A Comprehensive Review of Link Based Ranking Algorithms

Harleen Kaur[1] , Dr. Raman Maini[2]
[1] *Research Scholar, Department of Computer Engineering, UCOE, Punjabi University, Patiala.*
harleen786@gmail.com
[2] *Professor, Department of Computer Engineering, UCOE, Punjabi University, Patiala.*
research_raman@yahoo.com

Abstract - **World Wide Web (WWW) is the most useful source for Information Retrieval and Knowledge discovery. But due to the rapid increase in the size of the Web the users get easily lost in the rich hyper structure Web. The key goal of search engines is to provide relevant information to the users. Ranking has always been an integral component of any information retrieval system. In the context of the World Wide Web the role of ranking becomes all the more important. This work analyses three different link based ranking algorithms that are Page Rank , Weighted Page Rank and HITS. This paper discusses the strengths and weaknesses of these algorithms and provides a comparative anlaysis of these three algorithms.**

Keywords : **Web Mining, Link Based Ranking , PageRank, Weighted PageRank , HITS.**

I. INTRODUCTION

With the internet age the data and information explosion have resulted in vast amount of data. The enormous amount of data has made it difficult for users to evaluate and extract useful information. It has been seen that 65% - 70% users select the very first page of the returned results and about 20% - 25% of the users select the second page and very few of 3% - 4% users check the remaining results [1]. Therefore it becomes very important for the search engines to rank their results in an order which can satisfy the users request.

II. 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 content of Web pages or in link structure of WWW or in Web server logs. The complete process of extracting knowledge from Web data [2] is follows in Fig.1:

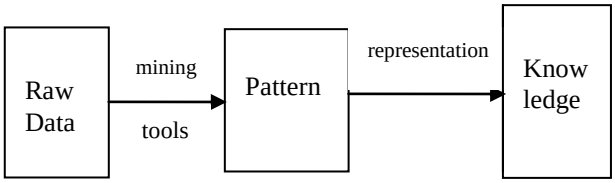

Figure 1.

Web mining is usually divided in three categories:

- Web content mining
- Web structure mining
- Web usage mining

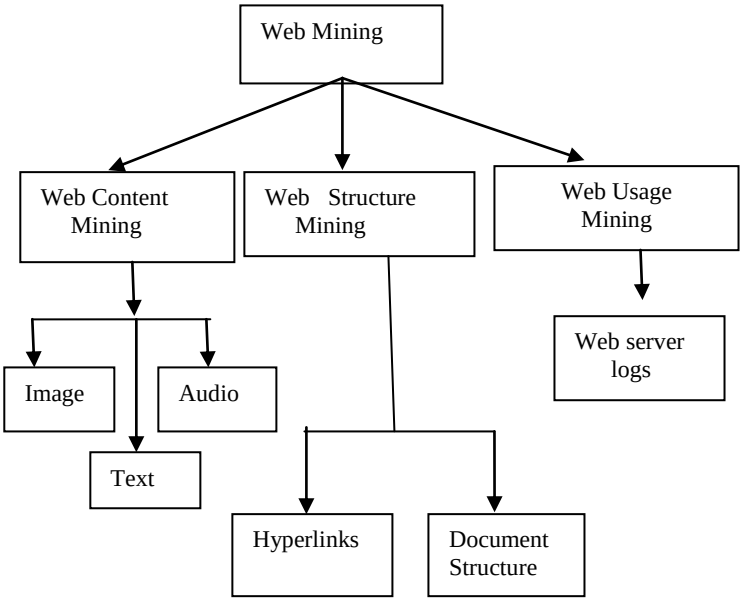

Figure 2.

A. Web Content Mining (WCM)

It is used to examine the contents of the web pages. Web Content mining focuses mainly on the structure within a document i.e. inner document level. It is related to Data Mining because many Data Mining techniques can be applied in Web Content Mining. Research in web content mining includes resource discovery from the web, document categorization , document clustering, and information extraction from the web pages [3]. The Web document usually contains several types of data such as text, image, audio, video and hyperlinks. Some of them are semi-structured such as HTML documents while others are structured data like the data organised in the tables. The unstructured characteristic of Web data force the Web content mining towards a more complicated approach.

B. Web Structure Mining (WSM)

It emphasizes on the data which describes the structure of the content. WSM 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.

