Page Ranking Algorithms for Web Mining: A Review

Charanjit Singh
Research Scholar
Guru Kasha University
Talwandi Sabo

S.K.Kautish, PhD
Professor & Dean Engineering
Gurur Kashi University
Talwandi Sabo

ABSTRACT

It is becoming very difficult for the web search engines to provide relevant information to the users with the growth of the WWW One of the Data Mining technique called Web mining is defined to extract the hidden information from the web documents and services. As per the information that is hidden, web mining can be divided into three different types: web content mining, web structure mining and web usage mining. The main application of web mining can be seen in the case of search engines. In order to rank their search results, they are using various page ranking algorithms that are either based on the content of the web pages or on the link structure of WWW. In this paper, a survey of page ranking algorithms based on both content and link structure of the web page and comparison of some important algorithms in context of performance has been carried out.

Keywords

WWW, Data mining, Web mining, Search engine, Page ranking.

1. INTRODUCTION

The World Wide Web is a popular segment of the Internet that contains billions of documents called Web pages includes documents can contain text, image, audio, video and metadata. With the rapid growth of information sources on the web world, it is becoming difficult to manage the information and satisfy the user needs. To retrieve the required information from the web matrix, numerous web search engines are used by the users. Some commonly used search engines are Google, msn, yahoo search etc.

A tool called Web Search engine is used to enable document search with respect to specified keywords, in the web and returns a list of documents where the keywords were found. Every search engine performs various tasks based on their respective architectures to provide relevant information to the users. Basic components of a web search engine are: Interface (user), Parser, Web Crawler, Database and Ranking Engine (see Fig. 1).

Web search engines work by sending out a spider or web crawler to visit and download all the web pages of the website and retrieve the information needed from them. Using the information gathered from the crawler, a search engine will then determine what the site is about and index the information. But before representing the pages to the user, search engine uses ranking algorithms in order to sort the results to be displayed. That way user will have the most important and useful results first.

In this review paper, various content based and link based page ranking algorithms has been done and a comparison is carried out. This paper is divided into different sections: in section 1, first introduce the concept of web search engines and explain its working. In section 2, present the web mining concepts, categories and technologies. As shown in section 3, present the detailed overview of some page ranking algorithms and section 4, includes the comparison of these algorithms in context of performance. Finally in section 5, conclude this review and discuss some future directions for the system.

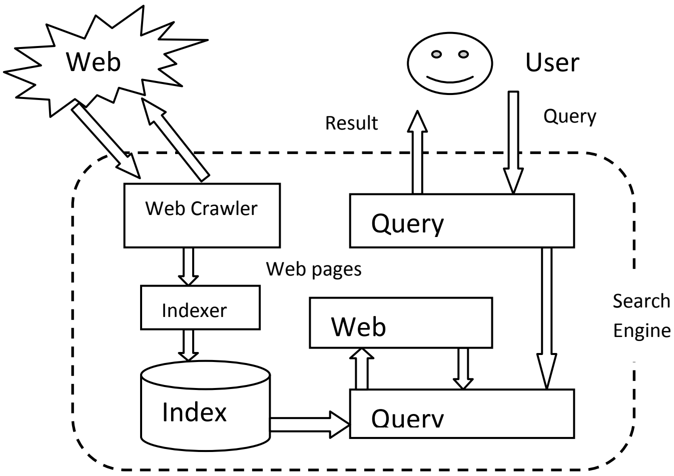

Figure 1: Search Engine Architecture

2. WEB MINING

Application of data mining technique such as web mining is used to discover automatically and take out information from Web data. Web mining data can be:

- Further Web Mining can be divided Web Content data- text, images, records, etc
- Web Structure data- hyperlinks, tags etc
- Web Usage data- http logs, app server logs, etc

into three categories [1] namely web content mining, web structure mining and web usage mining as shown in Fig. 2

2.1 Web Content Mining (WCM):

WCM technique is used to extract meaningful information from the web contents of web pages or documents. Content data, means to the collection of facts a web page was designed to convey to the users. It consists of text, images, audio, video, or structured records such as tables and lists which results search engine gives web pages itself or on the result pages. Web content mining is further categories in two methods such as web page content mining and search results mining.

