A Hybrid Page Rank Algorithm: An Efficient Approach

Madhurdeep Kaur
Research Scholar
CSE Department
RIMT-IET, Mandi Gobindgarh

Chanranjit Singh
Assistant Professor
CSE Department
RIMT-IET, Mandi Gobindgarh

ABSTRACT
As the web is escalating day by day, so the most concerned issue for the users would be how to collect the useful information and to find their genuine information effectively and quickly. With the tremendous growth of information available to end users through the web, search engines play a vital role in retrieving and organizing relevant data for various purposes. The ranking of the web pages for the web search engine is one of the significant problems at present. This leads to the important attention to the research community. In this paper, a page rank mechanism called Hybrid Page Rank Algorithm is proposed which is based on both content and link structure of the web pages. This algorithm is used to find more relevant information according to user's query. This paper also presents the comparison between SimRank Algorithm and the Hybrid Page Rank Algorithm.

Keywords
WWW; Data mining; Web mining; Search engine; Page ranking

1. INTRODUCTION
With the rapid development of internet, the number of internet users in the world is increasing very quickly and suddenly. Therefore, to manage the rapidly growing size of WWW and to retrieve only related web pages when given a searched query, current information retrieval approaches need to be modified to meet these challenges. The search engines are used to serve this purpose. They are used to collect, organize, index and serve results in its own unique way. However, all major search engines such as Google, Yahoo, MSN etc follow some general rules. When a user enters a query at a search engine site, then it returns a large number of pages in response to user query. Every search engine depends on its Ranking mechanism so that the user can find the most important and relevant results first. There are different types of algorithms developed; few of them are Page Rank, Weighted Page Rank, SimRank, HITS etc [3, 4, 5, 7, 8, 9, 14]. Most of the ranking algorithms proposed are either link or content based. Search engines are the main application of web mining. Web Mining is the extraction of appealing and latently useful patterns and implicit information from artifacts or movements associated to the World Wide Web. Web mining is the important technique of data mining. The absolute process of extracting knowledge from the web data [1] is follows in Fig 1. Web Mining is categorized into 3 main categories [1, 2]: Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM) [24]. Web Mining uses these categories to achieve information from the web. Every category of web mining has its own application areas including business intelligence, site improvement & modification, web personalization, ranking of pages etc.

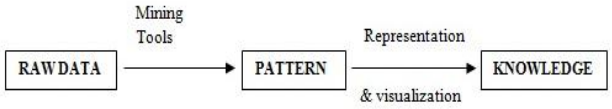

Figure 1 Web Mining Process

In this paper, a Hybrid Page Rank Algorithm is proposed which is based on both content and link structure of the web pages. In order to optimize the results of a search engine by retrieving the more relevant pages on top of the search result list, we have to consider both link structure as well as content of the web pages.

1.1 Organization of the Paper
The rest of the paper is organized as follows: we first introduce about some related work regarding Page Ranking approaches in section II. In section III, we present our proposed system in which we describe the overall architecture of the system and a framework that computes the rank of the web pages on the basis of content and link structure of the web pages. Section IV, includes the comparison of SimRank algorithm with the proposed algorithm in context of accuracy and average time of retrieval. Finally in section V, we conclude the paper and discuss some future directions for the system.

2. RELATED WORK
Kleinberg's [7] HITS algorithm, Brin & Page's Page Rank algorithm [4, 5] are the most important page ranking algorithms. Most of the search engines use these algorithms which can arrange the web documents in order of their relevance, importance and content score. The Alta Vista Search Engine implements HITS algorithm [25]. But HITS (Hyperlink Induced Topic Search) ignores the textual content and is purely link structure based computation. Page rank algorithm is used by the Google [6]. Google first retrieve a list of relevant pages to a given query based on factors such as title tags and keywords. Then it use Page Rank to adjust the results so that more important pages are provided at the top of the page list. In [8], the algorithm was proposed called Weighted Page Rank by Wenpu Xing and Ali Ghorbani, which is an extension of Page Rank algorithm. It is based on the popularity of the web pages and assigns the rank values accordingly rather than dividing it evenly. SimRank [14] is a new page rank algorithm which is based on similarity measure from the vector space model and is used to assign the more relevant score to the web pages.

