ISSN: 2229-6956(ONLINE)
DOI: 10.21917/ijsc.2012.0053

EVALUATION OF WEB SEARCHING METHOD USING A NOVEL WPRR ALGORITHM FOR TWO DIFFERENT CASE STUDIES

V. Lakshmi Praba[1] and T. Vasantha[2]
[1]*Department of Computer Science, Government Arts College for Women, India*
E-mail: vlakshmipraba@rediffmail.com
[2]*Department of Computer Science, Manonmaniam Sundaranar University, India*
E-mail: vasanthasankarganesh@gmail.com

Abstract

The World-Wide Web provides every internet citizen with access to an abundance of information, but it becomes increasingly difficult to identify the relevant pieces of information. Research in web mining tries to address this problem by applying techniques from data mining and machine learning to web data and documents. Web content mining and web structure mining have important roles in identifying the relevant web page. Relevancy of web page denotes how well a retrieved web page or set of web pages meets the information need of the user. Page Rank, Weighted Page Rank and Hypertext Induced Topic Selection (HITS) are existing algorithms which considers only web structure mining. Vector Space Model (VSM), Cover Density Ranking (CDR), Okapi similarity measurement (Okapi) and Three-Level Scoring method (TLS) are some of existing relevancy score methods which consider only web content mining. In this paper, we propose a new algorithm, Weighted Page with Relevant Rank (WPRR) which is blend of both web content mining and web structure mining that demonstrates the relevancy of the page with respect to given query for two different case scenarios. It is shown that WPRR's performance is better than the existing algorithms.

Keywords:

Web Structure and Content Mining

1. INTRODUCTION

Emerging field of *web mining* aims to find and extract relevant information from web. Like data mining, web mining is a multi-disciplinary effort that draws techniques from fields like information retrieval, statistics, machine learning, natural language processing. Search engines have significant role to play in the information retrieval and are used by many users to find information from the web.

Web mining is commonly divided into the following three sub-areas: Web Content Mining is the application of data mining techniques to unstructured and semi structured text typically HTML document. Web Structure Mining is the process of using hyperlink structure of the web as an information source. Web Usage Mining is about the analysis of user interaction with web server [1],[4],[8]. According to the survey by Google, the size of the World Wide Web is estimated to increase every day. In December 2011, the Indexed Web had around 7.78 billion pages, which had increased at an astounding rate to 8.07 billion pages [15].With the explosive growth of information sources available on the World Wide Web, finding relevant information is becoming more important for web users. The survey indicate that the, maximum number close to 26% of search query terms was three-worded. User queries issued to search engine is very short and the results of search engine are thousand of Web pages commonly retrieved in the form of ranked list. The problem is, more than 70% of users navigate either one or two pages and try to get required relevant documents within the navigated pages or ignore the rest of the pages without getting the needed information [13].

There are numerous approaches to tackle the problem of presenting the most relevant page on the top of search result. Most popular method to address the problem is by ordering the search result and presenting most relevant page on the top, this method called as page ranking. Nearly 100 factors are used to rank the search results [14].

According to Brin et al., the importance of any web page can be judged by looking at the pages that link to it, which is the key idea of the PageRank algorithm. It is the dominant link analysis model for Web search. A web page *a* include a hyperlink to the web page *b*, this means that page *b* important and relevant for topic. If there are a lot of pages that link to *b*, this means that page *b* is important. PageRank algorithm consider only structure of web [6, [8].

Ashraf Ali et al, exposed one of the major problems with the PageRank is link spamming, which gives websites higher rankings the more other highly ranked websites link to it. These techniques also aim at influencing other link-based ranking techniques [3].

Dangling pages were discussed by Crestani, et al in 2002. A page with no outgoing edge is a dangling page. In existing PageRank algorithm, all dangling pages are removed from the system, and then the rank score of the pages are calculated for the remaining Web pages. Finally, dangling pages are added back in a heuristic way [9]. Using Improved PageRank algorithm the dangling Page problem is overcome by adding outgoing links from it to each of all pages including itself. Computation of Improved PageRank algorithm also focuses on the hyperlink structure of web pages [9].

