Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 12, December 2014, pg.498 – 506

RESEARCH ARTICLE

IMPLEMENT A TOOL TO EXTRACT AND ANALYSE PATTERNS FROM WEB DATA

Ashima Miglani
Pursuing M.Tech from Royal College of Engg
Mr. Jitender Arora
(Head of Deptt, CSE)
Department of Computer Science, Dcrust, India
ashimamiglani2312@gmail.com

Abstract- This document emphasizes on research issues in web mining. Web mining provides high performance system to the users to search for the product and obtains information of a particular product by searching through the servers that contains the sources. Web Data Extraction (WDE) is an important problem that has been studied by means of different scientific tools and in a broad range of applications. Web usage mining (WUM) is the application of data mining techniques (DMT) to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. The proposed work will use the featured analysis based approach. The keyword extraction (KE) and analysis based approach will be done dynamically using clustered approach to perform the document match over web.

Keywords- Web Data Extraction (WDE), Web Usage Mining (WUM), Clustered Approach, Keyword Extraction (KE), Data Mining Techniques (DTM)

I. INTRODUCTION

The explosive growth and popularity of the world-wide web has resulted in a huge amount of information sources on the Internet. Data mining is the process of extracting interesting and meaningful information from web and analyze to discover some useful knowledge. The resulting growth in on-line information combined with the almost unstructured web data necessitates the development of powerful yet computationally efficient web data mining tools. The three categories used for web mining are given below:

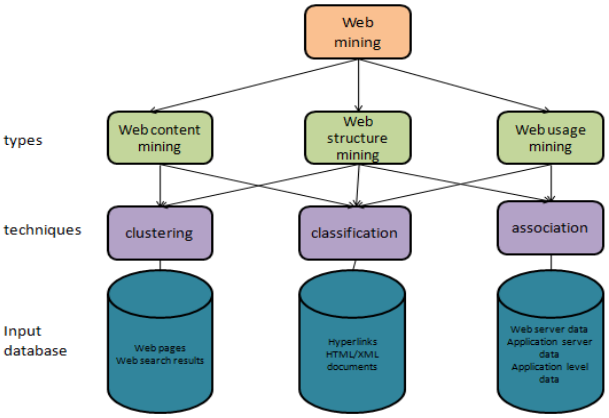

Fig.1 Categories of Web Mining

A. Web Content Mining

It is the process of extracting useful information from web with the generation of wrappers. Wrappers basically comprises of set of certain rules that either manually or automatically contributes in extracting data. Techniques used for web content mining are clustering, classification, document tree extraction, labelling attributes for result.

B. Web Structure Mining

It is the process of extracting useful information in the form of structures. Databases, however, require structured data and most Web data is unstructured and cannot be queried using traditional query languages. To attack this problem, various approaches for querying the Web have been suggested. The methods that are done in the web usage mining are Data cleaning, Transaction identification, Data integration, Transformation, Pattern Discovery, Pattern Analysis.

C. Web Usage Mining

It is the process of Discovering meaningful patterns from data generated by client-server transactions on one or more Web localities. In this process interesting usage patterns are discovered which includes server data, Application server data and Application level data.

Challenges of Web Data Mining Techniques:

1) Web Data Extraction techniques normally requires the help of human experts. The challenge comprises of providing a high degree of automation by reducing human efforts as much as possible. Main focus is on creating a balance between highly automated procedures and getting accurate performance.
2) Machine based learning approach requires large set of manually created labeled Web pages. But these pages requires lot of time and error-prone, so there existence is quite difficult.
3) These techniques must provide solid privacy guarantees, as users want their personal data to be secure. Therefore attempts to private data should be timely and adequately identified and counteracted.
4) The extraction process occurs routinely and evolve over time. But the problem is that web sources are changing continuously and these changes are unpredictable. So, these techniques must be capable of maintaining the systems and must be enough flexible to adapt to these structural changes.

II. RELATED WORK

Nanpoulos et al. [6] proposed a method for discovering access patterns from web logs based on a new type of association patterns. They handle the order between page accesses, and allow gaps in sequences. They use a candidate generation algorithm that requires multiple scans of the database. Their pruning strategy assumes that the site structure is known.

