International Journal of Engineering, Management, Humanities and Social Sciences Paradigms (IJEMHS)
Volume 31, Issue 02, Quarter 02 (2019)
Publishing Month and Date: June 27, 2019
An Indexed and Referred Journal with Impact Factor: 2.75
ISSN (Online): 2347-601X
www.ijemhs.com

Extract and Analyse Patterns from Web Data: A Review

Suman Devi
M. Tech. Student, IIET, Kinana, Jind, Haryana

Abstract

While browsing the web, the user has to go through many web pages of the Internet, filter the data and download related documents and files. This process of searching and downloading is time consuming. Sometimes the search queries call for specific option, say, limiting search to few links. To reduce the time spent by users, a web extraction and storage tool has been designed and implemented in Java, that automates the downloading task from a given user query. The Test Scenario has been presented with various keywords. The present work can be a useful input to Web Users, Faculty, Students and Web Administrators in a University Environment. 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, Keyword Extraction, Web User.*

Introduction

In this chapter, the introduction to the all the related concepts to data plagiarism is defined in detail. The plagiarism is about to search the duplicate data contents over the web, because of that the whole concepts is around the web content mining, search engines and the crawling concepts. In this chapter, the detailed introduction to all these concepts is defined. At the initial stage, the web architecture is defined along with search engine and the crawling concepts. Main stress is given to the search process and utilities available over the web.

Literature Survey

According to [11] Yaakov HaCohen-Kerner performed a work," Detection of Simple Plagiarism in Computer Science Papers". This research, Author developed software capable of simple plagiarism detection. Author have built a corpus (C) containing 10,100 academic papers in computer science written in English and two test sets including papers that were randomly chosen from C.

Objectives of the Study

The purpose of analysis is to find answer to queries through the appliance of scientific procedures. The main aim of dissertation is to find out the truth which is hidden and which has not been discovered as yet. The study under consideration is primarily devised to achieve the following objectives:

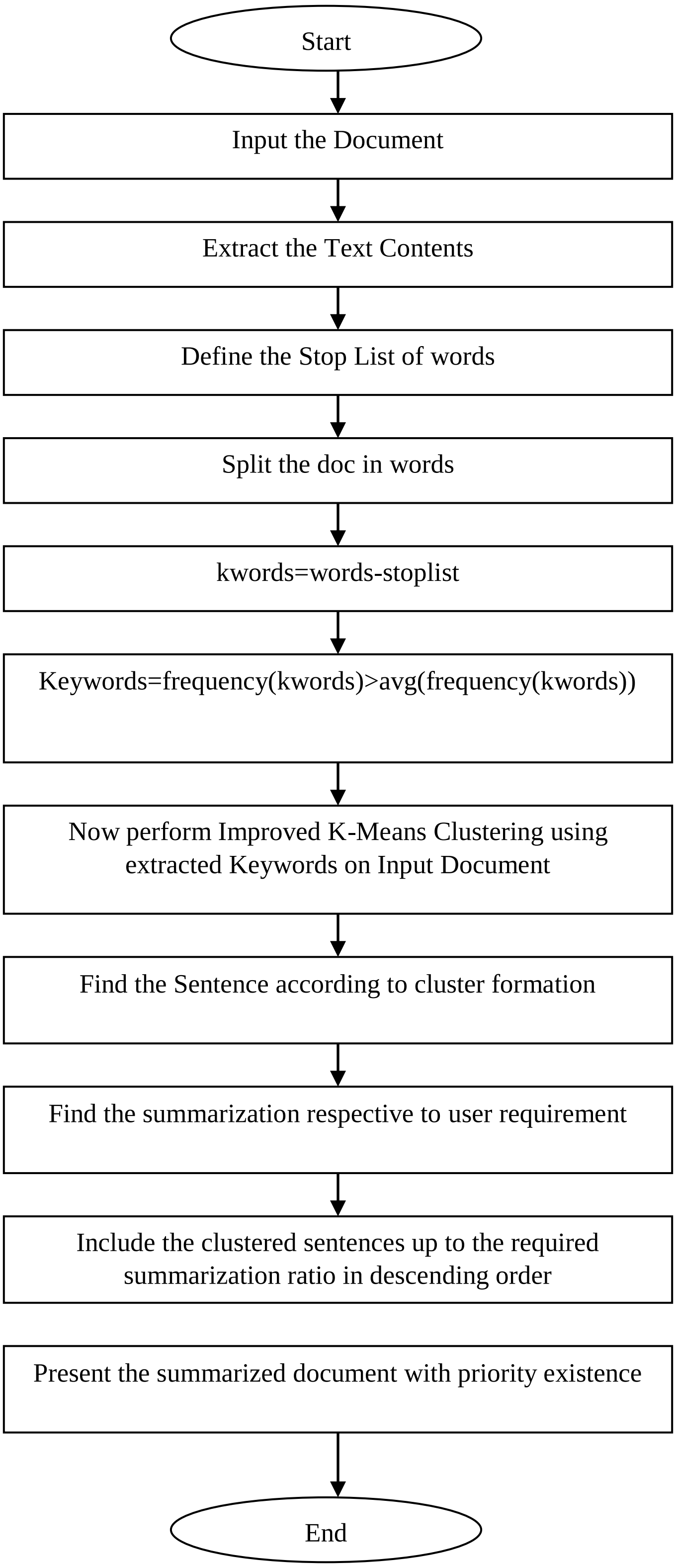
1. To produce a report containing all the original web pages without any duplication or near duplication in the real time application.
2. To study and manage a database of related URLs and the keywords
3. To design an approach to compare two Web Pages respective to duplicate
4. To perform indexing.

