International Journal of Computer Science and Engineering (IJCSE)
ISSN 2278-9960
Vol. 2, Issue 3, July 2013, 115- 122
© IASET

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

A SIMILARITY MEASURE ANALYSIS BASED IMPROVED APPROACH FOR PLAGIARISM DETECTION

MAULI JOSHI[1] & KAVITA KHANNA[2]

[1]Research Scholar, Department of CSE, PDM College of Engineering for Women, Bahadurgarh, Haryana, India

[2]Associate Professor, Department of CSE, PDM College of Engineering for Women, Bahadurgarh, Haryana, India

ABSTRACT

Web Mining is one most useful concept used by the each user knowingly or unknowingly. As user searches some text over the search engine, the web information retrieval is performed. When a user writes some document, research paper, report, most of the content part is taken from the web. In such case, it is required to analyze the effective working of a user. The concept of detection of user contents over the web is called plagiarism. Plagiarism is a kind of content based or the logic based theft. The presented work is about to detect these plagiarism contents over the web. To perform this detection we have defined a statistical approach. At the initial stage, the content filtration is performed to reduce the query size. After that the similarity measures are been used to perform the content match over the web. The obtained results show the effective detection of user contents over the web.

KEYWORDS: Web Mining, Statistical Measure, Plagiarism, Query Filtration

INTRODUCTION

Web content mining is one of the important concepts used by web that is used to perform a query search over the web. This query search is performed by the search engine with the help of web crawler to identify the content URL over the web. But search engine have some limitation respective to the query size. When we detect a particular paragraph or the document over the web for the duplicate contents, the search engine itself is not effective to perform this work. The presented work is in same direction to identify the duplicate web contents.

Search engines indexes millions of web pages involving a comparable number of distinct terms that could answer tons of queries. There is large amount of dependence over them still there is so much to explore with terms of research work [1]. To search across web two things are searched words within page and page location. While searching the words like a, an, the called stop words are removed, stop words are the most frequent used word that do not have much relevance while searching through web. With the advancement in technology and web proliferation, web search engine creation is changed a lot [2].

Every search engine somewhat works the same way and they all perform basic tasks like:

- They search the Internet to look for keywords.
- They index the words they find with their location to make it easier to search for next time, and where they find them. [3]

A Search engine searches for particular keywords and returns a list of documents that contains that keyword. These works on web via crawlers, spiders and robots. A web crawler is a program that works on urls so it creates a list or a queue of them so as to retrieving and prioritizing the urls of the web pages. From this queue crawler downloads the page and extract URL and again puts this URL in queue. This process is carried out again and again until crawler aborts the process. The pages collected are later used by web cache and process is repeated until the crawler decides to stop [4].

A search engine not only provides search performance but also the quality of the results, crawling and the indexing of web contents efficiently. It provides relevant and accurate search for the queries we input. Some well known search engines are Google, Yahoo etc. Most of search engines work on same principles [4].

QUALITY MEASURE

There are various measures to compute the similarity between web documents. Similarity measures which have been frequently used for plagiarism detection are discussed below:

- **Euclidean Distance:** is used to measure similarity and dissimilarity in two points. Distance metric calculated as:

  $$d_2(x_i, x_j) = \left(\sum_{k=1}^{d}\left(x_{i,j} - x_{j,k}\right)^2\right)^{\frac{1}{2}}$$

  $$= \|x_i - x_j\|_2$$

  Where xi and xj are represents document.

- **Minkowski Distance**: is generalization of Euclidean distance , measure for (p=2):

  $$d_2(x_i, x_j) = \left(\sum_{k=1}^{d}\left(x_{i,j} - x_{j,k}\right)^p\right)^{\frac{1}{p}}$$

  $$=\|x_i - x_j\|_p$$

- **Cosine Similarity Measure:** It computes the cosine of the angle between two vectors or in this case document. Cosine similarity then gives measure of how similar two documents are in terms of their subject matter

  $$cos\ (m_p,m_j\ )= \frac{m_p^t\, m_j}{|m_p||m_j|}$$

  Where $m^t{}_p m_j$ denotes the dot-product of the two document vectors; |.| indicates the Euclidean length of the vector.