Srikant and Agrawal [9] presented an algorithm for finding generalized sequential patterns that allows user-specified window-size and user-defined taxonomy over items in the database. This algorithm required multiple scans of the database to generate candidates.

Parthasarathy et al. [5] introduced a mining technique given incremental updates and user interaction. This technique avoids re-executing the whole mining algorithm on the entire data set. A special data structure called incremental sequence lattice and a vertical layout format for the database are used to store items in the database associated with customer transaction identifiers. Their performance study has shown that the incremental mining is more efficient than re-computing frequent sequence mining process from scratch. However, the limitation of their approach, as they point out, is the resulting high memory utilization as well as the need to keep an intermediate vertical database layout which has the same size as the original database.

Liu et al. [10] proposed a clustering method based on a mixture of Markov models to cluster users and capture the sequential relationships hidden in user web navigation histories. The performance of this method is higher than the traditional Markov models, the association rules, or clustering methods.

Fu et al. [13] were one of the early researchers to propose the idea of generalizing web data and integrated this with a clustering algorithm to extract web access patterns. A page hierarchy is used to generalize sessions by replacing actual page-clicks with their general URLs. For example, a page like /programs/ugrad/cs/ is replaced by /programs/ugrad/or /programs/ depending on a pre-determined generalization level. This level, which is critical to both the efficiency and effectiveness of the approach, is a user-specified parameter.

Yang et al. [12] presented an application of web log mining that combines caching and prefetching to improve the performance of internet systems. In this work, association rules are mined from web logs using an algorithm called *Path Model Construction* [11] and then used to improve the GDSF caching replacement algorithm. These association rules assumes order and adjacency information among page references.

III. PROPOSED APPROACH

A. Problem Statement

Before study is carried out it is necessary to state the problem in clear words so the problem can be understood by the reader. The problem of duplicate detection is known since long times and several communities have worked on it using different terminology. Numerous web mining applications depend on the accurate and proficient identification of duplicate and near duplicates. Document clustering, detection of replicated web collections, detecting plagiarisms, community mining collections

in a social network site, collaborative filtering and discovering large dense graphs are a notable few among those applications Check summing techniques can determine the documents that are precise duplicates (because of mirroring or plagiarism) of each other. Detection of duplicate web pages in a fast way has great importance; because users want to reach information as quickest as possible and if duplicate detection begins to slow down the access to the information. Therefore there is a need to develop an approach that can recognize the duplicate and near duplicates web pages in an efficient way. So, that. Reduction in storage costs and enhancement in quality of search indexes besides considerable bandwidth conservation can be achieved by eliminating the duplicate and near duplicate pages.

B. Objectives of the Study

The study under consideration is primarily devised to achieve the following objectives:

1. To produce a report containing all the original web pages without any duplication or near duplication in the real time application.
2. To study and manage a database of related urls and the keywords.
3. To design an approach to compare two WebPages respective to duplicate data.
4. To perform page indexing.

C. ***Scope of the Study***

Besides piracy one of the problems on the Internet these days is redundant information, which exist due to replicated pages archived at different locations like mirror sites. Since the amount of information available on the Internet increases on a daily basis, filtering redundant and similar documents becomes a more difficult task to the user. In order to use the information available on the Web many technologies emerged, information retrieval systems is one of them. But the presence of duplicate pages decreases both effectiveness and efficiency of search engines. Because duplicate results for user queries decrease the number of valid results of the query and this also decreases system effectiveness. Processing duplicate results is time-consuming and does not add any value to the information presented to the user. So, duplicate documents decrease the efficiency of a search engine. In this dissertation an approach that can detect the replication of web pages has developed so as to reduce the search time and reduce the memory space in the repository.

D. Research Design

A design enables us to summarize a complex design efficiently. Our approach for detecting duplicate web pages in web crawling use constructive, analytical and exploratory research design. Constructive research design to get the objectives clearly defined, analytical research design to use facts or information already available, and analyze these to make a critical evaluation for research. Exploratory research design is used for developing a search engine that can detect duplicate web pages and to come up with results which are capable of avoiding duplicate web pages.

