ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

A Survey on Search Systems for Extracting And Searching in Scholarly Big Data

Veena.G[#1], Jennie Mathew[#2], Jyothis Joseph.[*3]

M. Tech Student, Dept. of Computer Science, College of Engineering, Kidangoor, Kottayam, Kerala, India[#1]

M. Tech Student, Dept. of Computer Science, College of Engineering, Kidangoor, Kottayam, Kerala, India[#2]

Assistant Professor, Dept. of Computer Science, College of Engineering, Kidangoor, Kottayam, Kerala, India[*]3

ABSTRACT: Search systems that is designed to search for information. CiteSeer was a public search engine and digital library for scientific and academic papers, primarily in the fields of computer and information science, that has been replaced by CiteSeer[x] .CiteSeer[x] provides free access to over 4 million full-text academic documents and rarely seen functionalities. There is no platform available for scholars to automatically search, extract, analyse and discover algorithms, tables, figures, images, collaborators, acknowledgments, metadata, pseudo codes, semantic hierarchical sections and mathematical expressions in scholarly big data. Hence Special purpose search systems are used to extract and search from articles. The limitations of existing search engines make the search problem more challenging. In this paper a survey of search systems has been presented.

KEYWORDS: Scholarly Big Data, Extraction, Search Engines, Searching, Indexing, Ranking.

.

I. INTRODUCTION

Increasingly, special purpose search engines are used to discover document elements from scholarly documents. Academics and Researchers worldwide produce a large number of scholarly documents articles, scientific publications. Search engines such as Google Scholar, Cite Seer[x] is used to search academic documents but these systems are failed to search document elements. A document-element is defined as an entity, separate from the running text of the document that summarizes the information contained in the running text. Figures, tables and pseudo-codes for algorithms are the document-elements in scientific literature and are sources of valuable information.Several search systems are introduced to tackle these problems. Researchers and others who aim to find efficient information would have to actively search and analyse relevant new publications in their fields of study.

A significant challenge for the search systems is to deal with the scale of the data they collect, integrating information from different multiple sources and extracting meaningful information from the data. A way to automatically extract information related to document-elements from document text is by extracting information as a synopsis. Availability of a concise and relevant synopsis may help save the end user's time when they are examining search results to find something that satisfies their information needs.

II. LITERATURE SURVEY

Search Systems are used to search and discover information. Several search systems are used to extract and search information's from scholarly big data. Scholarly big data includes all academic/research documents, journal & conference papers, books and theses.

A. TABLESEER : EXTRACTING TABLES

Tables are mainly used to illustrate experimental results. This is a search engine which extracts table's metadata, indexes and rank tables which is proposed by Y. Liu, *et al*. [1]. Figure 1 shows the architecture of TableSeer. It ranks matched tables by using Table Rank algorithm. It calculates similarity measure using vector space model and includes term weighting schemes. TableSeer consists of five components. They are table crawler, table metadata extractor, table

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

metadata indexer, table ranking algorithm, table searching query interface. Figure 1 shows the architecture of the Table seer.

Table crawler crawls all the documents. Table metadata extractor consists of three parts they are text information stripper which strips out text information from PDF source and others are table box detector and table metadata extractor. Here a box-cutting method is introduced which defines page box as a rectangle of adjacently connected lines with uniform font size. Table metadata is divided into six categories they are: table environment metadata, table frame metadata, table affiliated metadata, table layout metadata, table type metadata, table cell-content metadata.The table environment metadata includes the information of the document where a table is located. The table frames metadata records whether there are frames in the four sides around a table. Table affiliated data includes table caption, table caption position, table footnote, table reference text. Table Layout Metadata helps to capture the visualization of the original table. Table cell content data refers to the values in each cell of a table and table cell type metadata records the type of a table.