Another new concept which was discussed by Chakrabarti, et al., in 1999 is HITS algorithm. HITS is entirely a link-based algorithm. It ranks the web pages by analyzing their in-links and out-links. In this algorithm, web pages pointed to by many hyperlinks are called *authorities* whereas web pages that point to many hyperlinks are called *hubs.* This algorithm receives search results returned by traditional text indexing techniques as input. Once these results have been assembled, the HITS algorithm ignores textual content and focuses itself on the structure of the Web only. Mahesh Chandra Malviya, et al., also discussed about this algorithm [2],[8],[10].

Weighted PageRank algorithm is yet another approach, which was used by many search engines retrieved large number of documents in the form of ranked list based on issued queries. Weighted PageRank algorithm provides important information

about a given query by using the structure of the web and not using the content of the web. Some pages irrelevant to a given query are included in the results, because it has many existing in-links and out-links [11],[12].

In this paper, Weighted Page with Relevancy Rank (WPRR) Algorithm that uses both content as well as structure of the web to represent the relevant documents on the top of the search results is being proposed. To avoid the noise results from irrelevant pages, we use Three Level Score method (TLS) to determine the relevancy of pages, which is used to categorize the pages into four classes based on their relevancy to the user queries. In the next section, Weighted PageRank algorithm is being reviewed. Section 3 explains Three-Level Scoring method, while in Section 4, architecture for WPRR has been proposed and WPRR algorithm is presented. In section 5 evaluation is being carried out with two different case scenarios and Section 6 concludes the paper with reference to the possible future work.

2. WEIGHTED PAGERANK

Weighted PageRank Algorithm was proposed by Wenpu Xing and Ali Ghorbani during 2004.It is an extended PageRank Algorithm, that assigns larger rank values to more popular pages and does not divide the rank of a page equally among its out-link pages. The popularity of a page is determined by observing the number of in-links and out-links. WPR determines the rank score based on the popularity of the pages by taking into account the importance of both the in-links and out-links of the pages. Each out-link page gets a value proportional to its popularity [5],[7],[12].

The popularity is assigned in terms of weight values to the incoming and outgoing links that are recorded as $W^{in}_{(m,n)}$ and $W^{out}_{(m,n)}$ are calculated as follows,

$$W^{in}_{(m,m)} = \frac{I_n}{\sum p \in R(m) I_p} \quad (1)$$

$$W^{out}_{(m,n)} = \frac{O_n}{\sum p \in R(m) O_p} \quad (2)$$

where,

$W^{in}_{(m,n)}$ is the weight of link(m, n) calculated based on the number of in-links of page n and the number of in-links of all reference pages of page m.

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

I_n and I_p is the number of incoming links of page n and page p respectively

O_n and O_p is the number of outgoing links of page n and p respectively

$R(m)$ is the reference page list of page m.

The importance of pages can be computed by using the following formula,

$$WPR(n) = \left(\frac{d}{n}\right) + (1-d) * \sum_{m \in B(n)} WPR(m) W^{in}_{(m,m)} W^{out}_{(m,n)} \quad (3)$$

3. RELEVANCY SCORING

Relevance score determines the relevance of a page with respect to query terms by counting the number of occurrences of the query terms within the web document.

3.1 THREE LEVEL SCORING METHOD (TLS)

In TLS method consider the following criteria are considered to assign relevance scores:

1. Relevant pages are pages which contain very important information about the given query.
2. Weak Relevant pages, which have relevant information but not important information about the given query. They are only partially related to a query.
3. Irrelevant pages are the pages that contain irrelevant information about the query, which consist neither the keywords nor the relevant information about the given query.

The relevancy of a Web page to a given query using TLS method is computed in the following way:

1. Initially the stop words are removed from the query phrase.
2. Given a query phrase q with n terms and a web page p, the initial score $A_{(q,p)}$ is computed as,

$$A(q, p) = \frac{t_n.k^{n-1} + t_{n-1}.k^{n-2} + ... + t_1}{k^{n-1}} \quad (4)$$

where, k is a constant, corresponding to the weight for longer sub-phrases t_i, $1 \le i \le n$ is the number of occurrence of the sub-phrases of length i. The order of the terms in the sub-phrases should be exactly the same as that in the original query phrase q. The benefit of TLS method lies in its giving higher scores to the occurrence of substrings with more distinct terms. It also considers the order of query terms. The change in the order of query terms may change the meaning of the phrase [16].

