International Journal of Engineering Sciences Paradigms and Researches (IJESPR)
Volume 48, Issue 03, Quarter 03 (July-Aug-Sept 2019)
An Indexed and Referred Journal with Impact Factor: 2.80
ISSN (Online): 2319-6564
www.ijesonline.com

Implement a Tool to Extract and Analyse Patterns from Web Data

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

Publishing Date: 27th July, 2019

Abstract

This document focus on the major factors related to web data mining. Web data mining is just like the data mining tool in database or warehouse. With the growth of Web-based applications, specifically electronic commerce, there is significant interest in analyzing data to better understand Web usage, and apply the knowledge to or high performance of search result. Web browser goes to all the server that contains related data relevant to the product of desire. 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 pre-processing, pattern discovery, and pattern analysis.

Keywords: *Web Data Extraction, Data Mining Techniques, World Wide Web, HTML, Web Usage Mining.*

Introduction

The data present on internet is vast and increasing very speedily. Hence there is a possibility of redundancy of related data which is called data plagiarism. 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.

The World Wide Web

The World Wide Web [1] is the open web system that is available publicly. It provides the simple handiness to any or all users over the planet. The web is that the metastasis for various text documents, image media and the video information. Each host over the web can access Thais information publicly and provide the resource access to the system with short and global identifiers. To access this information, the web server is identified a specific address or the domain name called Uniform Resource Identifier. This information access to the system is defined in a simple and effective way. The URL is actually the pool of HTML documents that connected through the Hyperlinks. These interlinked web pages actually represent a complete web site domain.

Web search engines get their information by web crawling from site to site. The "spider" checks for the standard filename robots.txt, addressed to it. The robots.txt file contains directives for search spiders, telling it which pages to crawl. After checking for robots.txt and either finding it or not, the spider sends certain information back to be indexed depending on many factors, such as the titles, page content, JavaScript, Cascading Style Sheets(CSS) ,headings, or its metadata in HTML meta tags After a certain number of pages crawled, amount of data indexed, or time spent on the website, the spider stops crawling and moves on. No web crawler may actually crawl the entire reachable web. Due to infinite websites, spider traps, spam, and other exigencies of the real web, crawlers instead apply a crawl policy to determine when the crawling of a site should be deemed sufficient. Some sites are crawled exhaustively, while others are crawled only partially.

Anatomy of Search

To identify the usage of web search engine and the web crawler we have to identify and understand the basic phenomenon behind the search. A search process is defined under the following characteristics.

1. A Search engine identify the information over the web
2. A search engine maintains a cache to keep track on retrieved URLs from the system

3. A Search engine is the user-friendly architecture in which user can pass the input effectively.
4. As the search is performed, the results are presented in the indexed form
5. The search engine returns the effective results from the search.

After search ranking will be given to the web pages which are the result of search. this is done with the help of optimization algorithm the professionals that can perform the allotment of the web sites or the web servers so that web information retrieval will be performed effectively. Page domain is then calculated by the number of clicks to visit a web page.

Data Detection Approach

This is done by using the following steps:

1. Copy detection approach
2. Named entity recognition
3. Rule based approaches
4. Machine learning based approaches
5. Supervised or semi-supervised learning

Literature Survey

In this chapter, the work performed by different authors is defined in this chapter. We have collected the work done by more than 20 authors. The work discussed in this chapter is basically related to the plagiarism concepts. Different observations and the approaches presented by different authors is discussed in this chapter. Some of the authors described the plagiarism concept under different application areas such as education etc. Some authors defined a survey based analytical work on existing plagiarism detection approaches. Some work is also presented to identify the challenges and the problems while working with plagiarism detection. A discussion is also presented on different available plagiarism detection tools and the evolution methods used in these tools. By analysing these all works, the conclusion is drawn about the strengths and weaknesses of different plagiarism detection approaches and identify the research gap. By analysing this research gap, the proposed work is formulated to present.

