Volume 4, Issue 12, December 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Review on Extracting and Analyzing Patterns from Web Data

Ashima Miglani,
M. Tech Student
Royal College of Engineering
Haryana, India

Mr. Jitendra Arora
HOD CSE
Royal College of Engineering
Haryana, India

Abstract- This document emphasizes on detecting plagiarism over web efficiently. Plagiarism is one of the serious forms of thefts done in professional areas where a professional or student use other person work without his permission and present it by his name .Web mining can be defined as extracting and analyzing useful information from web. Proposed work is similarity measure of user document to web document, but it is not possible to compare two documents completely. The proposed work will use the featured analysis based approach. The keyword extraction and analysis based approach will be done dynamically using clustered approach to perform the document match over web.

Keywords- Plagiarism ,Web Mining, Keyword Extraction, Featured Analysis, Clustered Approach

I. INTRODUCTION

Detecting duplicate pages slow down the retrieval of pages from web. Therefore there is a need to develop an approach that can recognize web pages efficiently and access information as quickest as possible.

This approach [1] has basically three phases :

A. Pre-processing

Before summarization begins ,pre-processing should be done on the documents which involves stemming of words and stop words removal (*like to, with, are, it*).Pre-processing comprises of converting the usage, content and structure information from different data sources into data abstractions which helps in pattern discovery.

B. Pattern Discovery

Different methods can be adopted for pattern discovery like Classification, Association, Rules Statistical Analysis, Clustering etc.

C. Pattern Analysis

Main Process executed here is filteration of unwanted information. Methods like SQL(*Structured Query Language*) and OLAP(*Online Analytical Processing*).

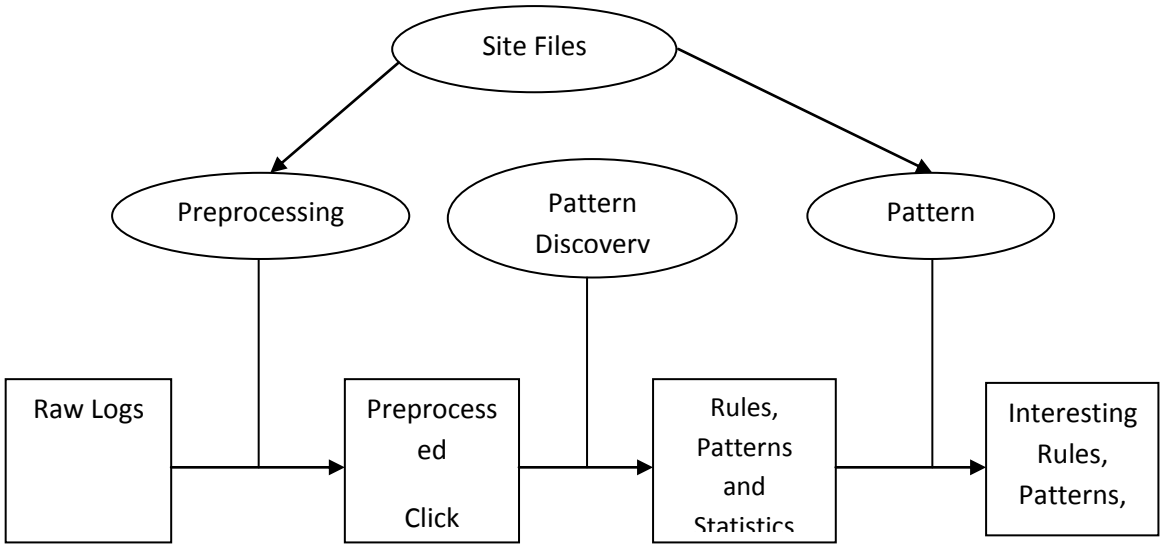

Fig.1 Overall Process

II. BACKGROUND

According to [2] Cem Kaner performed a work," A Cautionary Note on Checking Software Engineering Papers for Plagiarism". Different tools are available for detecting and preventing plagiarism. The article compares the performance

of two leading tools MyDropBox and TurnItIn. Plagiarized articles published in IEEE journals were considered for detecting plagiarism. Both tools do not perform upto mark because they do not compare writings to publications in IEEE database. Also ACM and other important scholary databases are not covered by these tools. Reports from these tools suggesting that a submission has "passed" can encourage false confidence in the integrity of a submitted writing