4. WEIGHTED PAGE WITH RELEVANT RANK ALGORITHM (WPRR)

Existing Weighted PageRank Algorithm works based on web structure mining techniques. Many search engines use this algorithm but the user does not get the required relevant documents on the top of the search results. To overcome the problem found in weighted page rank algorithm, a new algorithm has been proposed as Weighted Page with Relevant Rank Algorithm that makes use of both the web structure mining and the web content mining techniques. Web structure mining is used to calculate the popularity of the page based on the number of in-links and out-links of the page and the web content mining is used to find the relevancy of the web page by matching query term with the content of web page. Weighted Page with Relevant Rank Algorithm is proposed to improve the order of the search results.

4.1 PROPOSED ARCHITECTURE OF SEARCH ENGINE

Using the Information Retrieval search engine embedded in the web site, set of pages relevant to a given query is retrieved. This set of pages is represented as root set. Expanding the root set with pages that directly point to or are pointed to by the pages in the root set are called as base set. WPR algorithm is applied to the base set, hence it rely on the web structure. Using WPR algorithm, the popularity of the pages based on the hyperlink structure of web site can be calculated and it gives sorted order of the web pages according to the popularity values. This might not satisfy the users, because they might not get required relevant document on the top of search results.

Once a user fires query in the form of keywords on the interface of a search engine, it is retrieved by the query processor component and process it word by word. One of the important components of search engine is web spider also known as web crawler, a program that visit web and read their pages and other information in order to download the web pages. Using WPR algorithm, the popularity of the pages based on the hyperlink structure of web site can be calculated and it gives sorted order of the web pages according to the popularity values. This might not satisfy the users, because they might not get required relevant document on the top of search results.

The downloaded pages are routed to index module to build an index. The purpose of storing an index is to optimize speed and performance. After matching the query keywords with the index, it returns the URLs of the pages to the user.

The proposed architecture suggest two more modules such as Relevance estimator and Link based Weight estimator module, which are used to compute relevancy of web pages using TLS method and estimate popularity of web pages using WPR algorithm respectively. The outcomes of the two modules are combined to obtain most relevant web pages on the top of the search results by using proposed algorithm. WPRR algorithm suggests the proposed architecture of search engine to include the module for calculating the popularity and relevancy of web pages as shown in Fig.1.The entire process of Fig.1 is described in the following steps:

4.1.1 Compute Popularity of Web Pages:

Search engine look at the links of web pages to determine the importance of web pages. The number of in-links and out-links of web pages are used to find the rank of web pages. *Web spider* passes downloaded web pages and their link information to the *Indexer*. The link information is routed to the *Link Structure Generator* module for generating web graph. *Link Based Weight estimator* module compute the popularity of web pages by considering the importance of both the in-links and out-links of the pages.

4.1.2 Compute Relevancy of Web Pages:

Once a query is entered to the *Search Engine Interface* it analyzes the text and link within the web site to determine a site's relevancy. It is then retrieved by the *Query Processor* and processes it word by word, which is sent to the *Web Spider*, which is already mentioned. *Indexer* used to build an index to facilitate query processing and also it provides a more useful vocabulary for the search engine after removing stopping words at the prefix and suffix of the query and the sub phrase of query, sentence boundaries, punctuation marks and non-word etc matching the query term with the index. *Index* is usually built in an alphabetical order of query terms and contains extra information regarding the page such as its URL, frequency, position of terms etc.

4.1.3 Compute WPRR:

WPRR algorithm combines the output of *Link Based Weight estimator* module and *Relevance Estimator* module to determine the weighted Page with Relevant Rank of all the pages returned after determining the popularity of web pages and matching the query term with the index. WPRR algorithm returns the required relevant documents easily on the top few pages of search result.

