A Hybrid Approach for Relevant Information Retrieval from Web Pages

Er. Harleen Kaur[1] , Dr. Raman Maini[2]
[1] Research Scholar, Department of Computer Engineering, UCOE, Punjabi University, Patiala.
[2] Professor, Department of Computer Engineering, UCOE, Punjabi University, Patiala.

***Abstract*-The World Wide Web has to carry out number of tasks in order to retrieve information. Search engines may return millions of pages in response to a user query. It is not possible for a user to preview each and every result. Therefore the most important goal for search engines is to provide the most relevant information on the top of the list. Most of the ranking algorithms are based on Web Structure Mining or Web Content Mining. The proposed ranking method combines Web Structure Mining , Web Usage Mining as well as Web Content Mining . The results of the proposed method are accurate and are relevant with the user's search query.**

Keywords : Web Structure Mining , Web Content Mining , Web Usage Mining , Link Structure , Inlinks , Outlinks .

I. INTRODUCTION

II.

The World Wide Web has become a very powerful platform to store as well as retrieve information due to the tremendous growth of information available over the Internet. The main aim of Page Ranking algorithms is to display the desired and relevant results for the users at the top of the list [1]. The proposed method incorporates multiple factors such as Inlinks , Outlinks , Total number of visitors on the website , Visitor rate , Average time spent by the users , Visitor per day ,Domain Age , Domain Keyword and the time period of Last Updation . It gives different weightage to different parameters based on their importance.

III. INFORMATION RETRIEVAL

Information retrieval is defined as an process of locating information which is relevant to a user's query.It has been recorded that only 80-85% of the total pages available on the web provide useful information and the remaining 20-15% are either complete duplicates or somewhat duplicates of the original pages and some of them are completely irrelevant pages [4].

IV. WEB MINING

Web Mining is Data Mining technique which deals with the extraction of hidden information from the World Wide Web. This hidden information i.e. knowledge could be contained in the content of Web pages or in the Web server logs or in the link structure of WWW . Web Mining can be classified into following three categories:

- *Web Structure Mining* (WSM)

WSM is done at the hyperlink level [2]. It is used to create structural summary about the web pages in the form of web graph where the web pages act as nodes and on the other hand hyperlinks act as edges connecting two related pages.

- *Web Content Mining* (WCM)

Web Content Mining is the process of examining the content of the web pages. Research in web content mining usually includes resource discovery from the web, document categorization , document clustering, and information extraction from the web pages [3]. The web pages mostly include data in the form of text , audio , video , graphic , image etc.

- *Web Usage Mining* (WUM)

Web usage mining is the process of extracting useful information from server logs. It is the process of finding out what users are looking for on the internet.

V. LIMITATIONS OF EXISTING METHOD

Existing page ranking algorithms such Page Rank , Weighted page Rank , HITS etc have following limitations:

- The page rank computed using Page Rank and Weighted Page Rank algorithm is based only on the link structure of the web and it remains unaffected whether the page has been accessed by the users or not i.e they ignore the relevancy of the web page [5].
- It favours old pages and provides them a higher rank than the newer pages.
- Dead Ends often occur in Page Rank algorithm. Dead Ends are the pages that have no outlinks.

- Another problem in with existing algorithms is that of Spider Traps. A group of pages is a spider trap if there are no links from within the group to outside the group [6].

VI. PROPOSED RANKING TECHNIQUE

In this paper a new ranking technique has been proposed which takes into consideration many crucial factors for computing the rank score. Each factor is given different weightage based upon its significance in the computation the total rank score. The flowchart of the proposed technique is as follows:

