International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878
Volume 2 Issue 8, August 2014

Advanced Mining Technique for Improving the Order of Web Data Search in Online Communities

S. Nagarjuna Reddy[1], Dr. J. Janet[2]

[1]Pursuing M. Tech, Computer Science and Engineering, Sri Venkateswara College of Engineering and Technology, Chittoor-517127, Andra Pradesh, India
[2]Professor Computer Science and Engineering, Sri Venkateswara College of Engineering and Technology, Chittoor-517127, Andra Pradesh, India

Abstract: ***Advanced mining technique is for improving the order of web data search in online communities in different platforms, like business and enterprises, education communities. The Web database generally contain huge amount of data and thousands of blogs and millions of web pages. The scope of this technique is to decrease the noise of data and the unwanted evidence data list. The weighted Page Ranking Technique is proposing to overcome the huge amount of ambiguity information like as relevance and reputational of a keyword person name for a query topic. Weighted Page Ranking algorithm (WPR) is the modified algorithm of the Page Ranking algorithm. Weighted Page Ranking (WRP) gives ranking based on the popularity of the person name or pages by takes to consideration importance of In and Out links of pages and blogs etc. Weighted page ranking algorithm is mainly two classes "Web structure mining" and "web content mining" combination. Weighted Page Ranking Algorithm reduces search time and weighted content pages are tending to move upwards in the searching result list.***

Keywords: Web mining, Expert Search, Weighted Page Rank, Web Content and Structure mining.

1. Introduction

World Wide Web is most popular information source for all kinds of data like text, audio, video, and metadata. We appraise general searching problem, searching on web were millions of WebPages and huge amount of data in all pages and blogs. Web pages are varying full noise and unwanted data. Therefore, we use Advance Weighted Page Ranking algorithm for extracting the exact information on web mining. The Web mining techniques along with other important areas database, Natural Language Processing (NLP), Information Retrieval etc. There is many search engines are there as (Google, Yahoo, Bing, etc) used to extract information from the World Wide Web. Web mining consists of three important aspects that are Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM).

WCM is the process of obtaining appropriate information from the Web document contents. WCM connect to data mining because many data mining techniques are used in web content mining. WCM is also used to connect text mining because most of the web contents are text based. WCM is also used in different text mining, because the nature of web data is semi structure where as the nature of text mining is unstructured. Web content mining can be categorize into two different points of view:

Information Retrieval (IR) and Database (DB) views, Information Retrieval: The main aim of WCM from the information retrieval view is mainly to assist or to improve the information finding or filtering the information to the users usually based on either inference or seek user profiles.

Database view: The goal of WCM from the Database views is to model the data on the Web and to integrate them so that more advanced queries other than the keywords based search could be accomplished. Based on the kind of structure information old, it is subdivided into two kinds. They are Hyperlinks and Document Structure.

Web Usage Mining (WUM)

For the reason that Web is the connection between Web users and Web pages, we need to concern about the navigational behavior of web users during web mining. WUM is able to capture analysis and model the interaction between users and pages during browsing, which in turn provides complementary assistance for advanced web applications. Web usage Mining uses the data mining techniques to discover interesting usage patterns from web data, and understands and simply to better serves the needs of web-based application. Since it records the browsing behavior of site visitors, web server log plays an important role in performing web Usage Mining. The data recorded in server logs gives the information regarding the web site access by multiple users.

Web Structure Mining is the process of deducing knowledge from the WWW and links between references and referents in the Web. A typical web graph structure consists of web pages and hyperlinks. Web pages are called nodes and hyper links are called edges, which connects related web pages. WSM uses graph theory to analyze the node and connection structure of a web site. This is also used to discover structure information from the web. Based on the kind of structure information used, it is divided into two kinds. They are Hyperlinks and Document Structure.

2. Related Work

In Web mining process most the search engines are used Page ranking algorithm (PRA), which can place the documents in order of their relevance, co-occurrence

importance and content score(page rank). Some search engines are used Mining techniques such as classification, clustering (grouping), and association. Most of the page ranking algorithms proposed in their literature such as Page content rank, HITS, Clever, Page Rank, and Weighted Page Ranking. Researchers have notice using additional information to improve the retrieval performance, such as in degree, and Page Rank, person-person similarity, query expansion, and related feedback using person names, concurrence between occurrence of query words and person's names. Expert search is turn in to a hot research topic since start of the TREC enterprise track in 2005.