According to [1] Thomas S Dee performed a work," Rational Ignorance in Education: A Field Experiment in Student Plagiarism". Despite the priority that student plagiarism has become progressively common, there is relatively little objective data on the prevalence or determinants of this illicit behaviour. This study presents the results of a natural field experiment designed to deal with these queries. Over 1,200 papers were collected from the scholars in college man courses at a selective post-secondary establishment. Students in 1/2 the collaborating courses were every which way appointed to a demand that they complete associate anti-plagiarism tutorial before submitting their papers. Author found that assignment to the treatment cluster considerably reduced the probability of plagiarism, particularly among student with lower SAT scores who had the highest rates of plagiarism.

According to [2] Elizabeth Wager performed a work, "How should editors respond to plagiarism? COPE discussion paper". This paper aims to stimulate discussion concerning however editors ought to answer plagiarism. Different types of plagiarism area unit delineate in terms of their: extent, originality of the copied material, context, referencing, intention, author seniority, and language. The current COPE flowcharts advocate totally different responses to major and minor plagiarism. Possible, additional careful, definitions of these are proposed for discussion. Decisions concerning once to use text-matching package are printed.

Theoretical Development

The basic thinking or the requirement of the presented work is discussed in this chapter. As the Initial problem formulations are defined, the next work is to present the associated research design. The research design includes the theoretical discussion associated with the work along with risk categorization. At the end of this chapter, the advantages associated with the present work and the respective limitations are also been discussed.

Experimental Development

In this chapter, all these stages of research process are defined in detail. Along with this, the research methodology implemented in this research is also defined. At the final stage, the whole research is summarized in the form an algorithm as well as the flowchart. The chapter also includes the definition of the work model to extract the web contents. This

model is been discussed with the exploration of each stage. The stages include the pre-processing stage, filtration stages, web information extraction, matching process etc. The presented model shows the significant and the mathematical model to derive the results from the system.

Research Design

The complete research work will be performed in following steps:

Extract the Research Document

The first step of the research is to extract the Web Document. For the web document extract we will prefer some news site. We need to perform the web content mining to extract the document. The basic architecture followed by Web page extraction is given as:

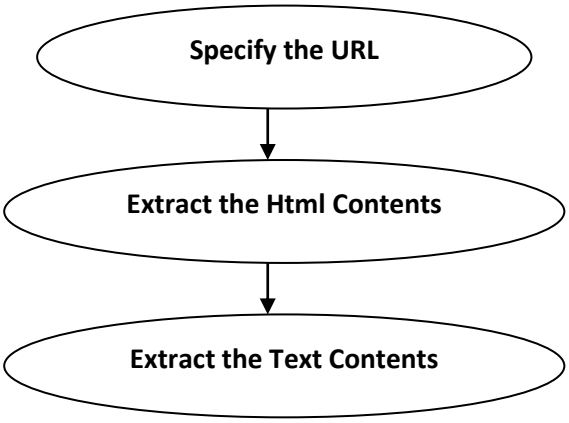


Figure 1: Basic Steps of Web Page Extraction

Document Summary Generation

To summarize a document, we need to study and analyse the document in terms of

- Prioritization of Keyword, Heading etc.
- The Frequency of the appearance
- The interval of appearance of word in the document.
- The basic position i.e. top bottom etc.
- The basic architecture of system will be

The basic architecture of system will be:

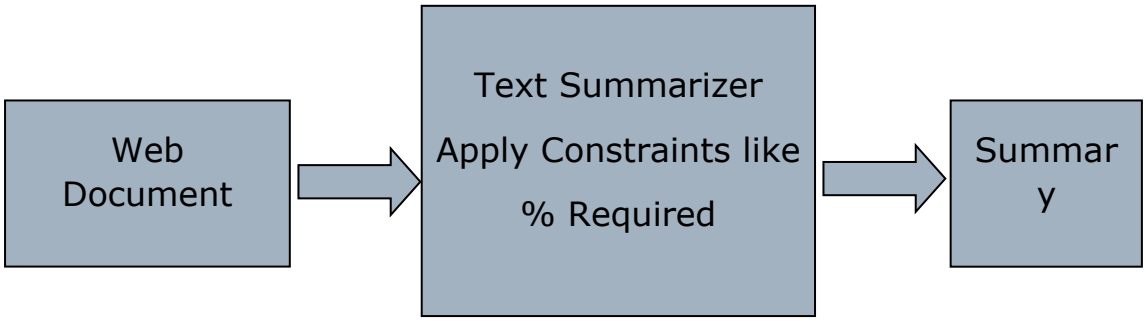
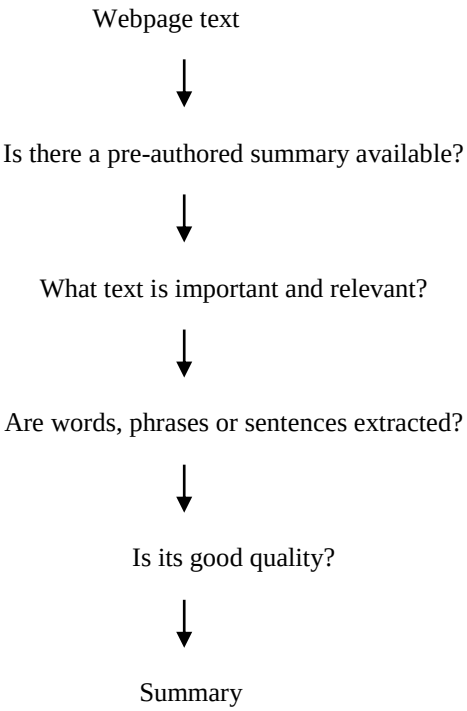


Figure 2: Summarization Architecture



Development Scheme

Overall development scheme consisting of following steps as depicted in figure 3

Figure 3: Development Scheme

Over time, the protocol has become the pattern of the net for net crawlers. The Robot protocol specifies that Web sites wishing to restrict certain areas or pages from crawling have a file called robots.txt placed at the root of the Web site:

Robot.txt file:

robots.txt for http://somehost.com/

User-agent: *

Disallow: /cgi-bin/

Disallow: /registration # Disallow robots on registration page

Disallow: /login

The first line of the sample file has a comment on it, as denoted by the use of a hash (#)