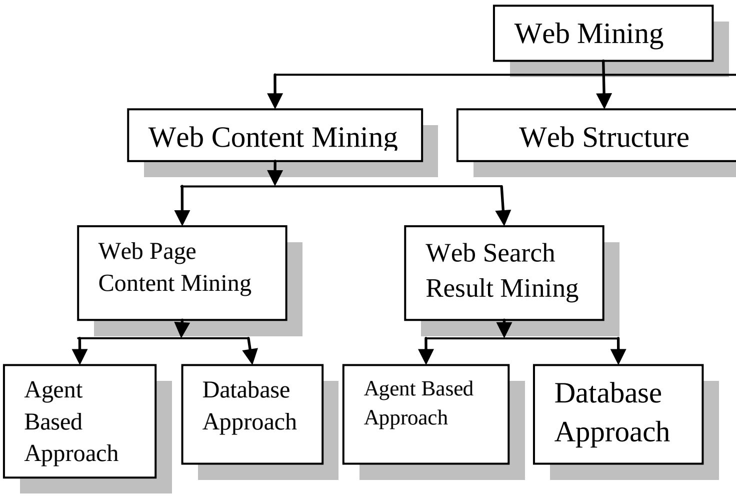

Figure 2: Taxonomy of Web Mining

Traditional searching called web page content mining of web pages using content and search result mining is a further search of pages found in previous search.

2.2 Web Structure Mining (WSM):

Using graph theory and Web Structure Mining (WSM) discover structure information from the web. Web structure mining can be performing either at the hyperlink (inter-page) or at document level (intra-page). The structure of a typical web graph consists of web pages as nodes of graph, and hyperlinks as edges of graph connecting between two related pages.

2.3 Web Usage Mining (WUM):

To find out the important patterns from data, generated by client-server transactions on one or more web localities web content mining is used. Web usage mining can be further divided in discovery the general access patterns or in discovery the patterns matching the particular parameters.

Site improvement and modification, Business intelligence, web personalization, ranking of pages are application area of alleged mentioned categories of web mining. A variety of ranking algorithms are used by search engine to sort the results to be displayed and to present appropriate information to the users to cater to their needs. There are a range of ranking algorithms developed; the minority of them has been discussed in the next section: Page Rank, Weighted Page Rank, HITS and SimRank.

3. PAGE RANKING ALGORITHMS

With the swift development of network techniques, huge information resources glut the whole web world. Web search engine is increasingly becoming the leading information retrieving approach.

Relevant information provides to the users is the primary goal of search engines. As a result, various Page Ranking Algorithms are used to rank the query results of web pages in an effective and efficient fashion.

Some algorithms rely only on the link structure of the document. i.e their popularity scores (web structure mining), whereas others look for the content in the documents (web content mining), while some use a combination of both i.e they use link as well as content of the document to assign a rank value to the concerned document. Some commonly used page ranking algorithms have been discussed as follows:

3.1 Page Rank Algorithm

Page Rank Algorithm was purposed by Surgey Brin and Larry Page [4, 5]. Page Rank was named after Larry Page also was cofounder of Google search engine. Usually used by the Google [6] web search engine to rank websites in their search engine results. Page Rank is used to measure the importance of website pages by counting the number and quality of links to a page.

This algorithm states that the Page Rank of a page is defined recursively and depends on the number and Page Rank metric of all pages that link to it (incoming links). If a page has some important incoming links to it than its outgoing links to other pages also become important. A page that is linked to by many pages with high Page Rank receives a high rank itself.

A Page Rank Algorithm considers more than 25 billion web pages on the www to assign a rank score [6]. A simplified version [4] of Page Rank is defined in Eq.1:

$$PR(u) = C \sum_{V \in B(u)} PR(v) / N_v \tag{1}$$

here ‘u’ represents a web page, B(u) is the set of pages that points to u, PR(u) and PR(v) are rank scores of pages u and v respectively, Nv denotes the number of outgoing links of pages v, C is a factor used for normalization. In Page Rank, the rank score of a page, p, is evenly divided among its outgoing links. The values assigned to the outgoing links of page p are in turn used to calculate the ranks of the pages to which page p is pointing as shown in Fig. 3

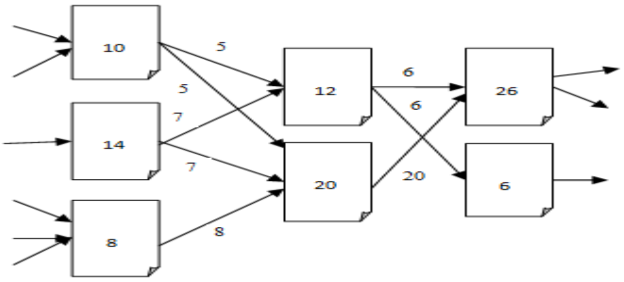

Figure 3: Distribution of page ranks

Later algorithm was modified, observing that not all users follow the direct links on WWW. The modified version is given in Eq. 2:

$$PR(u) = (1-d) + d \sum_{V \in B(u)} PR(v) / N_v \qquad (2)$$

Here'd' is a damping factor that is usually set to 0.85 and it can be thought of as the probability of users' following the links and (1-d) as the page rank distribution from non-directly linked pages.