E. Analysis

There is a need for developing an approach that can detect duplicates in web documents efficiently and effectively. Crawled web pages are preprocessed using document parsing which removes the HTML tags and java scripts present in the web documents followed by the removal of common words or stop words from the crawled pages. Stemming algorithm is applied to filter the affixes (prefixes and the suffixes) of the crawled documents in order to get the keywords. Finally, the similarity score between two web pages is calculated on basis of the extracted keywords. The pages with similarity scores greater than a predefined threshold value are considered as duplicate.

F. Flow Chart

Here the basic flow chart of proposed work is defined. As we can see in figure, Flow Chart of Text Summarization:

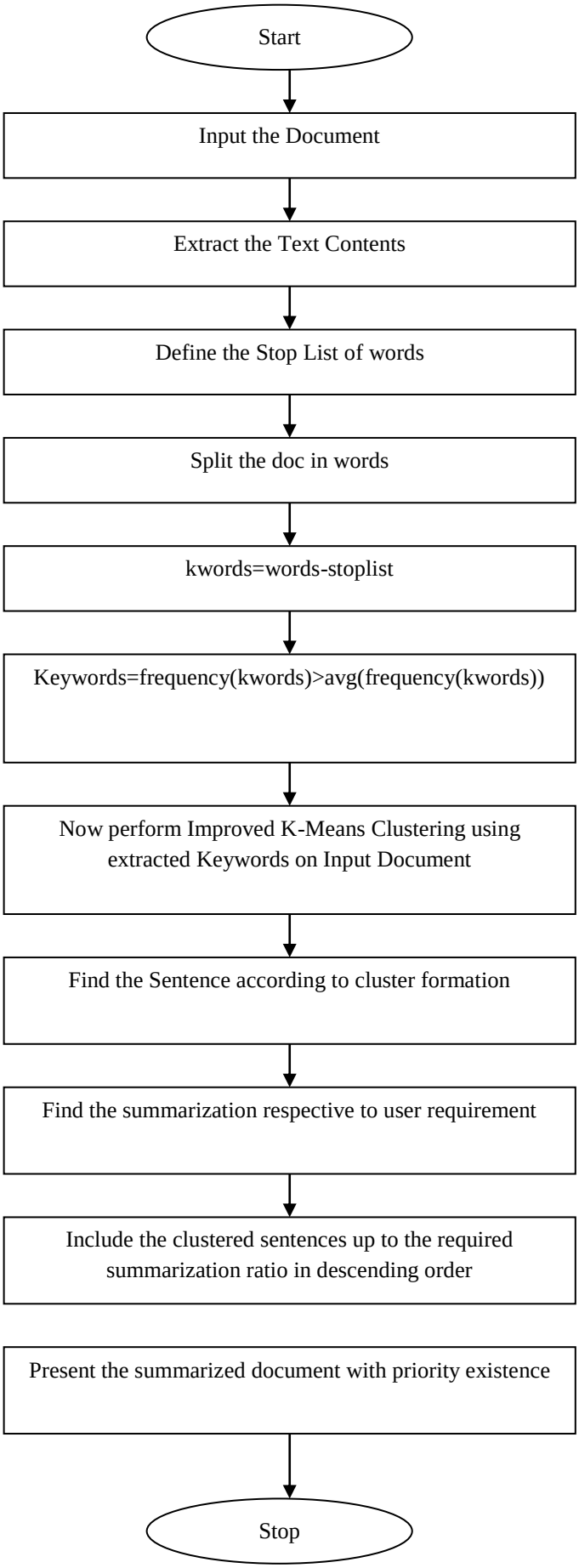

Fig.2 Text Summarization Procedure

The steps included in the research are given as

- The system will first parses the query in natural language and finds the major parts in the string.
- Then first it will look for the table name and then it parses the string .
- After parsing it will construct the parse tree of the abstracted symbols.
- Once the parse tree is generated will analyze the prioritization and the frequency of the abstracted symbols.
- All these symbols and keywords will be documented in a table.
- Now we will analyze the user requirement of summarization
- Finally we will extract all the sentences having the same keywords respective to the priority and the user requirement

G. Final Analysis.

A search engine finds information for its database by accepting listings sent in by authors who want exposure, or by getting the information from their "web crawlers," "spiders," or "robots," programs that roam the Internet storing links to and information about each page they visit. A web crawler is a program that downloads and stores Web pages, often for a Web search engine. Roughly, a crawler starts off by placing an initial set of URLs, So, in a queue, where all URLs to be retrieved are kept and prioritized. From this queue, the crawler gets a URL (in some order), downloads the page, extracts any URLs in the downloaded page, and puts the new URLs in the queue. This process is repeated until the crawler decides to stop. Collected pages are later used for other applications, such as a Web search engine or a Web cache. Fig. 3 represents the analysis of proposed approach.