Our work is also related to expert search using advanced weighted Page algorithm. Numerous models have been proposed rank to persons by their level of expertise with any respected to any topic. A factor is used to evaluate the strength of the association between a document and persons. All search engines 30% of search engines queries with person names creating ranking for persons for ambiguous queries. The exploration of structured scrutiny had increased and those effort result research area is called link mining, which is determine at the edge of the work in link analysis, hypertext and web mining. There are two basic algorithms have been proposed to lead that possible co relevant: Page ranking algorithms Google Page Ranking Algorithm (PRA) is proposed by Brain and Page in 1998, HITS (Hypertext includes topic selection) is proposed by Kleinberg in 1998. These two algorithms give equal weight to data or names or pages for all links for result's the rank score. The goal of this technique is to improve analyzing of that inner social accumulation of the web for link mining in conclave enlargement.

3. Proposed System

Wenpu Xing and Ali Ghorbani propose weighted Page Ranking Algorithm (WRP). This algorithm is modification of its original algorithm Page Rank. Weighted Page Rank algorithm is to gives ranks according to their popularity of pages, names by the impotence of both in and out links of pages. This algorithm gives top value of the rank to the most popular pages and it does not same, the rank of a page among it is out link page. Every outline page provides a rank score based on how much that page is popular. The main factor of the polarity is the count of number of in links and out links. The testing of WPR is using different blogs and websites and future task is to calculating the score for ranking by using more number of levels of a reference pages index and escalation the number of users to allocate the web pages.

4. Architecture

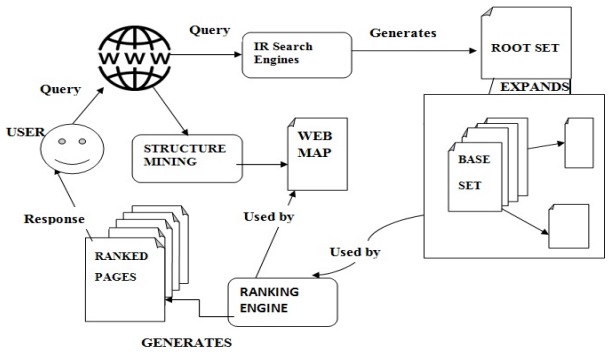

Figure 1: Architectural components of the system used to implement and evaluate the WPR algorithm

5. Algorithm

The important allocated in terms of weigh values for in links and out links are $W^{in}(m, n)$ and $W^{out}(m, n)$ respectively. (m,n) calculated base on the number of incoming links of a page *n and m* is denoted for the number of incoming likes of all reference pages of page *m.*

$$W^{in}(m, n) = In / \Sigma Ip \quad (1)$$

Where $P \in Re(m)$

$$W^{out}(m, n) = On / \Sigma Op \quad (2)$$

Where In and Ip denote the number of incoming links with respect to page n and page p. $Re_{(m)}$represents the all reference pages list of page m. Similarly computation performed for $W^{out}(m, n)$ as shown in equation (2) is the weight of link (m, n) which is depend on the number of outgoing links of page n and the number of outgoing links of all the reference pages of m. Whereas On and Op are the number of outgoing links with respect to page n and p. The formula as proposed for the WPR is as shown in equation (3) which is a modification of the Page Rank formula.

$$WPR(n) = (1-d) + d \Sigma WPR(m) W^{in}(m,n) W^{out}(m,n) \quad (3)$$

where $m \in B(n)$

WPR calculation calculated for the same hyperlink structure as shown in Figure 5. The WPR Equation for Page A, B, C, and D are as follows.

$$WPR(A) = (1-d) + d \Sigma WPR(B) W^{in}_{(B,A)} W^{out}_{(B,A)} + WPR(C) W^{in}_{(C,A)} W^{out}_{(C,A)} + WPR(D) W^{in}_{(D,A)} W^{out}(D,A) \quad (4)$$

So for getting the value of WPR (A), before it we will calculate the value of incoming links and Outgoing links weight as bellow:

$$W^{in}(B, A) = IA/(IA+IC)$$
$$= 3/(3+2)$$
$$= 3/5 \quad (5)$$
$$W^{out}(B,A) = OA/(OA+OC+OD)$$
$$= 2/(2+3+1)$$
$$= 1/3 \quad (6)$$
$$W^{in}(C,A) = IA/(IA+IB)$$
$$= 3/(3+2)$$
$$= 3/5 \quad (7)$$

