ISSN:0975-9646

An Improved Page Rank Algorithm based on Optimized Normalization Technique

Hema Dubey ,Prof. B. N. Roy
Department of Computer Science and Engineering
Maulana Azad National Institute of technology
Bhopal, India

***Abstract*— Page Ranking is an important component for information retrieval system. It is used to measure the importance and behavior of web pages. We review two approaches for ranking: HITS concept and Page Rank method. Both approaches focus on the link structure of the Web to find the importance of the Web pages. The Page Rank algorithm calculates the rank of individual web page and Hypertext Induced Topic Search (HITS) depends upon the hubs and authority framework. A fast and efficient page ranking mechanism for web retrieval remains as a challenge. This paper proposed a new page rank algorithm which uses a normalization technique based on mean value of page ranks. The proposed scheme reduces the time complexity of the traditional Page Rank algorithm by reducing the number of iterations to reach a convergence point.**

Keywords- Ranking, Page Rank, HITS, Hyperlink, Normalization.

I. INTRODUCTION

The web as we all know is the largest source of data. During the past few years the World Wide Web has become the foremost and most popular way of communication and information dissemination. It serves as a platform for exchanging various kinds of information, ranging from research papers, and educational content, to multimedia content, software and personal logs. Every day, the web grows by roughly a million electronic pages, adding to the hundreds of millions pages already on-line. So with the rapid growth of information sources available on the World Wide Web, it has become increasingly necessary for users to use automated tools to find the desired information resources, and to track and analyze their usage patterns. These factors give rise to the requisite of creating server side and client side intelligent systems that can effectively mine for knowledge. Web mining can be broadly defined as the extraction and mining of useful information from the World Wide Web.

Thus the Internet is an infinite source of information which includes massive collection of web pages and countless hyperlinks. These hyperlinks contain a huge amount of concealed human explanation that can be extremely valuable for automatically inferring concept of authority. Thus the structure of a typical Web graph (Figure 1) consists of web pages as nodes, and hyperlinks as edges connecting two related pages.

Web Structure Mining is the process of discovering information from the Web, finding information about the web pages and inference on hyperlink, finding authoritative web pages, retrieving information about the relevance and the quality of the web page. Thus Web structure mining focuses on the hyperlink structure of the web. We review two approaches: HITS concept and Page Rank method. Both approaches focus on the link structure of the Web to find the importance of the Web pages. Mainly In links to the pages and out links from the page can give idea about the context of the page. PageRank does not rank web sites as a whole, but it calculates the rank of individual web page and Hypertext Induced Topic Search (HITS) depends upon the hubs and authority framework.

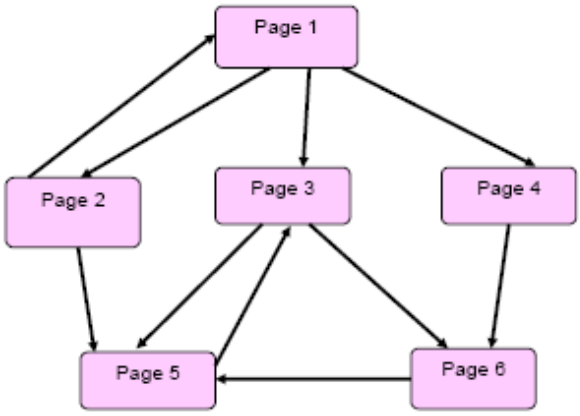

Figure 1, Web Graph

We provide here an overview of Recursive Data Mining. The rest of this paper is organized as follows: Section II introduces Background and Related Work; Section III describes about Traditional Page Rank Algorithm; Section IV shows the Proposed Page Rank Algorithm; Section V discusses the detailed overview of Experimental Results. Section VI summarizes the Conclusion and prospect. Finally references are given.

