INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

FAST AND EFFICIENT WEB SPIDER

Satwinder Kaur[1], Alisha Gupta[2]

[1]Research Scholar, [2]Lecturer
[1-2]Department of Computer Science and Engineering, HEC Jagadhri, Haryana, India

ABSTRACT: -PageRank (PR) algorithm ascertains the importance of web pages based on a PageRank metric. It state that if a page has important links to it , its link to other pages also contribute to their importance and a page with high PageRank is most relevant page to be downloaded. An improved version of PageRank, WPR, assigns more value to more important pages instead of dividing the rank value of a page evenly among all its outgoing links whereas TR combats spam pages on World Wide Web and ultimately aims at improving the efficiency of search results. These two variations of PR (WPR and TR) gives an insight of hybrid formula that distribute rank score according to popularity of links as well as ensures the crawling of only trusted pages.

KEYWORDS: Web crawling, PageRank metric, Web spamming PR, TR, WPR, and WPPR.

1. INTRODUCTION

The World Wide Web is the largest collection of data today and it continues increasing day by day. A web crawler is a program from the huge downloading of web pages from World Wide Web and this process is called Web crawling. To collect the web pages from www a search engine uses web crawler and the web crawler collects this by web crawling. Due to limitations of network bandwidth, time-consuming and hardware's a Web crawler cannot download all the pages, it is important to select the most important ones as early as possible during the crawling process and avoid downloading and visiting many irrelevant pages. This paper reviews help the researches on web crawling methods used for searching.. Some of the worth mentioning strategies are:

1. Page Rank algorithm: This algorithm determines the importance of web pages based on a PageRank metric. It state that if a page has important links to it , its link to other pages also contribute to their importance and a page with high PageRank is most relevant page to be downloaded

2. Weighted PageRank algorithm [13]: This algorithm is an improved version of PageRank which assigns more value to more important pages instead of dividing the rank value of a page evenly among its entire outgoing links.

3. Trustrank algorithm: To determine the importance of pages their PageRank was computed using personalized PageRank that assumed that a user goes to a trusted site rather than to a page of equal probability.

2. OBJECTIVE

The main stages of this research work are comparison of various page ranking formulas, designing an improved ranking formula, comparison of the existing and proposed improved ranking formula and finally designing an algorithm to rank the web pages effectively by crawling trusted pages. In the comparison one algorithm took maximum time in convergence to a stable value. In improving the algorithm, the algorithm which took maximum

number of iterations is improved and in comparison of existing and proposed improved algorithm, the proposed algorithm took less number of iterations and provide genuine ranking to all web pages

3. Research Methodology in Designing of Proposed Improved Pageranking Algorithm

Experiment methodology has been used in this objective. Firstly the WPPR ranking formula proposed in earlier section is used in an algorithm. The complete process of selecting seed pages and their ordering is explained in the algorithm. The proposed algorithm is implemented on an authorized website of Kurukshetra University. For a specific query "kurukshetra university", three search engines namely Google, Yahoo and Bing are scrolled as shown in figures below.