$W^{out}(C,A) = OA/(OA+OB+OD)$
$= 2/(2+3+1)$
$= 2/6$
$= 1/3$ (8)
W^{in}
$(D,A) = IA/(IB+IC)$
$= 3/(2+2)$
$= ¾$ (9)
W^{out}
$(D,A) = OA/OA$
$= 2/2$
$= 1$ (10)

Now these in links and out links weight, equation numbers (5, 6, 7, 8, 9, 10) are put in the Equation (4) to calculate the weighted rank of the nodes A, B, C, and D as following:

$WPR(B) = (1-d) + d \Sigma WPR(A) W^{in}(A,B) W^{out}(A,B) + WPR(C) W^{in}(C,B) W^{out}(C,B)$ (11)
$WPR(C) = (1-d) + d \Sigma WPR(A) W^{in}(A,C) W^{out}(A,C) + WPR(B) W^{in}(B,C) W^{out}(B,C)$ (12)
$WPR(D) = (1-d) + d \Sigma WPR(B) W^{in}(B,D) W^{out}(B,D) + WPR(C) W^{in}(C,D) W^{out}(C,D)$ (13)

For WPR(A) calculation the value of d is set to 0.85(standard value) and the initial values of WPR(B), WPR(C) and WPR(D) is considered 1, so calculation for 1st iteration as follows:

WPR (A) = (1- 0.85) + 0.85(1* 3 / 5 *1/ 3 +1* 3 / 5 *1/ 3 +1* 3 / 4 *1) = 1.127 (14)

$W^{in}(A,B) = IB/(IB+IC+ID)$
$= 2/(2+2+2)$
$= 2/6$
$= 1/3$ (15)
$W^{out}(A,B) = OB/(OB+OC)$
$= 3/(3+3)$
$= 3/6$
$= ½$ (16)
$W^{in}(C,B) = IB/(IA+IB)$
$= 2/(3+2)$
$= 2/5$ (17)
$W^{out}(C,B) = OB/(OA+OB+OD)$
$= 3/(2+3+1)$
$= 3/6$
$= 1/2$ (18)

Again now for calculation of WPR (B) these equations (15, 16, 17, and 18) are put in to equation (11).

In this the initial value of WPR(C) is set to 1.

WPR (B) = (1- 0.85) + 0.85(1.127*1/ 3*1/2 + 1*2/5 *1/ 2)
= (0.15) + 0.85(1.127*0.33*0.50+1*0.40*0.50)
= 0.4989 (19)
$W^{in}(A,C) = IC/(IB+IC+ID)$
$= 2/(2+2+2)$
$= 2/6$
$= 1/3$ (20)
$W^{out}(A,C) = OC/(OB+OC)$
$= 3/(3+3)$
$= 3/6$
$= 1/2$ (21)
$W^{in}(B,C) = IC/(IA+IB)$
$= 2/(3+2)$
$= 2/5$ (22)
$W^{out}(B,C) = OC/(OA+OC+OD)$
$= 3/(2+3+1)$
$= 3/6$
$= 1/2$ (23)

By substituting the values of equations (14), (19), (20), (21), (22) and (23) to equation (12), you will get the WPR of Page C by taking d as 0.85.

WPR(C) = (1 - 0.85) + 0.85((1.127 *1/ 3 *1/ 2) + (0.499 * 2 / 5 *1/ 2))
= (0.15) + 0.85((1.127*0.33*0.50) + (0.499 * 0.40 * 0.50))
= 0.392 (24)
$W^{in}(B,D) = ID/(IB+IC)$
$= 2/(2+2)$
$= 2/4 = ½$ (25)
$W^{out}(B, D) = OD/OA$
$= 2/2$
$= 1$ (26)
$W^{in}(C,D) = ID/(IA+IB)$
$= 2/(2+3)$
$= 2/5$ (27)
$W^{out}(C,D) = OD/(OA+OB+OD)$
$= 2/(2+3+1)$
$= 2/6$
$= 1/3$ (28)

Again by substituting the values of equations (19), (24), (25), (26), (27) and (28) to equation (13), you will get the WPR(D) by taking d as 0.85.