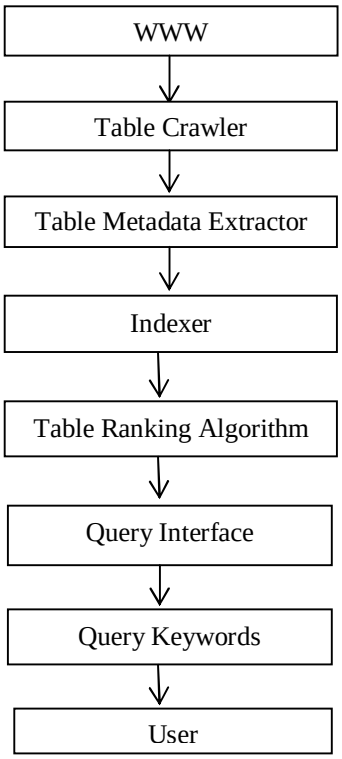

Fig.1. Architecture of TableSeer

B. ACKSEER: EXTRACTING ACKNOWLEDGMENTS

AckSeer is a search system which is proposed by M. Khabsa, *et al.* [2]. Acknowledgments are rich with information about indirect contributors and collaborators of the workHere introduced Named Entity Recognition Tools [NER] and propose a method to merge the outcome. Open Calais and Alchemy are used as NER tools. The passages are passed through multiples NERs to identify extract acknowledgment sections. The output of NER is merged together by removing duplicates. The passage along with entities and metadata are stored in a database and initiates disambiguation and clustering process inside the database. Figure 2 shows the diagram of AckSeer.

The acknowledgment extraction includes passages and entities extraction. Instance Merging and Disambiguation is the next step. Earlier work on acknowledgment indexing used dictionaries to match and merge instances of the same entity and it is a difficult task. OpenCalais and AlchemyAPI tag the extracted entities into three categories: person, company, and organization. OpenCalais is a web service that provides many knowledge extraction

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

services. AlchemyAPI is another system which provides text extraction services including text stripping, language detection, and last but not least Named Entity Extraction. And finally searching and ranking is performed.Totally AckSeer is domain independent and repository independent.

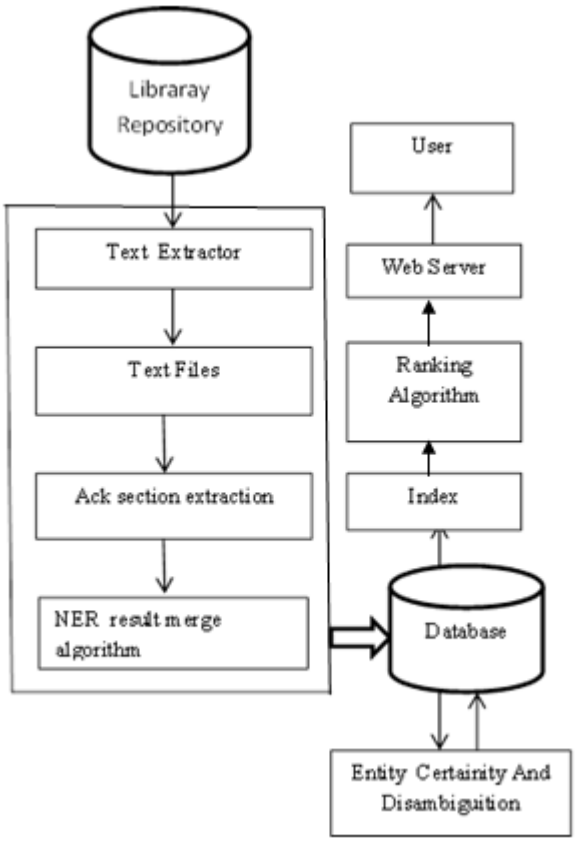

Fig.2. Architecture of AckSeer

C. COLLABSEER: EXTRACTING COLLABORATORS

CollabSeer discovers collaborators based on co-author network and is proposed by H-H Chen, *et al.* [3]. It uses three vertex similarity measures they are jaccard similarity, cosine similarity and relation strength similarity. The relation strength similarity is an asymmetric similarity which uses measures to be used in social network applications. Lexical similarity measure uses publication history to identify author's research interests. Figure 3 shows the architecture of the CollabSeer. The system consists of five components. They are co-author information analyser which retrieves data from CiteSeer^{x} dataset get a co-author network, vertex similarity to analyse the network structure, key phase extractor to extract key phrases of each article, lexical similarity integrates authors to the key phrases and the similarity integrator concatenates indexed vertex similarity score and indexed lexical similarity score to calculate collaborate recommendation.