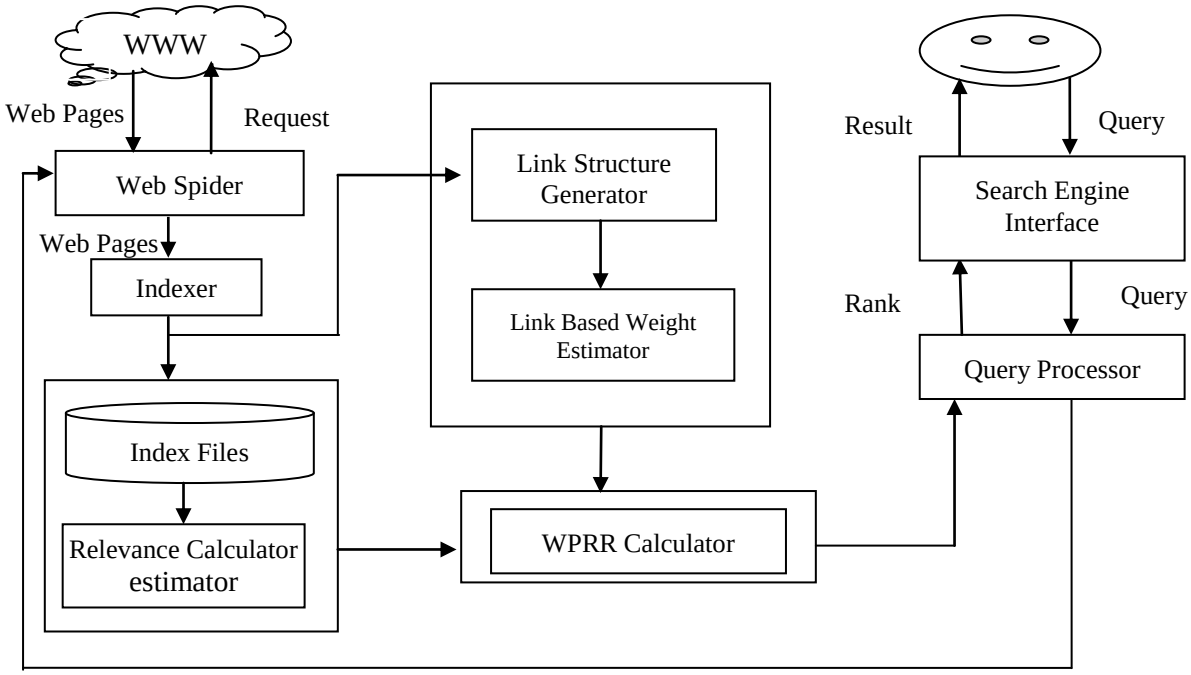

Fig.1. WPRR Architecture

Algorithm: ***WeightedPage_with_Relevant Rank***

Input: *User Query q, Set of pages p, Damping factor d =0.15, In_links and Out_links of pages p.*

Output : *Rank score of WPRR.*

I. Calculate Link Popularity

1. Build an adjacency matrix of the web graph

2. Initialize all pages with initial rank is 1

3. for(i ε p, 1 ≤ i ≤ p) //Set of pages *p*

{

4. for(j ε ref_i, 1 ≤ j ≤ ref_i)

{

Calculate ril_j and rol_j //where ril_j refer In_links of //reference page *j* and rol_j refer //Out_links of reference page *j*

$W^{in} = il_i / ril_j$ //W^{in} refer weight of In_links

$W^{out} = ol_i / rol_j$ //W^{out} refer weight of Out_ //links

$WPR_j += WPR_j * W^{in} * W^{out}$ //where WPR_j refer weighted //page rank of page *j*

}

$WPR_i = (d/p)+(1-d) * WPR_j$

}

II. Calculate Relevancy Score

1. Get meaningful query term of q

2. Decide phrase and sub phrases of query term q

3. Decide optimal value of k, δ and β

4. Get page info of p pages

5. for(i ε p, 1 ≤ i ≤ p)

{

6. Calculate frequency of phrase and sub phrases of query term q in i^{th} page

7. $$A(q,i) = \frac{t_n.k^{n-1} + t_{n-1}.k^{n-2} + ... + t_1}{k^{n-1}}$$

//*A(q,i)* denote initial score computation

III. Calculate WPRR

//Take summation of Link popularity and Relevancy Score

1. for(i ε p, 1≤ i≤p)

2. $WPRR_i = WPR_i + A(q,i)$

5. EVALUATION

In order to show that the WPRR algorithm satisfies the user requirements by retrieving the required relevant documents on the top of the search results, the algorithm was implemented and the performance was evaluated in C language. Two real time experiments were conducted on workstation (Intel(R) Pentium (R) Dual CPU (200 Ghz) machine with 2GB of RAM). The evaluation is carried out by analyzing the real time experiment of the following IT company web sites. The case studies considered are given below.