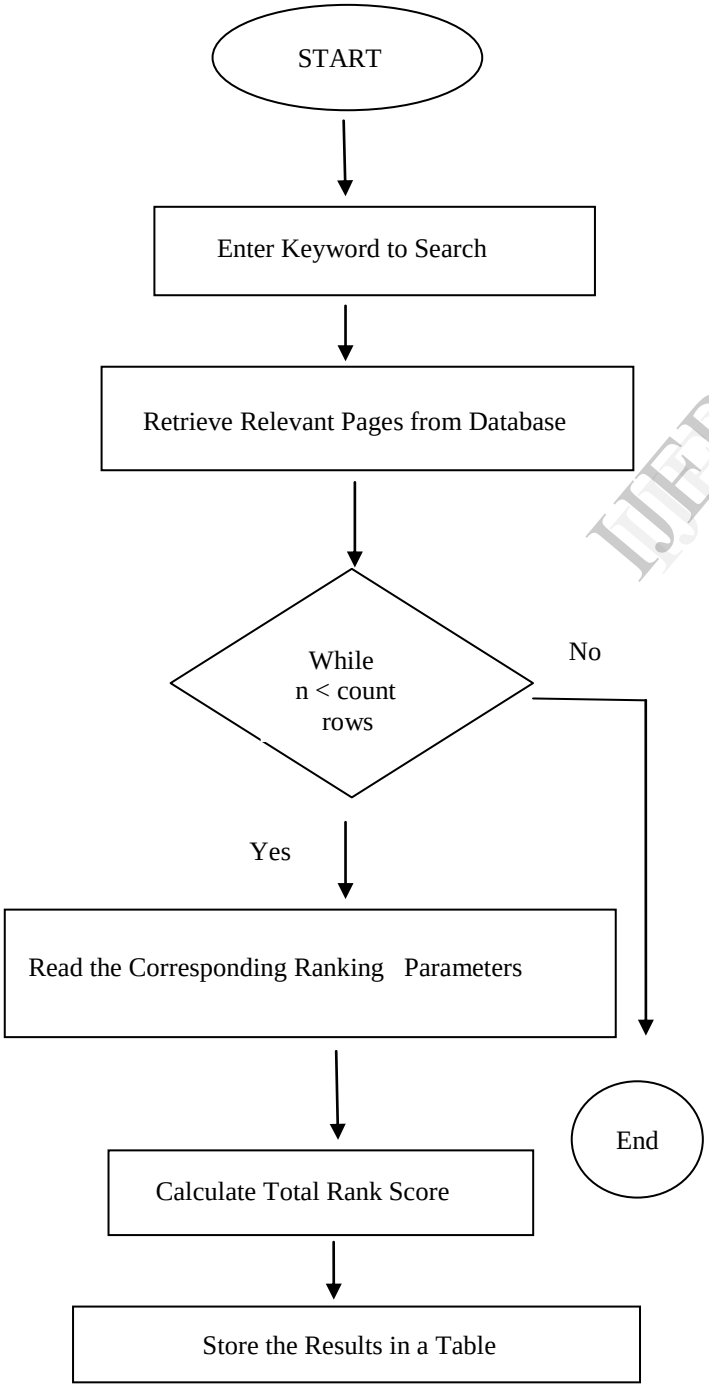

Figure 1. Flowchart of Proposed Method

The various factors used in the proposed technique are explained as follows:

1. *DOMAIN AGE* : Domain Age refers to the time the website became active and got itself registered. The domain age has a significant impact on ranking of a website because an older domain is usually considered to more authoritative. Another important reason for taking domain age into consideration is that spammers tend to register and drop domains quickly therefore a newly registered site always has a greater chance for being used for spam. Domain age score is calculated using the following formulae:

$$\text{Domain Age Score (DS)} = \frac{\text{Total outlinks on webpage} \times \text{outlink \% value}}{100}$$

2. *DOMAIN NAME* : This is another ranking factor used in the proposed method. The webmasters should carefully choose the domain name because choice of domain name has a huge impact on the success of the website. An ideal domain name is the one which is unique and has the keywords the website is trying to target. Exact match for domain names in the proposed method is given high weightage.

If the domain name is an exact match with the search query then

Rank =Value 1

Else

Rank =Value 2

Where Rank Value 2 << Rank Value 1.
The Domain name score is represented as PNS.