CollabSeer also allows users to select topics of interests in order to refine the recommendations. First, CollabSeer extracts the key phrases of each document and associates the authors of the document to the key phrases. It can also use the table to infer an author's research interests. CollabSeer considers both vertex similarity and lexical similarity to recommend the collaborators.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

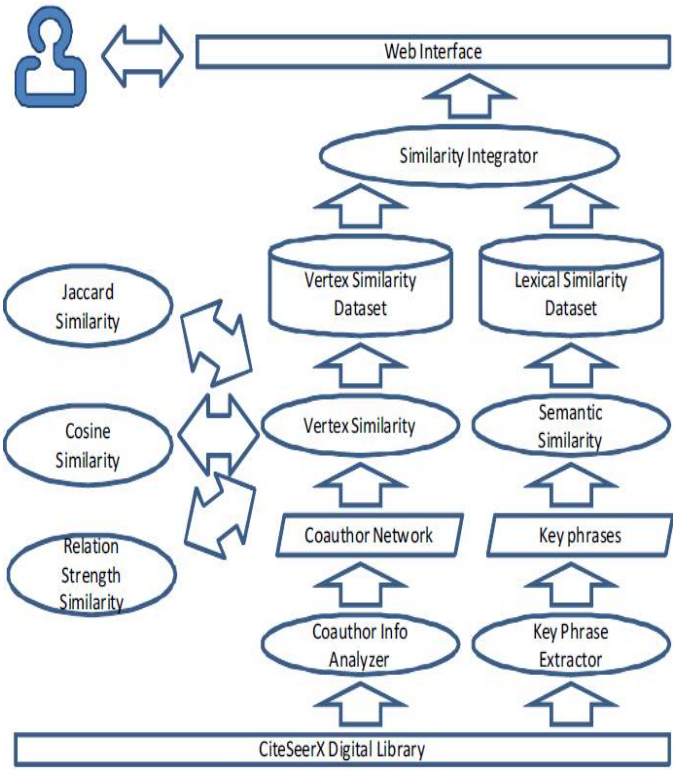

Fig. 3. Architecture of collabseer

D. CITESEER^X: EXTRACTING ACADEMIC DOCUMENTS

CiteSeer^x is a digital library and search engine for academic documents. Jian Wu *et al.*[4]. discussed the CiteSeer^x search engine.The key features of CiteSeer^x are that it automatically performs information extraction on all documents and automatic citation indexing. The backend to the system is made up of the crawler, extraction modules and backup data stores. CiteSeer^x has three main data stores, the index server that is used to enable fast searching; the master repository that stores the physical files; and the database that stores metadata, citations, and other relevant information. It make use of focused crawling while filtering out non-academic documents. CiteSeer^x currently incorporates several different information extraction modules header extraction, citation extraction and other information extraction. After that document clustering and entity linking is done. CiteSeer^x data is shared in through the Open Archives Initiative Protocol for metadata harvesting. A CiteSeerExtractor can be integrated into any application that needs to perform scholarly information extraction.

1. CiteSeerExtractor : API for extracting information

A Restful API for extracting information from scholarly big data and is proposed by Kyle Williams, *et al.* [5]. The web service is based on resource oriented architecture. Figure 4 shows the architecture of CiteSeer extractor. The Restful API is the entry point and communicates directly with the Python Web Server, which is responsible for handling the creation of resources. The web service is used for the creation and removal of the resources.

Four types of extractors are used they are text extractor, citation extractor, header extractor and body extractor. A file store is used to store documents and their associated text expressions. A NoSQL backend exists for matching near duplicates in order to avoid repetitive extraction. Since this is a major component of CiteSeerExtractor. The purpose of backend is to store the metadata.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

E. ALGORITHMSEER : EXTRACTING ALGORITHMS

A search engine for algorithms to identify and extract algorithm representations in a pool of scholarly documents. A set of hybrid techniques based on ensemble machine learning to discover pseudo-codes (PCs) and algorithmic procedures (APs) is proposed by Suppawong, *et al.* [6].First, scholarly documents are processed and identify algorithm representations. Then, textual metadata is extracted. It is then indexed and made searchable. PCs are detected by using three approaches: a rule based method (PC-RB), an ensemble machine learning based method (PC-ML), and a combined method (PC-CB).Figure 4 illustrates the architecture of the system.