All the above mentioned algorithms have been proposed in the literature. These are used to rank the query results of web pages in an efficient manner. Some algorithms depend only on the link structure of the document i.e their popularity score (web structure mining), some look for the content of the

document (web content mining), while other use a combination of both i.e they use links as well as the content of the web document to assign a rank value to the concerned document. In this paper, a Hybrid Page Rank algorithm is proposed which is based on both, the link structure as well as the content of the web documents.

3. PROPOSED FRAMEWORK

3.1 Outline of the Algorithm

Page Rank based on Links of the web pages [26]: Page Rank algorithm is a link based algorithm. The rank or score of the page is calculated on the basis of in-bound and out-bound links of the pages. Inbound links or commonly known as "backlinks" are the links pointing to our web page which decides the rank of the page. Outbound links are the links pointing to other pages from our web page. In our approach we are using a local web repository. A link structure is defined, showing how pages are linked to each other. This link structure will be used to calculate the page rank. Eq.1 is used to calculate the page rank on the basis of links.

$$PR(u) = (1-d) + d \sum_{v \in S(u)} PR(v)/TL(v) \quad (1)$$

where S(u) is the set of pages that points to u, PR(u) and PR(v) denotes the rank scores of page u and v respectively. *TL(v)* denotes total number of outgoing links of page *v*, d is a damping factor that is usually set to 0.85. d can be thought of as the probability of users following the links and (1-d) as the page rank distribution from non-directly linked pages.

Page Rank based on Content of the web pages [26]: Web Mining is basically extracting the information from the web. Retrieving the content of a web page is a process of web content mining. While searching for the pages according to user query, only using link based algorithm, cannot be considered an appropriate approach. User is searching the web for some relevant content that can never be found just listing pages on the basis of link structure. Therefore content based scoring algorithm is important. In our approach we use the modified SimRank algorithm [14] which is a content based algorithm to assign more relevant score to the pages using similarity measure. Pages are scored on the basis of user query and best scoring pages are listed further to create final list of pages. To calculate the rank on the basis of content of web pages, Eq.2 is used.

$$CPR(u) = occur(q,u) * d * \sum_{u \in D} wpc(u) * fieldscore(f,u) \quad (2)$$

where CPR(u) denotes content based page rank of page u. Occur(q, u) denotes query based score that is how many terms of the query are found in the page u. Fieldscore (f, u) defines a score on the basis of the field (f) of page (u) that contains the terms of the query. Note that the content of a crawled page contains two parts or fields: title and body. The weight values assigned to them are different and computed using Eq.3.

$$wpc(u) = tf(iu) * idf(i) \quad (3)$$

where tf(iu) denotes term frequency that is number of time term i appears in page u. And idf(i) is the inverse page frequency that is total number of pages divided by pages in which term i appear.

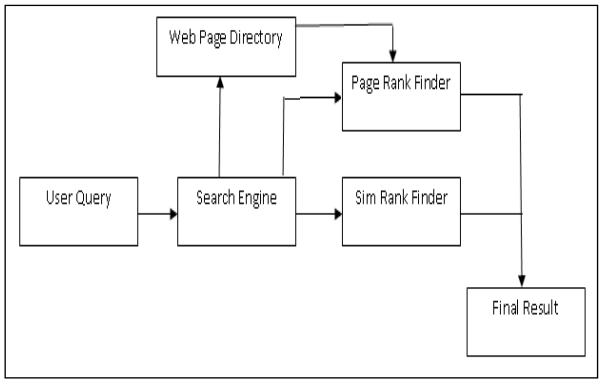

Figure 2 Frame work of our Proposed Model

3.2 Research Methodology

Input: Web document D = {D1, D2,……..DN}

Output: Relevant documents.

Step1: A repository (database) of web pages is created

Step 2: After creating the database a link structure will be created that will explain how pages are linked to each other. On the basis of links, page rank will be calculated for each page at the beginning.

Step 3: User will add a query and database will be searched for the pages related to user query.

Step 4: Pages will be searched for user query. Web Pages will be selected on the basis of their similarity content and those are similar to user search will be selected for user. Web Page similarity will be calculated using modified Sim-Rank technique i.e. content based rank.

Step 5: After having the web pages those are matched with user query, their page ranks will be compared. Pages with high page and content rank will be placed on top of the search result list. To build our final search list we will consider both web pages content and links. HPR denotes hybrid page rank.

$$HPR_{vol}(u) = x\%\ of\ PR(u) * y\%\ of\ CPR(u) \quad (4)$$

By using Eq.4 a Hybrid Page Rank is computed by combining the percentage of page rank and content based page rank to get a final score and to generate a final list of pages.