3. PAGE NAME : Another ranking factor used is page name . If the searching keyword matches the page name then it helps to get a higher rank score.
If the page name matches with the search query then

Rank =Value 1

Else

Rank =Value 2

Where Rank Value 2 < Rank Value 1.

The Page name score is represented as PNS

4. *INLINKS* : Inlinks are also known as backlinks , inward links or inbound links. They are considered to be incredibly important because they act like votes for a page. Therefore Search Engines often use the number of inlinks that a website has as an important factor for determining that website's ranking, popularity as well as importance [7]. Higher are the number of inlinks to a page higher will be its rank score. Inlink weightage is calculated using the given formulae :

$$\text{Outlink Score (OS)} = \frac{\text{Total inlinks on webpage} \times \text{inlink \% value}}{100}$$

5. *OUTLINKS*: Outlinks are defined as links that point your web page to another web page. They are also known as forward links. In the proposed method outlinks are given lesser weightage as compared to inlinks. Outlink weightage is calculated with the following formulae:

$$\text{Outlink Score (OS)} = \frac{\text{Total outlinks on webpage} \times \text{outlink \% value}}{100}$$

6. *TOTAL NUMBER OF VISITORS* : Visitors are true measure of a websites popularity The number of visitors visiting a page also has significant role in determining the page rank. If the content on a page is original as well as useful then it would automatically draw more visitors to the page as compared to web pages that are not useful. This factor takes into account total number of visitors who have till date visited the website.

$$\text{No. of Visitor Score (VS)} = \frac{\text{Total Visitors on webpage} \times \text{Visitor \% value}}{100}$$

7. *VISITORS PER DAY*= The number visitors visiting a site per day reflects the popularity of the website. It is an important factor for search engine ranking. If a page has more number of visitors visiting per day then it will be certainly ranked higher than a competitive page having lesser number of visitors on daily basis. Visitors per day score is calculated as :

$$\text{Visitors Day score (VDS)} = \frac{\text{Visitors / Day} \times \text{Visitor / Day \% value}}{100}$$

8. *VISITOR RATE* : Visitor Rate is a factor that is used to determine whether the number of visitors have increased or decreased. It takes into account whether the visitors have increased or decreased in comparison to last month. It is calculated as:

$$\text{Visitor Rate Score (VRS)} = \frac{\text{Visitor Rate} \times \text{Visitor Rate \% value}}{100}$$

9. *AVERAGE TIME SPENT BY VISITOR* : The time spent on a website is a traffic quality metric and is an indicative of user satisfaction. It is an important factor for improving the precison of the rank score. A long time spent on the website depicts strong interest in the content and services offered by the website. The ideal time spent varies depending upon the services offered by the website. Average time spent by the user is calculated as:

$$\text{Time Score (TS)} = \frac{\text{Time Spent} \times \text{Time Spent \% value}}{100}$$

10. *LAST UPDATE* : This is a significant factor for web ranking. Search Engines prefer new and unique content , so in the proposed method the website which is recently updated will tend to get a better rank score. This factor helps in assigning lower score to the websites containing old and irrelevant data.

If Last Update < n months

Then

Rank =Value 1

Else

Rank =Value 2

Where Rank Value 2 < Rank Value 1.

Last Update score is represented as LUS.

The total rank score of a website is calculated using the following formulae:

$$\text{Total Score} = \frac{DS+DNS+PNS+IS+OS+VS+VDS+VRS+TS+LUS}{10}$$

VII. RESULTS AND DISCUSSION

Inorder to evaluate the proposed method we have created our own database which includes information about different domains such as Banking , Education , Automobiles , Games , Fitness , Healthcare , Jobs , Matrimonials , Books etc.

