ISSN No. 2278-3091

Volume 1, No.5, November – December 2012
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://warse.org/pdfs/ijatcse02152012.pdf

Canvassing Ranking Algorithms

Rupal Bhargava[1], Rekha Jain[2]
[1]Banasthali Vidyapith, Jaipur , bhargava.rupal@gmail.com
[2]Banasthali Vidyapith, Jaipur, rekha_leo2003@gmail.com

ABSTRACT

With the tremendous growth and increasing demand of information on web it has become quite necessary to satisfy the user demand, up to the level of his/ her expectation. User always expects to get the most relevant results, which, with such complex structure and varying queries becomes hard to provide for a Search Engine. Hence different Ranking algorithms are used in different Search Engines to deal with such problems. This paper deals with the web mining, web mining taxonomy and different ranking algorithms used to satisfy user's demands in Information Retrieval. A comparative analysis of few algorithms like Page Rank Algorithm, HITS (Hypertext Induced Topic Selection algorithm), Weighted Page Rank algorithm, Distance Rank algorithm are given. Besides this some other proposed algorithms like Weighted Page Content Rank and Improvised Page Rank algorithm, Weighted Page Rank Algorithm Based on number of Visits of Links of Web Page and Weighted Page Rank algorithm using link attributes is also explained.

Keywords: HITS, Page Rank, Distance rank, WPR

1. INTRODUCTION

The World Wide Web is the universe of network-accessible information, the source of human learning. It is the most potential source of information and communication now days. Today whether it be any field, WWW is the prime knowledge source. It has so embedded in our lives that we can't think of surviving without it. It has become a need for humans which they depend on as it is a largest and most popular repository of information. Also it is a rapidly intensifying system of interlinked hypertext documents. Day by day the information keeps piling on in this massive structure. Hence it becomes necessary to structure this diverse and dynamic unstructured storage of data. For the purpose mentioned it is important to understand and analyze the underlying data structure of web for effective and efficient information extraction with the increasing demand of users. Hence it has become necessary for the search engines to give most specific and user need satisfying results. There are a lot of search engines but few like Google, Yahoo, etc. are famous because of their crawling and ranking methodology. Every day they solve and satisfy millions of queries. So, Ranking methodology becomes a very important aspect of web mining in all the three components of search engine (i.e. Crawler, Indexer, Ranking mechanism). Figure 1 shows the sample architecture *(David Hawkin et al. 2006)* of Search Engine that comprises of various components like Ranker, Indexer, Query Builder, Presenter etc.

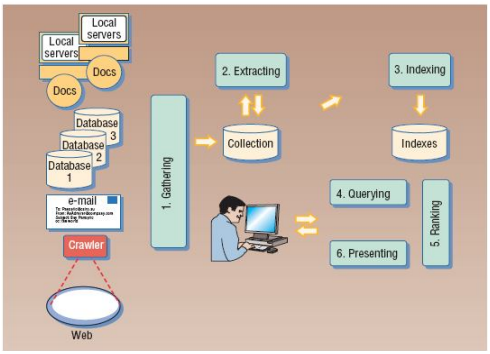

Figure 1: Sample architecture of search engine

2. WEB MINING

Web mining is a data mining technique used to extract information from World Wide Web. Also we can say that it is process of taking out knowledge from web. The absolute process of extracting knowledge from Web data *(Neelam Duhan et al. 2009)* is given in Figure 2:

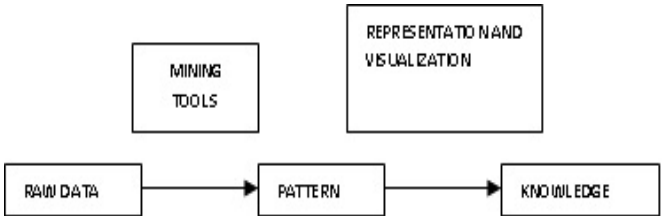

Figure 2: Process of Web Mining

3. WEB TAXANOMY

According to the usage web data, Web Mining can be categorized *(Cooley, R.et al. 1997)* into three categories namely Web Content Mining (WCM), Web Usage Mining (WUM), and Web Structure Mining (WSM) as shown in Figure 3. A comparative analysis is given by *(R. Kosala et al. 2000)* which is summarized in Table 1 as below:

:

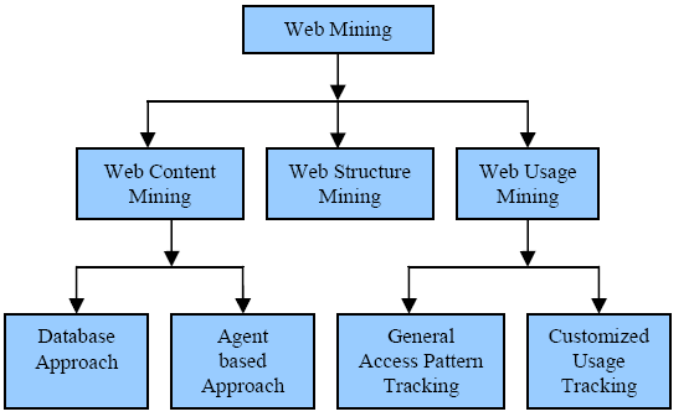

Figure 3: Web Mining Categories

Table 1: Comparative Analysis of Web Mining Categories

Web Mining				
	Web Content Mining		**Web Structure Mining**	**Web Usage Mining**
	IR View	**DB View**		
View of Data	• Unstructured • Structured	• Semi Structured • Website as DB	Link Structure	Interactivity
Main Data	• Text documents • Hypertext documents	Hypertext documents	Link Structure	• Server Logs • Browser Logs
Representation	• Bag of words, n-gram Terms, • Phrases, Concepts or ontology • Relational	• Edge labeled Graph, • Relational	• Graph	• Relational Table • Graph
Method	• Machine Learning • Statistical (including NLP)	• Proprietary algorithms • Association rules	• Proprietary • algorithms	• Machine Learning • Statistical • Association rules
Application Categories	• Categorization • Clustering • Finding extract rules • Finding patterns in text	• Finding frequent sub structures • Web site schema • discovery	• Categorization • Clustering	• Site Construction • Adaptation and • management • Marketing, • User Modeling

4. RANKING ALGORITHMS

Today with the rising demand of information on web, search engines have to adopt different techniques to prioritize different web pages. It has been a great deal of work to rank pages such that it gives user most appropriate results according to its requirement. To make it happen various algorithms have been designed and introduced with different perspective. Some algorithms use link structure of web pages whereas other use content to define relevancy of web pages to user queries. Here are some ranking algorithms discussed with their varying nature of web mining category, working, and input parameters etc.

4.1. Page Rank Algorithm

To rank web pages with their popularity, this algorithm uses number of pages that points to it, also known as in degree algorithm (since it ranks web pages according to their in degree). This concept was used and enhanced by *(S. Brin et al. 1998&1999)* during their PhD at Stanford University. This algorithm is used in most famous search engine 'Google' named as Page Rank Algorithm. It uses concept of citation analysis and treats incoming links as citations. But as only citation analysis was not giving efficient and relevant result, *(S. Brin et al. 1998&1999)* added a concept to citation analysis such that a link coming from an important page was given high weight whereas page which was not so important was given a low weight. Also they assumed links as votes.

Not only the total numbers of votes were important but relevancy and popularity of page casting vote was also considered.

(S. Brin et al. 1998&1999) proposed a formula to calculate Page rank of a Page 'A' where T1, T2…Tn are pages pointing to it. Formula is as follows:

$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right) \quad (1)$$

Where,

d, damping factor (whose value is generally 0.85).It is used to stop other pages having too much influence)

C (T_i), number of links going out of T_i

PR (T_i), Page rank of Page T_i

Page Rank forms probability distribution such that sum of page rank of all web pages will be 1. Page rank uses an iterative approach to calculate actual page ranks of web pages starting with page rank 1 of all web pages. Also it corresponds to Principal Eigen vector of normalized link matrix of web.

4.2. Weighted Page Rank Algorithm

Weighted page rank is an improvised or extended version of Page Rank. It divided the weight according to the importance of page rather than simply dividing rank value evenly among outgoing links. More the page is important, higher rank value it gets.

According to *(W. Xing et al. 2004)* Popularity of pages is calculated using Weight of in links ($W^{in}_{(v,\ u)}$) and out links ($W^{out}_{(v,u)}$)

$W^{in}_{(v,\ u)}$ is the weight of link (v,u) calculated based on the number of in links of page u and page p, respectively. R(v) denotes the reference page list of page v.

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{p \epsilon R(v)} I_p} \quad (2)$$

Where I_u and I_p represent the number of in links of page u and page p, respectively. R(v) denotes the reference page list of page v.

$W^{out}_{(v,u)}$ is the weight of link(v,u) calculated based on the number of out links of page u and the number of out links of all reference pages of page v.

$$W^{out}_{(v,u)} = \frac{O_u}{\sum_{p \epsilon R(v)} O_p} \quad (3)$$

Where O_u and O_p represent number of out links of page u and page p, respectively. R(v) denotes the reference page list of page v.