- **Jaccards Coefficient:** it is used to measure similarity between two sample sets. For two documents A and B the Jaccards Coefficient is computed as below:

  $$J(A, B) = \frac{|A \cap B|}{|AU\ B|}$$

Paper is organized as follows: Section 2 discusses literature survey. Section 3 discusses research methodology.

EXISTING WORK

To plagiarize means using another author's work without crediting him or without citing the text. Increasing use of computers and internet has made it easier to plagiarize the work of others. Plagiarism can be found in scientific paper, source code, articles etc. Plagiarism in scientific journal thus can be classified into theft of 'idea' and 'words'. When words or sentences are manipulated from original text it is known as plagiarism of words. Idea Plagiarism is when the original idea of the author is used as one's own [7]. According to Salha Alzahrani [5], Plagiarism has also been classified as 'literal' or 'intelligent' plagiarism. As in figure 1, we can see the taxonomy of plagiarism. Literal Plagiarism can also be denoted by cut and paste from the original source while intelligent plagiarism is manipulating data by paraphrasing or other techniques to make it look different, but it means the same. There are many tools that can be used to detect plagiarism in a document.

Plagiarism is using words of someone else's work without citing the author to be as their own. It falls under intentional plagiarism.

In 2013 Nwokedi Idika, Mayank Varia, and Harry Phan, in paper "The probabilistic provenance graph", IEEE security and privacy workshops, have presented a probabilistic provenance graph (PPG) model that could be used for similarity detection.

The algorithms for plagiarism detection have heuristic-based lower bounds on the plagiarism cases they will detect as it is impossible to know how long the plagiarism cases will be in the document, it is desirable to search for different sized plagiarism cases in the documents being compared and this allows for detection of plagiarism cases that would have been missed if only one algorithm were used.

In the paper more precise algorithms are placed in the front of the pipeline and the least precise algorithms are placed toward the back of the pipeline and the algorithms that require more time to process documents are placed toward the end of the pipeline. Algorithms requiring more running time are generally more sophisticated algorithms and these algorithms will be able to detect more complex forms of plagiarisms [6].

In 2012 Jonathan Y. H. Poon in paper "Instructor-Centric Source Code Plagiarism Detection" presented a plagiarism detection system that assists instructors with the tedious task of detecting and reporting source code plagiarism cases which includes Source-code reuse, direct Copying, Copying with modification minimal, moderate, and extreme or sometimes converting a source-code to another programming language. For detecting source code plagiarism by the plagiarism detection tools, each program is converted into token and Token or strings are compared to determine similar source-code or chunks of code. Tools for detecting source code plagiarism: JPlag, MOSS.

Plagiarism Incidents: There are several cases of plagiarism that are reported not only in field of academics but also in other fields like journalism and arts. In 2007 Sandra Nadelson performed a survey from 72 reports with 460 incidents, suspected or known misconduct by undergraduate students and over 110 suspected situations with graduate students. And the majority of incidents reported were "accidental/ unintentional plagiarism" [5].

AVAILABLE TOOLS

Turnitin: It is a iParadigms product and a web based service. In this user uploads the suspected source document to the system database and system creates a fingerprint of the document and compares it with documents in database. There are about 10 million documents already submitted to the Turnitin database. Turnitin gives the result within few minutes describing the level of plagiarism.

WCopyfind: is open source and windows based program for detecting words or phrases of specific length from the local repository. it is a single executable file that could be just run with no need to install it.

In the process each document is read one word at a time and letter case, punctuation, numbers, and other word characteristics are removed or changed and each word is converted into a 32-bit hash code for efficient storage and comparison. After all the documents have been loaded and hash coded, the comparison step occurs which looks for matching phrases .

JPlag: is used to detect similarities among program source codes and is used to detect software plagiarism. It supports Java, C#, C, C++ and natural language text. It is free to use and users upload the files to be compared and the system presents a report showing matches.

PlagTracker: it is an online based service that uses a licensed algorithm to scan a given document and compare it to the content sources across a local database and the Internet. At first step the text of the original document is transferred on the webpage and PlagTracker returns with a list of web pages that contain similar text.