According to [3] Xiaoqing Zheng ,"Data Extraction from Web Pages Based on Structural Semantic Entropy" .Here structural semantics entropy is used for extracting useful information on web. The concept of measurement of the density of occurrence of the relevant information is used. Due to the difficulty of writing and, maintaining the wrappers and blocks identification in the vision based extractors, this method has been introduced.For identifying the density of the specified product entropy measure is calculated.

Robert J. Youmans [4] performed a work," Does the Adoption of Plagiarism- Detection Software in Higher Education Reduce Plagiarism?" In two studies, students at California State University, Northridge wrote papers that were checked for plagiarism using plagiarism-detection software. In the first study, half of students in two classes were randomly selected and told by the professor that their papers would be scanned for plagiarism using the software. Students in the remainder of each class were not informed that the software would be used. The researcher predicted that those students who were informed that the software will be used would plagiarize less than students who were not, but it did not make any effect. In a second study, students wrote two papers in a series. Their knowledge about plagiarism-detection software was inversely correlated with plagiarism rates on the first paper, but no correlation was found between knowledge and plagiarism on the second paper. Instead, participants were discovered to draw repeatedly from the same sources of plagiarized material across papers.

Chen et al. [5] introduced the,"Data mining for path traversal patterns in a web environment. Here for characterizing and minimizing traversal patterns,concept of maximal forward reference is adopted.A maximal forward reference comprises of sequence of pages requested by a user up to the last page before backtracking occurs during a particular server session.

III. RELATED WORK

As in [6],Clustered approach is used which is grouping together set of items having similar characteristics. In web basically there are two kinds of clusters:

-Usage clusters

-Page clusters

In usage clustering,users exhibiting similar browsing patterns are grouped together.

In page clustering,pages having related content are grouped together. In both applications, permanent or dynamic HTML pages can be created that suggest related hyperlinks to the user according to the user's query or past history of information needs.

Types of clustering methods:-

1) Partitioning Methods
2) Hierarchical Agglomerative (divisive) methods
3) Density based methods
4) Grid-based methods
5) Model-based methods.

As in [7,9], algorithms rely on character based**,** word based and syntax based lexical features. Comparison is made on the query document dq with each candidate document dx"Dx. String matching can be exact or approximate. Exact string matching involves matching two strings x and y such that they have exactly same characters in same order. For example, the character 6 -gram string x="aaabbc" is exactly the same as "aaabbc" but differ from y="aaabbd". Approximate string matching includes two strings that are upto some degree similar/dissimilar. For example, the character 9-gram x="aaabbbccc" and y="aaabbbccd" are highly similar because all letters match except the last one.

According to [10] Neural network comprises of set of input and output units each having a particular weight associated with them. In the phase of learning, for predicting the correct class label of input tuples ,network learns to adjust its weights. These networks are specialized in derieving meaning from imprecise or complicated data. This technique is used to extract patterns and detect trends that are quite impossible with humans or other methods. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries.

As in [7,8] a spectrum is implemented consisting of similarity values that range from one(exactly matched) to zero(entirely different).Every word in the document is linked with a fuzzy set containing words with similar meanings. This approach works great in detecting statement based plagiarism, because it can detect similar, yet not necessarily same statements based on the degree of similarity between words in the document and fuzzy set.

According to [10,11] this approach, the entities are analyzed by experienced linguistics and certain specified rules are created. For extracting entities basically three phases are involved:

- Linguistic Preprocessing
-Named Entity Identification
-Named Entity Classification

Linguistic Preprocessing includes tokenizing, part of speech tagging, stemming and using the list of known names (database lookup). This includes the start and end structure of all the words that can be thought as named entity. In this possible named entities are generated by using punctuation marks or capitalization.