4. COMPARISON OF RESULTS BETWEEN SIMRANK AND HYBRID PAGE RANK ALGORITHM

4.1 On the basis of Accuracy:

When a user enters a query, then both SimRank and Hybrid page rank algorithm returns a large number of pages in response to user search. The accuracy is calculated on the basis of number of relevant pages or results retrieved by both algorithms in first Top 5 and Top 10 search results. As shown in Table I, the Hybrid page rank algorithm shows more accuracy than the SimRank algorithm by providing the more relevant web pages for different keywords like galaxy, rainforest etc.

Table 1 Comparison on the basis of Accuracy

SNo.	*Searched Keywords*	*No. of relevant web pages by SimRank algorithm*		*No. of relevant web pages by Hybrid page rank algorithm*	
		In Top 5 results	*In Top 10 results*	*In Top 5 results*	*In Top 10 results*
1.	andromeda galaxy	3	7	5	10
2.	galaxy	4	8	5	10
3.	computers	4	7	5	7
4.	socket programming	5	7	5	7
5.	rain forest	2	7	5	7
6.	rainforest	3	6	5	10
7.	environment	4	6	5	9
8.	global warming	4	8	5	8
9.	pollution	5	8	4	10
10.	air	4	8	5	9
11.	memory	2	2	5	9
12.	virus	0	0	5	7
13.	stars	3	6	5	7
14.	network	4	4	5	9
15.	programming	5	7	5	7
16.	network connection	5	8	5	9
17.	client-server	5	7	5	8
18.	UNIX	3	3	5	8
19.	web	2	2	3	7
20.	atmosphere	4	5	5	6
21.	pollution prevention	4	8	5	10
22.	health	3	4	5	8
23.	milky way	2	7	5	10
24.	light year	3	5	2	4
25.	tropical	3	8	5	9

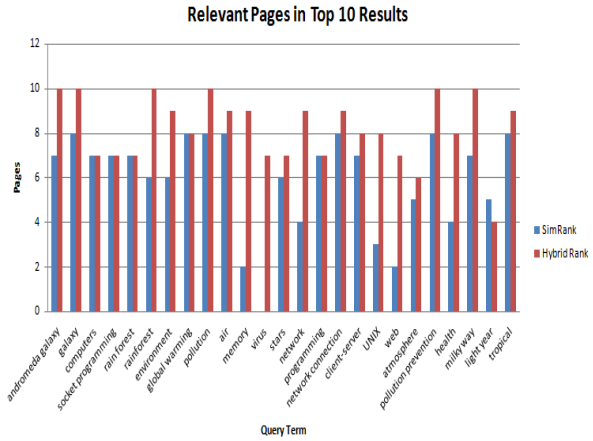

Figure 3 Comparison on the basis of Accuracy

4.2 On the basis of Average Time of Retrieval:

The time taken to retrieve the relevant results is the average time of search after running search engine 5 times for each query for both SimRank and Hybrid Page Rank algorithm. As shown in Table II, the average time of retrieval is calculated for both algorithms, but the Hybrid page rank algorithm provides the more relevant results in less time than the SimRank algorithm.

Table 2 Comparison on the basis of Average time of retrieval

SNo.	*Searched Keywords*	*Average Time of retrieval (in milli seconds)*	
		SimRank Algorithm	*Hybrid page rank Algorithm*
1.	andromeda galaxy	111.8	55.2
2.	galaxy	106.8	41.8
3.	computers	128	35.8
4.	socket programming	108	32.4
5.	rain forest	111	57
6.	rainforest	110	35
7.	environment	78	40
8.	global warming	115	45
9.	pollution	90	46
10.	air	98	35
11.	memory	130	37
12.	virus	105	36
13.	stars	117	60

14.	network	105	40
15.	programming	109	39
16.	network connection	117	46
17.	client-server	110	37
18.	UNIX	85	36
19.	web	105	43
20.	atmosphere	120	48
21.	pollution prevention	110	42
22.	health	80	40
23.	milky way	90	35
24.	light year	110	42
25.	tropical	105	36

