ISSN22773061

Improved PageRank Algorithm for Web Structure Mining

S.Sathya Bama[1], M.S.Irfan Ahmed[2], A.Saravanan[3]
Centre for Computer Applications, Sri Krishna College of Technology, Coimbatore, India
ssathya21@gmail.com
Department of MCA, Hindusthan College of Engineering and Technology
Coimbatore, India
msirfan@gmail.com
Centre for Computer Applications, Sri Krishna College of Technology, Coimbatore, India
a.saravanan21@gmail.com

ABSTRACT

The growth of internet is increasing continuously by which the need for improving the quality of services has been increased. Web mining is a research area which applies data mining techniques to address all this need. With billions of pages on the web it is very intricate task for the search engines to provide the relevant information to the users. Web structure mining plays a vital role by ranking the web pages based on user query which is the most essential attempt of the web search engines. PageRank, Weighted PageRank and HITS are the commonly used algorithm in web structure mining for ranking the web page. But all these algorithms treat all links equally when distributing initial rank scores. In this paper, an improved page rank algorithm is introduced. The result shows that the algorithm has better performance over PageRank algorithm.

Indexing terms/Keywords

Inlinks, Outlinks, PageRank, Web Page, Web Structure Mining.

Academic Discipline And Sub-Disciplines

Web Mining, Web structure mining

SUBJECT CLASSIFICATION

Computer Science

TYPE (METHOD/APPROACH)

Empirical analysis; Experimental Approach

Council for Innovative Research
Peer Review Research Publishing System
Journal: INTERNATIONAL JOURNAL OF COMPUTER AND TECHNOLOGY
Vol 10, No. 9
editor@cirworld.com
www.cirworld.com, member.cirworld.com

INTRODUCTION

The World Wide Web is a collection of Web sites and its contents. In this highly competitive world and with the broad use of the Web, providing useful information and fulfilling users' needs are the primary goals of website owners [1]. Since the web contains huge amount of information and provides an access to it at any time and in any place, efficient searching for web contents becomes more essential. Most of portal sites usually have their search engines, which are used to find relevant Web contents for users' queries. For efficient responses to users' queries, many portal sites have their indexed databases with indexed words pointing to positions in Web pages. Search engines find relevant Web contents by seeking indexed words related to the query strings given by users. Usually crawlers are used by the portal sites to update their indexed databases [2]

Since a Web site has its homepage and all the contents linked with this home page, crawlers usually fetch the homepage of the Web site first and then obtain other Web contents by traversing referenced links within Web pages. Commonly used techniques to traversing referenced links are breadth first search and depth first search. Regardless of which techniques are used to traverse Web contents of a Web site, it is necessary to avoid traversing a Web content that has already been visited and fetched which is inefficient [3], [4]. So the number of links in the web page both the numbers of inlinks (links to a page) and number of outlinks (links from a page) are valuable information in web structure mining. This paper provides a new improved page rank algorithm.

The rest of this paper is organized as follows. An introduction about web mining is presented in next section. The description about web graph is given in section 3. A brief background review of web structure mining is presented in section 4. Section 5 presents the PageRank algorithm, a commonly used algorithm in WSM. An improved PageRank algorithm is described in Section 6. Implementation, evaluation and experimental results of improved PageRank algorithm are presented in Section 7. Finally section 8 summarizes the conclusion of the present study.

WEB MINING

To provide useful information to the user, analyzing users' patterns of behavior becomes increasingly important. Web mining is used to discover the content of the Web, the users' behavior in the past, and the web pages that the users want to view in the future [5]. Web mining consists of the following tasks [6]:

- Resource finding: The task of retrieving intended Web documents
- Information selection and pre-processing: Automatically selecting and pre-processing specific information from retrieved Web resources
- Generalization: Automatically discovers general patterns at individual Web sites as well as across multiple sites.
- Analysis: Validation and interpretation of the mined patterns

According to analysis targets, web mining are broadly classified into three different categories [7],[8], [9] which are Web Usage Mining, Web Content Mining and Web Structure Mining.

Web usage mining is the process of extracting useful information from server logs and finding out what users are looking for on the Internet. Web content mining is mining, extraction and integration of useful data, information and knowledge from Web page contents. It mainly focuses on the structure within a document. Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. It tries to discover the link structure of the hyperlinks between documents. Many algorithms have been proposed for hyperlink analysis. PageRank is a commonly used algorithm in WebStructure Mining. It measures the importance of the pages by analyzing the links [8], [10], [11]. Hypertext Induced Topic Search (HITS) ranks webpages by analyzing their inlinks and outlinks. In this algorithm, webpages pointed to by many hyperlinks are called authorities whereas webpages that point to many hyperlinks are called hubs [12], [13], [14], [15]. Weighted Page Rank assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its outlink pages. [5].