II. BACKGROUND AND RELATED WORK

A web search engine typically consists of:

1. Crawler: used for retrieving the web pages and web contents
2. Indexer: stores and indexes information on the retrieved pages
3. Ranker: Measure the importance of Web Pages returned
4. Retrieval Engine: performs lookups on index tables against query

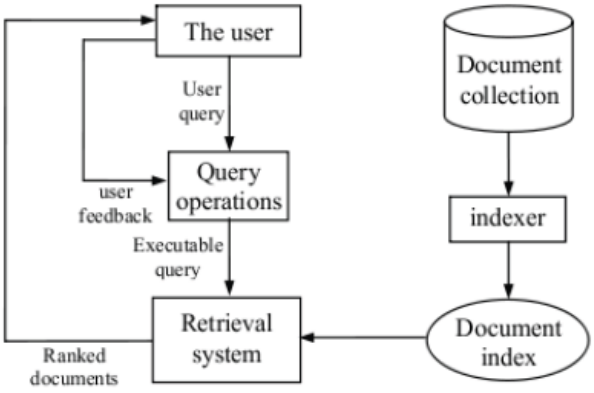

Figure 2, a search engine system during a search operation - A user issues a query which is first checked before it is forwarded and compared to documents indexes.

A. Ranking in Web Search

Nowadays searching on the internet is most widely used operation on the World Wide Web. The amount of information is increasing day by day rapidly that creates the challenge for information retrieval. There are so many tools to perform efficient searching. Due to the size of web and requirements of users creates the challenge for search engine page ranking [19]. Ranking is the main part of any information retrieval system Today's search engines may return million of pages for a certain query It is not possible for a user to preview all the returned results So, page ranking is helpful in web searching. Rankers are classified into two groups: - Content-based rankers and Connectivity-based rankers. Content-based rankers works on the basis of number of matched terms, frequency of terms, location of terms, etc. Connectivity-based rankers work on the basis of link analysis technique, links are edges that point to different web pages. There are two famous link analysis methods:- 1)PageRank Algorithm[18] and 2) HITS Algorithm[23].

PageRank has been developed by Google and is named after Larry Page, Google's co-founder and president [18]. PageRank ranks pages based on the web structure. PageRank uses global link information and is stated to be the primary link recommendation scheme employed in the Google search engine and search appliance. PageRank is designed to simulate the behaviour of a "random web surfer" [18] who navigate a web by randomly following links. If a page with no outgoing links is reached, the surfer jumps to a randomly chosen bookmark. In addition to this normal surfing behavior, the surfer occasionally spontaneously jumps to a bookmark instead of following a link. The PageRank of a page is the probability that the web surfer will be visiting that page at any given moment.

L. Page and S. Brin [21; 23] proposed the Page Rank algorithm to calculate the importance of web pages using the link structure of the web. In their approach Brin and Page extends the idea of simply counting in-links equally, by normalizing by the number of links on a page. The Page Rank algorithm is defined as [19]: "*We assume page A has pages T1...Tn which point to it (i.e., are* citations). The parameter d is a damping factor, which can be set between 0 and 1. We usually set d to 0.85. Also C (T1) is defined as the number of links going out of page A. The Page Rank of a page A is given as follows:

PR (A) = (1-d) + d (PR (T1)/C (T1) + ... + PR (Tn)/C (Tn))

Note that the Page Ranks form a probability distribution over web pages, so the sum of all web pages' Page Ranks will be one." And "The d damping factor is the probability at each page the "random surfer" will get bored and request another random page."

Google, one of the world's most popular search engines, state that PageRank is an important part of their ranking function [18]. In recent years there have been many studies of how PageRank might be improved [25], optimised [26] and personalised [16], but there have not been any detailed evaluations of its potential benefit to retrieval effectiveness. PageRank has been observed to be more resilient to small changes in the web graph than HITS [14]. This may be an important property when dealing with WWW-based search as it is difficult to construct an accurate and complete web graph, and the web graph is likely to be impacted by web server down-time [8]. PageRank has previously been observed to exhibit similar performance to non query-dependent HITS (global HITS) [7]. Wenpu Xing and Ali Ghorbani introduces an extended PageRank algorithm called the Weighted PageRank algorithm (WPR). Rungsawang and et al. introduce a pagerank computation to un-bias the link farm effect [5]. It is a good algorithm if the link farm can be identified efficiently, but it is a more complicate situation in the real world. Havellwala proposes a topic-sensitive pagerank algorithm to evaluate webpages with consideration of category relevance. This modification can approach more precise scores of web pages, but the computation complexity will be a heavy load to index worldwide documents and reduce the efficiency in query time. Al-Saffar and et al. [4] follow the Havellwala's idea and claim a new approach for personalization without relying on the web link structure.