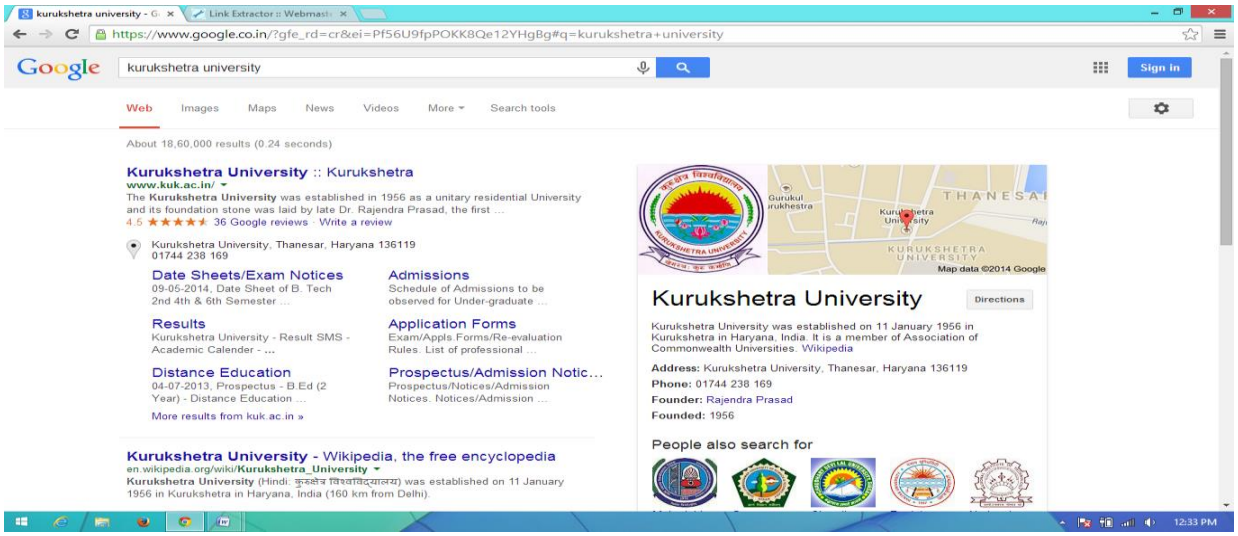

Figure: Search result of Google search engine

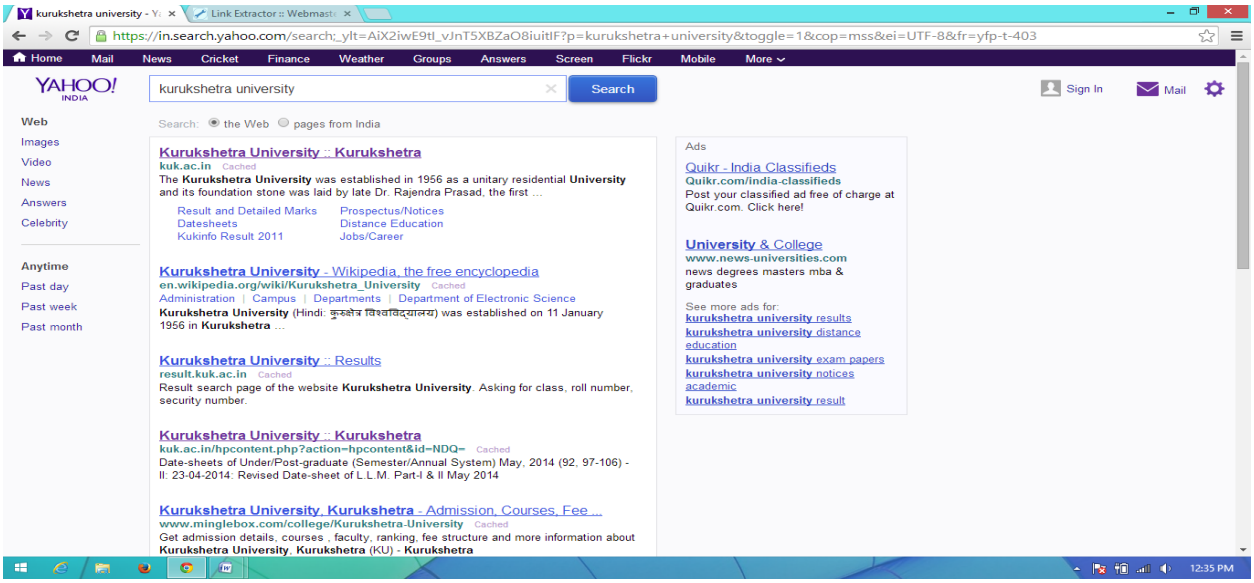

Figure: Search result of Yahoo search engine

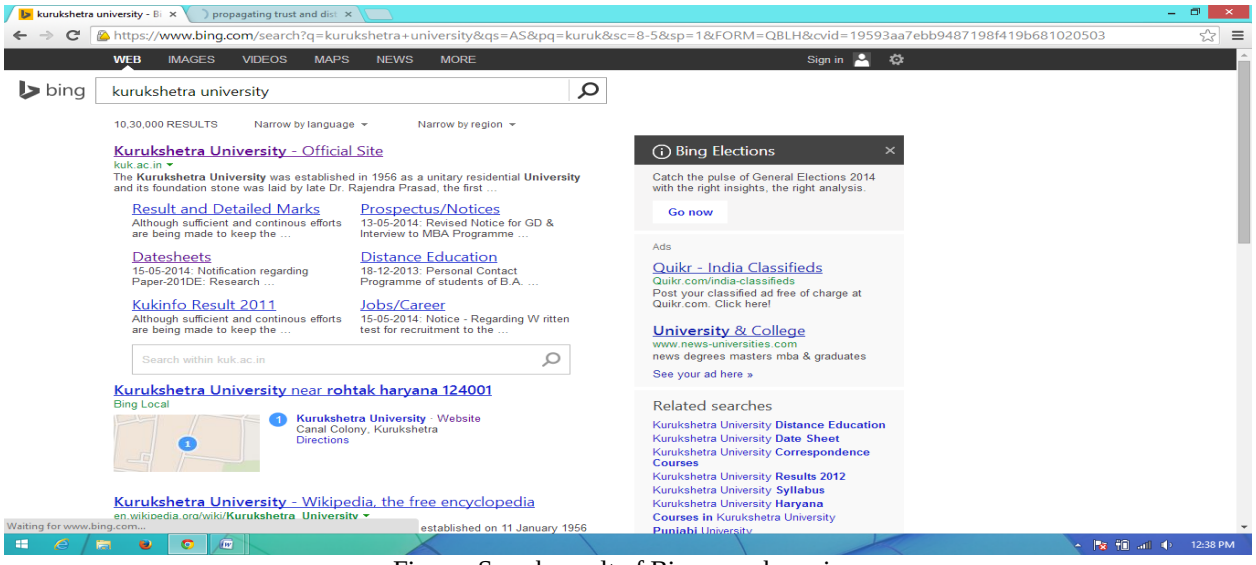

Figure: Search result of Bing search engine

The top search result pages are observed from search result. The link structure of university website is then observed in order to draw a small hypothetical graph for the implementation of proposed algorithm. Due to time limit and complexity of calculation only four pages of website are chosen, out of which one page is of high importance and two are of least importance. The web pages of the Kurukshetra website are Home Page, About Us Page, News Page and Datesheet Page. The graph thus generated after some unwanted eliminations is shown in figure .