Thus, Web mining has been developed into an independent research area. WCM and WUM have been studied by many researchers who have achieved valuable results.

WEB GRAPH

The pages and hyperlinks of the World Wide Web can be viewed formally as directed graph as nodes and arcs, where the nodes correspond to collection of HTML files having page contents and the arcs correspond to hyperlinks interconnected with those HTML files. The Web Graph is a starting point to generate a structure of the Web that can be used for Web site designers and owners, search engines, Web crawlers and analysts [3], [16]. Fig 1 depicts the structure of the World Wide Web.

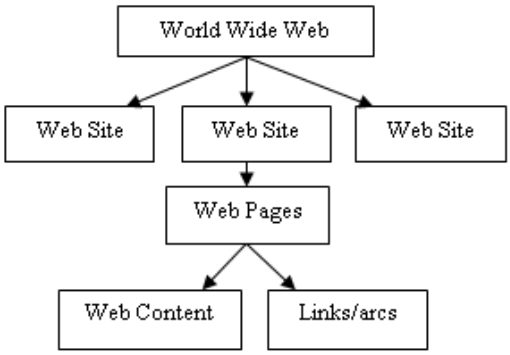

Fig 1: Structure of World Wide Web

Since a Web site consists of a homepage and many other Web contents linked each other, Web contents (including the homepage) of a Web site can be represented as a graph consisting of a set of nodes and associated arcs (links) [17], [18]. With this directed graph we can find a set of paths through which any Web page of the Web site can be accessed from the homepage. Fig 2 represents the web graph structure for a sample web site

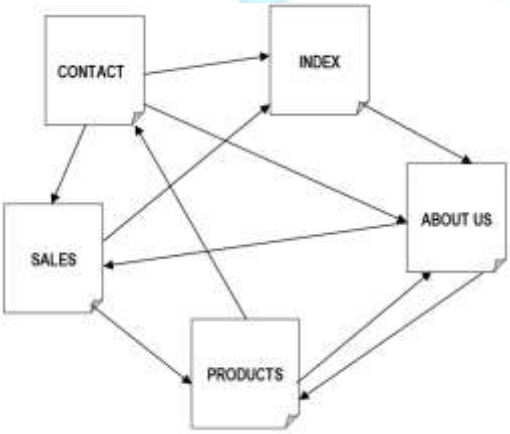

Fig 2: Web Graph structure for a sample web site

WEB STRUCTURE MINING

Web structure mining tries to determine the model underlying the link structures of the Web which 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 [19], [20]. 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 [6], [21]. There are two major link-based search algorithms, HITS (Hypertext Induced Topic Search) and PageRank. The basic idea of the HITS algorithm is to identify a small sub-graph of the Web and apply link analysis on this sub-graph to locate the authorities and hubs for the given query. The sub-graph that is chosen depends on the user query. The selections of a small sub-graph (typically a few thousand pages), not only focus the link analysis on the most relevant part of the Web, but also reduce the amount of work for the next phase. The main weaknesses of HITS are known to non-uniqueness and nil-weighting [22]. Google, which among search engines is ranked in the first place, uses the PageRank algorithm

PAGE RANKING ALGORITHM

The PageRank algorithm is one of the most extensively used ranking algorithms which states that if a page has important links to it, its links to other pages also become important. So backlinks are the number of pages that are linking to a particular web page. Therefore, PageRank takes the backlinks into account and propagates the ranking through links: a page has a high rank if the sum of the ranks of its backlinks is high [23][8].

Simplified PageRank

A simplified version of PageRank is defined as [8]:

$$PR_i = c\ (\sum_{j \in L_i} \frac{PR_j}{O_j}) \quad (1)$$

where PR_i, PR_j is the PageRank of page P_i and P_j, O_j is the number of outlinks from page P_j and L_i are the pages that link to page P_i. *c* is a factor used for normalization.

In PageRank, 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. The rank scores of pages of a website could be calculated iteratively starting from any webpage. Within a website, two or more pages might connect to each other to form a loop. If these pages did not refer to but are referred to by other webpages outside the loop, they would accumulate rank but never distribute any rank. This scenario is called a *rank sink* [5],[8].

PageRank

To solve the *rank sink* problem, the new PageRank algorithm had been proposed [8],[23].