WPR(D) = (1- 0.85) + 0.85((0.499 *1/ 2 *1) + (0.392 * 2 / 5*1/ 3))
= (0.15) + 0.85((0.499 *0.50 *1) + (0.392 * 0.40*0.33))
= 0.406 (29)

The values of WPR(A), WPR(B), WPR(C), WPR(D) are demonstrated in equations (14), (19), (24) and (29) consequence. The association between these are WPR(A) > WPR(B) > WPR(D) WPR(C).

Iteration	A	B	C	D
1	1	1	1	1
2	1.1275	0.47972	0.3912	0.19935
3	0.425162	0.27674	0.25727	0.18026
4	0.355701	0.244128	0.24189	0.177541
5	0.34580	0.247110	0.239808	0.17719
6	0.34454	0.23957	0.23953	0.17714
7	0.34438	0.23950	0.23950	0.17714
8	0.34436	0.23950	0.23949	0.17714

Iterative calculations values for weighted page rank

6. Experiments

To evaluate the WRP algorithm, we implemented WPR and the standard Page Rank algorithms to compare their results. The different Components involved in the execution and assessment of the WPR algorithm are illustrated in Figure 1.It consist of six major activities to be carry out in order to the perform simulation studies in this work.

1. *Finding a web site*: The standard Page Rank and the WPR Algorithm are relying on the web structures so it is significant to find out a website with rich hyperlinks. The Website of Saint Thomas University, in Fredericton, has been chosen after comparing the structures of the various websites.
2. *Building a web map*: A free spider software–J Spider–is used to generate the web map because the website does not consist of the required Web map.
3. *Finding the root set*: Using the IR search Engine, which is encapsulated in the website a set of the pages called root set has to be retrieved relevant to the given Query.
4. *Finding the base set*: By expanding the root set with pages which directly points to or pointed to the pages in the root set, then a base set is created
5. *Applying algorithms*: Applying of Standard Page Rank and WPR algorithm to the base set.
***6. Evaluating the results*:** Execute the algorithm by comparing their results.

7. Evaluation

To Evaluating the Standard Page Ranking and Weighted Page Ranking algorithm using the "Travel Agent" and "Scholarship" query topics. Travel agent" represents a non-focal point whereas "scholarship" represents a focal (popular) point in the Website of Saint Thomas University. The results of the evaluation are summarizing in the following subsections.

6.1. The determination of the relevancy of the pages to the given query

The Standard Page Rank and the WPR algorithms provide important information about a given query by using the structure of the website. We categorized the pages in the results into four classes based on their applicability to the given query:

• **Very Relevant pages (VR)**, which contain very important information about the given query,
• **Relevant pages (R)**, which have relevant but not important information about the given query,
• **Weak-Relevant pages (WR)**, which do not have relevant information about the given query even though they contain the keywords of the given query, and
• **Irrelevant pages (IR)**: which include neither the keywords of the given query nor relevant information about it.

6.2. The Calculation of the relevancy of the page lists to the given query

The performances of the WPR and the standard Page Rank algorithms have been evaluated to identify the algorithm that produces better results.

Table 1: the relevancy values for the query "travel agent" produced by PageRank and WPR using different page sets

	Number of Relevant Pages		Relevancy Value(k)	
Size of page set	PageRank	WPR	PageRank	WPR
10	0	1	0.1	0.5
20	4	3	13.1	16.8
30	4	4	47.1	49.8
40	4	4	82.1	84.8
50	4	4	117.1	119.8
60	5	5	159.6	162.3
70	7	7	211.7	214.4

6.3. Focused Topic Queries

This subsection evaluates the results obtained for the query "scholarship." The relevancy values of the results are shown in Table.2.Similar to the query "travel agent," (larger relevancy values) for the query "scholarship." Moreover, the two points derived from the query "travel agent" are shown more clearly in this case (see Table 2).

In conclusion, the results obtained from WPR and standard PageRank for the focused and non-focused topics show That WPR is superior to standard PageRank. In this utilise, we make a hierarchic tip of people's cant and leave the being identification difficulty to users. With a returned figure itemise, users can identify experts by searching their defamation together with the ask substance finished a web examine engine. We also use a set of calumny extracted from DBLP to conductor the repute extraction difficulty, which is certainly a main investigate job.