Copyscape: is multilingual online plagiarism detection service that checks similarity of text on the web. It was launched in 2004 by Indigo Stream Technologies, Ltd and is used by content owners to detect theft of content. It uses the Google Web API for searching. it is premium service that consist of two paid versions Copyscape Premium and Copy sentry.

RESEARCH METHODOLOGY

The proposed work is about to detect the duplicate pages over web. This detection process is name plagiarism detection. In this present work we have optimize the topic based web search process with the concept of similarity analysis for duplicate pages detection. For this a new architecture is proposed, this architecture will use the query filtration based keyword analysis.

Once the query is filtered, all the common words called stop list will be eliminated from the query. It will also remove the similar words. Once the filtration process is over the next work is the crawling of the web pages based on this text query.

The query is performed to the Google search engine and the web pages are retrieved. Now on these web pages the keyword based similarity analysis is performed. In this work a rule based similarity checker is implemented to identify the duplicate pages.

As the comparison is performed with these pages, the similarity ratio will be obtained. Now the pages are ordered respective to this similarity ratio and the relative result is presented to the user.

As we can see in this proposed architecture the user will interact to the Web pages with his topic based query to retrieve the Web pages. As the page is query performed it will perform request to the Web and generate the basic url list. Now it will retrieve the data from the Web.

For the url collection it will use some concepts like indexing and the similarity analysis based ranking. The most similar page will be placed at top of the list. . The indexing will provide a fast access to the Web pages where ranking will arrange the list according to the priority. The basic steps included in this work are given as under. The basic flow of the presented work is shown in figure 1.

- **Keyword Extraction from Query**

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.

- **Web Search**

Web search focuses on the effective keyword and phrases for effective retrieval.

- **Similarity Analysis**

Once the web pages will be retrieved, the next work is to perform the similarity match for these web pages. To perform this match different similarity analysis approaches are used. Once the match is performed, the pages are presented to the user with similarity match ratio.