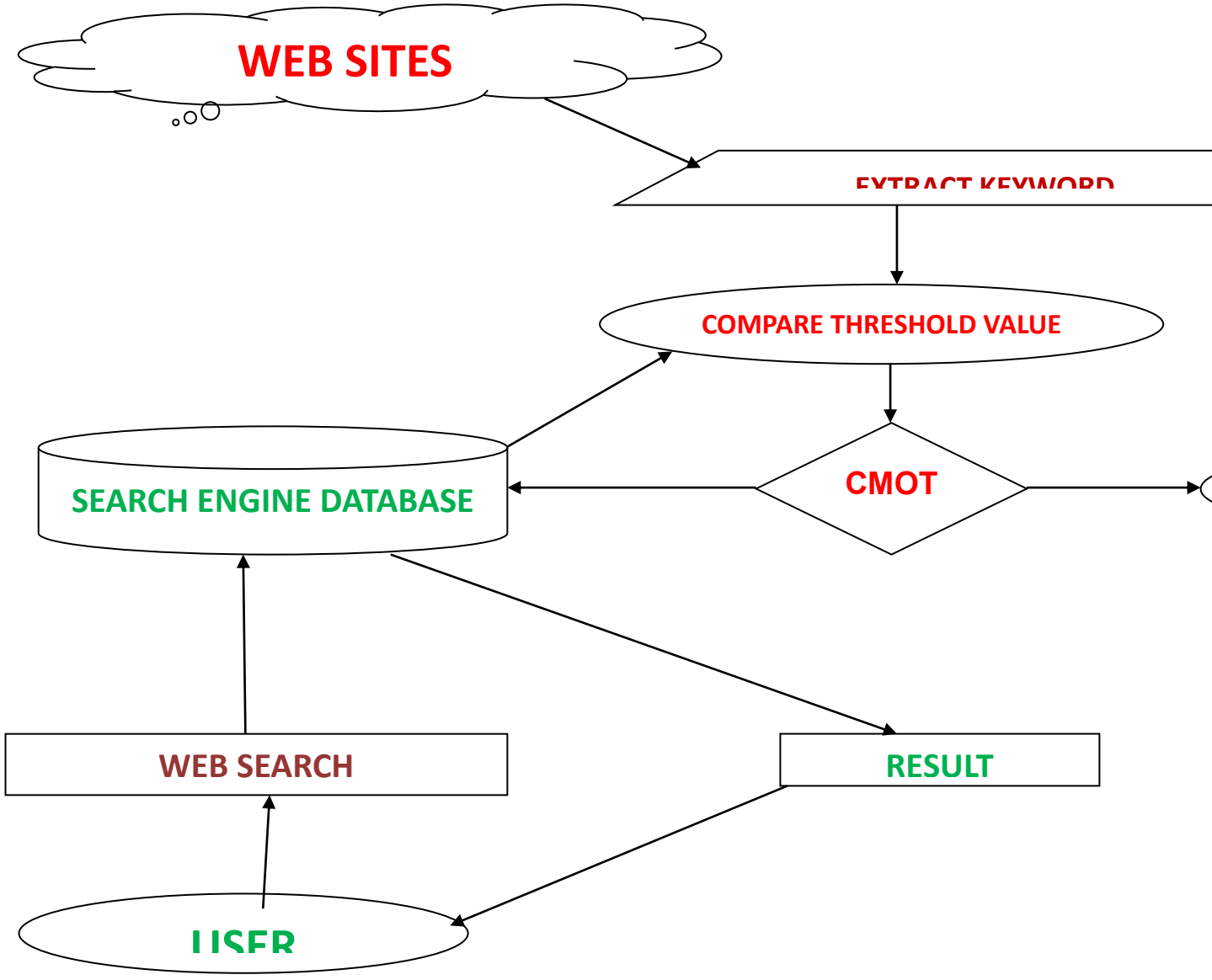

Fig.3 Analysis of proposed approach

IV. EXPERIMENTAL RESULTS

The results driven from the model and the algorithmic implementation of the proposed work. The model and the algorithm is implemented in the form of a web application. This application is defined in java by using the concepts of JSP and the servelets. At the final stage, the analysis work is defined in the form of graphs. The GUI of proposed work is designed in similar to