Application of Summarization Algorithm

The Summarization algorithm generates clusters based on similarity measure and data representation model. Numerous Summarization algorithms have been implemented.

Flow Chart

Here the basic flow chart of proposed work is defined. As we can see in figure 4.3 Flow Chart of Text Summarization.


Start
Input the Document
Extract the Text Contents
Define the Stop List of words
Split the doc in words
kwords=words-stoplist
Keywords=frequency(kwords)>avg(frequency(kwords))
Now perform Improved K-Means Clustering using extracted Keywords on Input Document
Find the Sentence according to cluster formation
Find the summarization respective to user requirement
Include the clustered sentences up to the required summarization ratio in descending order
Present the summarized document with priority existence
End

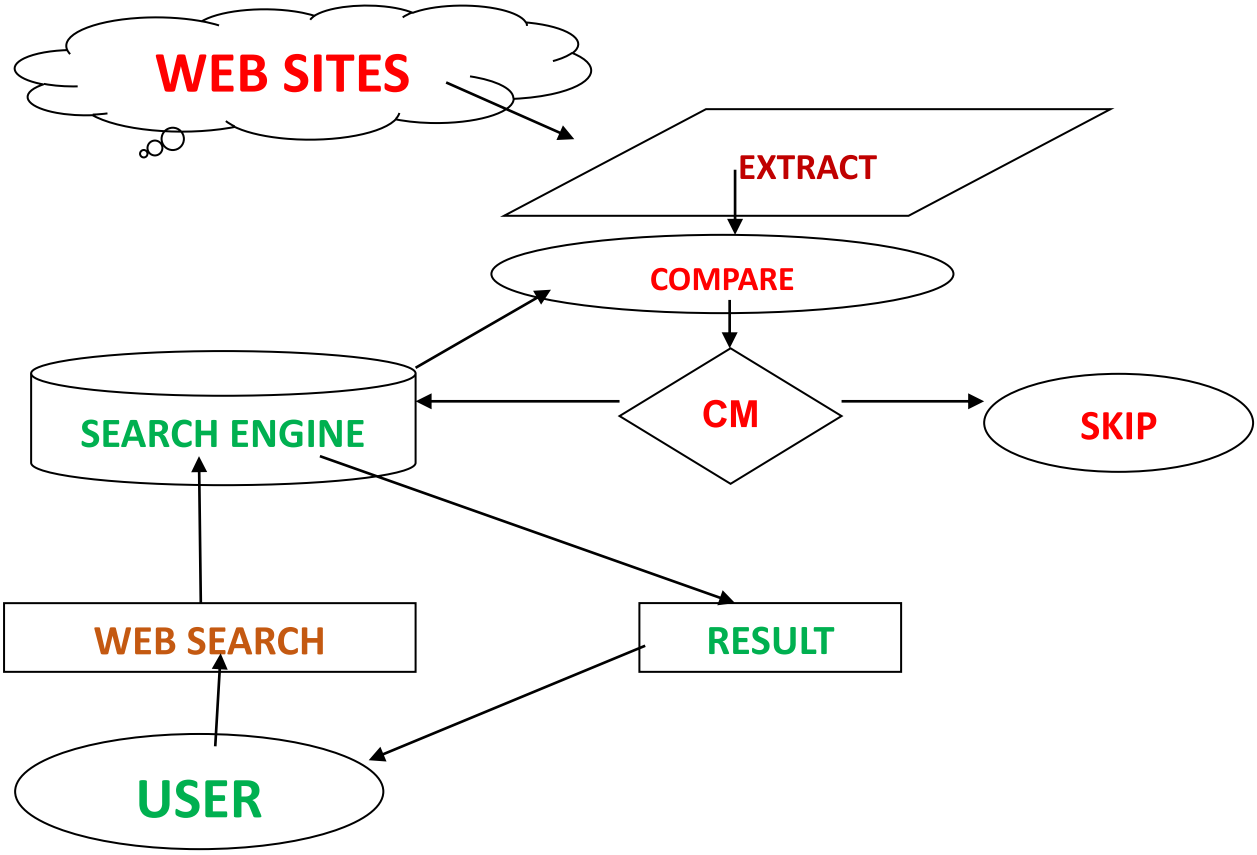


Figure 1: Analysis of Proposed Approach

The most necessary live for a quest engine is that the search performance, quality of the results and ability to crawl, and index the web efficiently.

Result and 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.