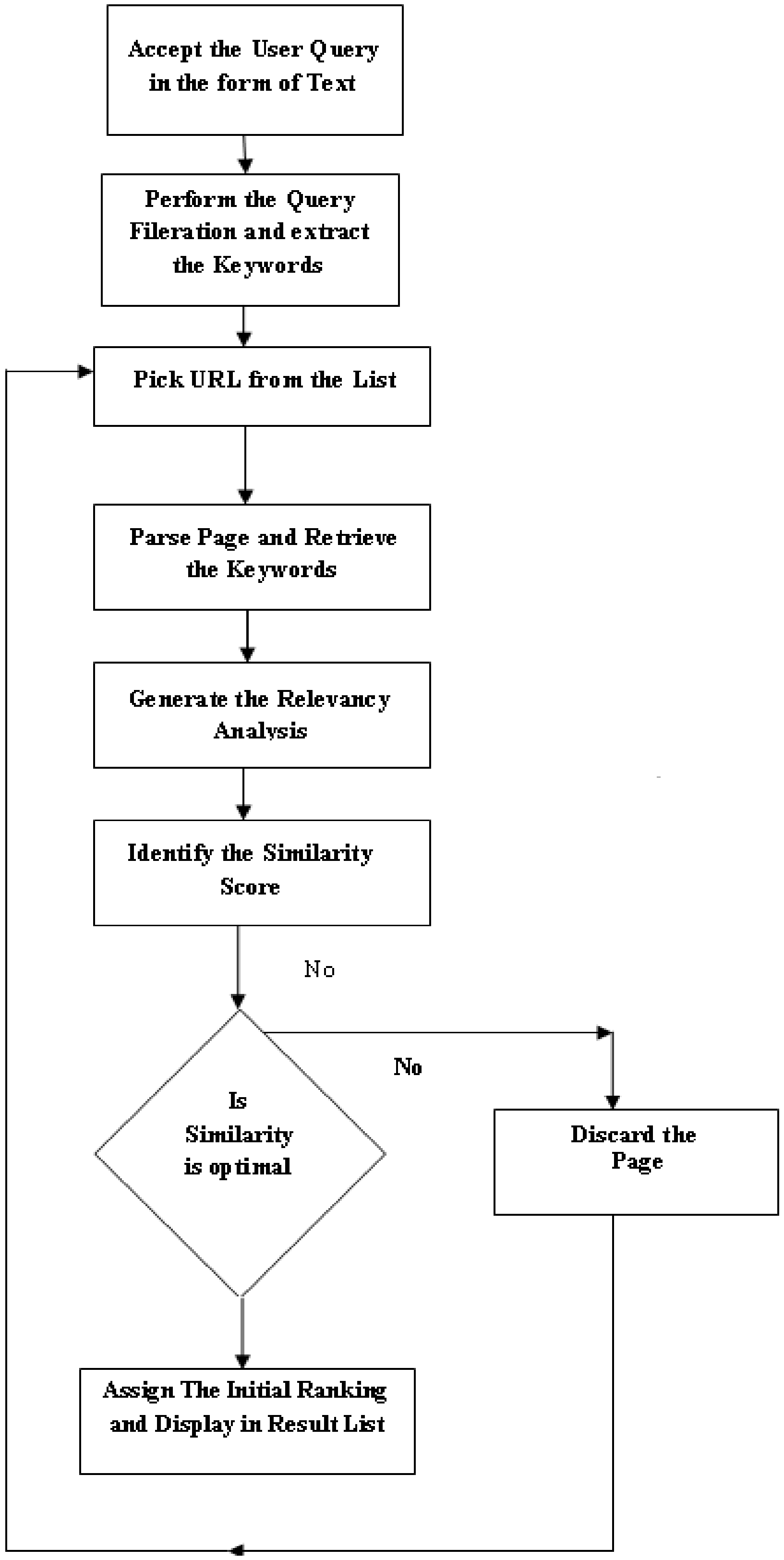
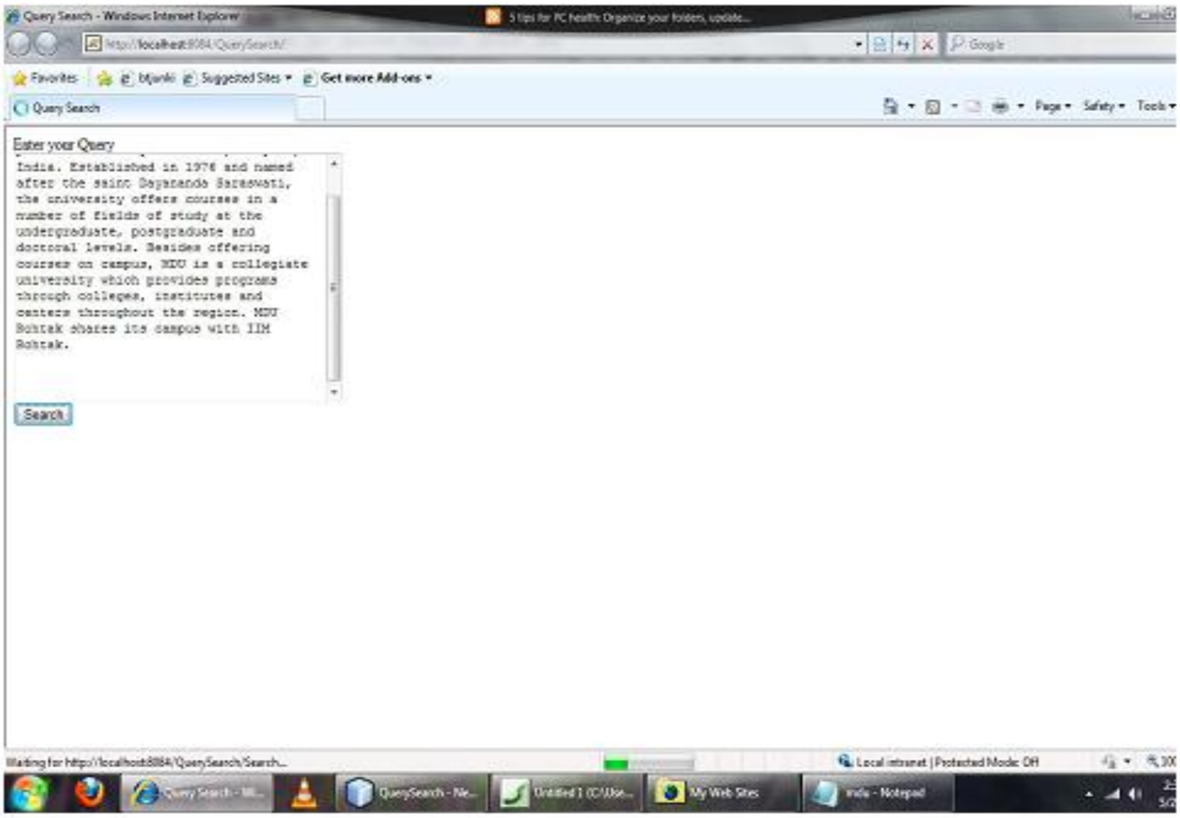