The basic idea of PageRank is that '*a page is important, if other important pages link to it*'. This idea can be seen as a way of calculating the importance of pages by voting for them. Link from one page to another page may be considered as a vote. However, not only the number of votes a page receives is considered important, but the "importance" or the "relevance" of the ones that cast these votes as well. Similarly, not all votes are equally important. A vote from a page with low importance (i.e. it has few inlinks2) should be worth far less than a vote from an important page (with thousands of inlinks) [9], [24].

For n pages P_i, i = 1, 2,.., n the corresponding PageRank is set to PR_i, i = 1, 2 ,.., n. The mathematical formulation for the recursively defined PageRank are presented in Eq. (2).

$$PR_i = (1-d) + d(\sum_{j \in L_i} \frac{PR_j}{O_j}) \quad (2)$$

where PR_i is the PageRank of page P_i, O_j is the number of outlinks from page P_j and L_i are the pages that link to page Pi and *d* is a damping factor which takes values between 0 and 1.

The sum of PageRanks of all web pages will be the number of web pages on the web. The PageRank of a page can be calculated without knowing the final value of PageRank of other pages. PageRank of a page depends on the number of pages pointing to a page. Page Rank algorithm starts with an arbitrarily guessed vector r (e.g. a vector of ones, all divided with number of pages present), that describes the initial PageRank value PRi for all pages Pi usually it is set to 1. Then it iterates the recursive formula until two consecutively iterated PageRank vectors are similar enough.

Assume any arbitrary page A has pages T1 to Tn pointing to it (incoming link). PageRank formula can be given as:

$$PR(A) = (1-d) + d(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) \quad (3)$$

here *PR(Ti)* is the PageRank of the Pages Ti which links to page A, *C(Ti)* is number of outlinks on page Ti and the value of the damping factor is 0.85.

To test the utility of the PageRank algorithm, Google applied it to the Google search engine [8]. In the experiments, the PageRank algorithm works efficiently and effectively because the rank value converges to a reasonable tolerance in the roughly logarithmic (log *n*) [8], [23]. The rank score of a web page is divided evenly over the pages to which it links. [5].

Weighted Page Rank Algorithm

The more popular webpages are, the more linkages that other webpages tend to have to them or are linked to by them. The Weighted PageRank Algorithm–assigns larger rank values to more important (popular) pages instead of dividing the rank value of a page evenly among its outlink pages. Each outlink page gets a value proportional to its popularity (its number of inlinks and outlinks).

The popularity from the number of inlinks and outlinks is recorded as $W^{in}_{(v,u)}$ and $W^{out}_{(v,u)}$ respectively.

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

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{p \in R(v)} I_p} \quad (4)$$

where I_u and I_p represent the number of inlinks 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 outlinks of page *u* and the number of outlinks of all reference pages of page *v*.

$$W^{out}_{(v,u)} = \frac{O_u}{\sum_{p \in R(v)} O_p} \quad (5)$$

where O_u and O_p represent the number of outlinks of page *u* and page *p*, respectively. *R*(*v*) denotes the reference page list of page *v*.

Considering the importance of pages, the original PageRank formula is modified as

$$PR_i = (1-d) + d\sum_{j \in Li} PR_j (W^{in}_{(i,j)})(W^{out}_{(i,j)}) \quad (6)$$

IMPROVED PAGERANK ALGORITHM

In page rank algorithm initially all the pages will have the same rank and it treat all links equally when distributing rank scores. This leads to large number of iteration to find the final page rank. So, instead of assigning the same value (usually one) to all the web pages, the proposed algorithm assigns the initial pagerank by making use of number of inlinks and outlinks of that page, as well as it gives importance to the inlinks than outlinks based on the fact that 'a page is important, if other important pages link to it'. This method reduces the number of iterations which leads to faster calculation.

The initial pagerank is given by the formula:

$$PR_{j(0)} = \frac{2(2I_j + O_j)}{\sum_{k \in R(Pj)} (I_k + O_k)} \quad (7)$$

Where $PR_{j(0)}$ is the page rank of the page P_j; I_j, O_j are inlink and outlink of P_j; I_k, O_k are inlink and outlink of P_k; *R*(P_j) denotes the reference page list of page P_j along with P_j.

Here the sum of initial page rank of web pages will be equal to the number of pages involved while calculating the page rank.

The Improved Page Rank Algorithm is given below.