HITS was used for the first time in the Clever search engine from IBM, and PageRank is used by Google combined with other several features such as anchor text, IR measures, and proximity. HITS provides an innovative methodology for Web searching and topics distillation. According to the definition by Google, a web page is an authority on a topic if it provides good information and is a hub if it provides links to good authorities. HITS uses the mutual reinforcement operation to propagate hub and authority values to represent the linking characteristic [10].
Bianchini [19] noted that HITS and PageRank are used as starting points for new solutions, and there are some extensions of these two approaches. There are other link-based approaches to be applied on the Web. For further information please refer to [19]. The CLEVER algorithm is an extension of standard HITS and provides an appropriate solution to the problems that result from standard HITS [3]. CLEVER assigns a weight to each link based on the terms of the queries and end-points of the link. It combines anchor text to set weights to the links as well. Moreover, it breaks

large hub pages into smaller units so that each hub page is focused on as a single topic. Finally, in the case of a large number of pages from a single domain, it scales down the weights of pages to reduce the probabilities of overhead weights [3]. Another major shortcoming of standard HITS is that 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. A *Probabilistic analogue of the HITS Algorithm* (PHITS) has been developed to solve this problem [3]. PHITS provides a probabilistic interpretation of term-document relationships and identifies authoritative documents. In the experiment on a set of hyperlinked documents, PHITS demonstrates better results compared to those obtained by standard HITS. The most important feature of the PHITS algorithm is its ability to estimate the actual probabilities of authorities compared to the scalar magnitudes of authority that are provided by standard HITS.
Several limitations of the HITS model, as presented by Kleinberg [10], were observed and addressed by Bharat and Henzinger [1]. These are: Mutually reinforcing relationships between hosts. This occurs when a set of Documents on one host point to a single document on a second host. Automatically generated links. This occurs when web documents are generated by tools and are not authored (recommendation) links. Non-relevant nodes. This arises through what Bharat and Henzinger termed topic drift. Topic drift occurs when the local sub graph is expanded to include surrounding links, and as a result, pages not relevant to the initial query are included in the graph, and therefore in the HITS calculation.

III. TRADITIONAL PAGE RANK ALGORITHM

PageRank ranks pages based on the web structure. Google, which among search engines is ranked in the first place, uses the PageRank algorithm. PageRank has been developed by Google and is named after Larry Page, Google's co-founder and president [18]. PageRank is a numeric value that represents how important a page is on the web. Google figures that when one page links to another page, it is effectively casting a vote for the other page. The more votes that are cast for a page, the more important the page must be. Also, the importance of the page that is casting the vote determines how important the vote itself is. Google calculates a page's importance from the votes cast for it.

The Page Rank algorithm is given by

1) Calculate page ranks of all pages by following formula:

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

Where

PR(A) is the PageRank of page A,

PR(Ti) is the PageRank of pages Ti which link to page A,

C(Ti) is the number of outbound links on page Ti and

d is a damping factor which can be set between 0 and 1, but it is usually set to 0.85

2) Repeat step 1 until values of two consecutive iterations match.