Accept the User Query in the form of Text
Perform the Query Fileration and extract the Keywords
Pick URL from the List
Parse Page and Retrieve the Keywords
Generate the Relevancy Analysis
Identify the Similarity Score
No
No
Is Similarity is optimal
Discard the Page
Assign The Initial Ranking and Display in Result List

Figure 1: Flow Chart of Proposed Work

RESULTS

The presented work is implemented in Java environment and the results obtained from the system are given in this section


Query Search - Windows Internet Explorer
5 tips for PC health: Organize your folders, update...
http://localhost:8084/QuerySearch/
Google
Favorites
btjunki
Suggested Sites
Get more Add-ons
Query Search
Page
Safety
Tools
Waiting for http://localhost:8084/QuerySearch/Search...
Local intranet | Protected Mode: Off
Query Search - Wi...
QuerySearch - Ne...
Untitled 1 (C:\Use...
My Web Sites
mdu - Notepad

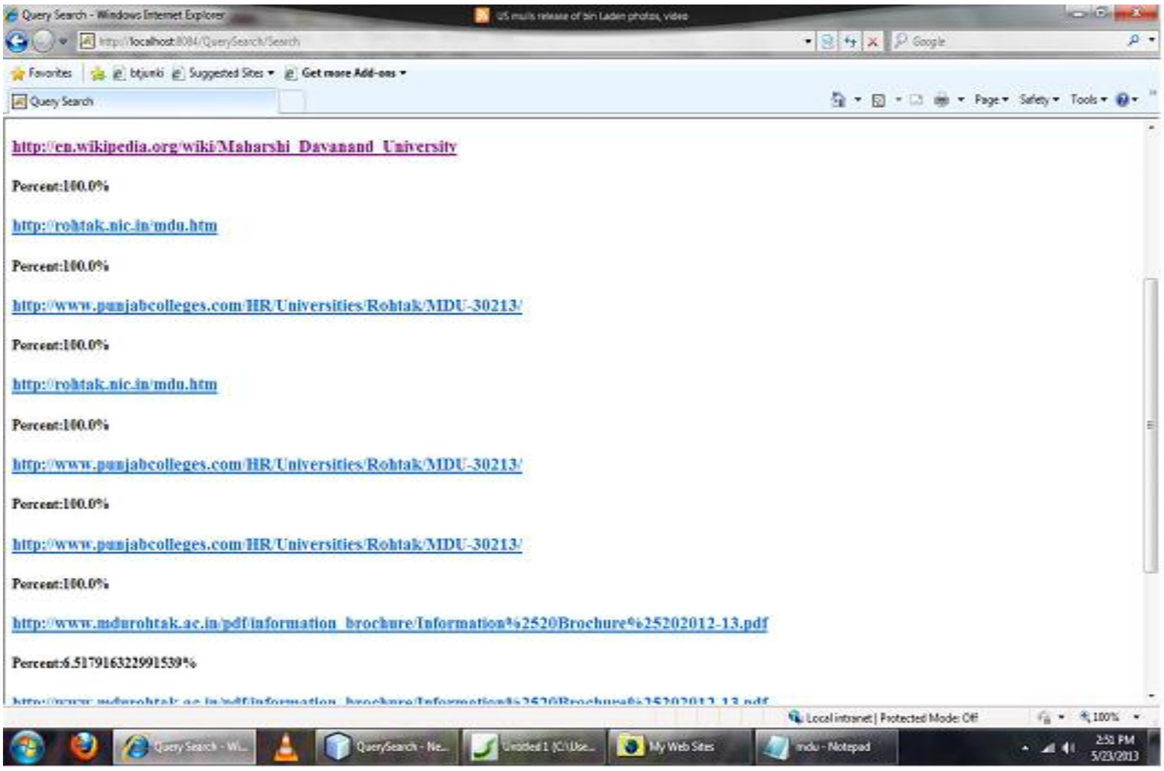
Enter your Query

```
India. Established in 1976 and named
after the saint Dayananda Saraswati,
the university offers courses in a
number of fields of study at the
undergraduate, postgraduate and
doctoral levels. Besides offering
courses on campus, MDU is a collegiate
university which provides programs
through colleges, institutes and
centers throughout the region. MDU
Rohtak shares its campus with IIM
Rohtak.
```