the search engine. This GUI is designed using html tags. Here the screen shot of the GUI is shown in fig.4 below:

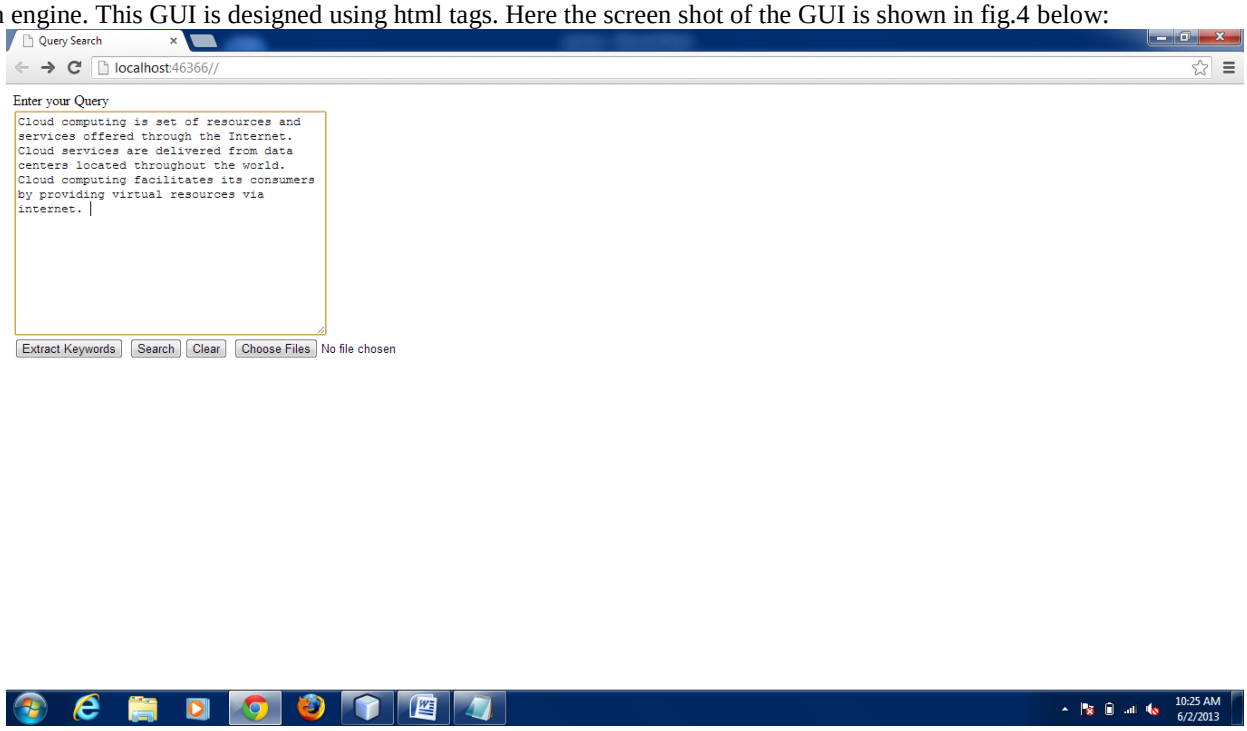

Fig.4 Gateway to plagiarism search

The figure is showing the basic gateway to the system. As we can see, the screen is having a text area to accept the user query text. The user can perform the query filtration by using extract keyword button and finally the query can be submitted to the system by using submit button.