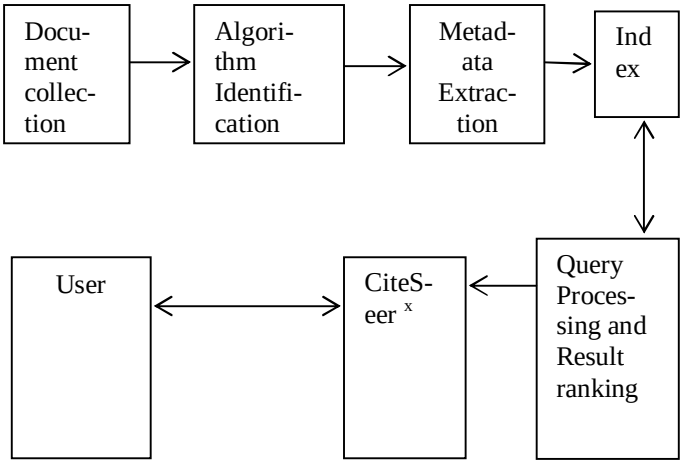

Fig.4. Architecture of AlgorithmSeer

The system specifically handles PDF documents since articles in digital libraries are in PDF formats. First plain text is extracted from PDF format .Then three sub-processes are done. The document segmentation process identifies sections in documents. PC detection process detects PCs and AP detector identifies Aps. The final step is linking together these algorithm representations. In detecting pseudo codes rule base method is used. The PC-ML detects and extracts sparse boxes and classifies each box whether it is a PC box or not. A PC box is a sparse box that contains at least 80% content of a PC. Here a feature selection for PC box classification is done. These features are font style based, content based, context based, structure based. Synopsis generation and metadata annotation are the two methods to generate additional textual metadata for PCs. Figure 5 shows the annotation process.

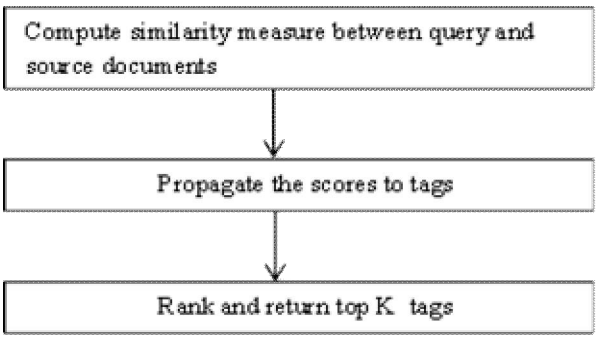

Fig.5. Diagram of annotation process.

1. A Topic Modeling Approach For Automatic Document Annotation:

To annotate documents firstly each metadata record as a tagged document, and then transform into the tag recommendation.Suppawong Tuarob *et al.*[7]. propose two algorithms for tag recommendation, one based on term frequency-inverse document frequency (TFIDF) and the other based on topic modeling (TM) using Latent Dirichlet

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 5, Special Issue 14, December 2016

3rd National Conference on Recent Trends in Computer Science and Engineering and Sustainability in Civil Engineering (TECHSYNOD'16)

19th, 20th and 21st December 2016

Organized by

Department of CSE, MCA &CE, Mohandas College of Engineering and Technology, Thiruvananthapuram, Kerala, India

Allocation (LDA).It is a generative model that allows a document to be represented by a mixture of topics. LDA model is used to find the topic distribution of document.

2. Algorithm Co-Citation Network for Algorithm Search:

Algorithm search engine only retrieves and ranks relevant algorithms. Suppawong Tuarob *et al*.[8] discussed a method for using the algorithm co-citation network to infer the similarity between algorithms. Algorithm co-citation network is an undirected, weighted graph where each node is a document that proposes some algorithms, and each edge weight is the frequency of algorithm co-citation. The algorithm citation detection returns the cited documents. The algorithm co-citation network captures the similarity between two algorithm-proposing documents, based on the conclusion that if two algorithms are cited together. Clustering the network result in groups of similar algorithm-proposing documents and could be applied in many applications to improve any algorithm search engine.

