Performance Evaluation and Implementation of Page Ranking Algorithm Based on Counts of Link Hits (PRCLH) for Interactive Information Retrieval in Web Mining

Zaved Akhtar, Saoud Sarwar

Abstract: While huge amount of data has become a highlighted buzzword since last some years, "big data mining", i.e., mining from big data, has almost immediately followed up as an emerging, interrelated research area. Extracting useful information has proven extremely challenging task. The Search engines generally return a large number of pages in response to user queries. To assist the users to navigate in the result list, ranking methods are applied on the search results. We have discussed about most of the page ranking algorithms based on link or content information retrieval in Web Mining**.** Here in this paper we have used page ranking Algorithms Based on Count of Link Hits (PRCLH) for calculation of page for interactive information retrieval in Web Mining.

Keywords: Web Mining, Data Mining, Page Rank ,WCM, Web Usage Mining, Web Structure Mining, PRCLH. .

1. INTRODUCTION

Web Mining is defined as the application of data mining techniques on the World Wide Web to find hidden information, This hidden information i.e. knowledge could be contained in content of web pages or in link structure of WWW [2, 16] or in web server logs. WWW is a vast resource of hyperlinked and heterogeneous information including text, image, audio, video, and metadata. With the rapid growth of information sources available on the WWW and growing needs of users, it is becoming difficult to manage the information on the web and satisfy the user needs. Actually, we are drowning in data but starving for knowledge. Therefore, it has become increasingly necessary for users to use some information retrieval techniques to find, extract, filter and order the desired information.

Search engine [8] receives users query, processes it, and searches into its index for relevant documents i.e. the documents that are likely related to query and supposed to be interesting then, search engine ranks the documents found relevant and it shows them as results. This process can be divided in the following tasks:

Crawler [14, 15] is in charge of visiting as many pages and retrieves the information needed from them. The idea is that this information is stored for the use by the search engine afterwards.

Indexing the information provided by a crawler has to be stored in order to be accessed by the search engine. As the user will be in front of his computer waiting for the answer of the search engine, time response becomes an important issue. That is why this information is indexed in order to decrease the time needed to look into it.

Searching: The web search engine represents the user interface needed to permit the user to query the information. It is the connection between the user and the information repository.

Sorting/Ranking Due to the huge amount of information existing in the web, when a user sends a query about a general topic (e.g. java course), there exist an incredible number of pages related to this query but only a small part of such amount of information will be really interesting for the user. That is why the search engines incorporate ranking algorithms in order to sort the results.

2. WEB MINING

Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from large databases is called Data Mining. Web Mining is the application of data mining techniques to discover and retrieve useful information and patterns (knowledge)

from the WWW documents and services web mining can be divided into three categories [1]:

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

2.1 Web Content Mining (WCM)

WCM describes the automatic search of information resources available online, and involves mining web data content. It is emphasis on the content of the web page not its links. It can be applied on web pages itself or on the result pages obtained from a search engine. WCM is differentiated from two different points of view: Information Retrieval (IR) View and Database View. In IR view, most of the researches use bag of words, which is based on the statistics about single words in isolation, to represent unstructured text. For the semi-structured data, all the works utilize the HTML structures insides the documents. For database view, Web mining always tries to infer the structure of the Web site to transform a Web site to become a database.

2.2 Web Structure Mining (WSM)

WSM is used to generate structural summary about the Web sites and Web pages. The structure of a typical Web graph consists of Web pages as nodes and hyperlinks as edges connecting two related pages. Technically, WCM mainly focuses on the structure of inner-document, while WSM tries to discover the link structure of the hyperlinks at the inter-document level.

Web structure mining tries to discover the model underlying the link structures of the Web. The model is based on the topology of the hyperlink with or without the link description. This model can be used to categorize the Web pages and is useful to generate information such as similarity and relationships between Web sites. And the link structure of the Web contains important implied information, and can help in filtering or ranking Web pages. In particular, a link from page A to page B can be considered a recommendation of page B by the author of A. Some new algorithms have been proposed that exploit this link structure not only for keyword searching, but other tasks like automatically building a Yahoo-like hierarchy or identifying communities on the Web. The qualitative performance of these algorithms is generally better than the IR algorithms since they make use of more information than just the contents of the pages. While it is indeed possible to influence the link structure of the Web locally, it is quite hard to do so at a global level. So link analysis algorithms that work at a global level possess relatively robust defenses against spamming.