So, first of all, we see that PageRank does not rank web sites as a whole, but is determined for each page individually. Further, the PageRank of page A is recursively defined by the PageRanks of those pages which link to page A. The PageRank of pages Ti which link to page A does not influence the PageRank of page A uniformly. Within the PageRank algorithm, the PageRank of a page T is always weighted by the number of outbound links C(T) on page T. This means that the more outbound links a page T has, the less will page A benefit from a link to it on page T. The weighted PageRank of pages Ti is then added up. The outcome of this is that an additional inbound link for page A will always increase page A's PageRank. Finally, the sum of the weighted Page Ranks of all pages Ti is multiplied with a damping factor d which can be set between 0 and 1.

Features of Page Rank Algorithm are:

- It is the query independent algorithm that assigns a value to every document independent of query.
- It is Content independent Algorithm.
- It concerns with static quality of a web page.
- Page Rank value can be computed offline using only web graph.
- Page Rank is based upon the linking structure of the whole web.Page
- Rank does not rank website as a whole but it is determined for each page individually.
- Page Rank of pages Ti which link to page A does not influence the rank of page A uniformly.
- More the outbound links on a page T, less will page A benefit from a link to it.
- Page Rank is a model of user's behavior

IV. PROPOSED PAGE RANK ALGORITHM

The proposed normalization technique for Page Rank algorithm is based on mean value of page ranks of all web pages with performance advantages over the traditional Page Rank algorithm. We present a novel approach for reducing the number of iterations performed in Page Rank algorithm to reach a convergence point.

A. The Proposed Page Rank Algorithm based on optimized normalization technique:

1) Initially assume PAGE RANK of all web pages to be any value, let it be1.

2) Calculate page ranks of all pages by following formula

PR(A) = .15 + .85 (PR(T1)/C(T1) + PR(T2)/C(T2) + ……. + PR(Tn)/C(Tn))

Where

T1 through Tn are pages providing incoming links to Page A

PR(T1) is the Page Rank of T1

PR(Tn) is the Page Rank of Tn

C(Tn) is total number of outgoing links on Tn

3) Calculate mean value of all page ranks by following formula :-

Summation of page ranks of all web pages / number of web pages

4) Then normalize page rank of each page

Norm PR (A) = PR (A) / mean value
Where norm PR (A) is Normalized Page Rank of page A and
PR (A) is page rank of page A
5) Assign PR(A)= Norm PR (A)
6) Repeat step 2 to step 4 until page rank values of two consecutive iterations are same.
The pages which have the highest page rank are more significant pages.