C. Web Usage Mining (WUM)

It refers to the discovery of user access patterns from the web usage logs. A Web server log is an important source for performing Web Usage Mining because it explicitly records the browsing behavior of site visitors. The data recorded in server logs reflects the access of a Web site by multiple users [2]. It focuses on various data mining techniques to understand and analyze search patterns. It analyses the interaction between the user and the web pages during browsing.

III. LINK BASED RANKING ALGORITHMS

A. Page Rank (PR)

PageRank was developed by Larry Page and Sergey Brin [4] at Stanford University. This algorithm is used by Google to prioritize the results obtained by keyword based search. PageRank is based on the principle that if a page has important incoming links to it, then its outgoing links to other pages are also considered to be important. Thus inorder to calculate the rank of web pages it takes backlinks into account. Although many factors are used find the rank of a Google search result but PageRank continues to provide the basis for all of Google's web search tools [5]. The PageRank is calculated by the formula given in equation (1).

$$PR(u) = c \sum_{v \in B(u)} \frac{PR(v)}{Nv} \qquad (1)$$

where u represents a web page , PR(u) and PR(v) represent the page rank of pages u and v respectively. B(u) is the set of web pages pointing to u , Nv represents the total numbers of outlinks of web page v and c is a factor used for normalization.
In Pagerank the score assigned to the outlinks of page p are in turn used to determine the ranks of pages to which p is pointing. Example of PageRank algorithm is illustrated in Figure 3.

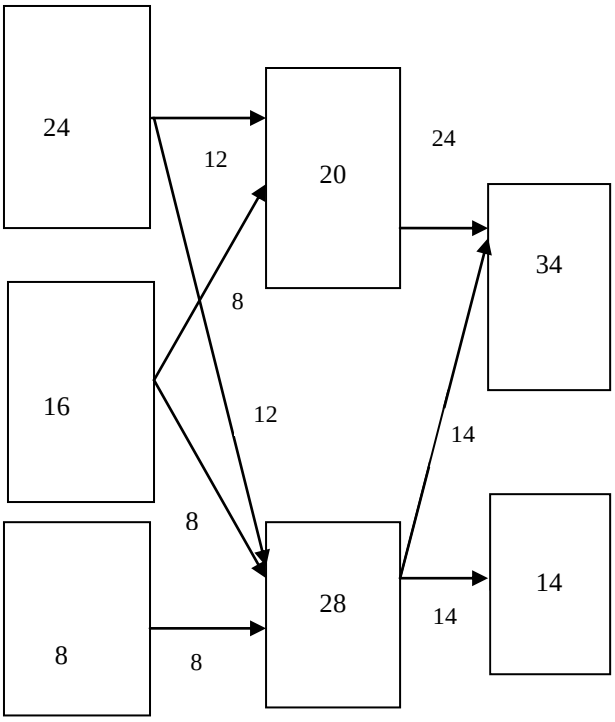

Figure 3

The PageRank formula has been modified and is given in equation (2).

$$PR(u) = (1-d) + d \sum_{v \in B(u)} \frac{PR(v)}{Nv} \qquad (2)$$

where d is the dampening factor. It represents the probability of user using direct links and the value of d can be set between 0 and 1.

1) Strengths of Pagerank

- The page rank score is generated using the entire web graph therefore the algorithm is lesser prone to localised link spams [12].
- PageRank algorithm is used by the most popular search engine i.e Google.
- It provides a very fast response to the user query.
- It has a good amount of efficiency associated with it as compared to other link based algorithms.

2) Limitations of PageRank

- It negotiates the content of the web since its completely based on the link structure of the web.
- In PageRank algorithm results are computed at index time rather than query time.
- In this algorithm search engines return results that are ordered according to their page rank but problem arises for polysemous words i.e words with several meanings [6].
- It favors old pages and provides them higher rank than the newer pages.

B. Weighted Page Rank

Wenpu Xing and Ali Ghorbani [9] proposed an extension to standard PageRank called Weighted PageRank (WPR). It works on the principle that more popular the web pages are, more linkages other web pages tend to have to them or are linked to by them. This algorithm assigns larger rank values to more important pages instead of distributing the rank value of a page evenly among its outgoing linked pages.

$W^{in}(m,n)$ and **$W^{out}(m,n)$** are the weight values of incoming and outgoing links respectively.

