International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 03 | Mar -2017 www.irjet.net p-ISSN: 2395-0072

ALGORITHM PROCEDURE AND PSEUDO CODE MINING

Jayshri Khirari, Nikita Barhate, Yogita Wani, Ashwini Zine

1.Student,Dept. Of Computer Engineering,KKWIEER college,Maharashtra,India.
2.Student,Dept. Of Computer Engineering,KKWIEER college,Maharashtra,India.
3.Student,Dept. Of Computer Engineering,KKWIEER college,Maharashtra,India.
4.Student,Dept. Of Computer Engineering,KKWIEER college,Maharashtra,India.

Abstract: *Algorithm Procedures(AP's) and Pseudo Codes(PC's) are important source of information for academicians, researchers, developers, scientists, innovators from various technology domain. 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 and efficient 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.*

Key Words: Extraction, Indexing, Regular expression, Analysis, PDFtoTEXT.

I.Introduction

In computer science, Algorithm Procedures (AP's) and Pseudo Codes (PC's) are important source of information for developing and analyzing various applications. This Algorithm Procedures and Pseudo Codes are used by academicians, researchers, developers, scientists, innovators from various technology domain. The AP's and PC's are effective method which contains finite set of instructions that produces the desired output within finite time and space complexity. Number of algorithm procedures and pseudo codes are being published every year in national and international journals. Standard algorithms and pseudo codes are available in different research papers, textbooks, encyclopedia, Wikipedia, etc.[1] As every year number of new algorithms and pseudo codes are being published, so each and every time searching for new relevant algorithms and pseudo codes is not feasible. So, user has to refer many research papers. Hence for searching 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. Yet the probability of acquiring relevant and efficient algorithm procedures and pseudo codes is less.

To search the algorithms and pseudo codes manually is a tedious task. Many algorithms and pseudo codes solutions are available for single problem.To address these issues, we would like to have a system which will automatically search algorithms and pseudo codes and gives relevant and efficient APs and PCs by applying appropriate mining techniques. 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, different techniques are applied like extraction of APs and PCs from web and available research papers from national and international journals and then 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 which results into easy and relevant search for users.

2. Implementation

2.1. Related work.

Aside from well-known web search engines such as Google5 and Microsoft's Bing6, various vertical search engines have been proposed. CiteSeer7, now CiteSeerX, was developed as a scientific literature digital library and search engine which automatically crawls and indexes scientific documents primarily in the field of computer and information science [2].
Liu, et al., presented TableSeer, a tool which automatically identifies and extracts tables in digital documents [3]. They used a tailored vector-space model based ranking algorithm, TableRank, to rank the search results. An implementation of TableSeer that extracts and searches for tables in the CiteSeerX document repository has been included in the CiteseerX suite. BioText8 search engine, a specialized search engine for biology documents, also offers the capability to extract figures and tables, and make them searchable [4]. Khabsa, et al., described AckSeer, an acknowledgement search engine that extracts, disambiguates, and

indexes more than 4 million mentioned entities from 500,000 acknowledgments from documents in CiteSeerX [5]. Chen, et al., emphasized the importance of scientific collaboration and introduced CollabSeer, a search engine for discovering potential collaborators for a given author or researcher by analyzing the structure of the coauthor network and the user's research interests [6].

2.2. Drawbacks of existing system.

1. In citeseerx, searching the relevant algorithms and pseudo codes is difficult as number of research papers are available.

2. As indexing mechanism is not applied so efficient APs and PCs are not displayed.

3. User has to refer thorough out research papers, to find appropriate APs and PCs. So it becomes a time consuming task.

4.The downloading facility is not available.

2.3. Proposed system

In this paper, system is developed in such a way that the APs and PCs are extracted from the research paper and web by. The extracted APs and PCs are stored in the database. Then the APs and PCs are analyzed with respect to application, time and space complexity. User has to give input request in the keywords form of APs and PCs. The system will display the relevant APs and PCs using the appropriate indexing mechanisms with downloading facility for APs and PCs.

2.4. Modules.

Basically there are five modules in the system:

1. PDF to Text Conversion.

2. Extraction of APs and PCs.

3. Analysis.

4. Indexing.

5. Display APs and PCs.

2.4.1 PDF to text Conversion.

As the research paper, are available in PDF form. So, PDF must be converted in text form for extraction. Extracting the AP's and PC's from PDF is tedious task so, initially the research papers is converted into text form using PDF TO TEXT libraries. Thus, the converted text is used for extracting the algorithm procedures and pseudo codes.

2.4.2 Extracting the APs and Pcs.

In Extraction, the APs and PCs are extracted using different techniques. This techniques includes regular expression and machine learning. In the research paper, using regular expression the keywords are matched. The keywords are like 'algorithm', 'pseudo codes', 'start', 'end', 'step', 'begin', initialize', etc. As soon as the keyword are matched the APs and PCs are marked in the text file. The marked APs and PCs are saved into the database.