Search

Figure 2: Input Query

Here figure 2 is showing the input query associated with the current system. As the query is input it is processed by the plagiarism server for the query processing and to retrieve the keywords from the query. Once the query is filtered the next work is performed by the web server to perform the similarity match and to retrieve the web pages based on similarity match. The results obtained are shown in figure 3.



```
http://en.wikipedia.org/wiki/Maharshi_Dayanand_University

Percent:100.0%

http://rohtak.nic.in/mdu.htm

Percent:100.0%

http://www.punjabcolleges.com/HR/Universities/Rohtak/MDU-30213/

Percent:100.0%

http://rohtak.nic.in/mdu.htm

Percent:100.0%

http://www.punjabcolleges.com/HR/Universities/Rohtak/MDU-30213/

Percent:100.0%

http://www.punjabcolleges.com/HR/Universities/Rohtak/MDU-30213/

Percent:100.0%

http://www.mdurohtak.ac.in/pdf/information_brochure/Information%2520Brochure%25202012-13.pdf

Percent:6.517916322991539%

http://www.mdurohtak.ac.in/pdf/information_brochure/Information%2520Brochure%25202012-13.pdf
```

Figure 3: Similarity Match Analysis

CONCLUSIONS

This paper is about to perform the detection of duplicate web contents over the web pages. In this work, a statistical analysis approach along with similarity measures is presented to retrieve the duplicate pages over the web. The obtained results show the extensive implementation of the proposed approach. It also describes about what is plagiarism and different similarity detection measures to detect suspicious documents. In all this approach could help to find more detailed results for search.

REFERENCES

1. Anatomy of a Search Engine (2010) [WWW Page]. URL from

   http://www.onecybertech.com/blog/2010/10/anatomy-of-a-search-engine/.

2. Curt Franklin. *Web Crawling.* [WWW Page]. URL from

   http://computer.howstuffworks.com/internet/basics/search-engine1.htm.

3. Methods of indexing (2007). URL from http://searchenginecrawler.blogspot.in/.

4. M. Bouville (2008), *Plagiarism: Words and ideas.* Science & Engineering Ethics, vol. 14, pp. 311-322.

5. Nadelson, S. (2007). *Academic Misconduct by University Students: Faculty Perceptions and Responses. Plagiary: Cross-Disciplinary Studies in Plagiarism*. Fabrication, and Falsification, 67-76.

6. Nwokedi Idika, Mayank Varia, and Harry Phan, (2013). *The probabilistic provenance graph*. IEEE security and privacy workshops.

7. P.S.Bhatia, Divya Gupta (2008). *Discussion on Web Crawlers of Search Engine*, Proceedings of 2nd National Conference on Challenges & Opportunities in Information Technology (COIT-2008).

8. S. M. Alzahrani, N. Salim, and A. Abraham (2011). *Understanding Plagiarism Linguistic Patterns, Textual Features and Detection Method.* IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews. vol., pp. 1-17.