# Enhancing Page Rank Algorithm

Ekta Bhardwaj[1], Shiv Kumar[2], Kuldeep Tomar[3]

[1]M.Tech scholar, Department of computer science & Engineering, NGFCET
[2]M.Tech scholar, Department of computer science & Engineering, NGFCET
[3]Research scholar, Department of CSE, MRIU, Faridabad, India
*Email: [1]ektabhardwaj88@gmail.com, [2]skrawat88@gmail.com , [3]kuldeep_karan@yahoo.com*

***Abstract:-*** World Wide Web (WWW) is a collection of web Pages. A web Page consists of audio, Images, video, text etc. Retrieving web pages from large collection of WWW is a challenge. Web mining is used in extracting data from WWW. Web structure mining and web content mining plays an effective role in this approach. Page rank and weighted page rank algorithms are commonly used in the web structure mining, whereas HITS and weighted page content rank algorithms are used in the web structure mining and web content mining techniques. In this paper, a new algorithm is proposed, WPUCR (Weighted Page User Content Rank) algorithm which is a combination of web Usage Mining, web content mining and Web structure mining. It is the extension of weighted page content rank algorithm and shows the relevancy of the pages to a given query in a much more refined manner and works on the user behavior.

***Index Terms:*** *—Web mining, PageRank, weighted PageRank, Weighted Page User Content Rank.*

*****

## I. INTRODUCTION TO WEB MINING

Web mining is used to discover the content of the web, the user's behavior in the past, and the web pages that the users want to view in the future. Web mining can be easily executed with the help of other areas like Database, Information Retrieval, NLP and machine learning. Web mining is the data mining technique that extracts the information from web documents.
The challenges in the web mining are:-

1. Web is huge.
2. Web pages are semi-structured.
3. Web information stands to be diverse in meaning.
4. Degree of quality of information extracted.
5. Conclusion of knowledge from information extracted.

**Web Mining Categorization**

Mining researchers focuses on discovering new information or knowledge in the data. Web mining is classified basically into Web Structure Mining, Web Content Mining, web Usage Mining.

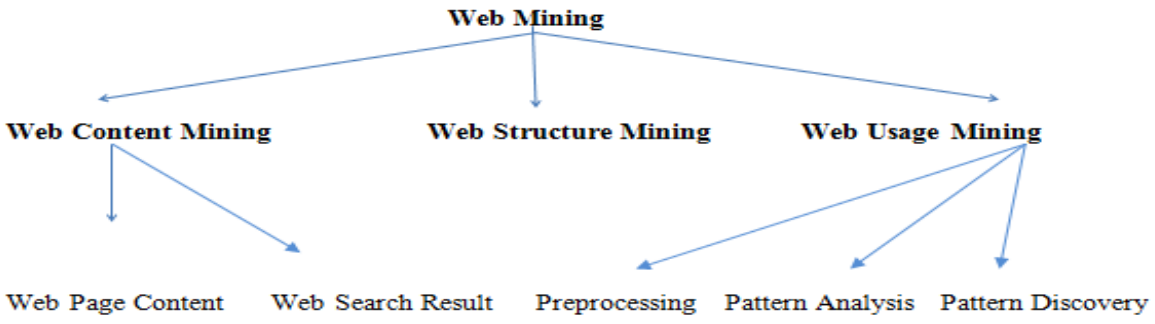



Figure 1: Web Mining Categories

**1. Web Content Mining (WCM)**

Web Content Mining is defined as the process of extracting information from the contents of web documents. Content data corresponds to the collection of facts a web page was designed to convey to the users. The technologies that are normally used in WCM are NLP and IR [3].

**2. Web Usage Mining (WUM)**

Web Usage Mining ascertains user profiles and user's behavior recorded inside web log files. It basically deals with discovering usage patterns from web data in order to understand and better serve needs of web based applications. One of the major challenges faced by web usage mining applications is that web server log data are anonymous, making it difficult to identify users and user session from the data [1] [3].

