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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 3, March 2017

Algorithm Procedures, Indexing, Mining and Pseudo Codes in Big Data

Vinod Jadhav[1], Rekha Rathore[2]

P.G. Student, Department of Computer Engineering, RKDF SOE Indore, University of RGPV, Bhopal, India

Associate Professor, Department of Computer Engineering, RKDF SOE Indore, University of RGPV, Bhopal, India

ABSTRACT: Efficient algorithms are extremely important and crucial for certain software projects. There are many Search Engines (SEs) are available but it is still hard to locate and concentrate only on materials relevant to a specific task. Recently, AlgorithmSeer, a search engine for algorithms, has been investigated as part of CiteSeerX with the purpose of providing a large database of algorithms. It has ability to automatically find and extract these algorithms in this increasingly large collection of scholarly digital documents would enable algorithm indexing, searching, discovery, and analysis. The System proposes a novel set of scalable techniques used by AlgorithmSeer to identify and extract algorithm representations *in a differerent pool of scholarly documents. Specifically, hybrid machine learning approaches are proposed to discover algorithm representations. Then, techniques for extracting textual metadata for each algorithm are discussed. And Finally, a different version of AlgorithmSeer that is built on Solr/Lucene open source indexing and search system is presented.*

KEYWORDS: AlgorithmSeer, Algorithm Procedures (Aps), Pseudo Codes(PCs),Machine Learning.

1. INTRODUCTION

Algorithms plays an Vital role in many software projects. For example, an efficient ranking algorithm helps to improve the performance of software used for indexing and searching billions of documents. Hence, it is important for software developers to keep abreast of latest algorithmic developments related to their projects[5]. There has been much work in past to help software developers search. for existing source code, however to the best of our knowledge, no effort has been made to develop tools and techniques that can help software developers search for literature about the algorithms used in source code or in documents that describe new algorithmic developments.

The relevant algorithm procedures and pseudo codes are not easily available with their analysis and it requires more efforts and time to search them. Number of algorithm procedures and pseudo codes are being published every year in national and international journals. So searching for efficient and relevant algorithm procedures and pseudo code become difficult as efforts are needed to compare and analyze them to determine the most efficient one. So, each and every research papers have to be referred. Yet the probability of acquiring relevant andefficient algorithm procedures and pseudo codes is less. To address these issues, appropriate mining techniques have to be applied. These techniques involve knowledge discovery from web and available research papers from national and international journals in order to obtain most suitable match for the input request from user. To achieve this, input request is accepted, indexing is done using proper mechanisms and most relevant algorithm procedures and pseudo codes are listed. Also,user is provided with downloading facility for algorithm procedures and pseudo codes. Hence, algorithm procedures and pseudo code extraction and analysis become an important part of this implementation.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 3, March 2017

II. BACKGROUND AND RELATED WORKS

Number of algorithm procedures and pseudo codes are being published every year in national and international journals[1]. So searching for efficient and relevant algorithm procedures and pseudo code become difficult as efforts are needed to compare and analyze them to determine the most efficient one.

A. ELEMENT EXTRACTION IN SCHOLARLY DOCUMENTS

Extractionby identifying informative entities such as mathematical expressions,, tables , figures, and tables of contents from documents have been studied.. They also proposed a method to index and search for the information, which is extracted . Sojka and Liska proposed Math Indexer and prescribed. The user may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and Searcher that interprets and collects mathematical expressions.

B. SEARCH SYSTEMS FOR SCHOLARLY INFORMATION

CiteSeer is a not only digital library but also search engine for all academic documents. Traditionally, the focus of CiteSeer_ has been on computer science, information science and related disciplines; however, in recent years there has been an expansion to include other academic fields also. The key features of this algorithm_are that it automatically performs information extraction on all documents added to the collection and automatic indexing, which allows for citations and other information among papers to be linked[1].

Fig. 1. Sample search results for the query “shortest path finding algorithm” on CiteseerX**.**

Though useful algorithms could be found in many search engines (e.g., CiteseerX9, Google Scholar10) and algorithm-related discussion forums (e.g., Stack Overflow11, Quora12), the search results can be contaminated with irrelevant items.

Fig. 1 illustrates an example of a search session (top five results) for shortest path finding algorithms using the query “shortest path searching algorithm” on CiteseerX search engine that hosts 5 million academic publications whose major proportion is computer science and related disciplines[4]. According to the figure, only two out of five top results are relevant (actual papers that describe shortest path algorithms), while the other three results are simply documents

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 3, March 2017

containing those search terms. These search engines would return the whole documents (papers or websites) as results, requiring the users to invest additional, unnecessary effort to read the entire documents to find the desired algorithms. This system is the first to explore the possibility of building a search engine specifically for algorithms extracted from scholarly digital documents.