Table 2: The relevancy values for the query "scholarship" produced by PageRank and WPR using Different page sets

	Number of Relevant Pages		Relevancy Value(κ)	
Size of the page set	PageRank	WPR	PageRank	WPR
5	2	3	2	5.5
10	2	4	9.5	22
20	4	4	34.5	57
30	8	5	87.5	99
40	10	8	158.5	159.3
80	16	15	624.8	655.3
100	22	19	999.2	1045.3
120	25	20	1470.4	1473.3

6.3 Rank Updater

In this module, rank score of the returned page is improved by applying the input of the query processor and matched documents of a user query. It operated online and applied the improvements to the concerned documents.

Step 1: Given an input query q and matched documents D collected from the query processor, the webpage is found to which the query q belongs.

Step 2: The level weight are calculated for every page X present in the sequential pattern.
Step 3: The rank are calculated for every page X present in the sequential pattern. The improved is calculated as the summation of pervious rank and assigned weight value.
Due to the optimization of the Search engine results, the rank will improve so that it will serve the user need by providing the popular and relevant pages upwards in the result list.

8. Experimental Result

Table 3: Experimental result

	Number of Relevant Pages		Relevancy Value(X)		New Relevancy Value
Size of the paper set	PageRank	WPR	PageRank	WPR	(PageRank + WPR)
10	0	1	0.1	0.5	0.6
20	4	3	13.1	16.8	29.9
30	4	4	47.1	49.8	96.9
40	4	4	82.1	84.8	166.9
50	4	4	117.1	119.8	236.9
60	5	5	159.6	162.3	321.9

9. Conclusion

Web mining is used to retrieve information from users' past behavior. In this approach, Web structure mining plays a major role. Two commonly used algorithms in web structure mining are HITS and Page Rank, which are used to rank the relevant pages. When distributing the rank scores both algorithms will be treat the links equally. To improve the performances of this method so many algorithms are introduce. This paper introduces the WPR algorithm, an extension to the Page Rank algorithm. Based on the importance of the in links and out links of the pages the rank scores are distributing using the popularity of the pages in the WPR algorithm. Saint Thomas University shows that WPR is able to identify a larger number of relevant pages to a given query compared to standard Page Rank. This algorithm is improving the order of the page in the result list so that the user gets the relevant and important pages in the list.

References

[1] N. Duran, A. K. Sharma and K. K. Bhatia, "PageRanking Algorithms: A Survey, Proceedings of the IEEE International Conference on Advance Computing, 2009.
[2] R. Kosala, H. Blockeel, "Web Mining Research: Survey", SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 15, 2000.
[3] R. Cooley, B. Mobasher and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web," Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, 1997.
[4] M. G. da Gomes Jr. and Z.Gong, "Web Structure Mining: An Introduction", Proceedings of the IEEE International Conference on Information Acquisition, 2005.
[5] A. Broder, R. Kumar, F Maghoul, P. Raghavan, S.Rajagopalan, R. Stata, A. Tomkins, J. Wiener, "Graph Structure in the Web", Computer Networks: The International Journal of Computer and telecommunications Networking, Vol. 33, Issue 1-6 2000.
[6] J. Kleinberg, R. Kumar, P. Raghavan, P. Rajagopalan and A. Tompkins, "Web as a Graph: Measurements, Models and methods," Proceedings of the International Conference on Combinatorics and Computing, 18, 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, 1999.
[8] W. Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proc. of the Second Annual Conference on Communication Networks and Services Research, IEEE.
[9] C. Ridings and M. Shishigin, "PageRank Convered", Technical Report, 2002.
[10] S. Brin and L. Page. The anatomy of a large-scale hypertextual.
[11] Naresh Barsagade, "Web usage mining and pattern discovery: A survey paper".CSE8331, Dec, 2003.
[12] 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.
[13] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In *Proceedings of 17th International Conference on Machine Learning*, pages 167–174. Morgan Kaufmann, San Francisco, CA, 2000.
[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] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, September 1999.
[16] R. Kosala and H. Blockeel. Web mining research: A survey. *ACM SIGKDD Explorations*, 2(1):1–15, 2000.
[17] 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.
[18] 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.
[19] S. Pal, V. Talwar, and P. Mitra. Web mining in soft computing.
[20] Framework: Relevance, state of the art and future directions. *IEEE Trans. Neural Networks*, 13(5):1163–1177, 2002