3.2 Weighted Page Rank Algorithm

This algorithm was proposed by Wenpu Xing and Ali Ghorbani [9] which is an extension of PageRank algorithm. Algorithm assigns rank values to pages according to their importance or popularity rather than dividing it evenly. The popularity is assigned in terms of weight values to incoming and outgoing links and are denoted as Win(v, u) and Wout (v, u) respectively. Win(v, u) is the weight of link (v,u) calculated on the basis of incoming links to page u and the number of incoming links to all reference (outgoing linked) pages of page v.

$$W^{in}_{(v,u)} = I_u / \sum_{p \in R(v)} I_P \qquad (3)$$

where Iu and Ip represent the number of incoming links of page u and page p, R(v) is the reference page list of page v. Wout(v,u) is the weight of link (v,u) calculated on the basis of the number of outgoing links of page u and the number of outgoing links of all the reference pages of page v.

$$W^{out}_{(v,u)} = O_u / \sum_{p \in R(v)} O_P \qquad (4)$$

Here Ou and Op represents, the number of outgoing links of page u and page p, respectively. Then the weighted Page Rank is given by a formula:

$$WPR(u) = (1-d) + d \sum_{V \in B(u)} WPR(v) W^{in}_{(v,u)} W^{out}_{(v,u)} \qquad (5)$$

3.3 HITS

This algorithm was developed by Jon Kleinberg [7] called Hyperlink- Induced Topic Search (HITS) [8] which gives two forms of web pages called as hubs and authorities. Hubs are the pages that act as resource lists. Authorities are having important contents.

A fine hub page for a subject points to many authoritative pages on that context and a good authority page is pointed by many fine hub pages on the same subject. HITS assumes that if the author of page p provides a link to page q, then p confers some authority on page q. Kleinberg states that a page may be a good hub and a good authority at the same time.

The HITS algorithm considers the WWW as a directed graph G(V,E) where V is a set of vertices representing pages and E is a set of edges that match upto links. Fig. 4 shows the hubs and authorities in web.

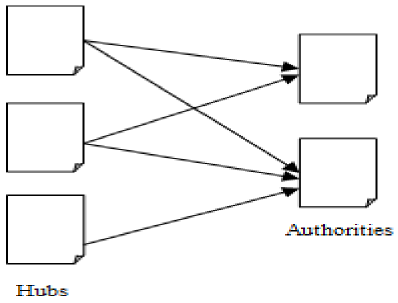

Figure 4: Hubs and Authorities

3.1.1 The HITS algorithm works in two major steps:

i. **Sampling step:** In this step, a set of relevant pages for the given query are collected.

ii. **Iterative step:** This step finds hubs and authorities using the output of sampling step. The scores of hubs and authorities are calculated as follows:

$$H_p = \sum_{q \in I(p)} A_q \qquad (6)$$

$$A_p = \sum_{q \in B(p)} H_q \qquad (7)$$

where Hq and Aq represents the Hub score and authority score of a page. I(p) and B(p) denotes the set of reference and referrer pages of page p. the page's authority weight is proportional to the sum of the hub weights of pages that it links to.

3.1.2 Constraints with HITS algorithm

The following are the constraints of HITS algorithms:

i. **Hubs and Authorities:** It is not simple to distinguish between hubs and authorities since many sites are hubs as well as authorities.

ii. **Topic drift:** Sometimes HITS may not produce the most relevant documents to the users queries because of equivalent weights.

iii. **Automatically generated links:** Some links are automatically generated and represent no human judgment, but HITS gives them equal importance.

iv. **Efficiency:** The performance of HITS algorithm is not efficient in real time.

HITS was used in a prototype search engine called Clever for an IBM research project. Because of the above constraints, HITS could not be implement in a real time search engine.

3.4 SimRank

A new page rank algorithm which is based on similarity measure from the vector space model, called SimRank [10]. In order to rank the query results of web pages in an effective and efficient manner, SimRank is used. Normally, traditional Page Rank algorithm only employ the link relations among pages to compute the rank of each page but the content of each page cannot be ignored completely. Actually, the accuracy of page scoring greatly depends on the content of the page. Therefore, SimRank algorithm is used to provide the most relevant information to the users. To calculate the score of web pages in

SimRank , a page in vector space model is represented as a weight vector, in which each component weight is computed based on some variation of TF (Term Frequency) ot TF-IDF (Inverse Document Frequency) scheme as follows:

3.4.1 TF scheme:

In TF scheme, the weight of a term ti in page dj is the number of times that ti appears in document dj, denoted as fij. The following normalization approach is applied [4]

$$tf_{ij} = \frac{f_{ij}}{\max\{ f_{1j}, f_{2j}, \ldots\ldots. f_{|V|j}} \quad (8)$$

Where fij is the frequency count of term ti in page j and |V| is the number of terms in page. The disadvantage of this scheme is that it does not consider the case that a term appears in several pages, which limits its application.

