A COMPARITIVE STUDY OF LINK BASED PAGE RANKING ALGORITHM

Shilpa Sethi[1], Ashutosh Dixit[2]

[1]Assistant Professor, [2]Associate Professor, Dept of CE,
YMCA University of Science and Technology, Faridabad , Haryana (India)

ABSTRACT

In order to assist the user to easily locate its information need, search engines use different page ranking algorithm to sort the pages based on some relevance factors. Most of the algorithms in this area are based on web mining concepts. Web mining discovers the useful pattern from the information repository This work in this paper conduct a comparative studies of various algorithm based on web mining techniques The study provides the benefits and limitations of these algorithms which can help researcher to discover new ideas for better page retrieval and page ranking.

Keywords: Search Engine, Page Ranking, Web Mining, Link Visits

I. INTRODUCTİON

The World Wide Web is a huge source of information which is growing at a rate of thousands of pages per day [10]. To access useful information such a huge repository, information retrieval tools such as search engines are used. The basic tasks of every search engine are crawling, indexing, ranking and presenting results to the user. Crawler downloads the web pages from different web server at some specified interval, indexer indexes these pages into search engine database and query processor fetches the documents based on term relevance with the user query and applies some ranking algorithms to sorts the matched documents. The basic architecture of search engine is shown in fig 1.

The paper is organized as follows: overview of web mining techniques is discussed in next section Section3 discuss various page ranking algorithms based on link structure mining with example illustrations. In Section 4, analytical results are provided. Concluding remarks are given in section 5.

`

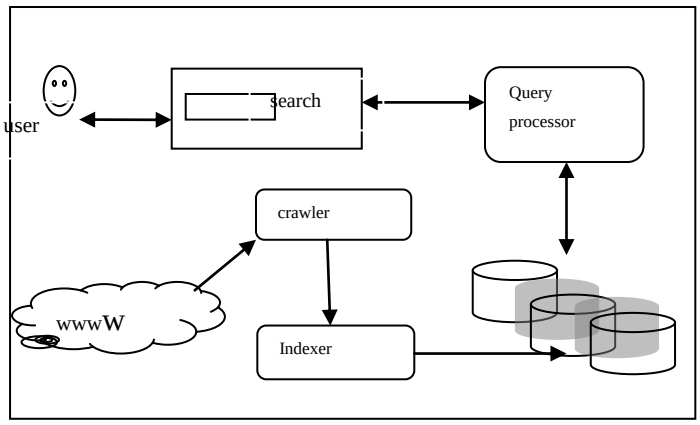

Fig1: Basic Architecture of Search Engine

II. RELATED WORK

This section describes an overview of web mining and its different categories with examples illustration.

Web mining means discovering the nontrivial, previously unknown and extremely useful patterns from WWW. It can be divided into three categories as shown in fig2.

1) web content mining
2) Link structure mining
3) Web usages mining.

These techniques are discussed in detail in subsequent sections

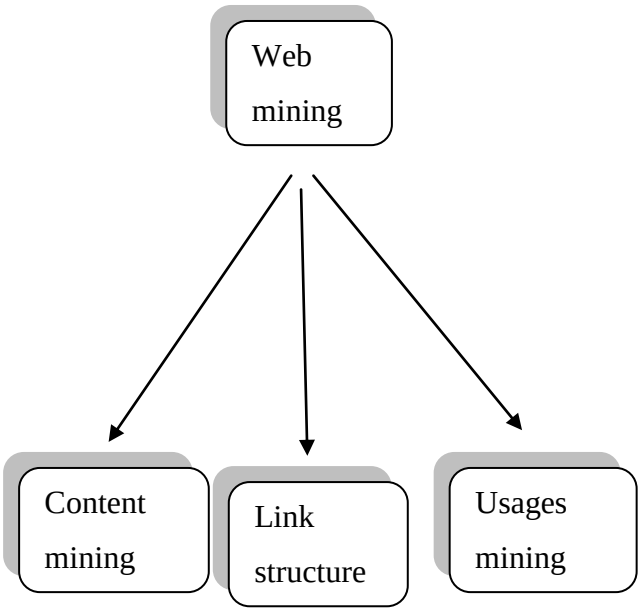

Fig2: Categories of Web Mining

2.1 Web Content Mining

It is the process of retrieving the information from WWW on the basis of content of pages [1,6] such as no. of keywords , frequency of keywords, the position of keywords in the document , their hyper textual information etc. the main focus is inner document structure. NLP, machine learning, association rules are some of methods used in this area.