Algorithm: Improved Page Rank

1. for each page j=0,1…n calculate the page rank
2. $PR_{j(0)} = \frac{2(2I_j + O_j)}{\sum_{k=1}^{n} (I_k + O_k)}$
3. end for
4. do
5. for each page k, k=0,1…n calculate the page rank recursively
6. $PR_{i(k)} = (1-d) + d(\sum_{j \in Li} \frac{PR_j}{O_j})$
7. end for
8. while $PR_{i(k)}=PR_{i(k-1)}$

With this initial Page Rank, the Page rank Algorithm can be carried out which minimizes the number of iterations

IMPLEMENTATION

Let us take an example of hyperlink structure of three pages A, B and C as shown in Fig 3.

ISSN22773061

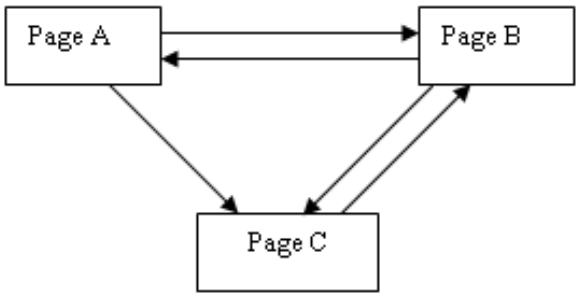

Fig 3: Hyperlink structure of three pages

To calculate the page rank of the web page the input i.e., the number of nodes and the number of outgoing links can be entered as input.

Here IA= 1, OA=2, IB=2, Ob=2, IC=2, OC=1, where IA, OA is the number of inlink and outlink to A. with this the Initial Page Rank for pages A, B and C are calculated by using Eq.(3).

$$PR_{A(0)} = \frac{2*((2*1)+2)}{(1+2+2+2+2+1)} = 8/10 = 0.8$$

$$PR_{B(0)} = \frac{2*((2*2)+2)}{(1+2+2+2+2+1)} = 12/10 = 1.2$$

$$PR_{C(0)} = \frac{2*((2*2)+1)}{(1+2+2+2+2+1)} = 10/10 = 1.0$$

Next the equation Eq. (2) calculates the page rank of the web page in successive iterations. The iteration is shown in the Fig. 3. The sample calculation is shown below.

PR(A) = (1-d) + d * (PR(B)/C(B)) =(1-0.85)+0.85(1.2/2) = 0.660

PR(B) = (1-d) + d * [(PR(A)/C(A)) + (PR(C)/C(A)]

= (1-0.85) + 0.85[(0.660/2)+(1/1)] = 1.281

PR(C) = (1-d) + d * [(PR(A)/C(A)) + (PR(B)/C(B)]

= (1-0.85) + 0.85[(0.660/2)+(1.281/2)] = 0.975

After doing many more iterations of the above calculation, the final Page Ranks will be arrived. The computation of the page rank is shown in Fig. 4. and the computation of Improved page rank is shown in Fig. 5.

Iteration	PR(A)	PR(B)	PR(C)
1	1	1	1
2	0.575	1.244	0.923
3	0.679	1.223	0.958
4	0.670	1.249	0.966
5	0.681	1.261	0.975
6	0.686	1.270	0.981
7	0.690	1.277	0.986
8	0.693	1.283	0.990
9	0.695	1.287	0.992
10	0.697	1.289	0.994
11	0.698	1.292	0.996
12	0.699	1.294	0.997
13	0.700	1.295	0.998
14	0.700	1.296	0.998
15	0.701	1.296	0.999
16	0.701	1.297	0.999

Fig 4: Iterations of PageRank Algorithm

Iteration	PR(A)	PR(B)	PR(C)
1	0.800	1.200	1.000
2	0.660	1.281	0.975
3	0.694	1.274	0.986
4	0.691	1.282	0.989
5	0.695	1.286	0.992
6	0.697	1.289	0.994
7	0.698	1.292	0.996
8	0.699	1.294	0.997
9	0.700	1.295	0.998
10	0.700	1.296	0.998
11	0.701	1.296	0.999
12	0.701	1.297	0.999

Fig 5: Iterations of Improved PageRank Algorithm

In the above table, Page Rank of B is higher than Page Rank of A and C. It is because Page B has 2 incoming links and 2 outgoing links as shown in Fig. 3. Page C has 2 incoming links and 1 outgoing link. Page A has the lowest Page Rank because Page A has only one incoming link and 2 outgoing links. So the link analysis becomes very important in ranking the page. From the Fig 4, after the iteration 16, the Page Rank for all the three pages gets normalized. But from Fig 5, i.e., in proposed algorithm, after the 12^{th} iteration the page rank gets converged. PageRank and the Improved PageRank Algorithms are implemented in an Intel Core 2 (2.40 Ghz) with 4GB RAM.