B. *Flow chart for proposed page rank algorithm*

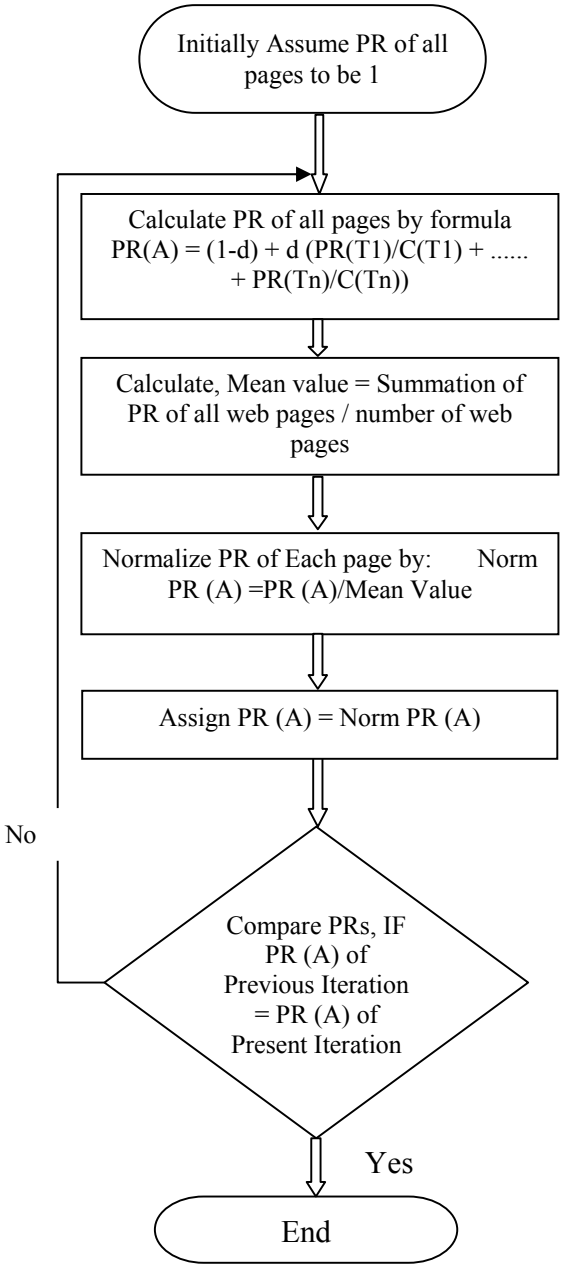

Figure 3, Flowchart for proposed Page Rank Algorithm

V. EXPERIMENTAL RESULTS

The implementation is performed on 3.06 GHz Pentium Dual Core PC with 3 GB RAM, running Windows 7. Java programming language is used; since it is an Object Oriented Language and has security packages. NetBeans IDE is an open source Integrated Development Environment which serves as a platform for implementation of Java based applications. In the implementation Java SE (Standard Edition) 6 Update 24 (released in February 15, 2011) and NetBeans IDE 6.9 (released in June 2010) has been used.

A. *Implementation details*

The proposed Page Rank algorithm is based on normalization scheme which uses mean value of page ranks. We have implemented the proposed algorithm on the www.icare.manit.ac.in website of MANIT.

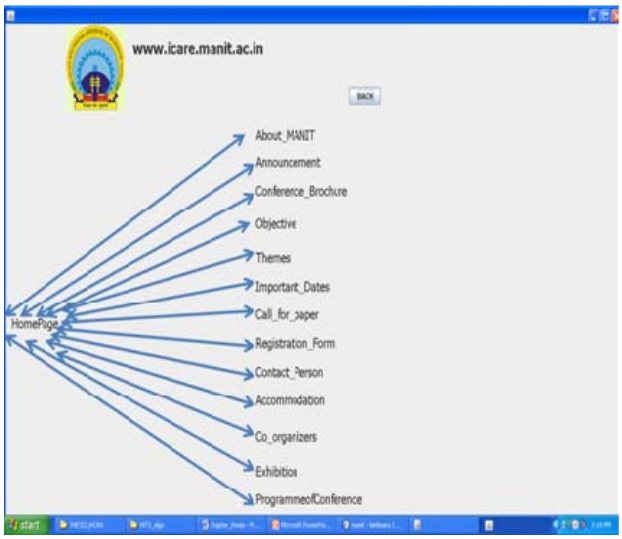

Figure 4, Web Graph for www.icare.manit.ac.in

B. *Result Analysis and Discussion*

The iterations recorded for the simulation of conventional PAGE RANK and the Proposed PAGE RANK algorithms to reach a convergence value is tabulated in Table 5.1. Since the number of iterations for calculating the page ranks in the Proposed Page Rank algorithm are reduced, therefore the time complexity of the proposed algorithm is less as compared to the conventional Page Rank algorithm.

No. of Iterations For Conventional PAGE RANK Algorithm	No. of Iterations For proposed PAGE RANK Algorithm
107	20

Table 1, Shows the no. of iterations performed, by the conventional PAGE RANK and the Proposed PAGE RANK algorithm.

Table 2, Shows the final Page Ranks, for different web pages.

