does not have a formal work to extract and to process the equations from scientific documents. Lipinski[3] evaluates the performance of tools from the extraction of metadata from scientific articles. This study is a guide for developers to integrate the most suitable and effective metadata extraction tool into their software. There are many approaches to extract metadata are being proposed and examined. The general methods to extract metadata are stylistic analysis, machine learning and the use of knowledge bases. It proposes some techniques to extract metadata such as support vendor machines (SVM), hidden Markov models (HMM) or conditional random fields (CRF). But the extraction is error prone without any standards to specify the metadata should be structured or formatted. Caragea[4] uses a search engine named Citeseer^x digital library to search and indexes the research articles related to Computer Science. This is proposed to make search easier for the users. Its purpose is to create a new dataset to research community. It provides an automated technique to extract metadata. It also provides a record linkage approach to use information to clear errors from Citeseer^x. The major drawbacks is of this paper is that many titles are extracted wrongly.

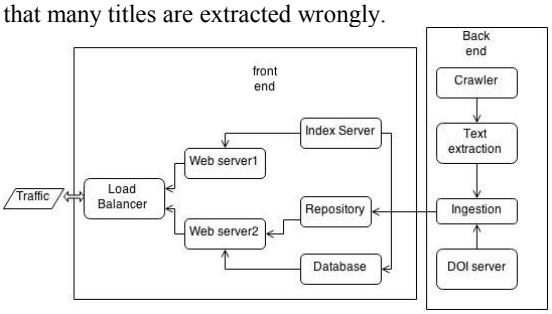

Fig.1 Digital Library Architecture and System Overview

III. DIGITAL LIBRARY

Digital Library is based on journal articles related to various fields. Its key features is that it automatically extract information from documents in the cites. These provide relatively a large number of documents. A challenge in this is the efficiency of extracting information from the documents and scaling.

3.1. Architecture and System Overview

The fig 1 shows the architecture and system overview of the library. The system overview of the digital library is that the backend is by crawler, extraction modules and backup data stores and the frontend by load balancers which passes the traffic to the Web servers. The Web servers are used to serve the user interface and interact with the repository, database and index servers.

3.2. Storage of Data

It uses three main data stores such as the index server for enabling fast searching, the master repository for storing physical files and the database for storing metadata, citations and other related information. The data stores are used often to respond to the requests for documents, thus they should be synchronized and linked. A unique identifier is used to link these data stores for the documents.

3.3 Methods to collect data

There are two methods to collect data such as focused crawling and document filtering. In focused crawling, the digital library uses two instances of Heritrix 1.14 crawler for performing focused crawling. The first crawler is used for scheduling crawls and the second is used as user submitted URL crawler. These two crawlers are to configure the PDF files. The main URLs of the crawler are taken from a whitelist. The whitelist is created according to the crawl history and it contains URLs with high quality. In document filtering, while crawling there is no means to tell whether the PDF that is retrieved is an academic paper or not. Hence the document filtering is performed when the PDF documents are crawled. In this the text are extracted from the documents and then the classification is done using the regular expression. This is a simple and error prone classification scheme. The documents which are identified as non-academic are filtered and the academic papers are used for extraction and ingestion.

3.4. Extraction of Information

Information extraction is the main in digital library. It also affects the use and quality of the service because the information which is used as metadata is automatically extracted and are used to search and interact with the site and the data collection. The extraction of information is robust and scalable, as the integration is on the academic papers which are from the Web and it is automatic. There are several information extraction modules.

3.4.1 Header Extraction

The most important type of information extraction is metadata in the form of information about papers. The extractions of information from the document are titles, authors, abstract, venue, volume and issue, page numbers, publisher and publisher address. The extraction tool used is SVMHeaderParse which is SVM-based header. Metadata Corrections and Improvement uses User Corrections which allows users for creating accounts in the system to provide metadata corrections and to add additional metadata. DBLP provides manually curated metadata from publications.

3.4.2 Citation Extraction

Citations is important in scholarly documents since they form a graph for mining to extract information based on evolution of ideas and importance of work. In each paper, the citations are extracted using Parscit citation string parsing tool. The context of the citation that is stored for analysis.

3.4.3 Extraction of Other Information

Header and citation extraction is the main extraction modules since they form the majority of metadata which the users use the most. In table extraction, the tables are automatically extracted from the documents and search the tables. The table functionality TableRank is integrated with the interface. In figure extraction, the figures and metadata are extracted from PDF files .In algorithm extraction, it includes the distributions of algorithms. It is similar to the table and figure extraction which is an extraction module with some application. Challenges for Scholarly information extraction are the accuracy, coverage and scalability. Accuracy means that the extracted information should be correct without errors. Coverage means that the classic tradeoff between precision and the recall in information retrieval. Scalability means to make the algorithms efficient and use parallel and distributed processing for map-reduce framework.

3.5. Document Clustering and Entity Linking

After the information is extracted from the document, one need to link the new data with the existing one. It takes three forms such as de-duplication and clustering, citation linking and matching and author disambiguation.

3.5.1. De-Duplication and Clustering

There are multiple versions of documents on Web. Even though it has minor differences and no bitwise identical we need to make near duplicates. Near duplicates are used to retain and cluster while the bitwise identical papers are identified by the hash algorithm such as SHA1 or MD5 and they are discarded. Clustering is performed on how the metadata information is extracted from the documents and it represents a set of near duplicates.