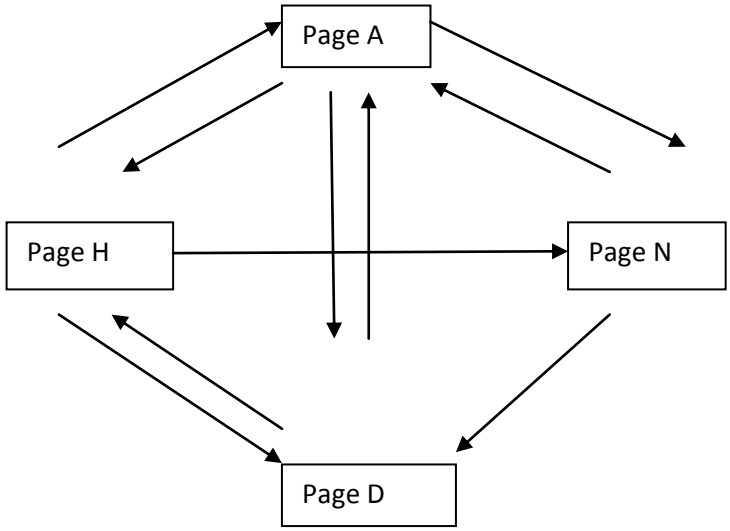

Figure: A small graph of four pages of kurukshetra university, website

The author's intention was to estimate the order of importance of different pages and match them with the search results of conventional search engines to prove the worthiness of proposed algorithm. The trust value associated with each page is calculated based on the inverse pagerank heuristic, which states that a page with more number of outlinks are more important pages. The number of links on all the considered pages is extracted using a link extractor tool. It is an online link extraction tool that requires URL of a web page as input and provides all the links available on that particular web page as shown below