5.1 CASE STUDY 1: http://www.tcs.com

In this case study, consider web graph of Fig.2, which is constructed by using the site http://www.tcs.com. From among the different modules, two modules have been chosen, which are related to IT Services along with hyperlink structure. Each module consists of seven and eleven pages respectively. The pictorial representation of web graph and its sub modules are as illustrated.

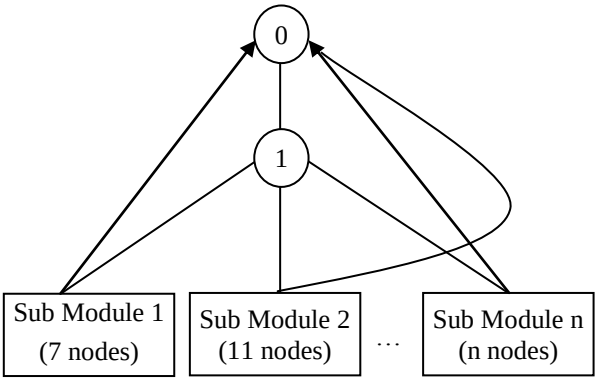

Fig.2. Web graph with twenty nodes

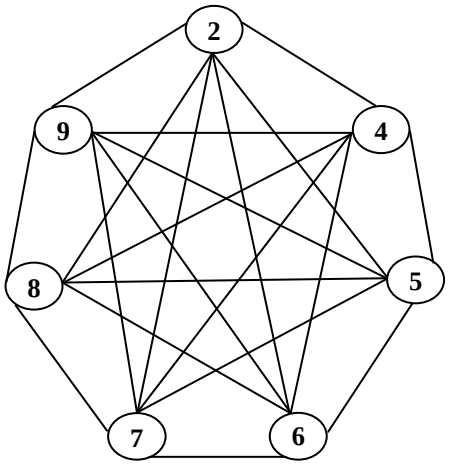

Fig.2.(a). SubModule1 with seven nodes

Fig.2(a) and Fig.2(b) illustrates modules which consists of N web pages with N-1 hyperlink structures for all considered nodes except sub node 19. Web page 2 in sub module1 has direct in-link from web page 1; similarly web page 3 in sub module2 also has direct in-link from web page 1. But all pages in sub module 1 and 2 have direct out-links to web page 1 and 0.

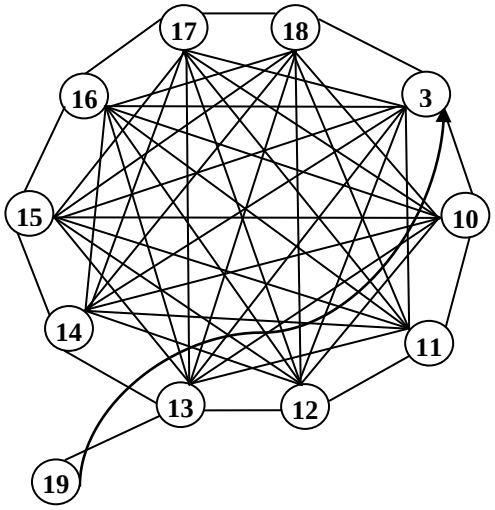

Fig.2(b). SubModule2 with eleven nodes

Table.1. Results WPRR values for case study I

Pages	WPR	TLS	WPRR
1	0.08122	0.02	0.10122
2	0.23222	0.15	0.38222
3	0.09567	0.23	0.32567
4	0.17707	0.15	0.32707
5	0.06082	0.15	0.21082
6	0.051	0.15	0.201
7	0.04107	0.15	0.19107
8	0.03104	0.17	0.20104
9	0.0209	0.14	0.1609
10	0.01066	0.13	0.14066
11	0.06157	0.05	0.11157
12	0.05534	0.04	0.09534
13	0.04908	0.14	0.18908
14	0.11451	0.41	0.52451
15	0.03706	0.02	0.05706
16	0.03068	0.05	0.08068
17	0.02425	0.07	0.09425
18	0.01778	0.03	0.04778
19	0.01126	0.04	0.05126
20	0.00753	0.07	0.07753

