Designing a Hybrid Page Ranking Algorithm for Semantic Web Search Engine

Sharmila Subudhi

Abstract**-** ***Web is the most important tool in now-a-days upon which people rely on to search their required information. In such a scenario it is the duty of service provider to provide proper, relevant and quality information to the internet where user can submit their query and find out the result. But it is a challenge for service provider to provide proper, relevant and quality information to the internet user by using the web page contents and hyperlink between the web pages. The next-generation Web architecture, represented by the Semantic Web, provides the layered architecture possibly allowing overcoming this limitation. Several search engines have been proposed, which allow increase in information retrieval accuracy by exploiting keywords and their relations. This paper deals with a hybrid approach of page ranking algorithm which simply based on the prediction and calculation of different numbers of back-links to a web page.***

Keywords**-** ***Semantic web, Page rank, HITS, Search engine, Back-link predictor.***

I. INTRODUCTION

In today's world the volume of information on the internet is increasing day by day so there is a challenge for website owner to provide proper and relevant information to the internet user. With the rapid growth of WWW, it is becoming more difficult to manage the information on WWW and satisfy the user needs. Therefore, the users are looking for better information retrieval techniques and tools to locate, retrieve and filter the necessary information. Most of the users use information retrieval tools like search engines to find information from the WWW.

There are many search engines available but the popular ones are Google, Yahoo, Bing etc., because of their crawling and ranking methodologies. The search engines download, index and store millions of web pages. They answer millions of queries every day. This kind of technique of information retrieval is called as web mining. Fig 1 [1] shows a working of a typical search engine, which shows the flow graph for a searched query by a web user.

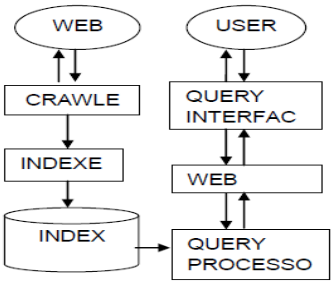


Fig.1 Architecture of a Search Engine

There are three important components in a search engine namely, Crawler, Indexer and Ranker. The crawler searches the web and downloads the web pages. The downloaded pages are being sent to an indexing function that parses the web pages and builds the index based on the keywords in those pages. When a user types a query using keywords on a search engine, the query processor will match the keywords with the index and return the URLs of the pages to the user. But before giving the result to the user, a ranking mechanism is done by the ranker of the search engine to present the most relevant pages at the top and less relevant ones at the bottom. It makes the search results navigation easier for the user. The ranking mechanism is explained in detail later in this paper. The remaining part of this paper is organized as follows: Related work is summarized in Section II. Section III describes Data Structure used for Web in particular the Web Graph. Section IV shows the implementation and simulation results and Section V concludes this paper.

II. RELATED WORK

Web mining is the technique to classify and segregate the web pages by taking into consideration the contents of the page and behavior of users in the past. Web mining consists of three sections i.e. web content mining (WCM), web structure mining (WSM) and web usage mining (WUM). WCM is responsible for exploring the proper and relevant information from the contents of web. WUM is responsible for recording the user profile and user behavior inside the log file of the web. WSM finds the structure of the hyperlink between different documents and classify the web pages. The number of out-links i.e. links from a page and the number of in-link i.e. links to a page are very important parameter in the area of web mining. The popularity of the web page is generally measured by a unit called as page rank which generally deals with the number of links that a particular page should be referred by other pages. So WSM becomes a very important area to be researched in the field of web mining [2].

A. Page Rank Algorithm

Page Rank algorithm [3] is the most commonly used algorithm for ranking the various pages. The working of the Page Rank algorithm depends upon the concept of back links how much a page contains. If the addition of the all the ranks of the back links is larger than the page then it is provided with a large rank. The PageRank is given by:

$$PR(A) = (1 - d) + d(PR(T1)\,/C(T1) + \ldots + PR(Tn\,/C(Tn\,)) \qquad (1)$$

Manuscript published on 30 October 2013.
***** Correspondence Author (s)

Sharmila Subudhi, Lecturer, Dept. of CSE/IT/MCA, Gandhi Institute For Technology, Bhubaneswar, Odisha, India.

The parameter d is the damping factor, usually sets to 0.85. The PageRanks form a probability distribution over the Web pages, so the sum of all Web pages' PageRank will be one.

B. HITS Algorithm

HITS algorithm [4] is a link based algorithm. This algorithm ranks the web page by processing in-links and out-links of the web pages. In this algorithm a web page is named as authority if the web page is pointed by many hyper links and a web page is named as HUB if the page point to various hyperlinks. Here the ranking of the web page is decided by analyzing their textual contents against a given query.

C. Weighted Page Rank Algorithm

Weighted Page Rank [5] Algorithm is proposed by Xing and Ghorbani. Weighted page rank algorithm (WPR) is the modification of the original page rank algorithm. It decides the rank based on the popularity of the pages by taking into consideration the importance of both the in-links and out-links of the pages.