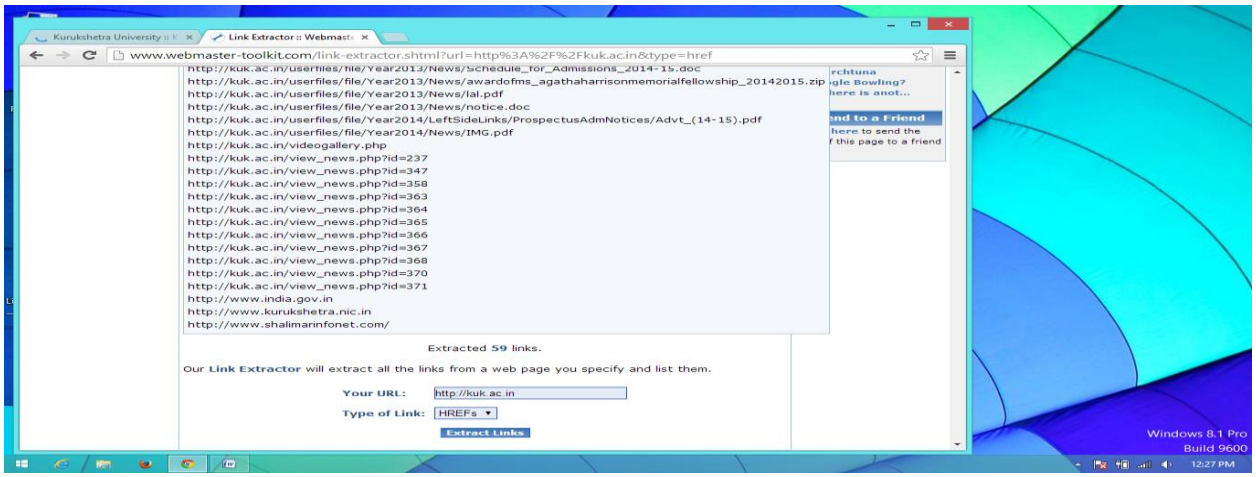

Figure 4.7: Link extractor tool

The number of links on each page contribute to its trust value through rank selection mechanism. In simpler form the share of importance of single pages in the global importance is calculated as follows:

Trustvalue(P) = Number of outlinks of Page P/ Total number of outlinks on all the pages

The final proposed algorithm is as follows:

Step 1: Extract the number of links on a page

Step 2: Calculate Ti (Normalized trust) using Rank selection method for each page.

Step 3: For i= 1 to N do

$Pr(b) = d \sum_{a \in B(b)} (PR(a).W^{in}(a,b).W^{out}(a,b)) + (1-d)\, Ti$

Step 4: Return PR(b);

4. Comparison of PR, WPR and TR Formula In Terms of Iterations Required For Convergence and Order of Preference

The three ranking formulas namely PR, WPR and TR is implemented on a two page graph shown in figure 4.1 at small initial value of Page (B) = 0. The results for PR and WPR are same due to simplicity of link structure of graph. These values are already shown in Table above. The values for TR considering Page A as good and Page B as bad page for crawling are shown in Table below. The trust value of page A is thus taken as 1 and of Page B is taken as 0 according to trust ignorant function discussed in literature.

Ranking Formula	No of iterations required	Order of importance	Remarks
PR	16	Same for both pages	Precision increases in followed iterations
WPR	16	Same for both pages	Precision increases in followed iterations
TR	14	A is more important than B being a good page	Precision remain almost stable

Table: Comparison of ranking algorithms for two page web graph

4.1. Proposed WPPR Formula

The two variations of PR namely WPR and TR gives an insight of hybrid formula that distribute rank score according to popularity of links as well as ensures the crawling of only trusted pages.The proposed formula thus has the following form:

$Pr(b)= d \sum_{a \in B(b)} (PR(a) Win(a,b) Wout(a,b)) +(1-d) ti$

Where PR(b) = pagerank of page b

d= damping factor

ti=trust score of page i

B(b) = reference pages of page b

Win(a,b) = weight of link(a,b) calculated based on the number of inlinks of page b and the number of inlinks of all reference pages of page a

Wout(va,b) = weight of link(a,b) calculated based on the number of outlinks of page b and the number of inlinks of all reference pages of page a

Next the proposed WPPR is used for calculating the same pages of the hypothetical graph. As the proposed formula uses trust vector, so again there are two different versions of the formula namely WPPR(i) and WPPR(ii). In this case, the trust score assignment used is given in Table.