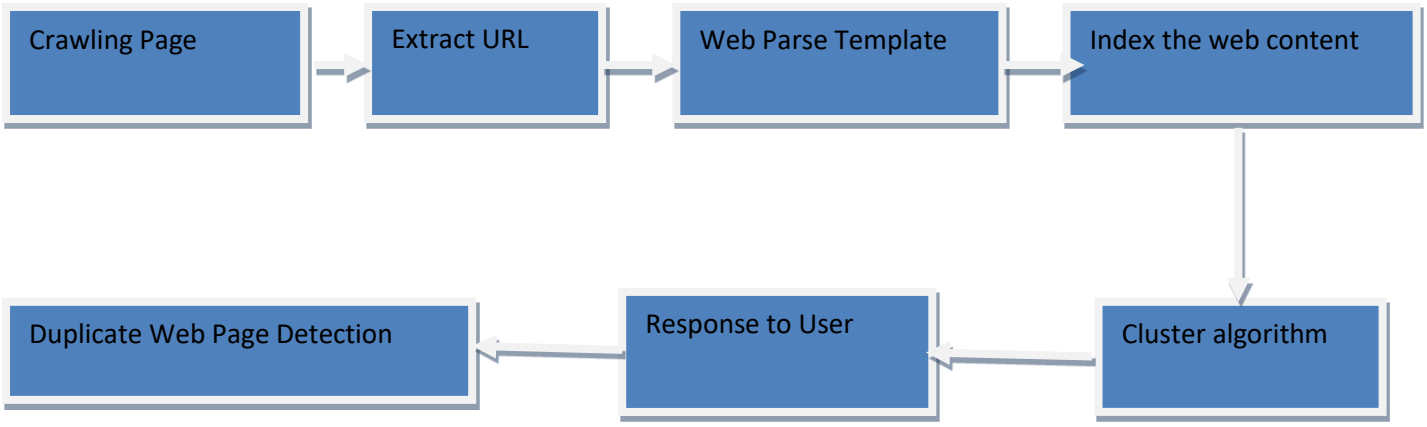


Figure 4: Representing the Development Schemes

The description and result of each step is presented in following sections.

Crawling the Web
Parsing the Web documents
Removal of stop words
Calculating similar score
My search engine
Result will be displayed

Figure 5: Represents Various Steps Involved in Detection of Duplicate Web Pages

Results

The result driven from the model and the algorithm implementation of the proposed work is discussed here.

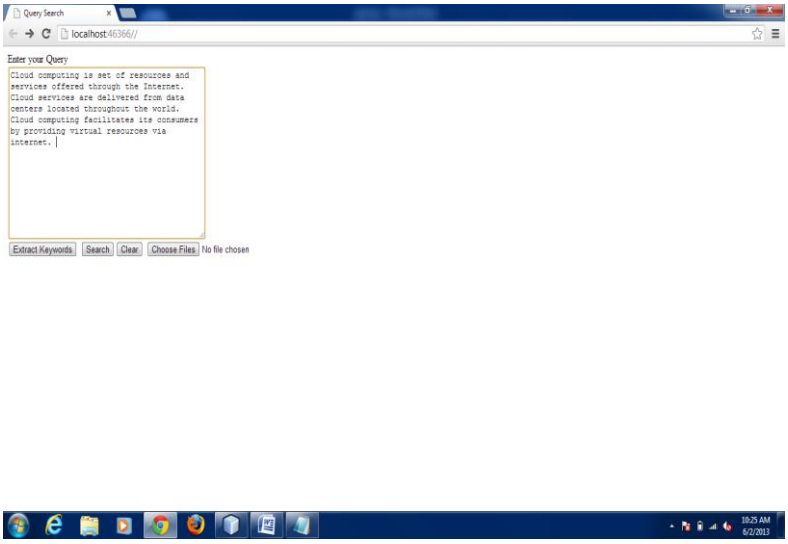


Figure 5.1: Gateway to plagiarisism

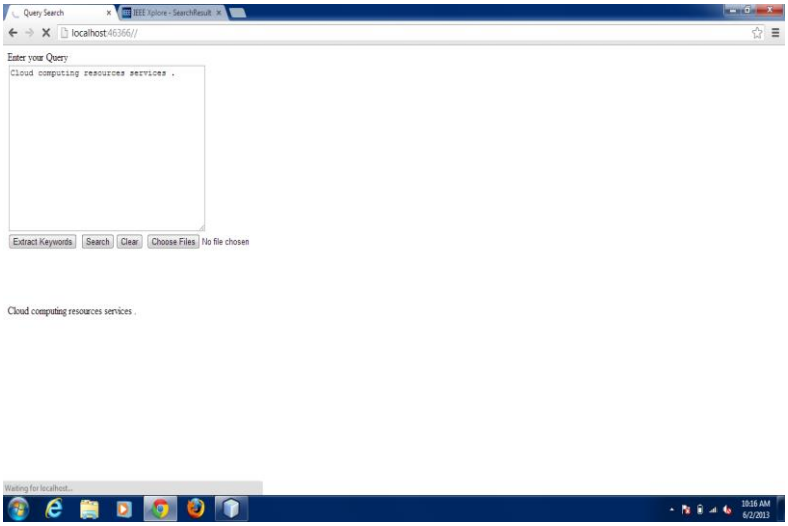


Figure 5.2 : Keyword Extraction

Figure 5.3 : Result

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.

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".