The weight of link (m,n) i.e $W^{in}(m,n)$ is calculated on the basis of number of incoming links to page n and the number of incoming links to all reference pages of page m. The formula is given in equation (3).

$$W^{in}(m,n) = \frac{In}{\sum_{p \in R(m)} Ip} \quad (3)$$

In is number of incoming links of page n, Ip is number of incoming links of page p, R(m) is the reference page list of page m.

On the other hand $W^{out}(m,n)$ is calculated on the basis of the number of outgoing links of page n and the number of outgoing links of all the reference pages of page m. the formula is given in equation (4).

$$W^{out}(m,n) = \frac{On}{\sum_{p \in R(m)} Op} \quad (4)$$

On is number of outgoing links of page n, Op is umber of outgoing links of page p. Therefore the weighted pagerank is given by following formula in equation (5).

$$WPR(n) = (1-d) + \sum_{m \in B(n)} WPR W^{in}(m,n) W^{out}(m,n) \quad (5)$$

1) Strengths of Weighted PageRank

- It takes into consideration both forward as well as backward links.

2) Limitations of Weighted PageRank

- Weighted PageRank algorithm is query independent.
- Weighted PageRank ignores relevancy which is the biggest limitation of this algorithm .

C. HITS

HITS stands for Hyperlink Induced Topic Search. This algorithm was proposed by Jon Klienberg. The webpages are ranked by analysing both their inlinks aswell as outlinks. It includes two forms of webpages Hubs and Authorities. Hubs are the pages that links to other important pages so they act as resource lists. Authorities are the pages containing important contents. A good hub page can be defined as a page which is having links pointing to many authoritative pages. Similarly, a good authority page is a page which is pointed by many good hub pages related to a given content. **[5].** The HITS algorithm treats WWW as directed graph G(V,E), where V is a set of vertices representing pages and E is set of edges corresponds to link. Figure 4 shows the hubs and authorities in web.

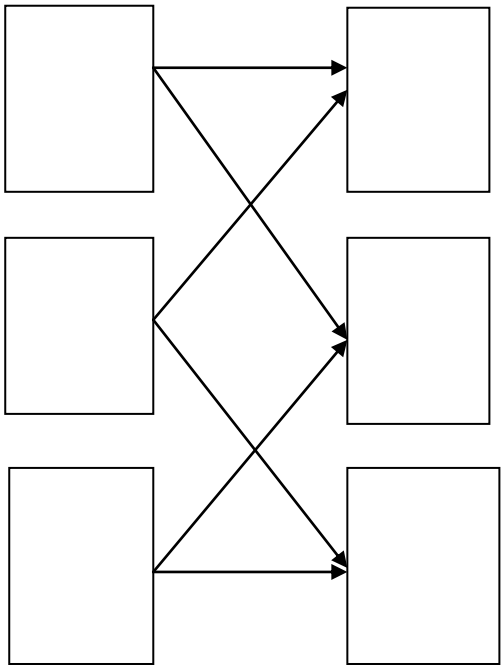

Figure 4.

HITS has two steps:

1. Sampling Step: In this step we collect a set of relevant pages for the given query.

2. Iterative Step: This step corresponds to finding out Hubs and Authorities pages. Following expressions given in equation 6 and equation 7 are used to calculate the weight of Hub (Hp) and the weight of Authority (Ap) respectively. [7] [8].

$$Hp = \sum_{q \in I(p)} Aq \quad (6)$$

$$Ap = \sum_{q \in B(p)} Hq \quad (7)$$

where Aq is the authority score of a page , Hq is defined as the hub score of a page and Ip and Bp are the set of reference pages of page p and set of referrer pages of page p respectively.

1) Strengths of HITS

- It provides relevant authorities and hubs.
- HITS algorithm is query dependent.
- With the use of hub and authority score HITS provides important pages.
- HITS can be combined with other ranking based information retrieval.

2) Limitations of HITS

- Many sites are hubs aswell as authorities which makes it difficult to distinguish between the both.
- It assumes that all links pointing to a page are of It assumes that all links pointing to a page are of equal weight and fails to recognize that some links might be more important than others.
- Topic drift often occurs in HITS.
- It is not efficient in real time.