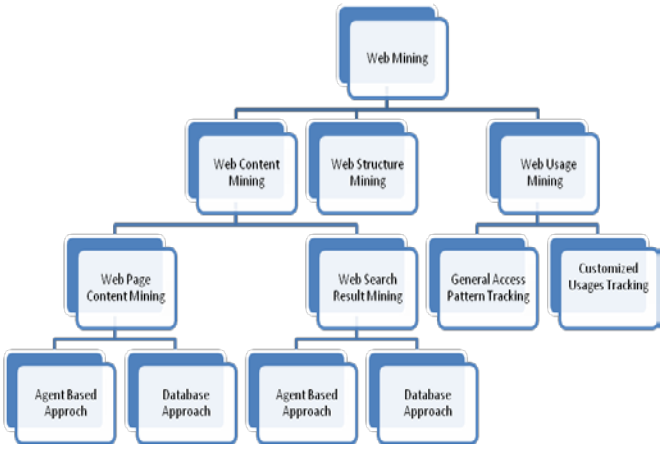

Fig 1. Taxonomy of Web Mining

2.3 Web Usage Mining (WUM)

Web Usage Mining (WUM) tries to discover user navigation patterns from web data and the useful information from the secondary data derived from the interactions of the users while surfing on the Web. It focuses on the techniques that could predict user behavior while the user interacts with Web. This type of web mining allows for the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web pages. This information is often gathered automatically into access logs via the Web server. CGI scripts offer other useful information such as referrer logs, user subscription information and survey logs. This category is important to the overall use of data mining for companies and their internet/ intranet based applications and information access.

The three categories of web mining described above have its own application areas including site improvement, business intelligence, Web personalization, site modification, usage characterization and page ranking etc. The search engines to find more important pages generally use the page ranking. Proposed PRNLV method use web structure and web uses mining technique to rank web pages.

3. RELATED WORK OF RANKING ALGORITHMS

The web is very large and diverse and many pages could be related to a given query. That is why a method/algorithm is used to sort the entire pages subject

to be interesting to a user's query. All the algorithms consider the web pages as a directed graph in which pages are denoted as nodes and links are denoted as edges.

3.1 PageRank Algorithm (PR)

Surgey Brin and Larry Page developed a ranking algorithm used by Google, named PageRank [8] after Larry Page (cofounder of Google search engine), that uses the link structure of the web to determine the importance of web pages [3, 17]. It takes back links into account and propagates the ranking through links. Thus, a page has a high rank if the sum of the ranks of its back links is high. A simplified version of page rank is defined as follows

$$PR(p) = (1-c)\sum_{q \in I(p)} \frac{PR(q)}{O(q)} \quad (3.1)$$

In the calculation of PageRank a factor c is used for normalization. Note that 0<c < 1 because there are pages without incoming links and their weight is lost .

Later PageRank was modified observing that not all users follow the direct links on WWW

$$PR(p) = (1-d) + d\sum_{q \in I(p)} \frac{PR(q)}{O(q)} \quad (3.2)$$

Where d is a dampening factor that is usually set to 0.85 (any value between 0 and 1), d can be thought of as the probability of users' following the links and could regard (1 − d) as the page rank distribution from non-directly linked pages .Consider the following directed graph[10]

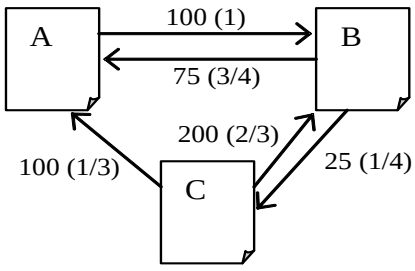

Fig 2 Example graph

The PageRanks for pages A, B, C are calculated by using (3.2) with d=0.5, the page ranks of pages A, B and C becomes:*PR(A)*=1.2, *PR(B)*=1.2, *PR(C)*=0.8