An algorithmic procedure is a set of descriptive algorithmic instructions and differs from a PC in the following ways:

1. Writing style. PCs are usually written in a programming style, with details omitted. Symbols, Greek letters, mathematical operators, and programming keywords (such as 'for', 'begin', 'end', 'return', etc.)
are usually used to compose PCs. On the other hand, APs are usually written in a listing style, with a descriptive manner. Each step usually begins with a bullet point, or a number. APs lack the power to express complex nested loops and are less concise than PCs, but they are easier to comprehend by general readers who do not have a programming background.

2. Location in documents. PCs are usually not part of the running text; they may appear anywhere in the documents. Because of this, most PCs have identifiers which the context in the document can refer to. These identifiers include captions, function names (e.g., 'APPROXMAX-SAT(g, S , p)'), and algorithm names (e.g., 'Algorithm BuildGalledNetwork'). On the contrary, algorithmic procedures mostly appear as part of the running text, and hence do not have unique identifiers. Hence, detecting APs would require a different set of techniques

2.4.3 Analysis of APs and Pcs.

In Analysis module, the APs and PCs which are stored in Database that are analyzed with respect to application, time and space complexity. Thus using this analysis phase, the APs and PCs are determine that which AP and PC is best, worst or average

2.4.4 Indexing of APs and PCs.

Using Indexing mechanism, the analyzed APs and PCs are indexed, so that the relevant APs and PCs are displayed to fulfill user request. And thus the APs and PCs are able to download.

3. System Architecture.

The System Architecture is given as below:

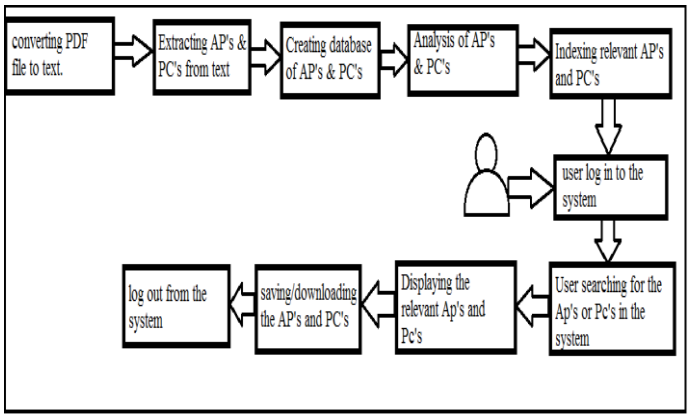

Fig.3.1 System Architecture

The architecture diagram gives the total representation of the modules and their execution flows. There are total five main modules to be designed in the system as:
1.Conversion of PDF document into text.
2. Extraction of algorithm procedures and pseudo codes.
3.Storing the extracted data into the database.
3. Analysis of algorithm procedures and pseudo codes.
4. Indexing of algorithm procedures and pseudo codes.
5. Displaying the most relevant algorithm procedures and pseudo codes.
6. Downloading the algorithm procedures and pseudo codes.

The APs and PCs that are detected from research papers and web are extracted and stored in the database. Then the APs and PCs are analyzed by calculating the time and space complexities and applications of APs and PCs. After that the APs and PCs are indexed to get the relevant results. Thus when the user will request for APs or PCs the list of algorithms or Pseudo codes will be displayed and user will able to download the required Algorithm or pseudo codes.

4 .Experimenting Results.

4.1 converting the pdf document to text.

In the system,the pdf document from which the algorithm procedures and pseudo codes are to be extracted is converted into the text. As the extracting from pdf document is very tedious task so, converting pdf into text is must.For converting the "PDFTOTEXT" libraries is used.The results are shown in the screen shot format as below:

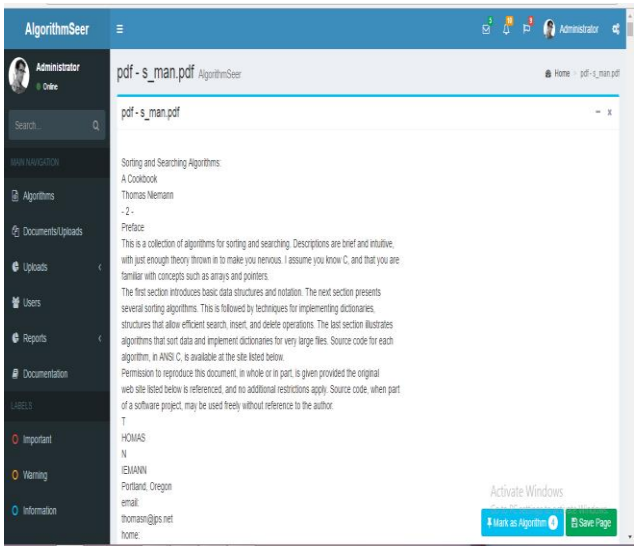

fig.4.1.PDF TO TEXT.