IV. COMPARISON OF RANKING ALGORITHMS

A detailed comparison between PageRank , Weighted PageRank and HITS algorithm has been shown in Table 1 [10][11]. The comparison is done on the basis of various parameters.

TABLE 1

Algorithm	PageRank	Weighted Pagerank	HITS
Web Mining technique used	Web Structure mining	Web Structure mining	Web structue mining and Web content Mining
Methodology	It computes the score for the pages at the time of indexing of the pages.	Weight of web page is calculated on the basis of incoming and outgoing links both .	It computes the hubs and authorities of the relevant pages.
Input Parameters	Backlinks	Backlinks and Forward Links	Content, Backlinks and Forward links
Complexity	O(logN)	<O(logN)	<O(logN)
Formulae	PR (u) = (1-d) + d ∑ PR(v) / Nv, v€B(u)	WPR (n) = (1-d) +∑ in WPRW(m,n). out W(m,n)	Hp = ∑Aq, q€I(p); Ap = ∑Hq, q€B(p)
Search Engine	Google	Research Model	Clever
Quality of Results	Medium	Higher than PR	Lesser than PR
Limitations	Results computed at index time	Relevancy is ignored	Topic drift and efficiency problem.
Invention and Year	Larry Page & Sergey Brin in 1998	Wenpu Xing And Ali Ghorbani 2004	Jon Kleinberg, 1998

V. CONCLUSION

Different page ranking algorithms assign different rank scores to different web pages. In this paper we have discussed three important page ranking algorithms , their strengths and weaknesses and we have also compared these three algorithms based on different parameters. It has been concluded that both PageRank as well as and Weighted PageRank algorithm give importance to links rather than the contents of the web pages. On the other hand HITS given importance to both the links as well as the content of the pages. The algorithms mentioned above have certain limitations associated with them in terms of relevancy of results , importance of web pages , efficiency etc. Therefore a

new algorithm is needed that gives significant importance to relevancy and accuracy of results in order to improve the quality of search result.

REFERENCES

1. Mercy Paul Selvan , A .Chandra Sekar , A.Priya Dharshin, "Survey on Web Page Ranking Algorithms", International Journal of Computer Applications, vol 41, no.19, March 2012.
2. Seifedine Kadry and Ali Kalakech , " On the Improvement of Weighted Page Content Rank" , Journal of Advances in Computer Networks, vol. 1, no. 2, June 2013.
3. Govind Murari Upadhyay, Kanika Dhingra, "Web Content Mining: Its Techniques and Uses", IJARCSSE, vol 3, issue 11, November 2013.
4. S. Brin, and L. Page, "The Anatomy of a Large Scale Hypertextual Web Search Engine", Computer Network and ISDN Systems, Vol. 30, 1998 pp. 107-117 .
5. Neel am Duhan , A.K Sharma , "A Novel Approach for Organizing Web Search Results using Ranking and Clustering", vol 5 , August 2010.
6. Rekha Jain, Sulochana Nathawat, Dr. G.N. Purohit, "Enhanced Retrieval of Web Pages using Improved Page Rank Algorithm", International Journal on Natural Language Computing (IJNLC) vol. 2, no.2, April 2013.
7. Ashish Jain . Dr.Gireesh Dixit, "Intelligent Search Method (ISM): A method to efficiently search authoritative web pages", vol.3 , issue 12, December 2013.
8. Ashish Jain, Rajeev Sharma, Gireesh Dixit, Varsha Tomar, " Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages", International Conference on Communication Systems and Network Technologies, 2013.
9. Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proceedings of the second annual conference on Communication Networks and Services Research (CNSR˝04), IEEE , 2004.
10. Bhanudas Suresh Panchabhai And Marathe Dagadu Mitharam , "A Comparative Analysis Of Web Page RankingAlgorithms" , Indian Streams Research Journal , vol - 3, issue-10, Nov-2013.
11. Sonal Tuteja , " Enhancement in Weighted PageRank Algorithm Using VOL" , IOSR Journal of Computer Engineering , volume 14, issue 5,Sep. - Oct. 2013.
12. Pooja Devi, Ashlesha Gupta, Ashutosh Dixit, "Comparative Study of HITS and PageRank Link based Ranking Algorithms" , InternationalJournal of Advanced Research in Computer and Communication Engineering vol. 3, issue 2, Feb 2014.