D. Distance Rank Algorithm

An intelligent ranking algorithm named as distance rank is proposed by Bidoki and Yazdani [6]. This algorithm is based on the distance between any pages. Here the ranking is done on the basis of the shortest logarithmic distance between two pages.

E. Relation Based Algorithm

Lamberti, Sanna and Demartini [7] proposed a relation based algorithm for the ranking the web page for semantic web search engine. This algorithm proposes a relation based page rank algorithm for semantic web search engine that depends on information extracted from the queries of the users and annotated resources.

III. DATA STRUCTURE FOR WEB

Web mining technique provides additional information through hyperlinks where different documents are connected through hyperlinks. The Web may be viewed as a directed labeled graph whose nodes are the hyperlinked web pages and the edges are the hyperlinks between them. This directed graph structure in the Web is called as Web Graph.

A graph G consists of two sets V and E and can also be expressed as G = (V, E). The set V is a finite, nonempty set of vertices and the set E is a nonempty set of edges. The notation V(G) and E(G) represent the sets of vertices and edges respectively of graph G. The web is represented in a directed graph where each edge is represented by a directed pair (u, v) along with a direction. Therefore, (v, u) and (u, v) represent two different edges. The graph in Fig. 2 is a directed graph with 3 Vertices and 6 edges.

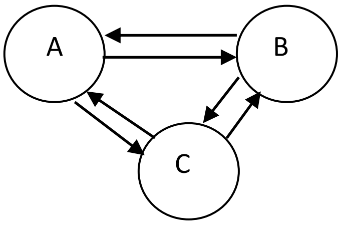


Fig.2 A Directed Graph G

The vertices V of G, V(G) = {A, B, C}. The Edges E of G, E(G) ={(A, B), (B, A), (B, C), (C, B), (A, C), (C, A)}. The degree of a vertex is the number of edges incident to that vertex. If G is a directed graph, we define the in-degree of a vertex v will be the number of edges coming to that vertex. The out-degree is defined to be the number of edges going out of v to other vertices. In Fig.2, the graph G, vertex B has in-degree 2, out-degree 2 and degree 4.

IV. IMPLEMENTATION & SIMULATION RESULT

Sparse Matrix

As the magnitude of the web graph is quite high, different efficient techniques are tried out for the rank computation. Sparse matrices are one of them. If most of the entries of a matrix are zero then the matrix is said to be sparse. There are various formats for the same. We would work on the following.

COO: Coordinate List. COO stores a list of (row, column, value) tuples.

$$A=\begin{pmatrix} 1 & 0 & 0 & 2 & 0 & 3 \\ 4 & 5 & 0 & 0 & 0 & 0 \\ 0 & 6 & 7 & 0 & 0 & 8 \\ 9 & 0 & 0 & 7 & 8 & 4 \\ 0 & 5 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix}$$

row: 0 0 0 1 1 2 2 2 3 3 3 3 4 4 5
col: 0 3 5 0 1 1 2 5 0 3 4 5 1 4 5
val: 1 2 3 4 5 6 7 8 9 7 8 4 5 1 2

Table.I COO Sparse Matrix

CSR: Compressed sparse row. CSR is (val, col_ind, row_ptr), where val is an array of the (left-to-right, then top-to-bottom) non-zero values of the matrix; col_ind is the column indices corresponding to the values; and, row_ptr is the list of value indexes where each row starts.

$$A=\begin{pmatrix} 1 & 0 & 0 & 2 & 0 & 3 \\ 4 & 5 & 0 & 0 & 0 & 0 \\ 0 & 6 & 7 & 0 & 0 & 8 \\ 9 & 0 & 0 & 10 & 11 & 12 \\ 0 & 13 & 0 & 0 & 14 & 0 \\ 0 & 0 & 0 & 0 & 0 & 15 \end{pmatrix}$$

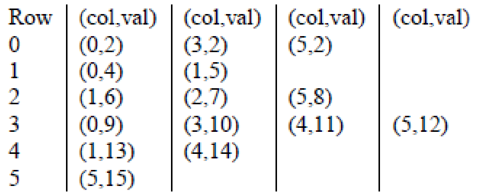
vals: 1 2 3 4 5 6 7 8 9 10 11 12 13
cols: 0 3 5 0 1 1 2 5 0 3 4 5 1
rows:0 3 5 8 12 14 15

Table.II CSR Sparse Matrix

LOL: List of Lists. LOL stores one list per row, where each entry stores a column index and value.

$$A=\begin{pmatrix} 1 & 0 & 0 & 2 & 0 & 3 \\ 4 & 5 & 0 & 0 & 0 & 0 \\ 0 & 6 & 7 & 0 & 0 & 8 \\ 9 & 0 & 0 & 10 & 11 & 12 \\ 0 & 13 & 0 & 0 & 14 & 0 \\ 0 & 0 & 0 & 0 & 0 & 15 \end{pmatrix}$$

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)



Row	(col,val)	(col,val)	(col,val)	(col,val)
0	(0,2)	(3,2)	(5,2)	
1	(0,4)	(1,5)		
2	(1,6)	(2,7)	(5,8)	
3	(0,9)	(3,10)	(4,11)	(5,12)
4	(1,13)	(4,14)		
5	(5,15)			