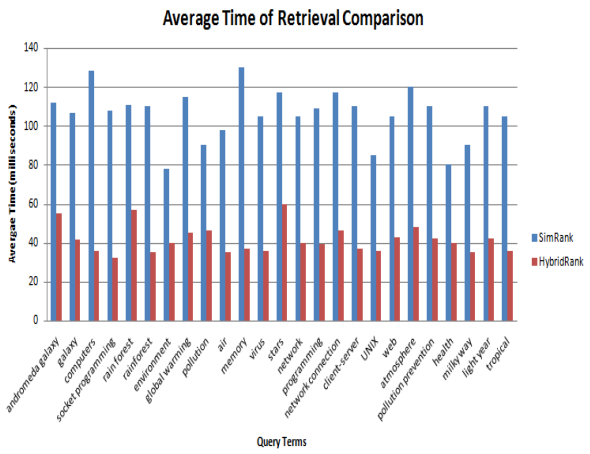

Figure 4 Comparison on the basis of Average Time of Retrieval

As shown in both the tables, our proposed algorithm (Hybrid page rank algorithm) is showing better results for most of the searched keywords in comparison to SimRank algorithm on the basis of accuracy and time taken to retrieve the web pages in response to user query.

4.3 Search Engine's Results:

A search engine is a platform which is used for searching the information which is relevant for users. Through this platform, user can add his query to search pages. The platform has provided an option to search on the basis of Page Rank, SimRank and Hybrid Page Rank Algorithm. After adding a query term in the search box user will have to choose a searching technique, then according to his chosen technique, the results will be displayed as shown in Fig 5, Fig 6, Fig7. The search engine also provides the searching time taken by both the algorithms that is displayed on the top left corner of the web page shown in Fig 6 and Fig 7.

4.3.1 Main interface of the search engine

Figure 5 Main interface of the search engine

4.3.2 Search engine results using SimRank Algorithm

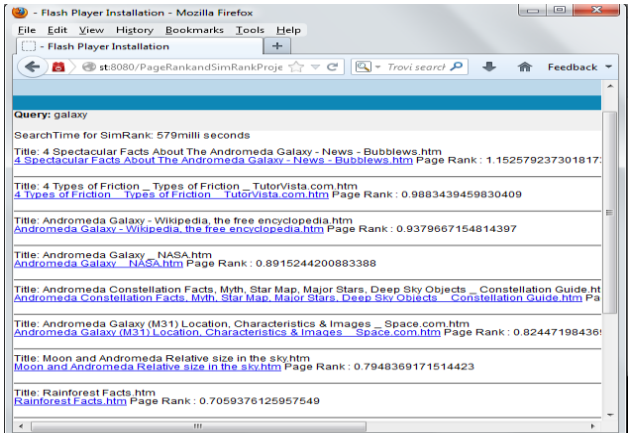

Figure 6 Search results using SimRank Algorithm

4.3.3 Search engine results using Hybrid Page Rank Algorithm

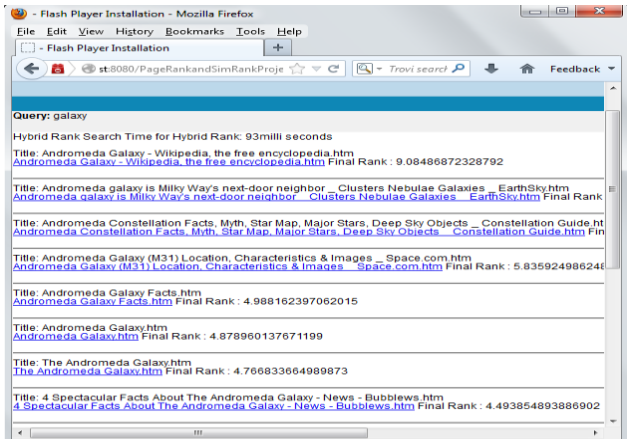

Figure 7 Search results using Hybrid Page Rank Algorithm

5. CONCLUSION

On the basis of this study it is concluded that both link and content based algorithms are important to calculate a final score or page rank of a web page. The usual search engines results in large number of pages in response to user's queries, while the user always want to get the best in a petite time. In order to rank massive web pages accurately and effectively, we propose a Hybrid Page Rank Algorithm which computes the score on the basis of content as well as link structure of the web pages. A comparison is made between the SimRank and Hybrid page rank algorithm on the basis of accuracy and time of retrieval. As we have conducted our research in a limited environment, in future we will expand it to a larger environment. We will embed a crawler that will directly index the pages from the web. Further concept of pre-fetching will be added where we will add an extra module, which will keep the track of most popular searches of users and when next time a user search a popular query, instead of searching index and repeating whole process again, we can display the list from the stored cache. This will reduce the searching time for few popular queries and will make search fast.