III. THE PROPOSED SYSTEM

In this system, a prototype of an algorithm search engine, AlgorithmSeer, is presented. Fig. 2 explains the high level of the proposed system. The proposed system analyzes a document to identify any algorithms that may be present in the document. If an algorithm is found in a document, the document text is further analyzed to extract additional information about the algorithm.

All the algorithms thus found with their associated metadata are indexed and made available for searching through a text query interface. For a given user query, the system utilizes evidence from multiple sources to assign relevance scores to algorithms and results are presented to the user in decreasing order of relevance.

First, scholarly documents are processed to identify algorithm representations . Then, the textual metadata that provides relevant information about each detected algorithm representation is extracted . The extracted textual metadata is then indexed, and made searchable.

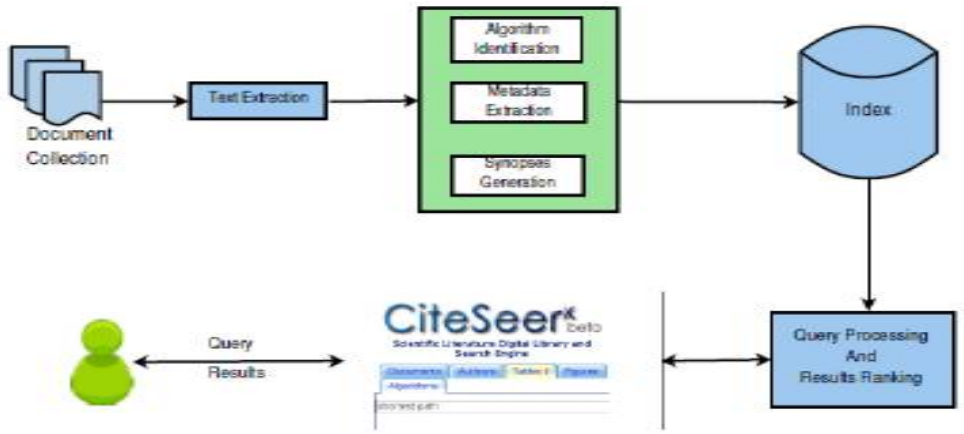

Fig. 2. Architecture of the proposed system.

IV. REFERENCE PAPER STUDY:

A.. AN ALGORITHM SEARCH ENGINE FOR SOFTWARE DEVELOPERS

In this paper, The initial effort towards developing such an algorithm search engine. The proposed system extracts and indexes algorithms discussed academic literature and their associated metadata. Users can search the index through a free text query interface. Even though popular academic literature search engines, such as Google Scholar2 and CiteSeerX3, offer the capability to search academic documents, these systems are not geared towards algorithm search. These document that does not discriminate between a document that contains an algorithm and. As a result a user searching for query shortest path will be offered many documents that contain the words shortest but do not discuss algorithmic aspects of the shortest path problem. The proposed system analyzes a document to identify any

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 3, March 2017

algorithms that may be present in the document.The initial efforts for developing a search engine for algorithms. The algorithms are extracted by proposed system and associated metadata from academic documents, offers free text query interface and presents a ranked list of results to the user by utilizing different sources of evidence. The source code for the proposed system will also be released as a part of SeerSuite toolkit.

B. ALGORITHM AND APPROACHES TO HANDLE LARGE DATA-A SURVEY

This paper presents a review of various algorithms from 1994-2013 necessary for handling such large data set. These methods implemented to handle Big Data and algorithms define various structures, also in the paper are listed various tool that were developed for analysing them. To extract this new information Data have hidden information in them; interrelationship among the data has to be achieved. Information may be retrieved from a hidden or a complex data set. Browsing through a large data set atime consuming andwould be difficult, we have to follow certain protocols , a proper algorithm and method is needed to classify the data, find a suitable pattern among them. The standard data analysis method such as clustering, factorial, exploratory analysis need to be extended to get the information and extract new knowledge treasure. Interesting association is to be found in the large data sets to draw the pattern and for knowledge purpose. The benefit of analysing the pattern and association in the data is to set the trend in the market, to understand customers, analyse demands, predict future possibilities in every aspect.