The Relevant, Weak Relevant and Irrelevant were set to 0.1 to 0.9, 0.01 to 0.09 and less than 0.01 respectively. For the considered web site, twenty web pages have been chosen randomly for the query string **"Business Application Development"** and relevance values have been calculated using WPR, TLS and WPRR algorithms. Table.1 presents the obtained results using WPR, TLS and WPRR algorithms.

From the analysis carried out it is evident that WPR algorithms provide significant information about a given query by using the hyperlink structure of the website only. Some of the pages irrelevant to a given query are included in the results as well. The home page of TATA CONSULTANCY SERVICES from the site http://www.tcs.com, is not related to the given query, it still is one among the top five ranks. The result of WPRR values are returned as a sorted ordered list with the most relevant pages on the top of the search result. It is observed that the irrelevant and weak relevant pages were automatically filtered and WPRR retrieves the most relevant pages for the users. Fig.3 illustrates the results of WPR, TLS and WPRR [17].

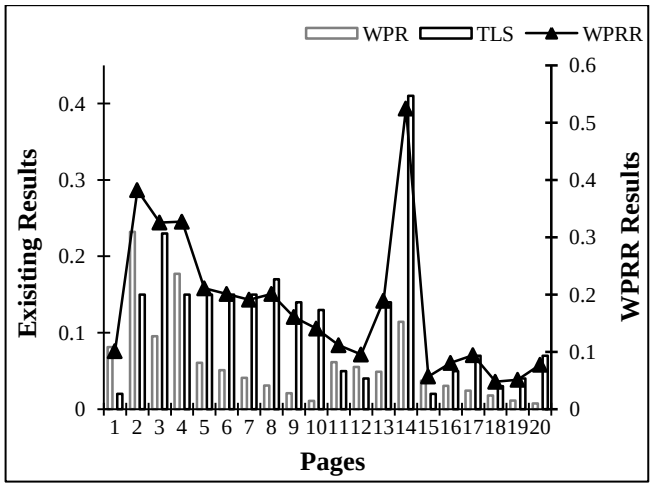

Fig.3. Result of case study I

5.2 CASE STUDY II: http://www.wipro.com

Fig.4 illustrates the web graph of the consulting services of WIPRO, which consists of fifteen pages with interrelated links as per WIPRO site specification given in http://www.wipro.com. Two sub modules were considered with four and eight pages respectively.

Fig.4(a) and Fig.4(b). Illustrates modules which consist of N web pages with N-1 hyperlink structures. Web page 3 in sub module1 has direct in-link from web page 2 , similarly web page 4 in sub module2 also has direct in-link from web page 2.But all pages in sub module 1 and 2 have direct out-links to web page 2,1 and 0.