When possible named entities are identified, classification begins. Classification is performed in three stages: application of rules, database lookup classification and considering the matching of classified named entities with the unclassified ones. Rules are generated by experienced linguists. Rules are formed considering appositives or certain keywords that can precede or succeed a possible name. By matching rules that are generated and named matching possible entity, classification begins. If no match is found with the rules, then database lookup is used.

IV. PROPOSED APPROACH

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. 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. The research method used in detection of duplicate web pages in Web Crawling is the constructive research method.

A. Extract The Research 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

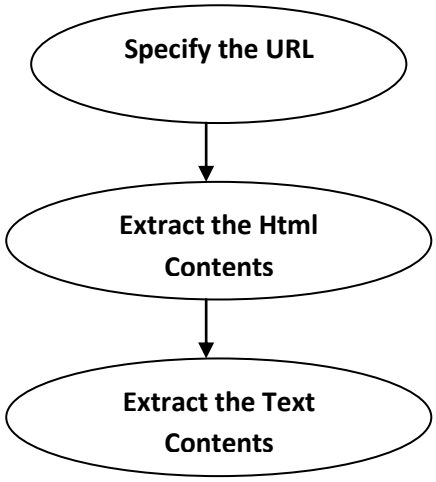

Fig.2 Basic Steps of Web Page Extraction

B. Document Summary Generation

To summarize a document we need to study and analyze 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.

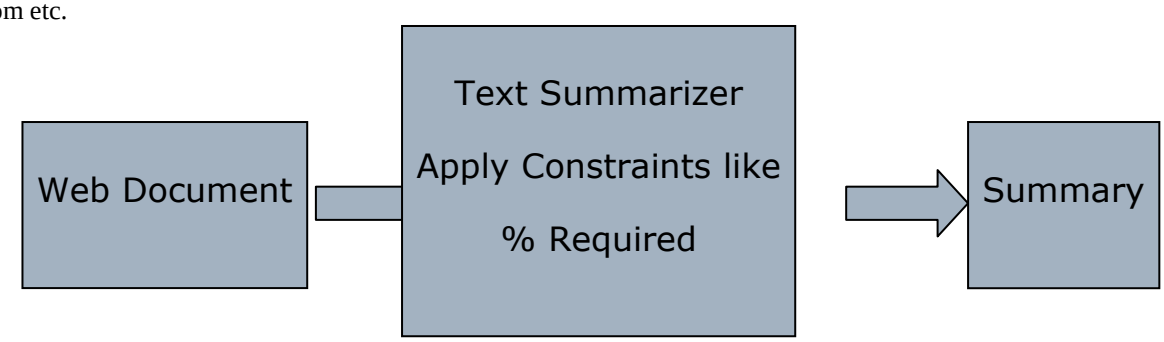

Fig.3 Summarization Architecture

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 priortization 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.

C. Final Analysis

Summarization of documents is a difficult task in text data mining owing to the high-dimensionality and sparse nature of text documents. It requires efficient algorithms which can address this high dimensional Summarization problem. Document Summarization plays an important role in web based applications and text data mining. Cluster Based Navigation is an interesting alternative to keyword searching, the standard information retrieval paradigm. This is extremely useful in cases where users prefer browsing over searching when they are unsure about which search terms to use. Its benefit is provision of alternate user interface i.e. 'search without typing'. The result of a query is now matched to a cluster rather than to each document thus reducing the search space. . 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.