2.2 İnk Structure Mining

link structure mining [2,6] provides the link summary of documents in the form of web graph by using the concept of hyperlink topology at inner as well as inter document level as shown in Fig 3.

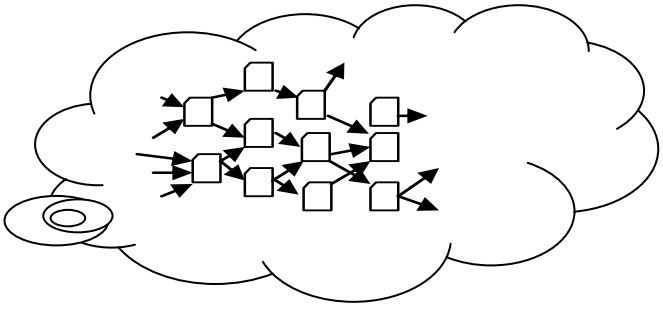

Fig 3: Web Graph

2.3 Web Usages Mining

This technique identifies the user browsing behavior by monitoring and storing information of user navigational patterns. It focuses on user browsing history and interested domains which can be identified either explicitly or implicitly [9].

In implicit approach query logs and server logs are used to discover user interest [5] whereas in explicit approach, user is asked to fill interested domain. Machine learning, personalization algorithms, association rules are used to mine the useful patterns.

Some of the popular algorithm based on link structure mining is discussed in next section.

III ANALYSIS BASED ON WEB STRUSUTE MINING

Link structure mining find out the link summary of pages in the form of web graph. A web graph is directed labeled graph having web pages as the nodes and hyperlinks as the edges between these nodes as shown in Fig 2 above. There are many algorithms based on link structure mining. Some of them which form the basis of comparative study are discussed in following sections.

3.1 Pagerank

The pageRank algorithm was developed Larry Page and S. Brim [4] . It is based upon citation analysis of web pages to find out the importance of a web page. According to this algorithm, if the incoming links of a page are important then its outgoing links also become important. The page rank of a page *P* is equally divided among its outbound links which further, propagated to their corresponding outgoing links. The page rank of a page n can be calculated by eqn (1) as given below.

$$PR(n)=\sum_{m \in I(n)} \frac{PR(m)}{N_m} \quad \ldots..(1)$$

Where:

- PR(m) and PR(n) represents the page rank of page m and page n respectively
- I(n) is set if incoming links of page n
- Nm represents the no. of outgoing links of page m
- d is the damping factor which is a measure of probability of user following direct link. Its value is usually set to 0.85

Example Illustrating Working of PR

To explain the working of page rank algorithm ,let us take an small hyperlinked structure shown in fig 4, consisting of three pages X, Y and Z. where page X links to the page Y and Z, page Y links to page Z and X, page Z links to page X and Y.

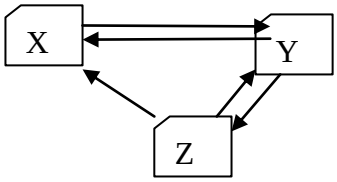

Fig 4 Sample Hyperlinked Structure

According to equation (1), Pagerank of page X, Y , Z can be computed as follows:

PR(X)=[1d)+d((PR(Z)/2] (1a)

PR(Y)-[(1-d)+d(PR(X)/1+PR(YZ)/2] (1b)

PR(Z)=[(1-d)+d(PR(X)/2 +PR(Y)/2] (1c)

Initially considering the page rank of each page equal to 1 and taking the value of d=0.5 , the new page rank of pages can be obtained as follows:

$$PR(X) = 0.5 + 0.5\left(\frac{1}{2}\right) = 0.75$$

$$PR(Y) = 0.5 + 0.5\left(1 + \frac{1}{2}\right) = 1.25$$

$$PR)Z) = 0.5 + 0.5\left(\frac{1}{2} + \frac{1}{2}\right) = 10$$

Calculating pagerank of each page by using iteration method as shown in table 1

Table 1 Calculation of Page Rank by PR Method

PR(X)	PR(Y)	PR(Z)
1.0	1.25	0.75
1.0	1.188	0.813
1.0	1.203	0.797
1.0	1.199	0.801
⋮	⋮	⋮

From the above table, it may be noted that PR(Y)>PR(X)> PR (Z) .