	Page A	Page B	Page C
WPPR(i)	1	0	½
WPPR(ii)	½	0	1

Table: Trust score assignment used for WPPR formula

The rank values given by WPPR (i) at different iterations are shown in Table

S.NO	PR(A)	PR(B)	PR(C)
1	1.0	.141666666	.478749999
2	.556937499	.078899478	.299863514
3	.404883987	.057358564	.238471908
4	.352701122	.0499965992	.217403077
5	.334792616	.047428953	.210172517
6	.32864664	.046558273	.20769108
7	.326537418	.046259467	.206839482
8	.325813559	.04615692	.206547223

9	.32556514	.046121727	.206446924
10	.325479885	.04610965	.206412503
11	.325450627	.046105505	.20640069

Table: Iterative values of WPPR (i) for three pages of hypothetical web graph

The rank values given by WPPR (ii) at different iterations are shown in Table

S.NO	PR(A)	PR(B)	PR(C)
1	1.0	.141666666	.553749999
2	.545687499	.077305728	.370321326
3	.389773127	.055217859	.307370899
4	.336265264	.047637578	.285767099
5	.317902034	.045036121	.278352945
6	.311600003	.044143333	.2758085
7	.309437225	.04383694	.274935279
8	.308694987	.043731789	.2746356
9	.30844026	.043695703	.274532754
10	.308352841	.043683319	.274497459

Table: Iterative values of WPPR (ii) for three pages of hypothetical web graph

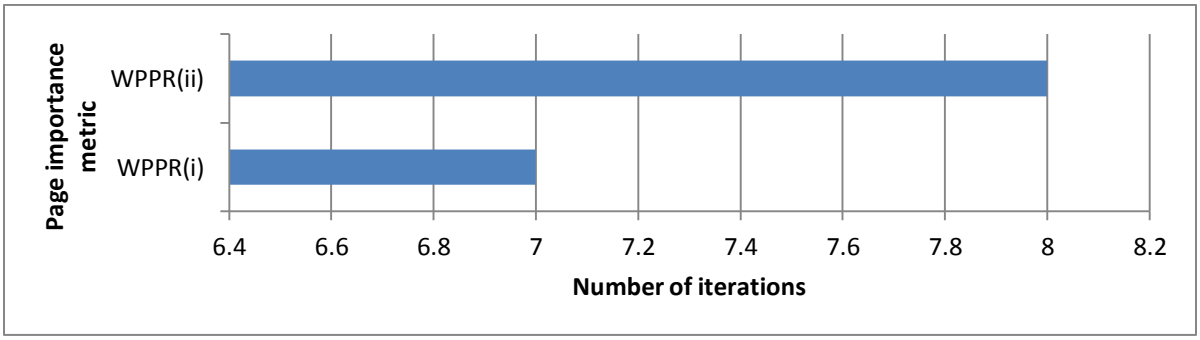

Figure: Graphical Representation of Number of iterations required by WPPR for convergence

5. Analysis and Interpretation of Design of Proposed Improved Crawling Algorithms

This section presents the result of implementation of proposed algorithm on a small web graph. In WPPR algorithm time taken to converge the rank value of web pages have been reduced. At the same time this algorithm ensures that more trusted web pages are ranked high in search results.

5.1. Calculation of trust score of four web pages

The graph shown in figure 4.7 is drawn by considering the explicit links and not all the links on individual page.The weighted personalize pagerank of various pages can thus be calculated by following equations:

$PR(H) = (1-d)\ T(H) + d\ [\ PR(A)\ W^{in}(A,H)\ W^{out}(A,H) + PR(D)\ W^{in}(D,H)\ W^{out}(D,H)]$

$PR(A) = (1-d)\ T(A) + d\ [\ PR(H)\ W^{in}(H,A)\ W^{out}(H,A) + PR(D)\ W^{in}(D,A)\ W^{out}(D,A) + PR(N)\ W^{in}(N,A)\ W^{out}(N,A)$

$PR(D) = (1-d)\ T(D) + d\ [\ PR(H)\ W^{in}(H,D)\ W^{out}(H,D) + PR(A)\ W^{in}(A,D)\ W^{out}(A,D) + PR(N)\ W^{in}(N,D)\ W^{out}(N,D)]$

$PR(N) = (1-d)\ T(N) + d\ [\ PR(H)\ W^{in}(H,N)\ W^{out}(H,N) + PR(A)\ W^{in}(A,N)\ W^{out}(A,N)\ ]$

To execute the proposed algorithm, initially trust score of all the web pages needs to be calculated . The trust values thus calculated are shown

	Links extracted	Normalized Trust	
Home	59	59/206	T(H)= 0.286
Datasheet	78	78/206	T(D)= 0.378
About Us	34	34/206	T(A)= 0.165
News	35	35/206	T(N)= 0.169

Table: Calculation of Normalized trust

5.2. Calculation of importance of links

Next, the calculation of weights of inlinks and oulinks of each page is done on the graph shown above. The use of equation