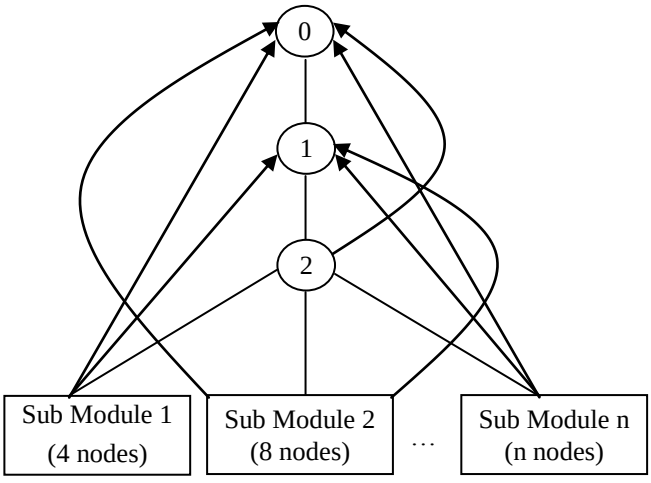

Fig.4. Web graph with fifteen nodes

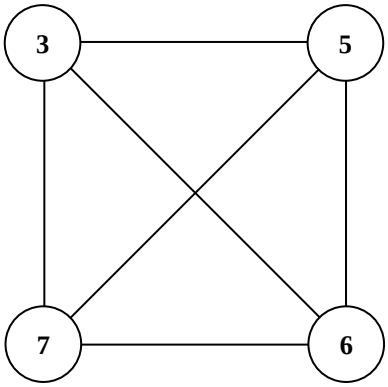

Fig.4(a). SubModule1 with four nodes

By using this site the WPRR values for web pages with query string as "**Consulting Service Management"** were calculated.

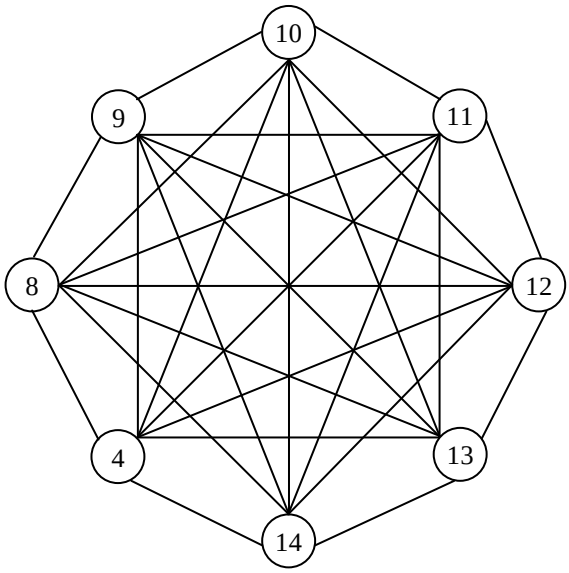

Fig.4(b). SubModule2 with eight nodes

Table.2 illustrates that the web page five was irrelevant to query and page one was weak relevant value to the query, but still receives highest link values, since these pages have huge number of in-links and out-links, it can also filter by using WPRR algorithm.

Table.2. Results of WPRR values for case study II

Page	WPR	TLS	WPRR
1	0.16493	0.04	0.20493
2	0.28376	0.04	0.32376
3	0.2928	0.05	0.3428
4	0.06586	0.05	0.11586
5	0.10413	0	0.10413
6	0.03481	0.03	0.06481
7	0.02322	0.01	0.03322
8	0.0115	0.04	0.0515
9	0.06184	0.08	0.14184
10	0.05388	0.05	0.10388
11	0.04584	0.19	0.23584
12	0.03774	0.15	0.18774
13	0.02957	0.11	0.13957
14	0.02133	0.06	0.08133
15	0.01302	0.11	0.12302

Fig.5 illustrates the WPRR algorithm filter the result of unrelated web pages and it is clear that the pages with most relevant information tops up [18].

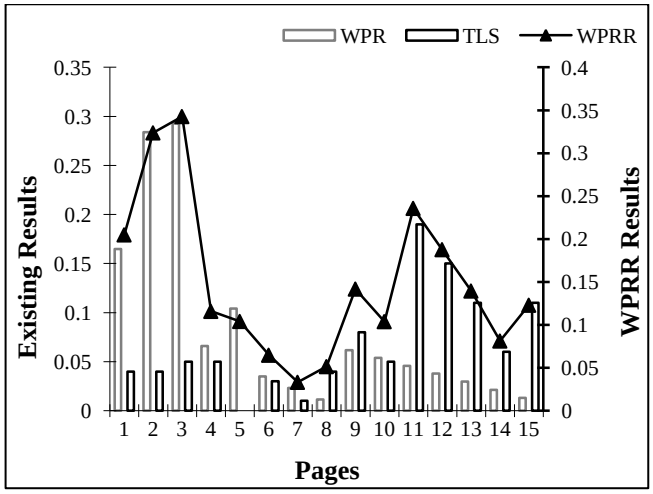

Fig.5. Result of case study II

6. CONCLUSION

Research in web mining tries to come out with most efficient web searching methods to retrieve relevant information from the web pages. Web Structure Mining and Web Content Mining play a vital role in achieving this. In this paper we proposed a new architecture which is a blend of these two techniques. A new algorithm WPRR has been implemented and has been evaluated for two different case study scenarios. From the obtained results it evident that WPRR algorithm explores most relevant pages on the top of search results.