3.2 Weighted Page Rank Algorithm (WPR)

Wenpu Xing and Ali Ghorbani [13] proposed an extension to standard PageRank called Weighted PageRank (WPR). It rank pages according to their importance not only consider link structure of web graph. This algorithm assigns larger rank values to more important pages instead of dividing the rank value of a page evenly among its outgoing linked pages. Each outlink page gets a value proportional to its popularity. The popularity is measured by its number of inlinks and outlinks.[13]

$$PR(p) = (1-d) + d\sum_{q \in I(P)} PR(q) W^{in}_{(q,p)} W^{out}_{(q,p)} \quad (3.3)$$

Where Win (q,p) and Wout (q,p), for inlinks and outlins is given as

$$W^{in}_{(q,p)} = \frac{I_p}{\sum_{v \in R(q)} I_v} \quad (3.3.1)$$

$$W^{out}_{(q,p)} = \frac{O_p}{\sum_{v \in R(q)} O_v} \quad (3.3.2)$$

Where Iv,Ip and Ov ,Op represent the number of inlinks and outlinks of page v and page p respectively. The Page Ranks for pages A, B, C are calculated by using (3.3) with d=0.5, the page ranks of pages A, B and C are PR (A)=0.65, PR (B)=0.93, PR(C)=0.60.

3.3 Page Content Rank Algorithm (PCR)

Jaroslav Pokorny and Jozef Smizansky[11] gave a new ranking method of page relevance ranking employing WCM technique, called Page Content Rank (PCR). This method combines a number of heuristics that seem to be important for analyzing the content of web pages. The page importance is determined on the basis of the importance of terms, which the page contains. The importance of a term is specified with respect to a given query q. PCR uses a neural network as its inner classification structure. The importance of a page P in PCR is calculated as an aggregate value of the importance of all terms that P contains. For a promotion of the significant term and a suppression of the others, the second moment is again used as an aggregate function [6]

Page_importance(P)=sec_moment({importance(t): t ∈ P}) (3.4)

3.4 Hyperlinked Induced Topic Search Algorithm (HITS) [14, 16]

This algorithm assumes that for every query topic, there is a set of "authoritative" or "authority" pages/sites that are relevant and popular focusing on the topic and there are "hub" pages/sites that contain useful links to relevant sites including links to many related authorities.

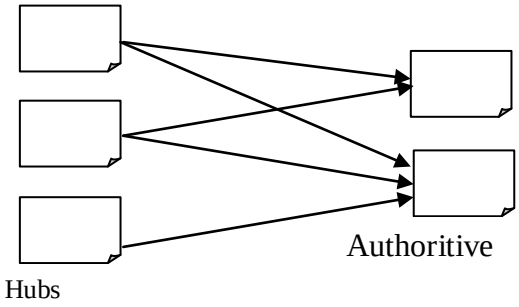

Fig 3. Hubs and Authorities

Working of HITS: The HITS works in two phases Sampling and Iterative in the Sampling phase a set of relevant pages for the given query are collected i.e. a sub-graph *S* of *G* is retrieved which is high in authority pages. The Iterative phase finds hubs and authorities using the output of the sampling phase using following equations.

$$H_p = \sum_{q \in I(p)} A_q \quad (3.4)$$

$$A_p = \sum_{q \in B(p)} H_q \quad (3.5)$$

Where H_p is the hub weight, A_p is the Authority weight, $I(p)$ and $B(p)$ denotes the set of reference and referrer pages of page p.