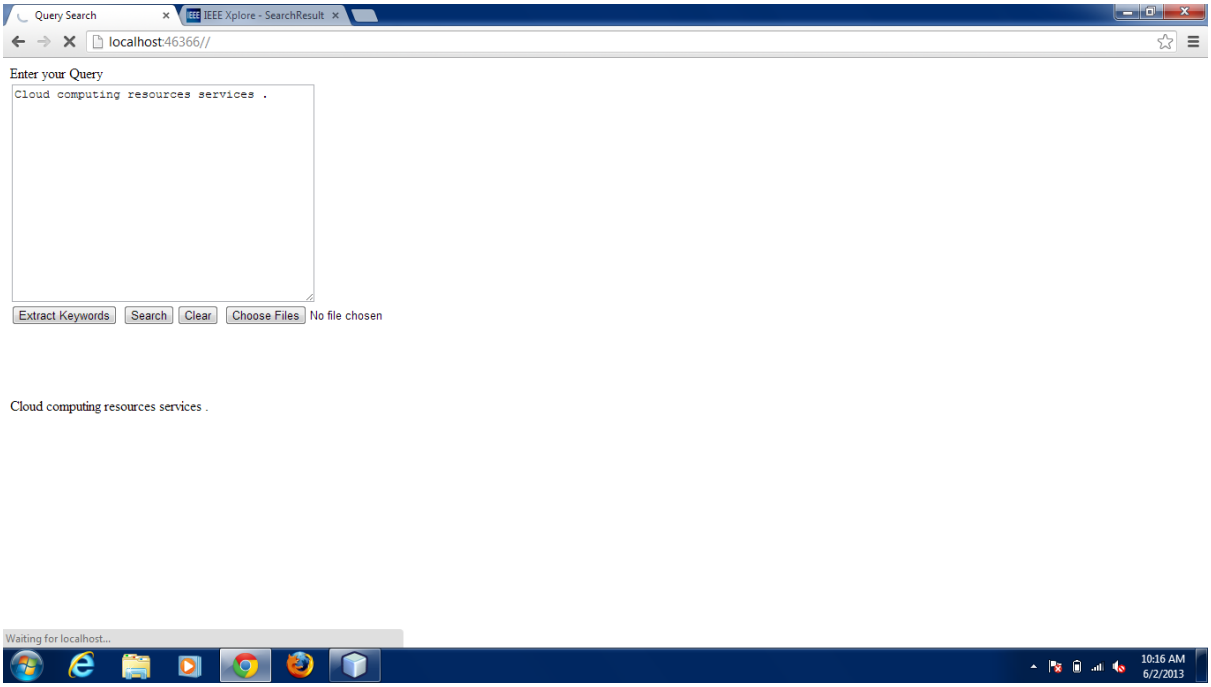

Fig 5. Keyword Extraction

Once the raw query is input by the user, the next work is of user to perform the filteration on this query. The filteration process is shown in fig 5.The filteration is here performed in terms of removal of stop list words. And performing the frequency analysis of these keywords. The high frequency words will be taken as the main query keywords.

Fig.6 Results

This figure is showing the results obtained from the system. The results are here presented in the form of url that is having the user input contents with the ratio specification in terms of duplicate data found over that web site.

V. RESULTS ANALYSIS

Analysis is done on the basis of different keywords. Here the screen shots of the results are obtained from different sites. The Blue Blocks represent the existing keywords on different sites and Red Blocks represent the keywords matched from the Proposed Work with the keywords existing on different sites.