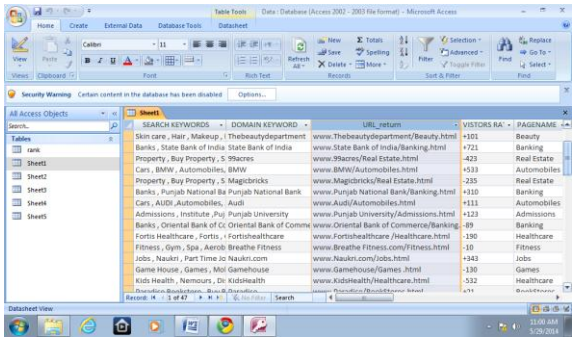

Figure 2. Snapshot of the Database.

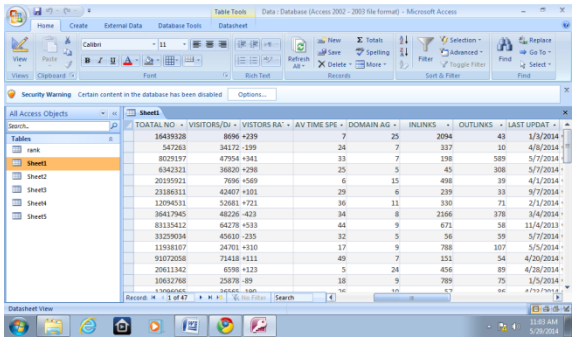

Figure 3 Snapshot of the Database

A Search Engine has been created using Visual Studio 2010. The results generated by the proposed method are listed in descending order based on their total rank score. Figure 4 shows the list of search results displayed corresponding to a Banking query.

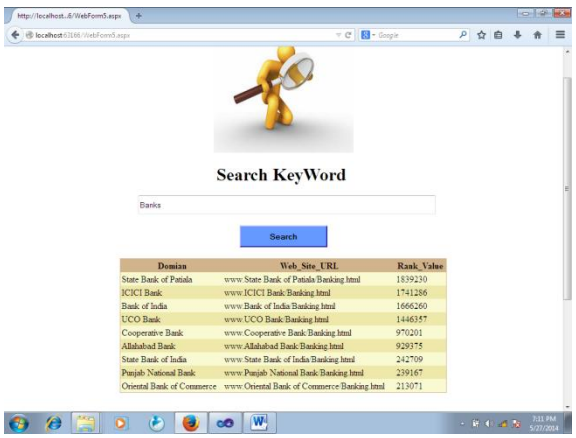

Figure 4 Snapshot of search results.

With the help of the proposed method more relevant results are displayed at the top of the list. The use of multiple crucial factors for ranking in the algorithm ensures that no website gets a higher score by any spamming activity.

VIII.CONCLUSION

This paper discusses a new ranking algorithm which consolidates Web Structure Mining , Web Usage Mining and Web Content Mining in order to produce more relevant search results. The multiple factors used in the method ensure precision and relevancy of results. In the near future , a thesaurus will be added to facilitate synonym discovery.

REFERENCES

1. Ranveer Singh and Dilip Kumar Sharma ,“ RatioRank: Enhancing the Impact of Inlinks and Outlinks ”, Third International Advance Computing Conference (IACC), IEEE , 2013.
2. Shital C. Patil, R. R. Keole , “ The Role of Web Content Mining and Web Usage Mining in Improving Search Result Delivery ”, IJCSMC, Vol. 3, Issue. 3, March 2014, pg.7 – 14
3. Govind Murari Upadhyay, Kanika Dhingra, “Web Content Mining: Its Techniques and Uses”, IJARCSSE, Vol 3, Issue 11, November 2013.
4. Deepak Garg and Deepika Sharma , “Information Retrieval on the Web and its Evaluation ”, International Journal of Computer Applications , February 2012.
5. Sonal Tuteja , “Enhancement in Weighted PageRank Algorithm Using VOL”, IOSR , Vol. 14, Issue 5 , Sep. – Oct. 2013.
6. Pooja Devi, Ashlesha Gupta , Ashutosh Dixit , “Comparative Study of HITS and PageRank Link based Ranking Algorithms ” , International Journal of Advanced Research in Computer and Communication Engineering , Vol. 3, Issue 2, February 2014.
7. http://en.wikipedia.org/wiki/Backlink