The comparison summary of algorithms discussed in Table 1

TABLE 1.COMPARISON OF PAGE RANKING ALGORITHMS

Algorithm	**Page Rank**	**WPR**	**PCR**	**HITS**	**SALSA,SimRank, Randomise HITS, etc**
Technique Used	Web Structure Mining	Web Structure Mining	Web Content Mining	Web Structure Mining, Web Content Mining	Web Structure Mining, Web Content Mining
Description	Computes scores at indexing time not on fly. Results are sorted according to importance of pages.	Computes scores at indexing time, Unequal distribution of score, pages are sorted according to importance.	Computes scores on the fly. Pages returned are related to the query i.e. relevant documents are returned.	Computes hub and authorities scores on the fly. Relevant as well as important pages are returned.	Computes hub and authorities scores on the fly. Relevant as well as important pages are returned.
I/P Parameters	Backlinks	Backlinks, forward Links	Content	Backlinks, forward Links	Backlinks, forward Links
Working levels	n	1	Nil	n	n
Complexity	O(log n)	< O(log n)	O(m*)	> O(log n)	> O(log n)
Relevancy	No	No	Yes	Yes	Yes
Importance	Yes	Yes	No	Yes	Yes
Quality of result	Low	High	Low	High	High
Stability in Results	High	High	Low	Low	High
Limitations	Computes scores at indexing time not on fly. Results are sorted according to importance of pages.	Relevancy is ignored. Method computes scores at a single level.	Importance of pages is totally ignored.	Topic drift and efficiency problems	Topic drift , complex, and efficiency problems

*n: number of web pages *m: number of terms in a page

4. PROBLEMS AND ISSUES OF THE PAGE RANKING ALGORITHMS

The main problems and issues of discussed page ranking algorithms are summarized as

4.1 Rank quality of PageRank

The discussed ranking algorithms have shown a really high quality and the proof is that success of Google (or they are still being used) successfully. However, some improvements can be done on it.

4.2 Data Mining Technique of PageRank

PageRank algorithm used only Web Structure Mining and Web Content Mining technique; it does not use Web Usage Mining, which may significantly improve the quality of rank of web pages according to users information needs.

4.3 PageRank is Static in Nature

In PageRank algorithm, the importance or rank score of each page are static in nature. The rank changes only with link structure of web.

5. PROPOSED PAGE RANKING ALGORITHM BASED ON COUNTS OF LINK HITS (PRCLH)

PRCLH (Page Ranking based on Counts of Link Hits) based on Web Structure Mining and Usage Mining; it takes the user visits of pages/links into account with the aim to determine the importance and relevance score of the web pages. To accomplish the complete task from gathering the usage characterization till final rank determination many subtasks are performed such as

- Storage of user's access information (hits) on an outgoing link of a page in related server log files.
- Fetching of pages and their access information by the targeted web crawler.
- For each page link, computation of weights based on the probabilities of their being visited by the users.
- Final rank computation of pages based on the weights of their incoming links.
- Retrieval of ranked pages corresponding to user queries

5.1 Calculation of Visits (hits) of links

If p is a page with outgoing-link set O(p) and each outgoing link is associated with a numerical integer indicating visit-count (VC), then the weight of each outgoing link connecting to page p to page o is calculated by[Proposed]

$$Weight_{link}(p,q) = \frac{VC(p,q)}{\sum_{q' \in O(p)} VC(p,q')} \qquad (5.1)$$

5.2 Page Rank based on Counts of Link Hits (PRCLH)

If p is a page having inbound-linked pages in set B(p), then the rank (PRCLH) is given by[Proposed]:

$$PRCLH(p) = (1-d) + d\left(\sum_{b \in B(p)} PRCLH(b).Weight_{link}(b,p)\right) \qquad (5.2)$$

where d is the damping factor as is used in PageRank, Weightlink() is the weight of the link calculated by (5.1).The iteration method is used for the calculation of page rank. Example fig 2 Taking d=0.5, these equations can easily be solved using iteration method the final results obtained are:

PRCLH(A)= 1.08, PRCLH(B)= 1.26, PRCLH(C)= 0.66

5.3 Experimental Result

For the experimental results assume a web graph

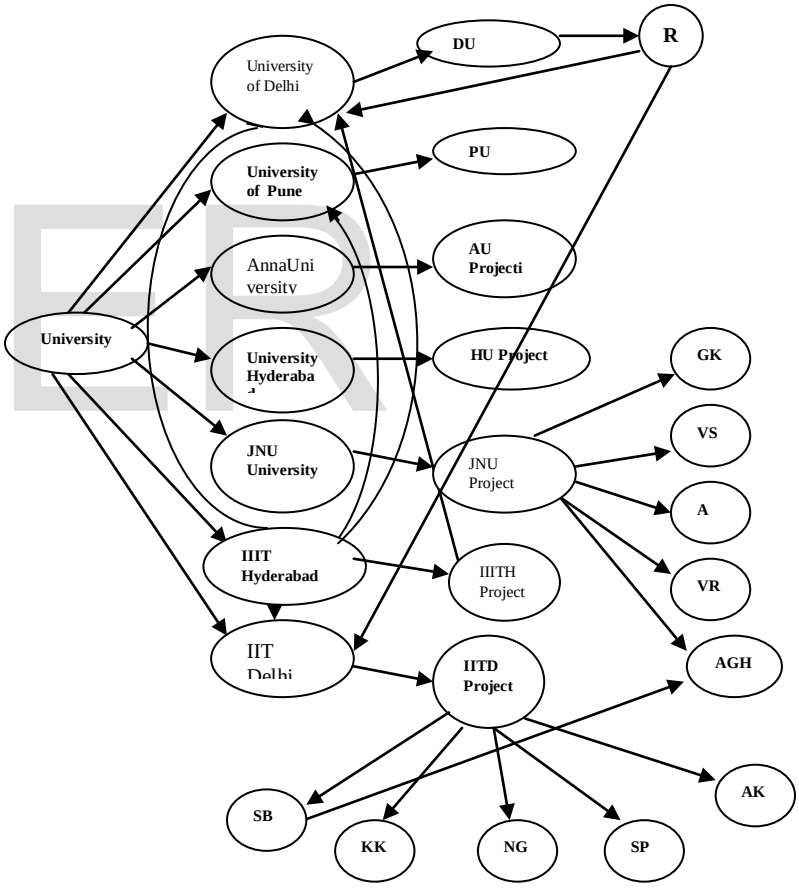

Fig 4. Web Graph

Page	URL	Hits
http://localhost/webgraph/university.html	http://localhost/webgraph/jnu.html	45
http://localhost/webgraph/Research Areas IIT Delhi...	http://localhost/webgraph/Prof K. K. Biswas.html	27
http://localhost/webgraph/university.html	http://localhost/webgraph/university of hyderabad	44
http://localhost/webgraph/university.html	http://localhost/webgraph/IIT DELHI.html	47
http://localhost/webgraph/university.html	http://localhost/webgraph/university of delhi.html	27
http://localhost/webgraph/university.html	http://localhost/webgraph/university of pune.html	7
http://localhost/webgraph/university.html	http://localhost/webgraph/anna university.html	23
http://localhost/webgraph/university.html	http://localhost/webgraph/IIIt Hyderabad.html	12
http://localhost/webgraph/Research Areas IIT Delhi...	http://localhost/webgraph/Sanjiva Prasad IIT Delhi	38
http://localhost/webgraph/Research Areas IIT Delhi...	http://localhost/webgraph/Saurav Bansal IIT Delhi.h	32
http://localhost/webgraph/jnu.html	http://localhost/webgraph/IIT DELHI.html	13
http://localhost/webgraph/Dn project.html	http://localhost/webgraph/Rakesh Kumar.html	23
http://localhost/webgraph/IIT DELHI.html	http://localhost/webgraph/Research Areas IIT Delhi	18
http://localhost/webgraph/jnu.html	http://localhost/webgraph/Project in JNU.html	4
http://localhost/webgraph/Sanjiva Prasad IIT Delhi...	http://localhost/webgraph/Indian Council of Agricu	10
http://localhost/webgraph/jnu.html	http://localhost/webgraph/university of delhi.html	30
http://localhost/webgraph/Research Areas IIT Delhi...	http://localhost/webgraph/Anshul Kumar IIT Delhi.h	5
http://localhost/webgraph/Research Areas IIT Delhi...	http://localhost/webgraph/Naveen Garg IIT DELHI.ht	21
http://localhost/webgraph/jnu.html	http://localhost/webgraph/Research Areas IIT Delhi	23
http://localhost/webgraph/anna university.html	http://localhost/webgraph/anna university project	15
http://localhost/webgraph/university of hyderabad....	http://localhost/webgraph/university of pune.html	21
http://localhost/webgraph/university of pune.html	http://localhost/webgraph/IIIt Hyderabad.html	40
http://localhost/webgraph/university of pune.html	http://localhost/webgraph/Vaibhav V. Kawade pune.h...	24
http://localhost/webgraph/university of delhi.html	http://localhost/webgraph/Dn project.html	13
http://localhost/webgraph/IIIt Hyderabad.html	http://localhost/webgraph/R & D of IIIT Hyderabad....	22
http://localhost/webgraph/Project in JNU.html	http://localhost/webgraph/Arun K. Attri jnu.html	20
http://localhost/webgraph/Project in JNU.html	http://localhost/webgraph/V.Rajamani jnu.html	3
http://localhost/webgraph/Project in JNU.html	http://localhost/webgraph/Indian Council of Agricu	40
http://localhost/webgraph/Project in JNU.html	http://localhost/webgraph/V.Subramaniam jnu.html	12
http://localhost/webgraph/V.Rajamani jnu.html		30

