Content Based and Link Based Page Ranking Algorithms: A Survey

Madhurdeep Kaur[1], Asst. Prof. Charanjit Singh[2]
[1]Research Scholar (Department of Computer Science), Rimt-Iet, Mandi Gobingarh
[2]Asst. Professor (CSE Dept), Rimt-Iet, Mandi Gobindgarh
[1]madhur8970@gmail.com
[2]sehgal_cs@yahoo.com

Abstract— **With the rapid growth of World Wide Web (WWW), it is becoming very difficult for the web search engines to provide relevant information to the users. Web mining is defined as the application of data mining techniques to extract the hidden information from the web documents and services. According to this hidden information, 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**

I. INTRODUCTION

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

Web Search engine is a tool enabling 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 number of tasks based on their respective architectures to provide relevant information to the users. Basic components of a web search engine are: User Interface, 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.

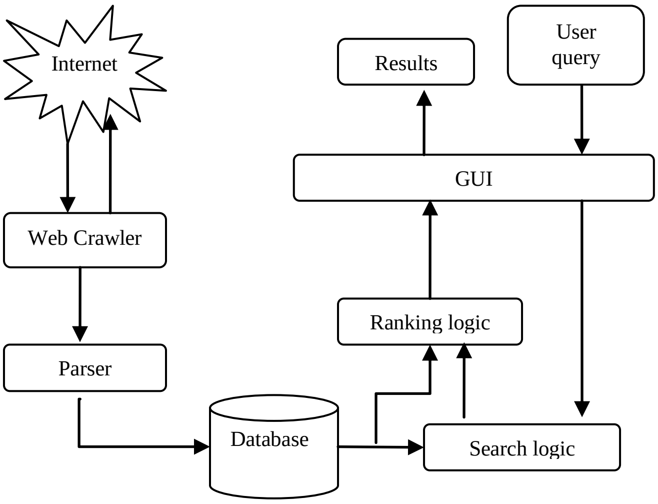

Figure 1 Architecture of a Search Engine

In this paper, a survey of 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, we first introduce the concept of web search engines and explain its working. In section 2, we present the web mining concepts, categories and technologies. As shown in section 3, we provide 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, we conclude this paper and discuss some future directions for the system.

II. WEB MINING

Web mining is the application of data mining techniques to automatically discover and extract information from Web data. Web data can be:

- Web Content- text, images, records, etc
- Web Structure- hyperlinks, tags etc
- Web Usage- http logs, app server logs, etc

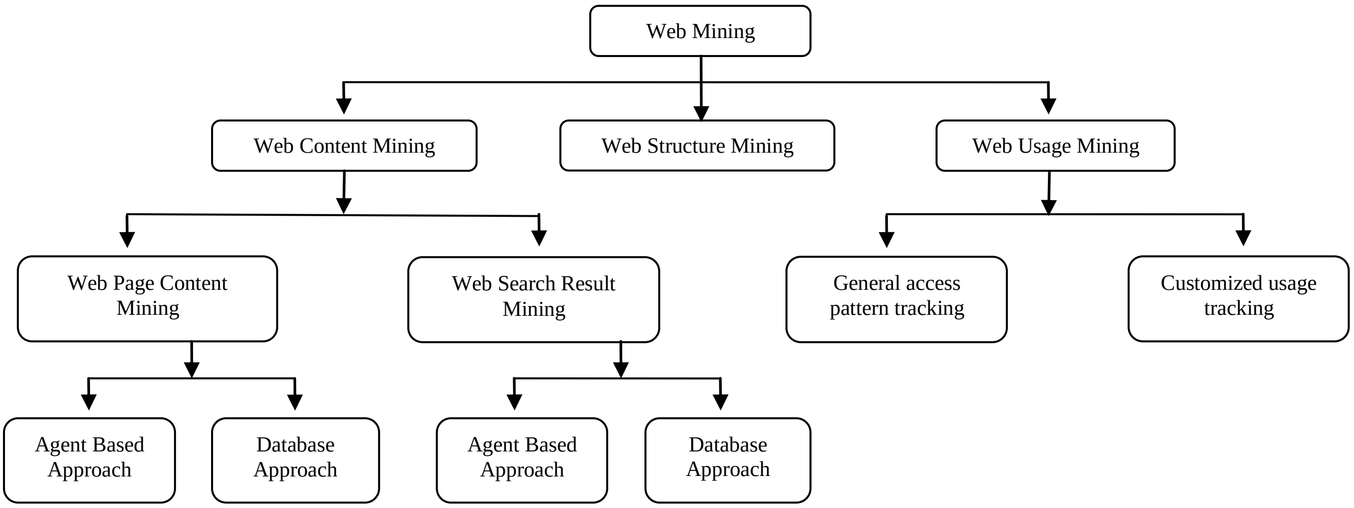

Figure 2. Taxonomy of Web Mining

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

Web Content Mining (WCM) is the process of extracting useful 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. It may consist of text, images, audio, video, or structured records such as lists and tables. It can be applied on web pages itself or on the result pages obtained from a search engine. Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). Web content mining is further divide into Web page content mining and Search results mining. Web page content mining is traditional searching of web pages with the help of content while search result mining is a further search of pages found in previous search.