WEB PAGES	PAGE RANKS
Home Page	6.51351351351351
About_MANIT	0.5758835758835756
Announcement	0.5758835758835756
Conference_Brochure	0.5758835758835756
Objective	0.5758835758835756
Themes	0.5758835758835756
Important_Dates	0.5758835758835756
Call_for_paper	0.5758835758835756
Registration_Form	0.5758835758835756
Contact_Person	0.5758835758835756
Accommodation	0.5758835758835756
Co_organizers	0.5758835758835756
Exhibition	0.5758835758835756
ProgrammeofConference	0.5758835758835756

C. Simulation Results

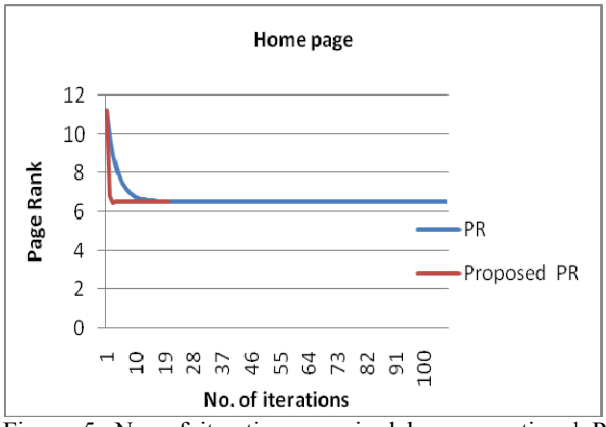

Figure 5, No. of iterations required by conventional Page Rank and Proposed Page Rank algorithm to reach a convergence value for Home Page.

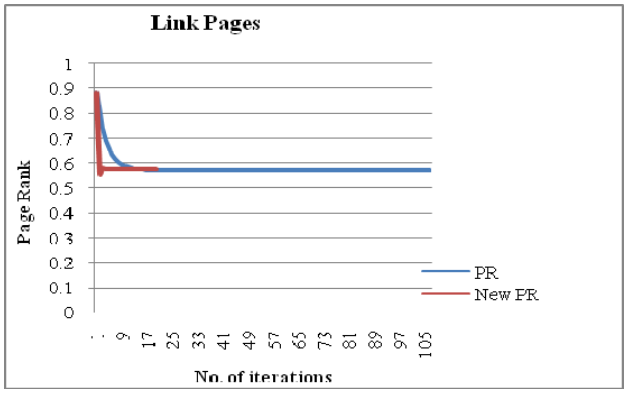

Figure 6, No. of iterations required by conventional Page Rank and Proposed Page Rank algorithm to reach a convergence value for Link Pages.

VI. CONCLUSION

In this paper an optimized page rank algorithm based on normalization technique has been proposed. In this proposed scheme the page rank of all web pages are being normalized by using a mean value factor, which reduces the time complexity of the conventional page rank algorithm. Comparative study of the computational characteristics of the proposed scheme with the previous works signifies that the proposed page rank algorithm is a better alternative to the previously introduced page rank algorithm as seen from the prospect of time complexity and the computational savings. In the future, the researchers can plan to explore more on the page rank algorithm based on damping factor to enhance the performance of the proposed scheme.

REFERENCES

[1] BHARAT, K., AND HENZINGER, M. Improved Algorithms for Topic Distillation in a Hyperlinked Environment. In Proceedings of ACM SIGIR'98 (Melbourne, Australia, 1998).

[2] A.K. Singh, Ravi Kumar and Alex Goh Kwang Leng "Efficient Algorithm for Handling Dangling Pages using Hypothetical node".

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

[4] S. Al-Saffar and G. Heileman, Experimental bounds on the usefulness of personalized and topic-sensitive pagerank, International Conference on Web Intelligence, pp. 671-675, 2007.

[5] A. Rungsawang, K. Puntumapon, B. Manaskasemsak, Un-biasing the link farm effect in pagerank computation, 21th International Conference on Advanced Networking and Applications, pp. 924-931, 2007.

[6] Cooper, C. Frieze A., "A general model of Web graphs", In ESA, 2001, pp. 500-511. CERN Common Log Format, http://www.w3.org/Daemon/User/Config/Logging.html# common-log file-format.

[7] CALADO, P., RIBEIRO-NETO, B., ZIVIANI, N., MOURA, E., AND SILVA, I. Local Versus Global Link Information in the Web. ACM Transactions on Information Systems 21, 1 (January 2003), 42–63.