These PR values are extracted by crawler while downloading a page from web server and these values will remain same till the web link structure will not change. In order to obtain the overall page score of a page , the query processor add the pre computed pagerank(PR) value associated with the page with text matching score of page with the user query before presenting the results to the user.

3.2 Weighted Pagerank Algorithm

Weipu Xing et.al [2] proposed an algorithm which is extension of basic page rank algorithm. It assigns the page rank on the basis of link popularity of incoming and outgoing links. The page rank of page n is computed by eqn (2) given below.

$$PR(n) = (1-d) + d\sum_{m \in I(n)} PR(m) * \frac{In}{\sum_{P \in R)m)} I_p} * \frac{On}{\sum_{P \in R)m)} Op} \quad \text{..... (2)}$$

Where :

- PR(m) and PR(n) are page rank of page m and n
- d is damping factor
- R(m) denotes the reference list of page m
- I_n, I_p denotes the no. of incoming links to page n and page p respectively.
- O_n , O_P denotes the no. of outgoing links of page n and page p respectively.

Example Illustrating Working of WPR

By considering the same hyperlinked structure as shown in fig 4 and initially taking weighted page rank of each page equal to 1 and d=0.5, the new weighted page rank of pages X. Y and Z can be computed by using eqn. 2

$$PR(X) = 0.5 + 0.5\left(\left(1 * \frac{2}{3} * \frac{2}{3}\right) + \left(1 * \frac{1}{2} * \frac{1}{3}\right)\right) = 0.53 \ldots\ldots (2a)$$

$$PR(Y) = 0.5 + 0.5\left((1 * 1 * 1) + \left(1 * \frac{1}{2} * \frac{1}{3}\right)\right) = 1.08 \ldots\ 2(b)$$

$$PR(Z) = 0.5 + 0.5\left(1 * \frac{1}{3} * \frac{2}{3}\right) = 1.11 \qquad \ldots\ldots\ 2(c)$$

Calculating weighted page rank of each page by iteration method as shown in table 2

Table 2 Calculation of Page Rank by WPR Method

PR(X)	PR(Y)	PR(Z)
0.53	1.08	1.11
0.8264	0.853	0.6188
0.7371	0.9627	0.5938
0.600	0.9160	0.6998
⋮	⋮	⋮

From the above table, it may be noted that PR(Y)>PR (Z)> PR(X). The order of page rank is different from PR method.

3.3 Pageranking Algorithm Based on Link Visit

Duhan et al [5] identified the limitation traditional PR method that it evenly distributes the page rank of page among its outgoing links whereas it may not be always the case that all the outgoing links of a page holds equal importance. S0, they proposed a method which assigns more rank to an outgoing link that is more visited by the user. For this purpose a client side agent is used to send the page visit information to server side agent. A database of log files is maintained on the server side which store the URLs of the visited pages its hyperlinks and IP addresses of users visiting these hyperlinks. The visit weight of a hyperlink is calculated by counting the distinct IP addresses clicking the corresponding page. The page rank of page 'm' based upon visit of link is computed by the eqn (3).

$$PR(n) = (1-d) + d \sum_{m \in I(n)} \frac{PR(m) * LV(m}{TV(m, O(m))} \qquad \ldots.(3)$$

Where:

- PR(m) and PR(n) is page rank of page m and n respectively.
- I(n) set of incoming links of page n
- LV (m, n) is no. of link visits from m to n.
- TV(m ,O(m)) total no. of user visits on all the outgoing links of page m

Example illustrating the working of PRLV

Consider the same hyperlinked structure as shown in fig 4 above. Let the no. of visits from page X to page Y are 100; the no. of visits from page Y to X are 45 and the no. of visits from page Y to Z are 15; the no. of visits from page Z to Y are 50 and the no. of visits from page from Z to X are 25.the PR based on link visit can be easily calculated using eqn (3). Initially taking page rank of each page equal to 1 and d=0.5

$$PR(x) = 0.5 + 0.5\left(\left(1 * \frac{45}{45+15}\right) + \left(1 * \frac{25}{25+50}\right)\right) = 1.0416 \ (3a)$$

$$PR(Y) = 0.5 + 0.5\left(\left(1 * \frac{100}{100}\right) + \left(1 * \frac{50}{25+50}\right)\right) = 1.33 \ (3b)$$

$$PR(z) = 0.5 + 0.5\left(\left(1 * \frac{15}{15+45}\right)\right) = 0.625 \ \ldots (3c)$$