4.2 Evaluating Algorithm Procedures.

The AP's are detected using the writing styles such as, "start", "begin", "end", etc. Using the regular expression the detected algorithm procedures are extracted from the file. The algorithm procedures are searched by the user using the keywords of algorithms.

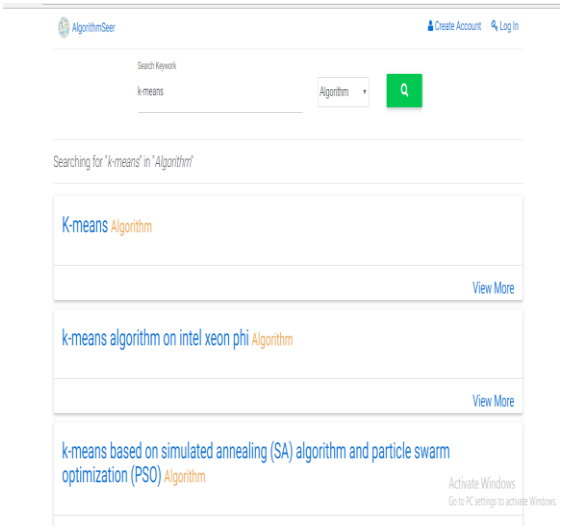

Fig.4.2.algorithm procedure

4.3 Evaluating Pseudo codes.

The PC's are detected using the writing styles such as, "int", "begin", "for", "end" etc. Using the regular expression the detected pseudo codes are extracted from the file. The pseudo codes are searched by the user using the keywords of pseudo codes.

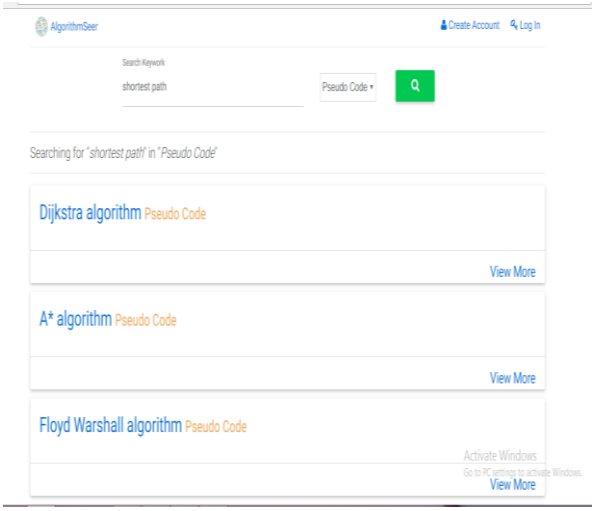

Fig.4.3.Pseudo codes

5. Conclusion.

Algorithm procedures and pseudo codes plays vital role in research and development centers, academicians, innovators, scientists, etc. This system is a search technique to extract and analyze relevant and efficient algorithm procedures (APs) and pseudo codes(PCs) from research papers and web. As number of options are available, choosing most appropriate is challenging task. So the system uses , indexing mechanism to display the most relevant algorithm procedures and pseudo codes. Thus, efforts required in terms of time and manpower are reduced. Hence,system proposes a efficient and relevant search engine to search the appropriate algorithm procedures and pseudo codes with its time complexity,space complexity and application.

References.

[1]"AlgorithmSeer: A System for Extracting and Searching for Algorithms in Scholarly Big Data.",Suppawong Tuarob, Sumit Bhatia, Prasenjit Mitra.
[2] H. Li, I. Councill, W.-C. Lee, and C. L. Giles, "Citeseerx: An architecture and web service design for an academic document search engine," in Proc. 15th Int. Conf. World Wide Web, 2006, pp. 883–884.
[3] Y. Liu, K. Bai, P. Mitra, and C. L. Giles, "Tableseer: Automatic table metadata extraction and searching in digital libraries," in Proc. 7th ACM/IEEE-CS Joint Conf. Digital Libraries, 2007, pp. 91–100.
[4] M. A. Hearst, A. Divoli, H. Guturu, A. Ksikes, P. Nakov, M. A. Wooldridge, and J. Ye, "BioText search engine: Beyond abstract search," Bioinformatics, vol. 23, no. 16, pp. 2196–2197, 2007.
[5] M. Khabsa, P. Treeratpituk, and C. L. Giles, "Ackseer: A repository and search engine for automatically extracted acknowledgments from digital libraries," in Proc. 12th ACM/IEEE-CS Joint Conf. Digital Libraries, 2012, pp. 185–194.
[6] H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles, "Collabseer: A search engine for collaboration discovery," in Proc. 11th Annu. Int. ACM/IEEE Joint Conf. Digital libraries, 2011, pp. 231–240.