C. ENABLING INTEROPERABILITY FOR AUTONOMOUS DIGITAL LIBRARIES: AN API TO CITESEER SERVICES

CiteSeer-API, a public API to CiteSeer-like services. CiteSeer-API is SOAP/WSDL based and allows for easy programatical access to all the specific functionalities offered by CiteSeer services, including full text search of documents and citations and citation-based document discovery. CiteSeer-API is currently showcased on SMEALSearch a digital library search engine for business academic publications. Digital Libraries (DL) systems remain strongly proprietary in the way they index, store collect and present their document collections. Efforts for access normalization, such as that of the OAI-PMH community address the issue of presenting collections in a standard format that allows, if notinteroperation of those systems, creation of meta-systems are able to aggregate virtually many heterogenous collections. Theinteroperation of DL systems is however not addressed by those efforts. In this paper we consider the case of CiteSeer-like servers how they can be made more interoperable. We introduce CiteSeer-API, a SOAP/WSDL based API to CiteSeer like servers which we will serve as the corner stone for seamless interoperability with and between CiteSeer services. We provide an overview of the current functionalities supported by CiteSeer-API and outline what we feel are the next necessary steps for enabling CiteSeer services on the Semantic Web.

D. SCHOLARLY BIG DATA INFORMATION EXTRACTION AND INTEGRATION IN THE CITESEER_ DIGITAL LIBRARY

Cite Seeris a digital library that contains approximately 3.5 million scholarly documents and receives between 2 and 4 million requests per day. How we: describe 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 research and collaboration. The variety of big scholarly data is evident from the wide variety of examples given above, such as articles and lecture slides. This data is of significant interest to groups involved in decision making in funding, education and government, scientists, businesses and the general public. Given the scale of scholarly big data as well as the interest in accessing and making use of it, a number of services have arisen that collect, analyse and provide access to this data, such as Google Scholar1, PubMed2, the ArXiv3 and CiteSeer_4.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 3, March 2017

Another challenge still lies in sharing data. there are issues related to intellectual property and copyright that may limit the copying and sharing of data among different groups. Furthermore, the size of the data may also be a prohibiting factor.

E.AN ALGORITHM FOR CLASSIFYING ARTICLES AND PATENT DOCUMENTS USING LINK STRUCTURE

Studying link structure of the World Wide Web is an area which has attracted a lot of interest in recent times.Several papers published on structural analysis of hyperlinked environments such as the World Wide Web. The World Wide Web can be modeled as a graph and valuable information can be derived by analyzing links between the web-pages primarily for the purpose of building better search engines. Many novel methods have been presented to discover communities from the World Wide Web and discover authoritative web-pages. Citation analysis is a branch of information science on which plenty of research has been done. Citation analysis pertains to analysis of articles and research paper scholarly field and deriving useful information from it. It has primarily been used as a useful tool to quantify and judge the impact of a paper or a journal. Organizations like universities, government research labs and corporate research labs produce intellectual property assets in the form of research papers and patents. A collection of all or a subset of the published papers and patents by an organization needs to archived and analyzed for a variety of reasons. For example, a government research lab would like to analyze its patent portfolio to gain insights about the areas where it has filed a large number of patents versus the ones where not much patent has happened within the organization

V. RESULTS AND CONCLUSION

Algorithms are very important in solving research problems. Scientificpublications host a tremendous amount of such high-quality algorithms developed by professional researchers. In digital libraries, being able to extract and catalogue these algorithms will introduce a number of exciting applications including algorithm searching, discovering, and analysing. Specifically, proposed system is a set of scalable machine learning based methods to detect algorithms in scholarly documents. The annotation methods for documents are extract textual metadata for pseudo-codes are discussed, and how algorithms are indexed and made searchable.

VI. FUTURE WORK

Future work would be the semantic analysis of algorithms, their trends, and how algorithms influence each other over time. Such analyses would give rise to multiple applications that could improve algorithm search. We can add document files also; Robustness of the system will increase.

REFERENCES

[1] Suppawong Tuarob, Member, IEEE, Sumit Bhatia, PrasenjitMitra, Senior Member,"AlgorithmSeer: A System for Extracting and Searching for Algorithms in Scholarly Big Data" IEEE TRANSACTIONS ON BIG DATA, VOL. 2, NO. 1, JANUARY-MARCH 2016.
[2] Sumit Bhatia, Suppawong Tuarob, Prasenjit Mitra and C. Lee Giles"An Algorithm Search Engine for Software Developers" The Pennsylvania StateUniversity Park, PA-16802, USA.
[3] Yves Petinot1, C. Lee Giles1,2,3, Vivek Bhatnagar2,3, Pradeep B. Teregowda1, Hui Han1"Enabling Interoperability For Autonomous.
[4] Suppawong Tuarob, Prasenjit Mitra and C. Lee Giles" Improving Algorithm Search Using the Algorithm Co-Citation Network"
[5] Suppawong Tuarob, Sumit Bhatia, Prasenjit Mitra, C. Lee Giles " Automatic Detection of Pseudocodes in Scholarly Documents Using Machine Learning".
[6] Kyle Williamsz, Jian Wuy, Sagnik Ray Choudhuryz, Madian Khabsay, C. Lee Gilesy"Scholarly Big Data Information Extraction and Integration in the CiteSeer_ Digital Library"