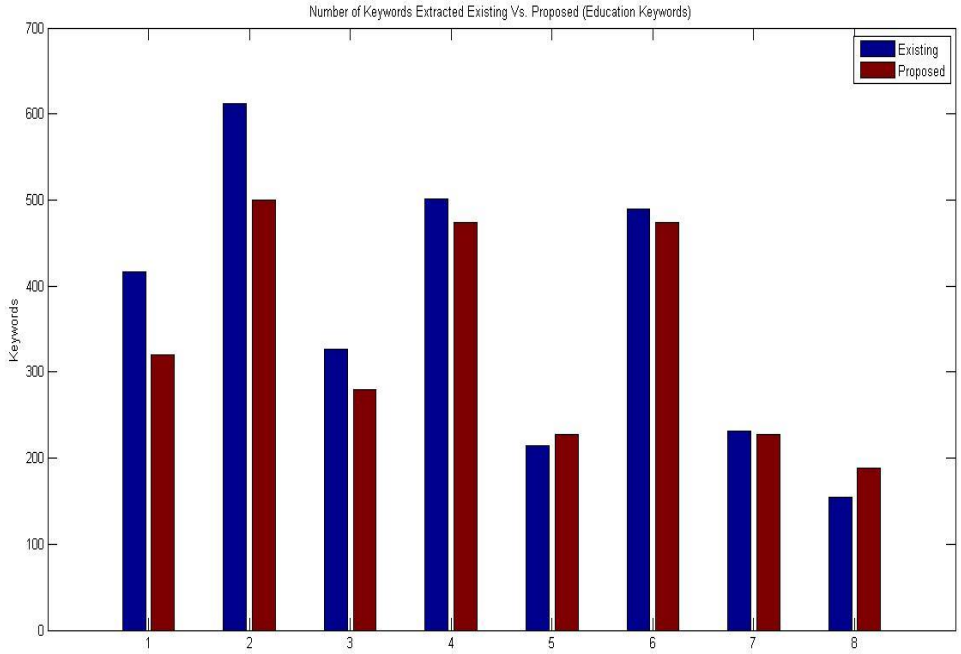

Fig.7 Education Keywords

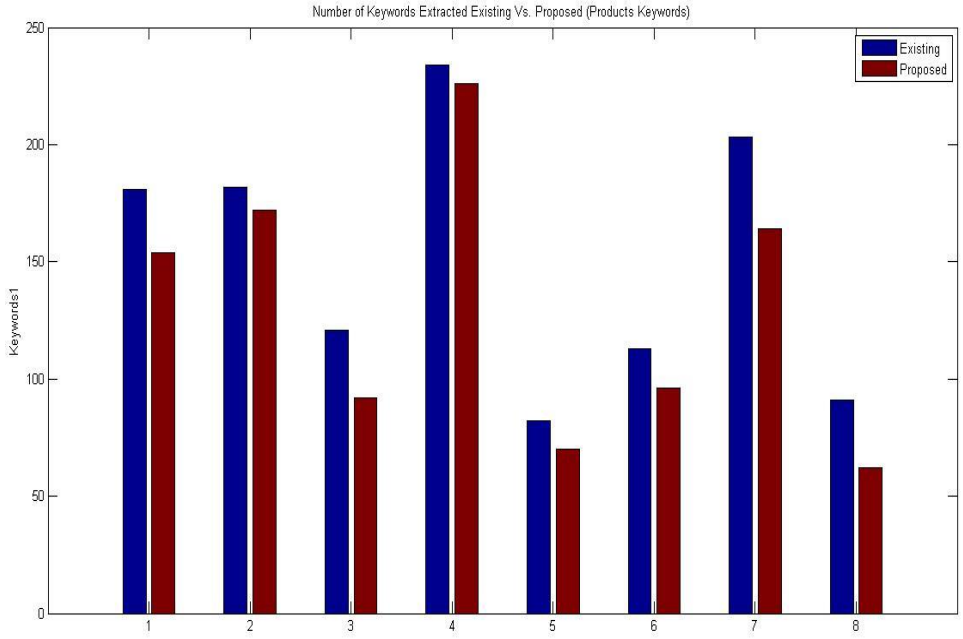

Fig. 8 Product Keywords

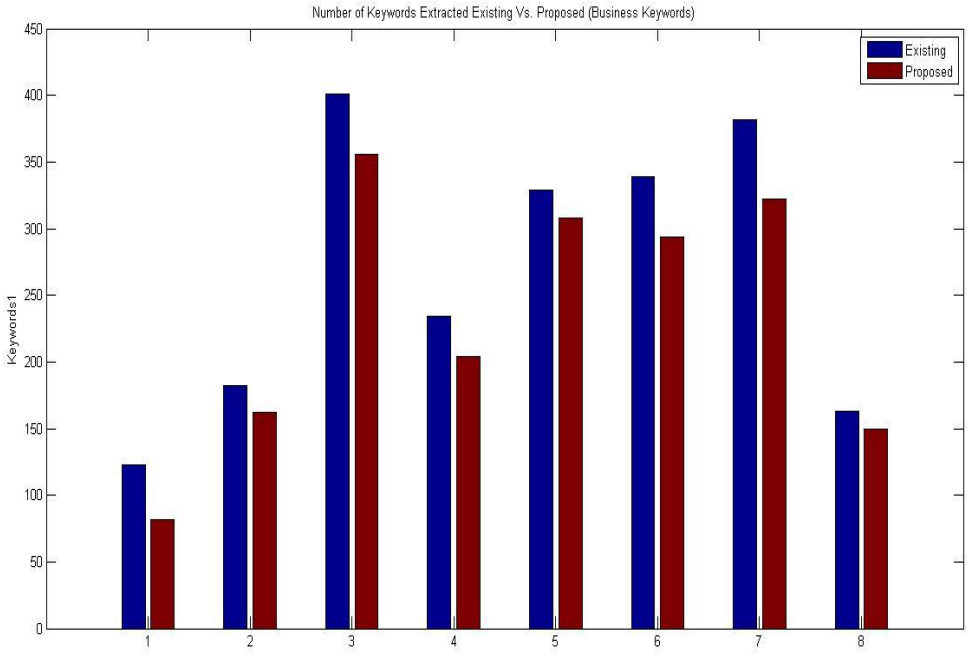

Fig. 9 Business Keywords

VI. CONCLUSION

In this present work, we have defined a statistical summarization based approach to detect the plagiarism on some user document. In this present work we have first extract the user text and find the most frequent keywords from the document. Now find the sentences that support these keywords. Once we get the summarized input text, same operation is performed on server side web pages. On server side the web crawling is performed to retrieve the web document. From these documents the text is extracted and summarized in same way. Finally documents having the maximum match are presented as the copied documents.

REFERENCES

[1] D.W. Embley, D.M. Campbell, Y.S. Jiang, Y.-K. Ng, R.D. Smith, *A Conceptual-Modeling Approach to Extracting Data from the Web*, Department of Computer Science, Brigham Young University, Provo, Utah 84602, U.S.A.

[2] Ananthi.J, *A Survey Web Content MiningMethods and Applications for Information Extraction from Online Shopping Sites*, Department of Computer Science,In IJCSIT, Vol.5,2014.

[3] Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, Robert Baumgartner,*Web Data Extraction, Applications and Techniques: A Survey*, Center for Complex Networks and Systems Research, Indiana University, Bloomington, USA