Web usage mining techniques contains 3 processes namely:

a) Preprocessing: - reformat the log files and retrieve the web pages to the local space.

b) Pattern analysis: - means filter the irrelevant information and to visualize and interpret interesting pattern to the users.

c) Pattern discovery: - uncover the patterns in server logs but is often carried out only on sample of data .

**3. Web Structure Mining (WSM)**

Web Structure Mining generates the structural summary about the websites and web pages. It tries to discover the link structure of hyperlinks at inter document level [6]. As it is very common that the web documents contain links and they use both the real or primary data on the web so it can be concluded that Web Structure Mining has a relation with Web Content Mining[7]

**Table: Comparison of all the Existing Algorithms**

| Algorithm/criteria | PAGE RANK | WEIGHTED PAGE RANK | HITS | WEIGHTED PAGE CONTENT RANK |
|---|---|---|---|---|
| Mining Techniques | WSM | WSM | WSM AND WCM | WSM AND WCM |
| Working Process | Computes values at index time and results are sorted on the priority of pages. | Computes values at index time and results are sorted on the basis of page importance. | An highly relevant pages are computed and find values on the fly. | Gives sorted order to the web pages returned by search engines as a numerical value in response to a user query. |
| I/P parameter | Backlinks | Backlinks and forward links | Backlinks, forward links and content | Backlinks, forward links and content |
| Complexity | O (log N) | <O (log N) | <O (log N) | <O (log N) |
| Limitations | Query independent | Query independent | Topic drift and efficiency problem | No limitation |
| Search engine | Google | Google | Used in IBM search engine Clever | Research model |

## II. PROPOSED ARCHITECTURE

Search engines are defined as the key to finding specific information on the vast expanse of the World Wide Web. There are at least three elements which contains important: information for a search engine: discovery of the database, the user search, presentation and ranking of results. With the proposed WPUCR, the search engine architecture is modified so as to add the components for calculating importance and relevancy of pages

The modified architecture is displayed in Figure 2.