CONCLUSION

The main purpose of this study is to explore the hyperlink structure and understand the Web graph in a simple way. The Page Rank computation results shows that the incoming links and the outgoing links play an important role in ranking of Web pages using link analysis. The further work on this area will be sustained by analyzing the problems that are faced by these algorithms and to identify the solution for those problems in an efficient way.

REFERENCES

[1] Rekha Jain , Dr. G. N. Purohit: Page Ranking Algorithms for Web Mining, International Journal of Computer Applications (0975 – 8887), Volume 13– No.5, January 2011.

[2] Wookey LEE, Hierarchical Web Structure Mining, Proceedings of Data Engineering Workshop(DEWS), 2006.

[3] [Pandurangan, G., Raghavan, P. and Upfal, E. Using PageRank to Characterize Web Structure. COCOON, 330-339, 2002.

[4] Wookey Lee, Seung Kim, Sukho Kang: Structuring Web Sites Using Linear Programming. EC-Web, 328-337, 2002

[5] Wenpu Xing and Ali Ghorbani, Weighted PageRank Algorithm, Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04), 0-7695-2096-0/04, IEEE, 2004.

[6] R. Kosala and H. Blockeel. Web mining research: A survey. ACM SIGKDD Explorations, 2(1):1–15, 2000.

[7] S. Madria, S. S. Bhowmick, W. K. Ng, and E.-P. Lim. Research issues in web data mining. In Proceedings of the Conference on Data Warehousing and Knowledge Discovery, pages 303–319, 1999.

[8] 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.

[9] S. Pal, V. Talwar, and P. Mitra. Web mining in soft computing framework : Relevance, state of the art and future directions. IEEE Trans. Neural Networks, 13(5):1163–1177, 2002.

[10] Brin, S. and L. Page, The anatomy of a large scale hypertextual web search engine. Comput. Network ISDN Syst., 30: 107-117. DOI: 10.1016/S0169-7552(98)00110-X, 1998.

[11] Erik Andersson, Per-Anders Ekström, Investigating Google's PageRank algorithm, Report in Scientifc Computing, advanced course - Spring 2004.

[12] J. Wang, Z. Chen, L. Tao, W. Ma, and W. Liu. Ranking user's relevance to a topic through link analysis on web logs. WIDM, pages 49–54, 2002.

[13] Kleinberg, J., Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604–632, September 1999.

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

[15] 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.

[16] Faloutsos, M., Faloutsos, P., Faloutsos, C. On Power-law Relationships of the Internet Topology. In SIGCOMM '99. pp.251-262. 1999

[17] Garofalakis, J., Kappos, P. and Mourloukos, D. Web Site Optimization Using Page Popularity, IEEE Internet Computing, 3(4): 22-29, 1999

[18] Wookey, L., Geller, J. Semantic Hierarchical Abstraction of Web Site Structures for Web Searchers, Journal of Research and Practice in Information Technology, 36(1), pp.71-82, 2004.

[19] 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

[20] Gyongyi, Z., Garcia-Molina, H., Pedersen, J. Combating Web Spam with TrustRank. VLDB, pp.576-587, 2004

[21] Wookey Lee: Discriminating Biased Web Manipulations in Terms of Link Oriented Measures. In ISCIS, 585-594, 2005.

[22] Haveliwala, T. Topic-Sensitive PageRank: A Context- Sensitive Ranking Algorithm for Web Search, IEEE TKDE.15(4), pp.784-796, 2003

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

[24] P. Ravi Kumar and Ashutosh Kumar Singh, Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval, American Journal of Applied Sciences 7 (6): 840-845, 2010.

Author' biography

S.Sathya Bama received Master of Computer Applications Degree from Anna University, Chennai, India. Currently she is working as an Assistant Professor in the Department of MCA in Sri Krishna College of Technology, Coimbatore, India. She has 4 years of experience in both teaching and research. She has presented and published research papers in varies international journals & conferences.

Dr. M.S.Irfan Ahmed holds his PhD degree in computer science from the Alagappa University, Karaikudi, India. He has more 2 years of industry experience, 16 years of teaching experience and 10 years of research experience. Currently he is working as Director in the Department of MCA in Hindusthan College of Engineering and Technology, Coimbatore, India. He received the best faculty award in the year 2012. He has presented and published more than 20 research papers in international journals & conferences.

A.Saravanan received his Master of Computer Applications Degree and M.Phil Degree from Bharathiar University, Coimbatore, India. Currently he is working as an Assistant Professor in the Department of MCA in Sri Krishna College of Technology, Coimbatore, India. He has 14 years of experience in teaching and 4 years in research. He has presented and published research papers in various international journals.