Table.III LOL Sparse Matrix

A major application of PageRank is searching. In general, the PageRank is a predictor for back-links. I have implemented PageRank in the following manner. First I need to make an initial assignment of the ranks based upon their number of in-links and out-links. Then I found results on tests of different web pages that PageRank is a better predictor of future citation counts. The experiment assumes that the system starts out with only a single URL and the goal is to try to gather the pages as more as possible as per the relevant links. For the purpose here, I have taken one function that will evaluate the rank as following manner i.e. simply the number of links present in the webpage, the number of clicks on a web page and the future probability of using that corresponding web page. The main important point is that all the information to calculate the function is not available until all the documents have been crawled. The benefits of PageRank are the greatest for underspecified queries. For example, a query for “matrix” may return any number of web pages which mention “matrix” on a conventional search engine, but using PageRank, the main page is listed first.

After implementing the hybrid algorithm in search engine, I have performed some comparative experiments and provided some sample results in this paper. The final page ranks are calculated as the combination of the prediction and the averaging of the results of COO and CSR matrix formats of different pages. I have taken a graph of 10 arbitrary nodes to find the result.

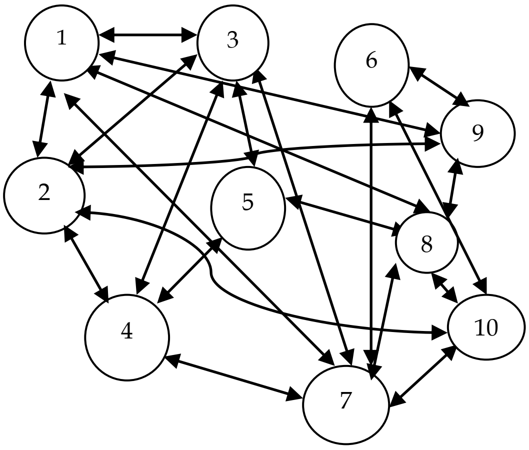


Fig.6 A graph of 10 nodes

PAGE RANKS are:-

Table.IV Page Ranks using COO Matrix

NODE	RANK
1	0.175500
2	0.752036
3	0.950620
4	0.150000
5	0.457001
6	0.814507
7	0.764011
8	0.653210
9	0.456321
10	0.950761

Table.V Page Ranks using CSR Matrix

NODE	RANK
1	0.752508
2	0.693493
3	0.756392
4	0.606766
5	0.450181
6	0.866437
7	0.703172
8	0.722522
9	0.822889
10	0.150000

Table.VI Page Ranks using Hybrid Algorithm

NODE	RANK
1	2.687866
2	3.378383
3	4.853506
4	2.733592
5	3.453591
6	1.639605
7	4.464004
8	2.840472
9	1.550380
10	4.722764

V. CONCLUSION

In this work, I have taken on the task of condensing few pages on the Web into a single number, its PageRank. PageRank is a global ranking of all web pages based purely on their number of in-links and out-links. Using the hybrid PageRank approach, I am able to order search results so that more important web pages are given higher preference. In experiments, this turns out to provide higher quality search results to users. Furthermore, back-links from important pages are more significant than back-links from average pages and the number of hits on a page and the future probability of using a page gave an interesting result.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Despite the promising results in terms of accuracy, further efforts will be requested to control the future prediction towards the usage of the web pages and the scalability of pages into Semantic Web repositories.

REFERENCES

[1] N. Duhan, A. K. Sharma and K. K. Bhatia, “Page Ranking Algorithms: A Survey, *Proceedings of the IEEE International Conference on Advance Computing*, 2009.

[2] M. G. da Gomes Jr. and Z.Gong, “Web Structure Mining: An Introduction”, *Proceedings of the IEEE International Conference on Information Acquisition*, 2005.

[3] S. Brin, L. Page, “The Anatomy of a Large Scale Hypertextual Web search engine,” *Computer Network and ISDN Systems*, Vol. 30, Issue 1- 7, pp. 107-117, 1998.

[4] J. Kleinberg, “Authoritative Sources in a Hyper-Linked Environment”, *Journal of the ACM* 46(5), pp. 604-632, 1999.

[5] Wenpu Xing and Ali Ghorbani, “Weighted PageRank Algorithm”, In proceedings of the 2nd Annual Conference on Communication Networks & Services Research, PP. 305-314, 2004.

[6] Ali Mohammad Zareh Bidoki and Nasser Yazdani, “DistanceRank: An Intelligent Ranking Algorithm for Web Pages”, Information Processing and Management, 2007, pp-22.

[7] “Relation-Based Page Rank Algorithm for. Semantic Web Search Engines”, In IEEE Transaction of KDE, Vol. 21, No.1, pp-20-30, Jan 2009.

AUTHOR PROFILE

Sharmila Subudhi has completed M.Tech(IT) from College of Engineering and Technology, Bhubaneswar, Odisha and is currently working as lecturer in the Dept. of CSE/IT/MCA at Gandhi Institute For Technology, Bhubaneswar, Odisha.