 ISSN: 2321-8169
3381 – 3385

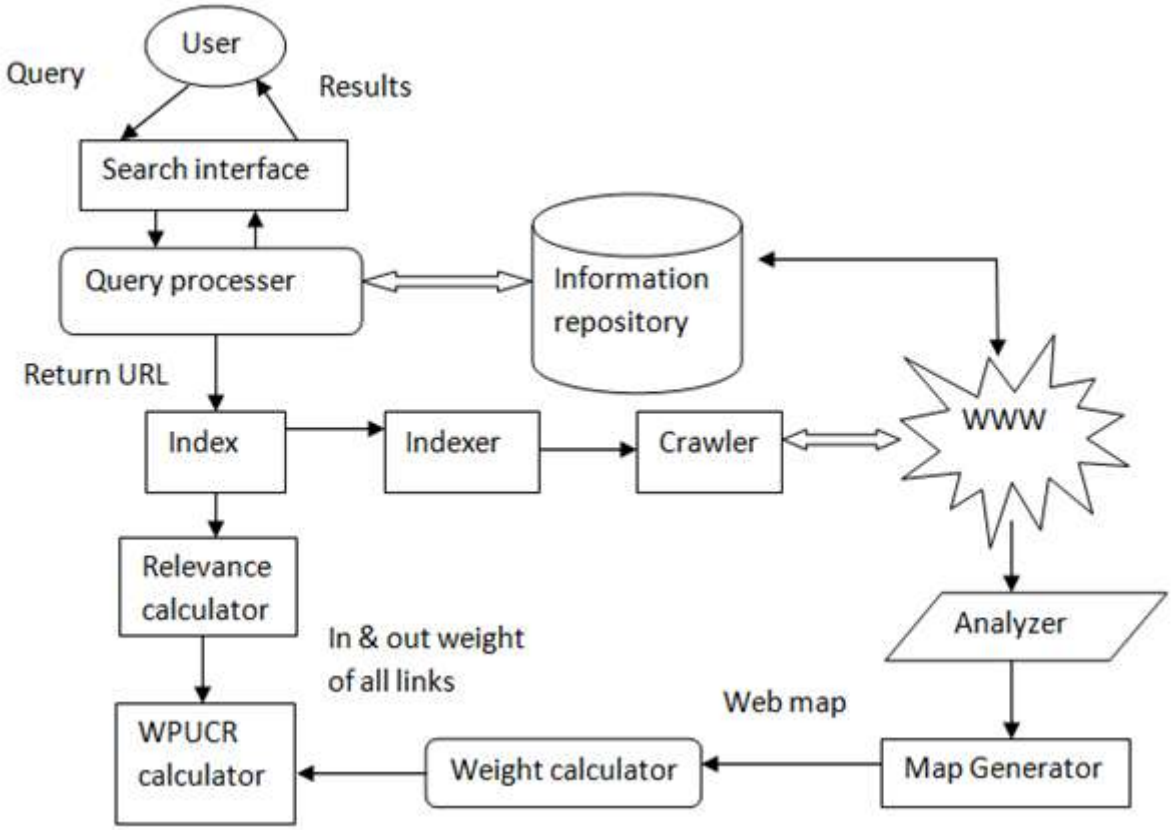



Figure 2: Modified Search Engine Architecture.

## III. PROPOSED ALGORITHM

Algorithm: WPUCR calculation
Input: Page P, In link and Out link Weights of all backlinks of P, Query Q, d (damping factor), Hit Frequency F of P.
Output: Rank score
Step 1: Relevance calculation:
a) Find all meaningful word strings of Q (say N)
b) Find whether the N strings are occurring in P or not?
Z= Sum of frequencies of all N strings.
c) S= Set of the maximum possible strings occurring in P.
d) X= Sum of frequencies of strings in S.
e) Content Weight (CW) = X/Z
f) C= No. of query terms in P
g) D= No. of all query terms of Q while ignoring stop words.
h) Probability Weight (PW)= C/D
i) Hit frequency = number of times the page P is traversed.

**Step 2: Rank calculation:**

a) Find all backlinks of P (say set B).
b)

$$\mathrm{WPUCR(n)} = \mathrm{F} * \{(1-\mathrm{d}) + \mathrm{d} \sum_{\mathrm{m} \in \mathrm{B(u)}} \mathrm{PR(m)} \mathrm{W}^{\mathrm{in}}_{\mathrm{(m,n)}}\, \mathrm{W}^{\mathrm{out}}_{\mathrm{(m,n)}} * (\mathrm{Cw} + \mathrm{Pw})\}$$

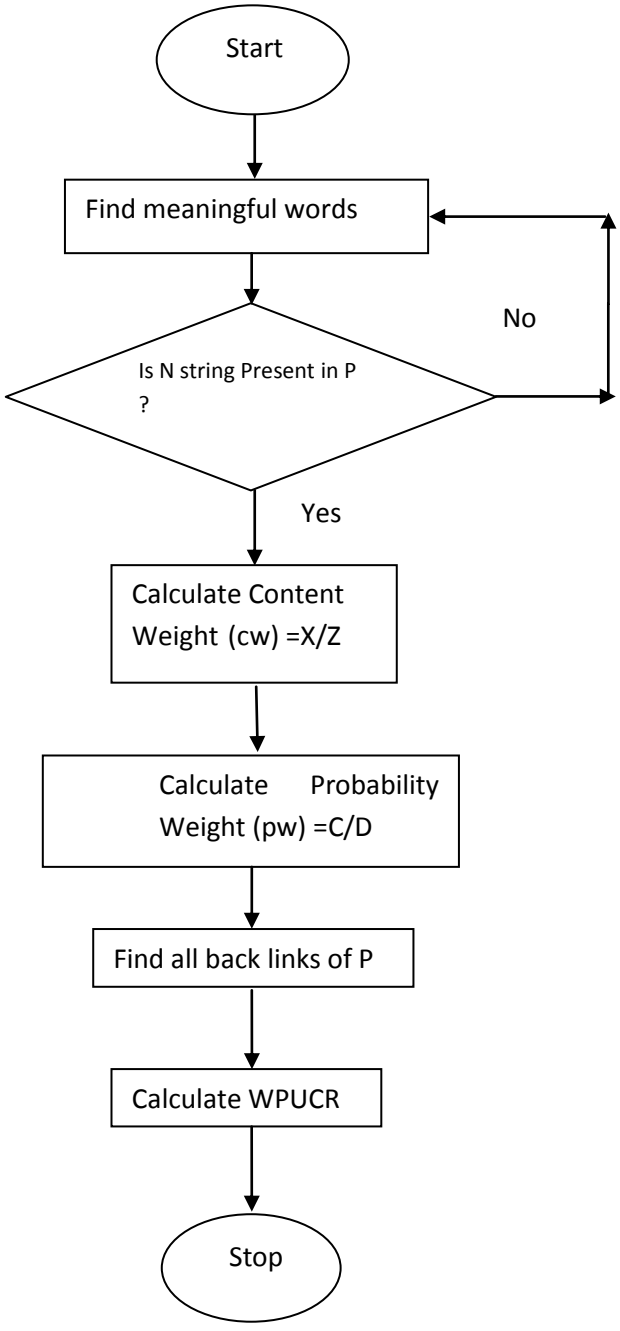
c) Output) WPUCR (P) i.e. the Rank score
It can be seen from the formula to calculate the Weighted Page Content Rank of a page U is