Fig 5. Numbers of the Hits of the links

WEB PAGE	PR	PRCLH
anna university projet.html	0.2929	0.12
anna university.html	0.1686	0.452
Anshul Kumar IIT Delhi.html	0.2499	0.87
Arun Kumar Attri jnu.html	0.1876	0.287
Du Project.html	0.675	0.321
IIIT Hyderabad.html	0.135	0.34
IITDelhi.html	0.4053	0.2344
Indian Council of Agricultural Research.html	0.1698	0.76
jun.html	0.123	0.1453
KG Saxena JNU.html	0.564	0.15
Naveen Garg IIT DELHI.html	0.187	0.198
Prof. KK Biswas.html	0.345	0.2679
Project in JNU.html	0.453	0.209
R & D of IIIT Hyderabad.html	0.122	0.673
Rankesh Kumar.html	0.2897	0.2897
Research Areas IIT Delhi.html	0.4642	0.4298
Sanjiva Prasad IIT Delhi.html	0.242	0.546
Saurabh Bansal IIT Delhi.html	0.1674	0.1897
Univesity of Delhi.html	0.134	0.456
Univesity of Hyderabad.html	0.1276	0.1986
Univesity of Pune.html	0.1342	0.564
Univesity.html	0.286	0.78
V.Rajamani jnu.html	0.242	0.342
V.Subramaniami jnu.html	0.675	0.765
Vaibhav V Kaware pune.html	0.1765	0.76

Fig 6. Page Rank Using (PR and PRCLH)

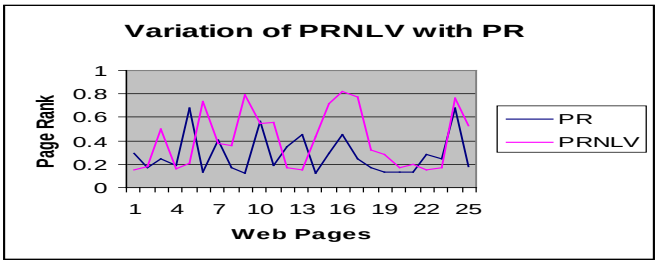

Fig 7. Variation of PRNLV with PR

TABLE 2. COMPARISON OF PRCLH with PR AND WPR

Algorithm Parameter	*PageRank (PR)*	*Weighted Page Rank (WPR)*	*Page Rank based on Numbers Link-Visit (PRCLH)*
Description	Computes scores at indexing time. Results are sorted according to importance of pages.	Computes scores at indexing time. Results are sorted according to importance of pages.	Computes scores at indexing time. Pages are sorted according to importance and relevance.
Mining Technique Used	Web Structure Mining	Web Structure Mining	Web Structure Mining, Web Usage Mining
Rank Distribution	Ranks are equally distributed to outgoing links.	Ranks are equally distributed to outgoing links.	Ranks are unequally distributed among outgoing links according to their probabilities of visit.
I/P Parameters	Inbound links of pages	Inbound links and Outbound links of pages	Inbound links, Outbound links, Visit Counts of links.
Working levels	n*	n*	n
Complexity	O(log n)	O(log n)	> O(log n)
Nature of Rank	Less dynamic (rank changes with link structure)	Less dynamic (rank changes with link structure)	More dynamic (rank changes with visit counts & structure of links)
Relevancy of pages	no	no	yes
Importance of pages	yes	yes	yes

Quality of result	Low	High	High
Advantages	Computation of ranks with minimum effort and less complexity.	Computation of ranks with minimum effort and less complexity.	Pages returned are of high quality and relevancy as user feedbacks are taken into account. Search space can be very much pruned as pages are sorted according to users' information needs.
Limitations	No relevancy of pages is considered in rank computation. All links are considered equally important.	No relevancy of pages is considered in rank computation. All links are considered equally important.	Extra effort on crawlers to fetch the visit counts of pages from web servers. Extra calculations to find the weights of links.