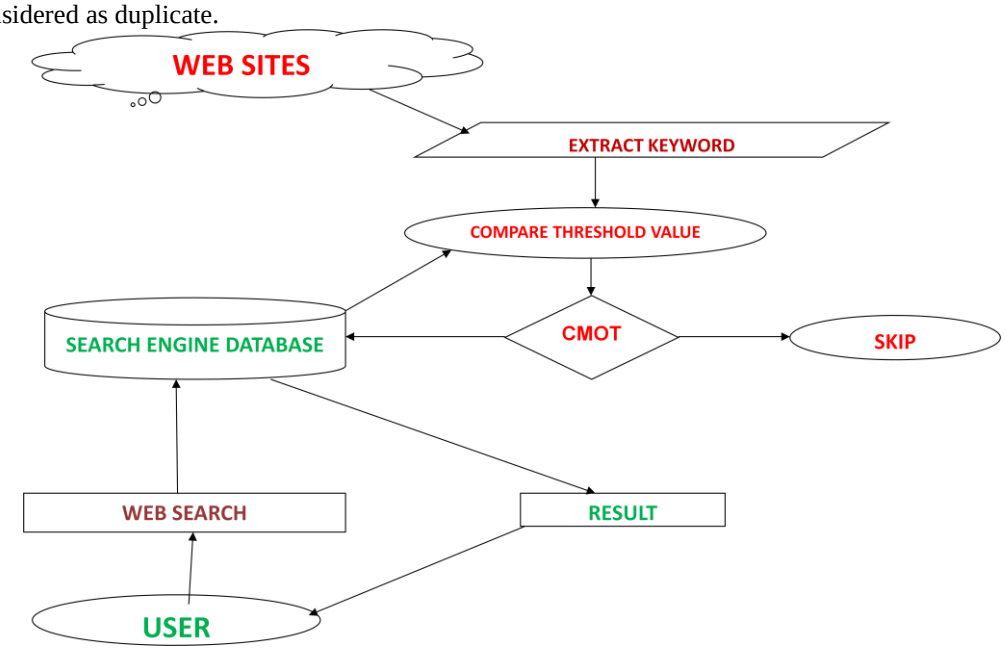

Fig.4 Analysis of proposed approach

V. 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.

ACKNOWLEDGEMENT

The first Author is thankful to Mr.Jitender Arora,Royal College of Engineering, Dcrust for his useful discussion and support.

REFERENCES

[1] Robert Cooley and Jaideep Srivastava, *Discovery of Interesting Usage Patterns from Web Data*, Department of Computer Science,University of Minneapolis, May 1999.

[2] Cem Kaner, *A Cautionary Note on Checking Software Engineering Papers for Plagiarism*, In IEEE Education society,May 2008.

[3] Xiaoqing Zheng, Yiling Gu, Yinsheng Li, *Data Extraction from Web Pages Based on Structural Semantic Entropy*,International world wide web conference committee(IW3C2),April 2012.

[4] Robert J.Youmans ,*Does the Adoption of Plagiarism Detection Software in Higher Education Reduce Plagiarism?*,University of California(2011).

[5] M.S. Chen, J.S. Park, and P.S. Yu, *Data mining for path traversal patterns in a web environment*. In 16th International Conference on Distributed Computing Systems, 1996.

[6] Jaideep Srivastava, *Web Usage Mining:Discovery and Applications of Usage Patterns from Web Data*,2000.

[7] Sindhu.L, Bindu Baby Thomas, Sumam Mary Idicula,*A Study of Plagiarism Detection Tools and Technologies*,In IJART,Vol.1,Issue 1,2011.

[8] Ceska Z, *The Future of Copy Detection Techniques*, In Young Researchers Conference,Pilsen,2000.

[9] C.De Stefano, C.Sansone, and M.Vento,*Evaluating Competitive Learning Strategies for Handwritten Character Recognition*,In IEEE Int. Conf. On Systems,Oct. 1994.

[10] Mrs. Bharati, M.Ramageri, *Data Mining Techniques and Applications*, Department of computer Application,Maharashtra,In Indian Journal of Computer Science and Engineering, Vol.1 No. 4 31-305.

[11] Salha Alzahrani, Naomie Salim, and Ajith Abraham, *Understanding Plagiarism Linguistic Patterns, Textual Features and Detection Methods,* In IEEE,2011.

[12] Ujwala Manoj Patil, J.B. Patil, *Web Data Mining Trends and Techniques,* Department of Computer Science,Maharashtra.

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

[14] Sergey Butakov, *On the number of search queries required for Internet plagiarism detection*, In Ninth IEEE International Conference on Advanced Learning Technologies, 2009.

[15] Martin Potthast," *Overview of the 1st International Competition on*