[4] Maged El-Sayed, Carolina Ruiz, and Elke A. Rundensteiner, *F-S Miner:Efficient and Incremental Mining of Frequent Sequence Patterns in Web logs*,Department of Computer Science,Worcester .
[5] S. Parthasarathy, M. J. Zaki, M. Ogihara, and S. Dwarkadas, *Incremental and interactive sequence mining*, In *CIKM*, 1999.
[6] A. Nanopoulos, D. Katsaros, and Y. Manolopoulos., *Effective prediction of web-user accesses: A data mining approach*, In WEBKDD Workshop, San Francisco, CA, Aug. 2001.
[7] Sang T.T. Nguyen, *Efficient Web Usage Mining Process for Sequential Patterns*,Faculty of Engineering and Information Technology,Sydney,Australia,2007.
[8] Tahira Hasan, Sudhir Mudur, Nematollaah Shiri, *A Session Generalization Technique for Improved Web Usage Mining*, Department of Computer Science and Software Engineering,Canada.
[9] R. Srikant and R. Agrawal, *Mining Sequential Patterns:Generalizations and Performance Improvements*, In EDBT,1996.
[10] Liu., Huang. X, *Personalized Recommendation with Adaptive Mixture of Markov Models*, The American Society for Information Science and Technology,2007.
[11] Z. Su, Q. Yang, Y. Lu, and H. Zhang, *Whatnext : A Prediction System for Web Request Using n-gram Sequence Models*, In *WISE*, 2000.
[12] Q. Yang, H. H. Zhang, and I. T. Y. Li, *Mining web logs for prediction models in WWW caching and prefetching*, In KDD,2001.
[13] Y. Fu, K. Sandhu, and M. Shih., *A generalization-based approach to clustering of web usage sessions*. In Intl. WEBKDD Workshop,1999.