 ISSN: 2321-8169
3381 – 3385

$$PR(U) = F * \{(1-d) + d \sum_{v \in B(U)} WPCR(V) W^{in}_{(U,V)} W^{out}_{(U,V)} * (Cw + Pw)\}$$

Where,
PR (U) =PageRank of page U,
B (U) = Set of all pages referring to page U,
D= Damping factor which can be set between 0 and 1,
$W^{in}$ (U, V) = in weight of link (U, V),
$W^{out}$ (U, V) = out weight of link (U, V),
Cw=Content weight of page U
Pw=Probability weight of page U,
F=Hit frequency of P

## IV. PROPOSED ALGORITHM FLOWCHART



**Conclusion**

Weighted Page User Content Rank algorithm (WPUCR) is the modification of the original Page Rank (PR) algorithm followed by Weighted Page Content Rank (WPCR). WPUCR decides the rank score based on the popularity of the pages by taking into consideration the importance of both the in links and out links of the pages along with the user behaviour. This algorithm provides high value of rank to the more popular pages and does not equally divide the rank of a page among it's out link pages. Every out-link page is given a rank value based on its popularity. Popularity of a page is decided by observing its number of in links and out links along with user behaviour.

## V. REFERENCES

[1] Pooja Sharma and PawanBhadana, "Weighted Page Content Rank For Ordering Web Search Result", International Journal of Engineering Science and Technology, Vol 2, 2010.

[2] Wenpu Xing and Ali Ghorbani, Weighted Page Rank Algorithm, Proceedings of the the Second Annual Conference on Communication Networks and Services Research (CNSR' 04), IEEE, 2004.

[3] Laxmi Choudhary and Bhawani Shankar Burdak,"Role of Ranking Algorithms for Information Retrieval", International Journal of Artificial Intelligence and Applications(IJAIA), Vol.3, No.4, July 2012.

[4] Dilip Kumar Sharma and A.K. Sharma," A Comparative Analysis of Web Page Ranking Algorithms", International Journal on Computer Science and Engineering, Vol. 02, No. 08, 2010, 2670-2676.

[5] DebajyotiMukhopadhyay, PradiptaBiswas and Young-Chon Kim," A Syntatic Classification based Web Page Ranking Algorithm", 6th International Workshop on MSPT Proceedings, MSPT 2006.

[6] Tamanna Bhatia," Link Analysis Algorithms for Web Mining", IJCST Vol. 2, Issue 2, June 2011.

[7] Rekha Jain and Dr G.N.Purohit, "Page Ranking Algorithms for Web Mining," International Journal of Computer application,Vol 13, Jan 2011.

[8] N. Duhan, A. K. Sharma and K. K. Bhatia, "Page Ranking Algorithms: A Survey, Proceedings of the IEEE International Conference on Advance Computing, 2009.

[9] M. G. da Gomes Jr. and Z. Gong, " Web Structure Mining: An Introduction", Proceedings of the IEEE International Conference on Information Acquisition, 2005.