3.5.2. Citation Clustering and Linking

The citations of the same paper are gathered together to perform citation clustering. The method used is same as that of the document clustering method. The metadata is extracted from the document while citation string parsing takes place. There is a flag to indicate whether it contains the version of paper. The use of linking papers and citations to clusters is to make metadata for best use.

3.5.3. Author Disambiguation

It has a page to each disambiguated author who has different variations in their names, their affiliation and homepage, their index and the list of their publications.

3.6. Sharing of Data

For foster collaboration and research we need to store the data. To facilitate content dissemination, a standard is proposed by the Open Archive Initiative through we use to share the data. This is used to download the papers easily and it is also used to access the data easily. The library is open sourced. The source code can shared from the web. The extractor used in this is a stand-alone web service and is open source available freely on the web. Therefore the researchers can use it easy to extract the metadata from document.

Table.1 Extraction time and Data Size for Citations and Header Extracted from the Documents

	Citations	Header
Mean Time(std dev)	01.11 sec	2.86 sec
Total Time	111.31 sec	286.40 sec
Size	1.4 MB	152 KB

IV. Experimental Results

The following experimental setup is used. The experiments run on a machine with the following hardware and software specifications: CPU: 24 x Intel(R) Xeon(R) CPU X5650 @ 2.67GHz; RAM: 48GB; OS: Windows. The texts from millions of documents are extracted from web for the experimental purpose using PDFLib TET. The documents are submitted to the API using GPU parallel tool with 24 threads. The header and citation metadata are extracted from the document with submitted document. At last a filter is used to make the documents with less than 100 words. A experiment is conducted to evaluate its performance. A hundred of documents are submitted and we measure the time taken to process these documents. The mean time taken to extract from those files is 4.26 seconds without any duplication. It takes 0.1 sec for one file. To extract without duplication a baseline method is used which is to scale linearly. The extraction process is analyzed to improve the performance. From Table.1, the citation extraction is faster than header extraction. Compression is used for the size reduction using zlib compression library with level 3. The documents are verified successfully according to their timing to extract the header and citations.

Conclusion

In this paper, we proposed how data are integrated from Web and how it performs automatic extraction, entity linking, clustering and name disambiguation in the data. We also described how the data, code and services are shared. We have migrated from physical architecture to a virtual architecture but it is not discussed in this paper. We have described how the data is extracted from the document such as header and citation extraction and also table and figure extraction. There are about thousands PDF files are crawled per day that are scholarly data and it potentially useful for research opportunities.

Future Work

Future research can be done on the extraction methods which makes the search easier for the users such as figure extraction. The quality of data that are automatically extracted can be improved for the new sources. The extraction algorithms can be improved for better extraction automatically without any errors. The provision services and design algorithms can be made better as there are a number of opportunities to analyse log access.

References

[1] S.R.Choudury, S.Tuarob, P.Mithra, L.Rokach, A.Kirk, S.Szep, D.Pellegrino, S.Jones, and C.L.Giles,"A figure search engine architecture for a chemistry digital library," in Proceedings of the 13th ACM/IEEE-CS joint conference on Digital Libraries.ACM,2013, pp.369-370.

[2] S.Carman, "Algseer: An Architecture for Extraction, Indexing and Search Algorithms in Scientific Literature," MScThesis, The Pennsylvania State University, 2013.

[3] M.Lipinski, K.Yao, C.Breitinger, J.Beel and B.Gipp, "Evaluation of header metadata extraction approaches and tools for scientific PDF documents," Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries-JDCL'13,pp.385-486, 2013.

[4] C. Caragea, J. Wu, A. Ciobanu, K. Williams, H. Ferandez-Ramirez, Juan Chen, Z. Wu, and C.L.Giles, "CiteseerX: A Scholarly Big Dataset," in 36th European Conference on Information Retrieval(To Appear), 2013.

[5] E.Archambault,D.Amyot,P.Deschamps,A.Nicol, L.Rebout and G.Roberge, "Proportion of Open Access Peer-Reviewed Papers at the European and WORLD Levels- 2004-2011," European Commission DG Research & Innovation. August, 2013.

[6] S.R.Choudury,P.Mithra, A.Kirk,S.Szep,D.Pellegrino, S.Jones, and C.L.Giles, "Figure metadata extraction from digital documents," in the 12th International Conference on Document Analysis and Recognition (ICDAR), 2013, pp. 135-139.

[7] M.Khabsa,P.Treeratpituk, and C.L.Giles, "Ackseer: a repository and search engine for automatically extracted acknowledgements from digital libraries," in Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries. ACM, 2012, pp. 185-194.

[8] M.Khabsa,P.Treeratpituk, and C.L.Giles, "Entity resolution using search engine results," in Proceedings of the 21st ACM International conference on Information and knowledge management. ACM, 2012, pp. 2363-2366.

[9] K.Wiiliams and C.L.Giles, "Near duplicate detection in an academic digital library," in Proceedings of the 2013 ACM symposium on Document engineering- DocEng'13, 2013, pp. 91-94.