Calculating page rank based on link visit of each page by iteration method as shown in table 3

Table 3 Calculation of Page Rank by PRLV Method

PR(X)	PR(Y)	PR(Z)
1.0416	1.3333	0.625
1.1041	1.2229	0.6666
1.0719	1.2743	0.6536
1.0869	1.2537	0.6593
⋮	⋮	⋮

From the above table, it may be noted that PR(Y)>PR(X)> PR (Z) .The order of page rank is different from above two methods.By comparing the above methods results, it is found that page Y is obtaining highest precedence over other pages.

IV. ANALYTICAL RESULTS

The analysis is done using an online web log analyzer [7]. The analysis is done on a sample server log files of http://www.smsync.com from 1/4/2015 to 15/4/2015. It has been found that the page which is getting more hits directly or indirectly by the user, placed higher in the result set as shown in fig 5. The graph is potted between dates of observation and number of user hits. It is observed that a user cannot intenti0nly, increase the rank of a page as it depends on the ranks of its back links. Considerable improvements are observed in the ordering of search results by taking user's feedback on the result set.

Fig 5: Graph Between No. of Hits and Dates

V. CONCLUSION

Web Mining plays an important role in providing relevant information to the user. The work in this paper evaluated the page rank of set of pages by using three popular page ranking algorithms. By comparing the result set of this algorithm and verifying it with the help of online tool, it is concluded that PRLV performs

considerably good in terms of user satisfaction as compared to traditional page ranking algorithms. Comparison of PR, WPR, PRLV based on different parameters are given in the table 4

Table 4 Comparison of PR, WPR and PRLV

Attributes	*PR*	*WPR*	*PRLV*
Mining technique	WSM	WSM	WSM and WUM
İnput	Hyperlinks	Hyperlinks	Hyperlinks User navigational patterns
Complexity	O(Log n)	< O(Log n)	>O(Log n)
Relevancy of pages	Less	Less	More
Distribution of rank to outgoing links	Equal	Unequal	Unequal
Search engine	Own(Google)	Research model	Research model
Limitation	-No focus on user query - evenly distribution of rank to outbound links	No focus on user query	Extra effort of client and server side agent

In future an efficient page ranking algorithm based on link visits and other factors of user behavior such as time spent on a page , degree of user interest in different domains can be incorporated to PRLV so that more user centered results can be presented to the user.

REFERENCES

[1] N. Duhan, A. K. Sharma and Bhatia K. K., "Page Ranking Algorithms: A Survey, Proceedings of the IEEE International Conference on Advance Computing, 2009, 978-1-4244-1888-6

[2] WenpuS Xing and Ali Ghorbani,"weighted page rank algorithm" Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04)0-7695-2096-0/04 $20.00 © 2004

[3] S. Pal, V. Talwar, and P. Mitra. Web mining in soft computing framework : Relevance, state of the art and future directions.IEEE Trans. Neural Networks, 13(5):1163–1177,2002.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd, "The Pagerank Citation Ranking: Bringing order to the Web". Technical report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999.

[5] Gyanendra Kumar, Neelam Duahn, and Sharma A. K., "Page Ranking Based on Number of Visits of Web Pages", International Conference on Computer & Communication Technology (ICCCT)-2011, 978-1-4577-1385-9.

[6] Zdravko Markov and Daniel T. Larose, "Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage Data". Copyright 2007 John Wiley & Sons, Inc.

[7] www.weblogexpert.com

[8] Tyagi Neelam et.ai "Weighted page rank algorithm based on number of visits of web page" International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012

[9] Sethi Shilpa, Dixit Ashutosh" Design of personalized search system based on user interest and query structuring" Proceedings of the 9th INDIACom; INDIACom 2nd International Conference on "Computing for Sustainable Global Development", 11th – 13th March, 2015

[10] Www. Worldwidewebsize.com