Also by *(W. Xing et al. 2004)*, Modified formula of Page Rank for Weighted Page rank is

$$PR(u) = (1-d) + d \sum PR(v) W^{in}_{(v,u)} W^{out}_{(v,u)} \quad (4)$$

4.3. Hypertext Induced Topic Selection (HITS)

HITS was used in research based search engine of IBM called CLEVER. But was not implemented because of its constraints. *(J. Kleinberg et al. 1999)* introduced two very important terms used in this algorithm, Hub and Authority. A good hub is one which links to may authority pages containing content of the query. Similarly, a good authority is one which is being pointed by too many good hubs having the same subject.

HITS has two major stages, Sampling and Iteration. In sampling stage a set of relevant pages for the query are obtained starting from the root set R, a set S is obtained such that it is relatively smaller than R and contains a large amount of good authority pages. Whereas in iterative stage, it finds hubs and authorities using eq. given by *(J. Kleinberg et al. 1999).*

$$H_p = \sum_{q \epsilon I(p)} A_q \quad (5)$$

$$A_p = \sum_{q \epsilon B(p)} H_q \quad (6)$$

Where Hp is hub weight of p

A_p is authority weight of p

I (p) is set of reference pages

B (p) is set of reference pages

4.4. Distance Rank

(Ali Mohammad Zareh Bidoki et al. 2007), proposed Distance Rank algorithm which is based on reinforcement learning. In this algorithm distance is considered as punishment and we try to minimize this distance. *(Ali Mohammad Zareh Bidoki et al. 2007)* considers distance between two pages i &j as logarithm of the number of i's o/p link when I points to j. This algorithm is for random surfers. The learning rate of surfer is used to model behavior of user in each state. Distance rank converges to static value by recursively iterating and then sorting the vector obtained in descending order. Page with low distance get high rank.

4.5. Weighted Page Content Rank

To resolve the problems faced in Page rank algorithm and Weighted Page Rank algorithm *(Pooja Sharma et al. 2010)* proposed a new algorithm Weighted Page Rank algorithm which implies both Web Structure Mining as well as Web Content Mining Techniques to give results. Web structured mining helps calculating the importance of page whereas Web Content Mining calculates relevancy of the page to the query. To make it happen, *(Pooja Sharma et al. 2010)* modified the Search engine architecture a little. By adding Weight Calculator, Relevancy Calculator, and WPCR

calculator. Results of Weight calculator and Relevancy Calculator are given to WPCR calculator to calculate final score of page. Modified architecture of Search engine and used by *(Pooja Sharma et al. 2010)* is shown in figure 4.

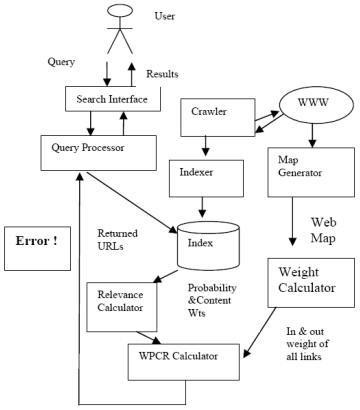

Figure 4: Modified architecture of Search Engine

4.6. Improved Page Rank Algorithm based on Optimized Normalization Technique

(Hema Dubey et al. 2011) proposed an Improved Page Rank Algorithm in which they initially calculated the page rank of all the web pages and then normalized page rank of all the web pages by dividing each page rank by mean value of all the page rank. By this they reduced number of iteration for calculating the page ranks as compared to conventional page rank algorithms, which then reduced the time complexity.

4.7. Weighted Link Rank

(Ricardo Baeza Yates et al. 2004) proposed this algorithm. In this algorithm they assigned page rank to web pages on the basis of following equation:

$$R(i) = \frac{q}{T} + (1-q)\sum_j \frac{W(i,j)R(j)}{\sum_k W(j,k)} \quad (7)$$

$$W(j,i) = L(j,i)\big(c + T(j,i) + AL(j,i) + RP(j,i)\big) \quad (8)$$

Where given a link from page j to page i,

L (j,i) is 1 if the link exists, or 0 otherwise; and c is a constant that gives a base weight to every link,
T (j,i) is a value that depends on the tag where the link is inserted,
AL (j,i) is the length of the anchor text of the link divided by a constant d
RP (j,i) is the relative position of the link in the page weighted by a constant b

The best attribute seemed was anchor text. Whereas relative position was not so effective as logical position and physical position are not always same.

4.8. Weighted Page Rank Algorithm based on Number of Visits of Links of Web pages

(*Neelam Tyagi et al. 2012*) proposed an extension of WPR, WPR (VOL). Weighted Page Rank based on Visit of Link calculates page rank value on the basis of visit of incoming links of a page as well as the popularity of in links. Not only this it also takes into consideration the user browsing behavior to provide relevant results according to the user needs. According to (*Neelam Tyagi et al. 2012*) links with high probability of visit contribute towards rank of its out linked pages. Also the main advantage of WPR (VOL) is that user cannot intentionally increase rank of web pages.