[8] CRASWELL, N., CRIMMINS, F., HAWKING, D., AND MOFFAT, A. Performance and cost tradeoffs in web search. In ADC'04 (Dunedin, New Zealand, January 2004), pp. 161–170. http://es.csiro.au/pubs/craswell adc04.pdf.

[9] A. L. Barabasi and R. Albert, Emergence of scaling in random networks, Science Magazine, Vol. 286. no. 5439, pp. 509-512, 1999.

[10] Kleinberg, J. M. Authoritative sources in a hyperlinked environment, Journal of the ACM, Vol.46 (5). (Sept. 1999). 604-632.

[11] Lerman, K., Getoor, L., Minton, S., and Knoblock, C.Using the Structure of Web Sites for Automatic Segmentation of Tables. SIGMOD (2004) 119-130.

[12] Eiron, N. McCurley, K., and Tomlin, J. Ranking the web frontier. Proceedings of the international conference on World Wide Web, (WWW'04). Pp.309-318, 2004.

[13] C. Guo and Z. Liang, An improved BA model based on the pagerank algorithm, 4th WiCOM International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1-4, 2008.

[14] NG, A. Y., ZHENG, A. X., AND JORDAN, M. I. Link analysis, eigenvectors, and stability. In Proceedings of IJCAI'01 (Seattle, USA, 2001), ACM Press.

[15] Deng Cai, Shipeng Yu, and et al. Block-based Web Search. In Proceedings of ACM SIGIR'04, July 25-29, 2004, Sheffield, South Yorkshire, UK. 465-463.

[16] HAVELIWALA, T. H. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. In IEEE Transactions on Knowledge and Data Engineering (July 2003).

[17] Ziv Bar-Yossef and Sridhar Rajagopalan. Template Detection via Data Mining and its Applications. In: Proceedings of WWW2002, May 7-11, 2002, Honolulu, Hawaii, USA. 580-591.

[18] L. Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank citation ranking: Bring order to the web. *Technical report*, Stanford University, 1998.

[19] B. Liu. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer, 2006.

[20] L. Getoor, Link Mining: A New Data Mining Challenge. *SIGKDD Explorations*, vol. 4, issue 2, 2003.

[21] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, Link analysis: Hubs and authorities on the world. *Technical report: 47847*, 2001.

[22] J. Wang, Z. Chen, L. Tao, W. Ma, and W. Liu, Ranking user's relevance to a topic through link analysis on web logs, Proceedings of the Second Annual Conference on Communication Networks and Services research (CNSR'04), 2002, pp. 49–54.

[23] Xianchao Zhang, Hong Yu, Cong Zhang, and Xinyue Liu "An Improved Weighted HITS Algorithm Based on Similarity and Popularity", 2007 IEEE

[24] Sekhar Babu Boddu, V.P Krishna Anne, Rajesekhara Rao Kurra, Durgesh Kumar Mishra, " Knowledge Discovery and Retrieval on World Wide Web Using Web Structure Mining", 2010 Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation 978-0-7695-4062-7/10 $26.00 © 2010 IEEE DOI 10.1109/AMS.2010.108 532

[25] ABITEBOUL, S., PREDA, M., AND COBENA, G. Adaptive On-Line Page Importance Computation. In Proceedings of WWW2003 (Budapest, Hungary, May 2003).

[26] ARASU, A., NOVAK, J., TOMKINS, A., AND TOMLIN, J. PageRank Computation and the Structure of the Web: Experiments and Algorithms. In Proceedings of WWW2002 (Hawaii, USA, May 2002).

[27] LEMPEL, R., AND MORAN, S. (SALSA) the stochastic approach for link-structure analysis. ACM Transactions on Information Systems (2001).

[28] TIAN Chong "A Kind of Algorithm For Page Ranking Based on Classified Tree In Search Engine", 2010 International Conference on Computer Application and System Modeling (ICCASM 2010).

[29] Shiguang Ju, Zheng Wang, Xia Lv "Improvement of Page Ranking Algorithm Based on Timestamp and Link", 2008 International Symposiums on Information Processing.