In this work, WPR algorithm and TLS algorithm have been implemented for web structure and content mining respectively. In similar line other existing algorithms could be analyzed for efficient Information retrieval.

REFERENCES

[1] Blockeel. H. and Kosala, R. "Web mining research: A survey", *ACM SIGKDD Explorations Newsletter*, Vol. 2, No. 1, pp.1-15, 2000.

[2] Chakrabarti S, Dom B. E, Kumar S. R, Raghavan P, Rajagopalan S, Tomkins A, Gibson D and Kleinberg J,

"Mining the Web's link structure", *Computer*, Vol. 32, No. 8, pp. 60-67, 1999.

[3] Ashraf Ali and Israr Ahmad. "Information Retrieval Issues on the World Wide Web", *International Journal of Computer Technology and Applications*, Vol. 2, No. 6, pp. 1951-1955, 2011.

[4] G. Poonkuzhali, R. Kishore Kumar, R. Kripa Keshav, K. Thiagarajan and K. Sarukesi, "Effective Algorithms for Improving the Performance of Search Engine Results", *International Journal of Applied Mathematics and Informatics*, Vol. 5, No. 3, pp. 216-223,2011.

[5] Dilip Kumar Sharma *et. al.*, "A Comparative Analysis of Web Page Ranking Algorithms", *International Journal on Computer Science and Engineering*, Vol. 20, No. 8, pp. 2670-2676, 2010.

[6] Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd, "The pagerank citation ranking: Bringing order to the web", *Technical report, Stanford InfoLab*, 1999.

[7] Rekha Jain and G. N. Purohit, "Page Ranking Algorithms for Web Mining", *International Journal of Computer Applications*, Vol. 13, No. 5, pp. 22-25, 2011.

[8] Ashutosh Kumar Singh and Ravi Kumar P, "A Comparative Study of Page Ranking Algorithms for Information Retrieval", *International Journal of Electrical and Computer Engineering*, Vol. 4, No. 7, pp. 469-480, 2009.

[9] Sung Jin Kim and Sang Ho Lee, "An Improved Computation of the PageRank Algorithm", *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval*, pp.73-85, 2002.

[10] Rakesh Kumar Malviya et al, "Survey of Web usage Mining", *International Journal of Computer Science and Technology*, Vol. 2, No. 3, pp. 661-619, 2011.

[11] Ying Ding, "Topic-based PageRank on author cocitation networks", *Journal of the American Society for Information Science and Technology* , Vol. 62, No. 3, pp. 449- 466, 2011.

[12] Ghorbani, A and Xing W, "Weighted PageRank Algorithm," *Proceedings of the Second Annual Conference on Communication Networks and Services Research*, pp. 305-314, 2004.

[13] Ben Choi and Sumit, "Ranking Web Pages Relevant to Search Keywords", *IADIS International Conference WWW/Internet*, pp.200-205, 2009.

[14] Vaughn's Summaries, "Google Ranking Factors", Available at: http://www.vaughns-1 pagers.com/internet/google-ranking-factors.htm, Accessed 23 January 2012.

[15] World Wide Web Size, "The size of the World Wide Web", Available at: http://www.worldwidewebsize.com. Accessed20 December 2011.

[16] Relevance Scoring Methods, "Three-Level Scoring Method", Available at: http://www2002.org/CDROM/refereed/643/node8.html Accessed 18 May 2011.

[17] Tata Consultancy Services, "IT Services and IT Infrastructure Services", Available at: http://www.tcs.com/offerings/Pages/default.aspx, Accessed 21 June 2011.

[18] WIPRO," Wipro Consulting Services", Available at: http://www.wipro.com/services/consulting-services, Accessed 13 February 2011.