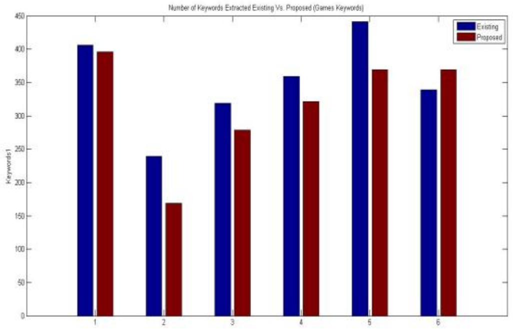


Figure 2: Education Keywords

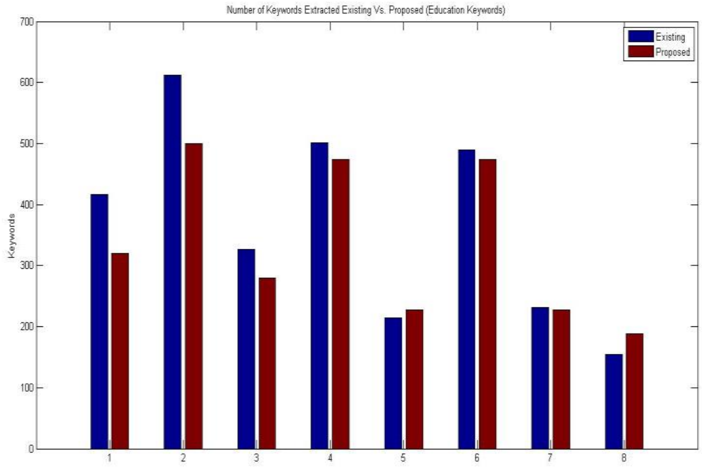


Figure 3: Entertainment Keywords

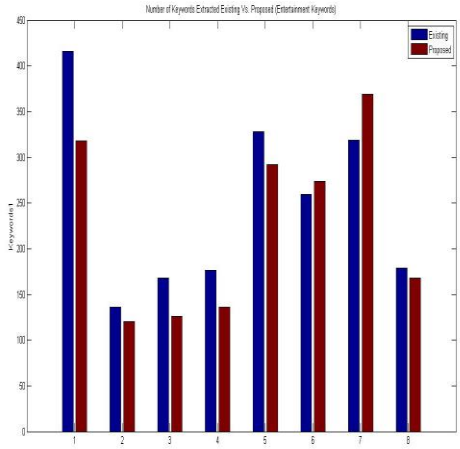


Figure 4: Games Keywords

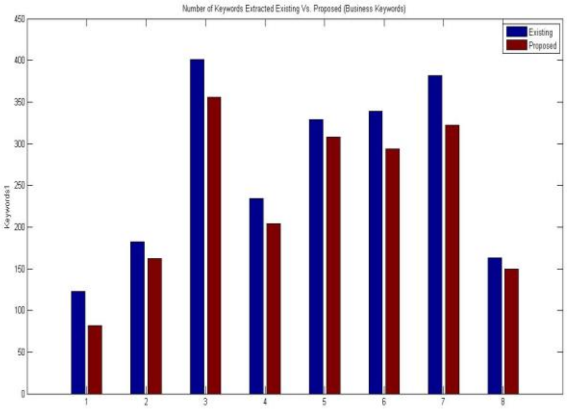


Figure 5: Business Keywords

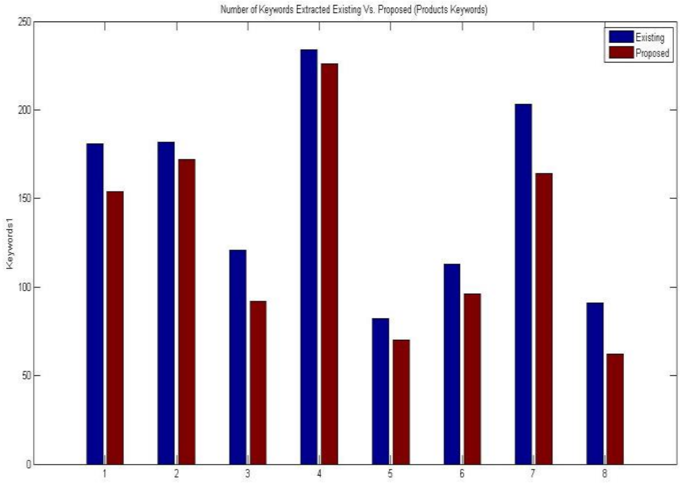


Figure 6: Product Keywords

Conclusion

An important research area in web mining is web usage mining which mainly aims on the discovery of patterns in the browsing and navigation data of web users. There are several techniques and tools proposed by deferent researchers for the web usage mining.

Future Work

In this present work we have used the summarization approach to detect the plagiarism over the web. This work is implemented for text pages and the html pages over the web. We can extend this work to detect text from different file formats also such as pdf files, doc files etc.

References

[1] Thomas S Dee, “Rational Ignorance in Education: A Field Experiment in Student Plagiarism”.

[2] Elizabeth Wager, “How should editors respond to plagiarism? COPE discussion paper”.

[3] Rebecca Moore Howard,” Understanding “Internet plagiarism”.

[4] Nick Fox, “Plagiarism: An Educational Approach”.

[5] Tracey Bretag, “Implementing plagiarism policy in the internationalized university”

[6] Tracey Bretag, “Self-Plagiarism or Appropriate Textual Re-use?”

[7] Cem Kaner, “A Cautionary Note on Checking Software Engineering Papers for Plagiarism”.

[8] Thomas E. Payne,” How to protect yourself from committing plagiarism”.

[9] Michele O"Dwyer, “Entrepreneurship Education and Plagiarism: Tell me lies, tell me sweet little lies’”.

[10] Vladislav Shcherbin in, “Using Microsoft SQL Server platform for plagiarism detection”

[11] Yaakov HaCohen-Kerner, “Detection of Simple Plagiarism in Computer Science Papers”.

[12] Martin Potthast, “Overview of the 1st International Competition on Plagiarism Detection”.

[13] Dr Joanna Bull, “Technical Review of Plagiarism Detection Software Report”.