$Wout\ (m,n) = On\ /\ \Sigma\ Op\ ,\ \ Win(m,n) = In\ /\ \Sigma\ Ip$

gives the following values as shown in Table. These values of weights define the popularity of links and help in estimating the actual weightage of various links

Weights of inlinks		Weight of outlinks	
Win(A,H)	2/7	Wout(A,H)	3/7
Win(D,H)	2/5	Wout(D,H)	½
Win(H,A)	3/8	Wout(H,A)	3/7
Win(D,A)	3/5	Wout(D,A)	½
Win(N,A)	½	Wout(N,A)	3/5

Win(H,D)	3/8	Wout(H,D)	2/7
Win(A,D)	3/7	Wout(A,D)	2/7
Win(N,D)	½	Wout(N,D)	2/5
Win(A,N)	2/7	Wout(A,N)	2/7
Win(H,N)	¼	Wout(H,N)	2/7

Table 5.14: Weights of links

5.3. Calculation of rank values of all the pages of concerned graph

On applying the proposed algorithm on graph shown in figure the following values are retrieved at different iterations as shown in Table

No of iterations	PR(H)	PR(D)	PR(A)	PR(N)
1	1	1	1	1
2	.3166	.6713	.0567	.1553
3	.7268	.2191	.2194	.0846
4	.1028	.1032	.0867	.0375
5	.0694	.0783	.0637	.0346
6	.0619	.0746	.0610	.0331
7	.0618	.0740	.0604	.0330

Table 5.15: Iterative values of rank given by WPPR algorithm

RESULT

It shows that the time taken by the improved WPPR algorithm to converge to a suitable precision value is lesser than TR algorithm alone. As the size of the web graph increases, the number of iterations required in convergence is however reduced in proposed formula. This proves that the improved WPPR algorithms shows better results than the individual ranking formulas available in literature.

FUTURESCOPE

The algorithms can be proposed on different basis such as number of visitors on a particular web page, global trust values of particular domain etc. In the comparison the researcher have considered ranking formulas to determine the importance of web pages, other criteria such as query impact, word density of pages, update patterns etc can also be used. The results can be compared with the existing results, and the facts can be explored.

REFERENCES

[1] A. Dasgupta, A. Ghosh, R. Kumar, C. olston, S. Pandey and A. Tomkins, The Discoverability of the Web. In Proc. WWW, 2007.

[2] A. Heydon and M. Najork. Mercator: A scalable, extensible web crawler, World Wide Web Conference, 2(4):219–229, April 1999.

[3] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting Spam Web Pages through Content Analysis, In Proc. of the 15th Intl. World Wide Web Conference (WWW'06), pp. 83–92, 2006.

[4] A. Patterson, why writing your own search engine is hard, ACM Queue, April 2004.

[5] A. da Silva, Eveline A. Veloso, P. Golgher, , A. Laender, and N. Ziviani, Cobweb - a crawler for the brazilian web, In Proceedings of String Processing and Information Retrieval (SPIRE), pages 184–191, Cancun, Mxico, 1999. IEEE CS Press.

[6] B. liu, "web data mining", from Chapter 1, 7, 8, Springer-Verlag Berlin Heidelberg, 2007, pages 4-5.

[7] B. liu, "web data mining", from Chapter 6, 7, 8, Springer-Verlag Berlin Heidelberg, 2007, pages183-235,237-270,273-318.

[8] B. Pinkerton. Finding what people want: Experiences with the WebCrawler. In Proceedings of the first World Wide Web Conference, Geneva, Switzerland, May 1994.

[9] C.Castillo and R. Baeza-Yates, A new crawling model, In Poster proceedings of the eleventh conference on World Wide Web, Honolulu, Hawaii, USA, 2002.

[10] D. Cai, S. Yu, J.-R. Wen and W.-Y. Ma, Block-Based Web Search. In Proc. of the ACM SIGIR Research and Development in Information Retrieval (SIGIR'04), pp.456–463, 2004.

[11] Ds Fetterly, Nick Craswell, VishwaVinay, The impact of Crawl Policy On web Search Effectiveness, in Proceeding of SIGIR,july 2009.

[12] G. kaur et al, Improving the Efficiency of Weighted Page Content Rank Algorithm using Clustering Method International Journal of Computer Science & Communication Networks, ISSN:2249-5789,Vol 3(4),231-239 231,2014.