F. CHEM$_X$SEER: AN ECHEMISTRY WEB SEARCH ENGINE .

Chem$_X$Seer offering chemists automatic access to information.Chem$_X$Seer intends to or currently does provide the following search features they are data search, chemical formula search, figure search, and table search.For chemical formula search C. Lee Giles *et al*.[9]. proposed the search engine which first extracts chemical entities from text, performs indexing suitable for chemical formulae and names, and supports different query models and develop a model of hierarchical conditional random fields (HCRFs) for entity tagging. An algorithm of independent frequent subsequence mining (IFSM) is used to discover sub-terms of chemical names and estimate their probabilities of occurrence.An unsupervised hierarchical text segmentation (HTS) method is used to represent a sequence with a tree which is based on discovered independent frequent subsequences. Finally, create formulae ranking algorithms for formulae search.

III. CONCLUSION

This survey paper is concentrated on several search systems to extract and search document elements. CollabSeer discovers collaborators based on the structure of co-author network and user's research interests. Table seer extracts tables from documents and provides a user friendly search interface. Table seer ranks matched tables using Table Rank algorithm.CiteSeer^x and Chem$_x$Seer are the other search systems.AckSeer for extracting acknowledgment sections. AlgorithmSeer to extract and search algorithms from large collection of documents by using ensemble machine learning based method. This is an efficient user friendly approach.

REFERENCES

[1] Y. Liu, K. Bai, P. Mitra, and C. L. Giles.*Tableseer:automatic table metadata extraction and searchingin digital libraries*.In Proceedigs of the 7th ACM/IEEE-CS joint conference on Digital libraries, JCDL '07, Pages 91-100, New York, NY, USA, 2007.ACM

[2] M. Khabsa, P. Treeratpituk, and C. L. Giles. *AckSeer:a repository and search engine for automatically extracted acknowledgments from digital libraries*.In Proceedings of the 12th ACM/IEEE-CS joit conference on Digital Libraries,JCDL '12, Pages 185-194, New York, NY, USA, 2012. ACM.

[3] H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles. *CollabSeer: searchengine for collaboration discovery*. In Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, JCDL'11, pages 231–240, New York, NY, USA, 2011. ACM.

[4] Kyle Williamsz,Jian Wuz,Sagnik Ray Choudhuryz, Madian Khabsay and C. Lee Gilesy.*Scholarly Big Data Information Extraction and Integration in the CiteSeer Digital Library*,ACM.

[5] Kyle Williams, Lichi Li, Madian Khabsa, Jian Wu, Patrick C. Shih and C. Lee Giles. *Towards Automated restful Web Service Composition*. In IEEE International Conference on Web Services, pp. 189–196, July 2009.

[6] Suppawong Tuarob, Sumit Bhatia, Prasenjit Mitra and C. Lee Giles. *AlgorithmSeer: A System for Extracting and Searching for Algorithms in Scholarly Big Data*.IEEE Transactions on Big data,pages 3-17,April 2016.

[7] S. Tuarob, L. Pouchard, P. Mitra, and C. Giles. *A generalized topic modeling approach for automatic document annotation*.International Journal on Digital Libraries, pages 1–18, 2015.

[8] S. Tuarob, P. Mitra, and C. L. Giles. *Improving algorithm search using the algorithm co-citation network*.In Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, JCDL '12, pages 277–280, New York, NY, USA, 2012. ACM.

[9] C. Lee Giles, Prasenjit Mitra, Karl Mueller, James Z. Wang, Bingjun Sun, Levent Bolelli, Xiaonan Lu, Ying Liu, Isaac Councill, William Brower, Qingzhao Tan, Anuj Jaiswal, James,Kubicki, Barbara Garrison, Joel Bandstra.*ChemXSeer: An eChemistry Web Search Engine and Repository*,2007.IEEE.