3.4.2 TF-IDF scheme:

The inverse document frequency (denoted by idfi) of term ti is computed by [4].

$$idf_i = \log \frac{N}{df_i} \quad (9)$$

where N is the total no of pages in a web database, dfi is the number of pages in which term ti appears atleast once, and fij is the frequency count of term ti in page dj. The term weight is computed by:

$$W_{ij} = tf_{ij} \times idf_i \quad (10)$$

Note that the TF-IDF scheme is based on the intuition that if a term appears in several pages, it is not important. SimRank algorithm is based on the similarity measure for computing the rank of each page. The main content of a crawled page contains two parts: title and body. The SimRank algorithm works on two distinct weight values that are assigned to the title and body of a page, respectively. The formula for calculating the SimRank is as follows [simrank paper]:

$$SimRank(p_j) = tconst * W_{ij}^{title} + bconst * W_{ij}^{body} \quad (11)$$

Where pj, could be denoted as (w1j, w2j,…….,wmj), Wij is the term weight, ‘t const’ and ‘b const’ are some constants between 0.1 to 1.

4. COMPARISON OF VARIOUS ALGORITHMS

On the basis of literature analysis, a comparison of certain Web Page Ranking Algorithms is shown in Table 1. The comparison is performed on the basis of some vaults such as Mining technique use, Methodology, Input parameters, Relevancy, Working levels, Quality of results, Importance and Limitations. On the basis of these parameters, we can check the performance of each algorithm.

5. CONCLUSION

An application of web mining called Page Ranking Algorithms, play an important role in making the user navigation easier in the results of a search engine. Paper described proposed algorithms like Page Rank algorithm, Weighted Page Rank algorithm, HITS, SimRank, etc. and all algorithms are able to provide satisfactory results in required cases. In some cases it is not able not provide respective answer due to considering or calculating ranks only, several considering content only and certain consider only links relation of respective web pages.

6. FUTURE SCOPE

A new technique can be proposed that will consider more than one aspect such as Ranks, Web Content, links of pages to illustrate the more accurate query result in search engines

7. REFERENCES

[1] Kaur,M., Singh,C., 2014. Content based and Link based page ranking algorithm: A Survey. International Journal of Advanced and Innovative Research (IJAIR), ISSN: 2278-7844,Vol 3, Issue – 4, pp. 250-255.

[2] R.Cooley, B.Mobasher and J.Srivastava, 1997. Web Mining: Information and Pattern Discovery on the World Wide Web". In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, (ICTAI'97).

[3] Dr. M. H. Dunham, 2002 Data Mining:Introductory and Advanced Topics, Prentice Hall.

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

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

[6] Kleinberg J., 1998. Authorative Sources in a Hyperlinked Environment". Proceedings of the 23rd annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[7] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, 2001. Link Analysis: Hubs and Authorities on the World. Technical report: 47847.

[8] Bing Liu., 2006. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications). Springer-Verlag NewYork, Inc., Secaucus, NJ, USA.

[9] Lizorkin, D., Velikhov, P., Grinev, M., Turdakov, D., 2008 Accuracy estimate and optimization Techniques for Simrank Computation. Published in ACM, Print ISBN No: 978-1-60558-305-1, on 24-30 Aug 2008, pp. 422-433.

[10] Li, C., Han, J., He, G., Jin, X., Sun, Y., Yu, Y., Wu, T., 2010. Fast Computation of SimRank for Static and Dynamic Information Networks. Published in ACM, Print ISBN No: 978-1-60558-9045-9, on 22-26 March 2010.

Table 1. Comparison of Page Ranking Algorithm

Algorithm	**Page Rank**	**Weighted Page Rank**	**SimRank**	**HITS**
Mining Technique used	Web Structure Mining	Web Structure Mining	Web Content Mining	Web Structure Mining, Web Content Mining
Description	Computes scores at indexing time not ay query time. Results are sorted according to the importance of pages.	Computes scores at indexing time, unequal distribution of score, pages are sorted according to importance.	Computes scores at query time. Results are calculated dynamically.	Computes hub and authority scores of n highly relevant pages on the fly.
I/P Parameters	Backlinks	Backlinks, forward links	Content	Backlinks, forward links, content
Working levels	N^*	1	1	<N
Relevancy	Less	Less (higher than PR)	More	More
Importance	More	More	Less	Less
Quality of results	Medium	Higher than PR	Approx equal to WPR	Less than PR
Limitations	Query Independent	Query Independent	Importance of page links is totally ignored	Topic drift and efficiency problems