5. COMPARISION OF VARIOUS WEB PAGE RANKING ALGORITHMS

A Comparative analysis of few algorithms is given in Table2

Table 2: Comparative Analysis of Ranking Algorithms

ALGORITHM	PAGE RANK	HITS	WEIGHTED PAGE RANK	DISTANCE RANK
MAIN TECHNIQUE	Web Structure Mining	Web Structure Mining, Web Content Mining	Web Structure Mining	Web Structure Mining
METHODOLOGY	It computes the score for pages at the time of indexing of the pages	It computes hub and authority of relevant pages.	Weight of web pages is calculated on the basis of input and outgoing links.	Based on reinforcement learning which consider the logarithmic distance between the pages.
INPUT PARAMETER	Back links	Content, Back and Forward Links	Back links and Forward links	Forward links
RELEVANCY	Less as pages are ranked at indexing time	More because it uses hyperlinks and considers content of the page	Less as weight of pages are calculated at indexing time	Moderate because of the use of hyperlinks
QUALITY OF RESULTS	Medium	Less than PR	Higher than PR	High
IMPORTANCE	Back links are considered	Hubs and Authorities scores are used	The pages are sorted according to importance	It s based on distance between the pages
LIMITATION	Results come at indexing time not at query time	Topic drift and efficiency problem	Relevancy is not considered	If new page is inserted, crawler will have to perform large amount of calculation.

6. CONCLUSION

Web mining is a field which now a days have become an important part of human life. All the search queries and the information can be extracted from web. Ranking algorithms are also an important part of search engine. In this paper we have discussed about Web Mining and its taxonomy, beside this we have mentioned methodology of different ranking algorithms and different aspects it undertake. Also we have compared few algorithms on the basis of different parameters.

REFERENCES

[1] Ali Mohammad Zareh Bidoki and Nasser Yazdani, **"Distance Rank: An Intelligent Ranking Algorithm for Web Pages"**, Information Processing and Management, 44 , pp.877–892, 2007.

[2] Cooley, R., Mobasher, B., and Srivastava, J**. "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), Newposrt Beach, CA, 1997.

[3] David Hawkin,**"Web Search Engines"**, CSIRO ICT Center, pp. 86-90, June 2006

[4] Hema Dubey et al**. "An Improved Page Rank Algorithm based on Optimized Normalization Technique"**, International Journal of Computer Science and Information Technologies, Vol. 2(5), pp. 2183-2188, 2011

[5] J. Kleinberg, **"Authoritative Sources in a Hyper-Linked Environment"**, Journal of the ACM 46(5), pp. 604-632, 1999.

[6] J. Kleinberg, **"Hubs, Authorities and Communities"**, ACM Computing Surveys, 31(4), 1999.

[7] 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, pp. 1-17, 1999.

[8] Neelam Duhan, A.K Sharma and Komal Kumar Bhatia,**"Page Ranking Algoithms: A Survey"**, In proceedings of the IEEE International Advanced Computing conference(IACC),2009.

[9] Neelam Tyagi and Simple Sharma**, "Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page"**, In International Journal of Soft Computing and Engineering(IJSCE), ISSN:2231-2307, Volume-2, Issue-3, pp. 441-446, July2012

[10] Pooja Sharma et al. **"Weighted Page Content Rank for ordering Web Search Result"**, International Journal of Engineering Science and Technology Vol.2(12), pp. 7301-7310, 2010

[11] R. Kosala, and H. Blockeel, **"Web Mining Research: A Survey"**, SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.

[12] Ricardo Baeza-Yates and Emilo Davis**, "Web page ranking using link attributes"**, In proceedings of the 13th International World Wide Web Confrence on Alternate track papers and posters, pp. 328-329, 2004.

[13] S. Brin, and L. Page, **"The Anatomy of a Large Scale Hypertextual Web Search Engine"**, Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.

[14] W. Xing and Ali Ghorbani**, "Weighted PageRank Algorithm"**, Proc. Of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.

ABOUT THE AUTHORS

Rupal Bhargava is pursuing her M.Tech in Computer Science from Banasthali Vidyapith, Rajasthan. She is undergoing the training of her M.Tech in supervision of Mrs. Rekha Jain. Her current research interest includes Web Mining, Semantic Web and Data Mining. She has published various papers in the conferences and journals.

Rekha Jain completed her Master Degree in Computer Science from Kurukshetra University in 2004. Now she is working as Assistant Professor in Department of "Apaji Institute of Mathematics & Applied Computer Technology" at Banasthali University, Rajasthan and pursuing Ph.D. under the supervision of Prof. G. N. Purohit. Her current research interest includes Web Mining, Semantic Web and Data Mining. She has various National and International publications and conferences.