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

Web Usage Mining (WUM) is used to discover the meaningful patterns from data generated by client-server transactions on one or more web localities. It can be further categorized in finding the general access patterns or in finding the patterns matching the specified parameters.

All the above mentioned categories of Web Mining have its own application areas including business intelligence, site improvement and modification, web personalization, ranking of pages etc. Search engine uses ranking algorithms in order to sort the results to be displayed and to provide relevant information to the users to cater to their needs. There are various ranking algorithms developed, few of them have been discussed in the next section: Page Rank, Weighted Page Rank, SimRank and HITS [3, 4, 5, 7, 8, 9, 15]

III. PAGE RANKING ALGORITHMS

With the rapid development of network techniques, huge information resources glut the whole web world. Web search engine is increasingly becoming the dominant information retrieving approach. The primary goal of these search engines is to provide the relevant information to the users. Therefore, 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:

A. Page Rank Algorithm

Page Rank Algorithm was developed by Surgey Brin and Larry Page [4, 5]. Page Rank was named after Larry Page (cofounder of Google search engine). It is 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:

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

where 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, N_v 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

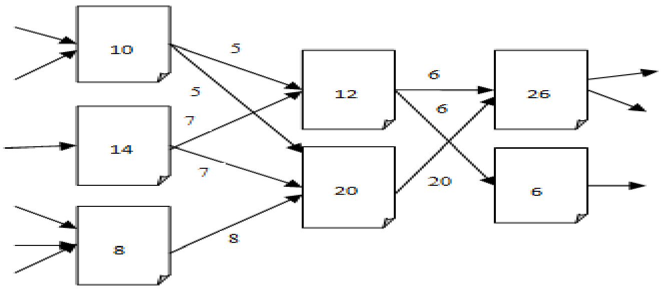

Figure 3 Distribution of page ranks

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

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

where d is a damping factor that is usually set to 0.85. d 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.

B. *Weighted PageRank Algorithm*

This algorithm was proposed by Wenpu Xing and Ali Ghorbani [9] which is an extension of PageRank algorithm. This 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 $W^{in}(v, u)$ and $W^{out}(v, u)$ respectively. $W^{in}(v, u)$ 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.

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

where I_u and I_p represent the number of incoming links of page u and page p, R(v) is the reference page list of page v. $W^{out}(v,u)$ 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.

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

where O_u and O_p represents the number of outgoing links of page u and page p, respectively. Then the weighted Page Rank is given by a formula:

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

C. *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.

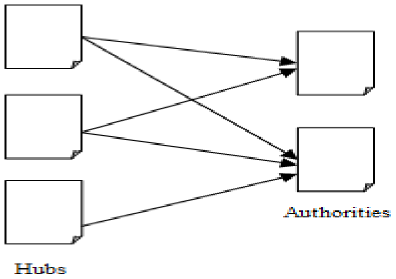

Figure 4 Hubs and Authorities

The HITS algorithm works in two major steps:

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

- **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 H_q and A_q 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.

Constraints with HITS algorithm

The following are the constraints of HITS algorithms:

- Hubs and Authorities: It is not simple to distinguish between hubs and authorities since many sites are hubs as well as authorities.
- Topic drift: Sometimes HITS may not produce the most relevant documents to the users queries because of equivalent weights.
- Automatically generated links: Some links are automatically generated and represent no human judgment, but HITS gives them equal importance.
- 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 implemented in a real time search engine.

D. SimRank

A new page rank algorithm which is based on similarity measure from the vector space model, called SimRank [15]. In order to rank the query results of web pages in an effective and efficient manner, SimRank is used.

Generally, 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. 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 [4]

- **TF scheme:** In TF scheme, the weight of a term t_i in page d_j is the number of times that t_i appears in document d_j, denoted as f_{ij}. The following normalization approach is applied [4]

$$tf_{ij} = \frac{f_{ij}}{\max\{f_{1j}, f_{2j}, \dots, f_{|V|j}\}} \qquad (8)$$

where f_{ij} is the frequency count of term t_i 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.

- **TF-IDF scheme:** The inverse document frequency (denoted by idf_i) of term t_i is computed by [4].