7. CONCLUSION

Mining of knowledgeable data from a huge amount of data is very complex task, World Wide Web information play a vital role for information collection and sharing. The ranking algorithms are used to search the relevance information in very
efficient manner. Different page ranking algorithms are used in different techniques. The PRCLH uses the user's browsing information in consideration to calculate rank of a documents rather than link structure. Due to browsing information in consideration PRCLH system is more dynamic than other ranking algorithms

REFERENCES

[1] Neelam Duhan, A. K. Sharma, Komal Kumar Bhatia, "Page Ranking Algorithms: A Survey", 2009 IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009.

[2] Romit D. Jadhav and Ajay B. Gadicha" A Novel Efficient Reivew Report on Google Page Rank Algorithm" International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 3, pp 393-397, March 2013.

[3] Hema Dubey and Prof. B. N. Roy "An Improved Page Rank Algorithm based on Optimized Normalization Technique" International Journal of Computer Science and Information technology, Volume 2(5), 2183 – 2188, 2011.

[4] Sweah Liang Yong Markus Hagenbuchner Ah Chung Tsoi, "Ranking Web Pages using Machine Learning Approaches", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008.

[5] Ja-Hwung Su, Bo-Wen Wang and Vincent S. Tseng "Effective Ranking and Recommendation on Web Page Retrieval by Integrating Association Mining and PageRank" IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008.

[6] Glen Jeh and Jennifer Widom. Simrank: A measure of structural-context similarity. Technical report, Stanford University Database Group, 2001.

[7] Jiawei Han and Micheline Kamber . Data Mining Concepts and Techniques. Second Edition, Morgan Kaufmann Publishers, 2006.

[8] Andrei Broder. A taxonomy of web search. Technical report, IBM Research, 2002.

[9] G. Jeh and J. Widom. Simrank: A measure of structuralcontext similarity, 2002.

[10] C. Ridings and M. Shishigin, Pagerank uncovered. Technical report, 2002.

[11] Jaroslav Pokorny, Jozef Smizansky, Page Content Rank: An Approach to the Web Content Mining.

[12] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, and R. Stata. Graph structure in the web. In In Proceedings of the 9th International World Wide Web Conference, 2000.

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

[14] Junghoo Cho, Hector Garc´ýa-Molina, and Lawrence Page. Efficient crawling through URL ordering. Computer Networks and ISDN Systems, 30(1–7):161–172, 1998.

[15] Tim Berners-Lee, Robert Cailliau, Ari Luotonen, Henrik Frystyk Nielsen, and Arthur Secret. The World-Wide Web. Communications of the ACM, 37(8):76–82, 1994.

[16] Alan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, and Panayiotis Tsaparas. Finding authorities and hubs from link structures on the world wide web. In World Wide Web, pages 415–429, 2001.

[17] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg, Mining the Web's link structure. Computer, 32(8):60–67, 1999.

[18] Andrew Y. Ng, Alice X. Zheng, and Michael I. Jordan. Stable algorithms for link analysis. In Proc. 24th Annual Intl. ACM SIGIR Conference. ACM, 2001.

[19] Brian Pinkerton. Finding what people want: Experiences with the web crawler. In The second Internation WWW Conference Chicago, 1994.

Zaved Akhtar is M. Tech (Computer Engineering) research scholar at Al – Falah School of Engineering & Technology, Dhauj, Faridabad, Haryana, India, PH- +91-9811160042
E-mail: javed.gkp@rediffmail.com

Saoud Sarwar is currently working as a Professor & Head in Computer Science & Engineering Department at Al – Falah School of Engineering & Technology, Dhauj, Faridabad, Haryana, India, PH- +91- 9953706957
E-mail: saoud.hod.cse@gmail.com