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

where N is the total no of pages in a web database, df_i is the number of pages in which term t_i appears atleast once, and f_{ij} is the frequency count of term t_i in page d_j. The term weight is computed by:

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

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]:

$$SimRank(p_j) = t\ const * W_{ij}^{title} + b\ const * W_{ij}^{body} \qquad (11)$$

where p_j could be denoted as $(w_{1j}, w_{2j},..,w_{mj})$, W_{ij} is the term weight, t const and b const are some constants between 0.1 to 1.

IV. COMPARISON OF VARIOUS ALGORITHMS

On the basis of literature analysis, a comparison of various web page ranking algorithm is shown in Table1. 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.

V. CONCLUSION

The Page Ranking Algorithms, which are an application of web mining, play an important role in making the user navigation easier in the results of a search engine. This paper described several proposed algorithms like Page Rank algorithm, Weighted Page Rank algorithm, SimRank , HITS, etc. all algorithms may provide satisfactory performance in some cases but many times the user may not get the relevant information. This is because some algorithms calculate the rank by considering only the content of web page but others compute it by focusing on link relations among pages. Therefore, a new technique can be proposed that will consider both content and link relation of a web page.

ACKNOWLEDGEMENT

I am extremely grateful and remain indebted to all the people who have given their intellectual support throughout the course of this work. And a special acknowledgement to the authors of various research papers and books which help me a lot.

TABLE I. COMPARISON OF PAGE RANKING ALGORIHMS

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
Quality of results	Medium	Higher than PR	Approx equal to WPR	Less than PR
Importance	More	More	Less	Less
Limitations	Query Independent	Query Independent	Importance of page links is totally ignored	Topic drift and efficiency problems

*n: number of pages chosen by the algorithm, N: number of web pages , m: Total number of occurrences of query terms in n pages

REFERENCES

[1] R.Cooley, B.Mobasher and J.Srivastava, "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), 1997.

[2] Companion slides for the text by Dr. M. H. Dunham, "Data Mining:Introductory and Advanced Topics", Prentice Hall, 2002

[3] Jaroslav Pokorny, Jozef Smizansky, "Page Content Rank: An Approachto the Web Content Mining".

[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] C. Ridings and M. Shishigin, "Pagerank Uncovered". Technical report,2002.

[6] http://WWW.webrankinfo.com/english/seo-news/topic-16388.htm. January 2006, Increased Google index size.

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

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

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

[10] http://www.google.com/technology/index.html, Our Search: Google Technology.

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

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

[13] Lizorkin, D., Velikhov, P., Grinev, M., Turdakov, D., "*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.

[14] Li, C., Han, J., He, G., Jin, X., Sun, Y., Yu, Y., Wu, T., "*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.

[15] Qiao, S., Li, T., Li, H., Zhu, Y., Peng, J., Qin, J., "*SimRank : A Page Rank Approach based on Similarity Measure*", Published in IEEE, Print ISBN No: 978-1-4244 -6793-8, 2010, pp. 390-395.

[16] Taneja, H., Gupta, R., "*Web Information Retrieval using Query Independent Page Rank Algorithm*", International Conference on Advances in Computer Engineering, Published in IEEE, Print ISBN No: 978-0-7695-4058-0, 2010, pp. 178-182.

[17] Ma, H., Chen, S., WANG, D., "*Research of PageRank Algorithm Based on Transition Probability",* International Conference on Web Information Systems and Mining, Published in IEEE, Print ISBN No: 978-0-7695-4224-9, 2010, pp. 153-155.

[18] Cailan, Z., Kai, C., Shasha, Li., "*Improved PageRank Algorithm Based on Feedback of User Clicks*", Published in IEEE, Print ISBN No: 978-1-4244-9763-8,2011,pp. 3949-3952.

[19] Kumar, G., Duhan, N., Sharma, A.K., "*Page Ranking Based on number of Visits of Links of Web Page*", International Conference on Computer

& Communication Technology (ICCCT), Published in IEEE, Print ISBN No: 978-1-4577-1386-6,2011, pp. 11-14.

[20] Zhao, C., Zhang, Z., Li, H., Xie, X., "*A Search Result Ranking Algorithm Based on Web Pages and Tags Clustering*", Published in IEEE, Print ISBN No: 978-1-4244-8728-8, 2011, pp.609-614.

[21] Sharma, R., Kandpal, A., Bhakuni, P., Chauhan, R., Goudar, R.H., Tyagi, A., "*Web Page Indexing through Page ranking for Effective Semantic Search*", 7th International Conference on Intelligent Systems and Control (ISCO), Published in IEEE, Print ISBN No: 978-1-4673-4603-0, 2012.

[22] Jain, A., Sharma, R., Dixit, G., Tomar, V., "*Page Ranking Algorithm in Web Mining, Limitations of existing methods and a new method for Indexing Web Pages*", Published in IEEE, Print ISBN No: 978-0-7695-4958-3,2013, pp. 640-645