6. ACKNOWLEDGEMENT

I am extremely grateful and remain indebted to all the people who have given their intellectual support throughout the course of this work. And a special acknowledgement to the authors of various research papers and books which help me a lot.

7. REFERENCES

[1] R.Cooley, B.Mobasher and J.Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web". In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence(ICTAI'97), 1997.

[2] Companion slides for the text by Dr. M. H. Dunham, "Data Mining:Introductory and Advanced Topics", Prentice Hall, 2002

[3] Jaroslav Pokorny, Jozef Smizansky, "Page Content Rank: An Approachto the Web Content Mining".

[4] 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.

[5] C. Ridings and M. Shishigin, "Pagerank Uncovered". Technical report,2002.

[6] http://WWW.webrankinfo.com/english/seo-news/topic-16388.htm.

[7] January 2006, Increased Google index size. Kleinberg J., "Authorative Sources in a Hyperlinked Environment". Proceedings of the 23rd annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998.

[8] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm",Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04), 2004 IEEE.

[9] http://www.google.com/technology/index.html, Our Search: Google Technology.

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

[11] Bing Liu. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications). Springer-Verlag NewYork, Inc., Secaucus, NJ, USA, 2006.

[12] Lizorkin, D., Velikhov, P., Grinev, M., Turdakov, D., "Accuracy estimate and optimization Techniques for Simrank Computation", Published in ACM, Print ISBN No: 978-1-60558-305-1, on 24-30 Aug 2008, pp. 422-433.

[13] Li, C., Han, J., He, G., Jin, X., Sun, Y., Yu, Y., Wu, T., "Fast Computation of SimRank for Static and Dynamic Information Networks", Published in ACM, Print ISBN No: 978-1-60558-9045-9, on 22-26 March 2010.

[14] Qiao, S., Li, T., Li, H., Zhu, Y., Peng, J., Qin, J., "SimRank : A Page Rank Approach based on Similarity Measure", Published in IEEE, Print ISBN No: 978-1-4244 -6793-8, 2010, pp. 390-395.

[15] Taneja, H., Gupta, R., "Web Information Retrieval using Query Independent Page Rank Algorithm", International Conference on Advances in Computer Engineering, Published in IEEE, Print ISBN No: 978-0-7695-4058-0, 2010, pp. 178-182.

[16] Ma, H., Chen, S., WANG, D., "Research of PageRank Algorithm Based on Transition Probability", International Conference on Web Information Systems and Mining, Published in IEEE, Print ISBN No: 978-0-7695-4224-9, 2010, pp. 153-155.

[17] Cailan, Z., Kai, C., Shasha, Li., "Improved PageRank Algorithm Based on Feedback of User Clicks", Published in IEEE, Print ISBN No: 978-1-4244-9763-8,2011,pp. 3949-3952.

[18] Kumar, G., Duhan, N., Sharma, A.K., "Page Ranking Based on number of Visits of Links of Web Page", International Conference on Computer & Communication Technology (ICCCT), Published in IEEE, Print ISBN No: 978-1-4577-1386-6,2011, pp. 11-14.

[19] Zhao, C., Zhang, Z., Li, H., Xie, X., "A Search Result Ranking Algorithm Based on Web Pages and Tags Clustering", Published in IEEE, Print ISBN No: 978-1-4244-8728-8, 2011, pp.609-614.

[20] Sharma, R., Kandpal, A., Bhakuni, P., Chauhan, R., Goudar, R.H., Tyagi, A., "Web Page Indexing through Page ranking for Effective Semantic Search", 7th International Conference on Intelligent Systems and Control (ISCO), Published in IEEE, Print ISBN No: 978-1-4673-4603-0, 2012.

[21] Jain, A., Sharma, R., Dixit, G., Tomar, V., "Page Ranking Algorithm in Web Mining, Limitations of existing methods and a new method for Indexing Web Pages", Published in IEEE, Print ISBN No: 978-0-7695-4958-3,2013,pp.640-645.

[22] Hyperlink Analysis: Techniques and Applications Prasanna Desikan, Jaideep Srivastava, Vipin Kumar, and Pang-Ning Tan, Department of Computer Science, University of Minnesota, Minneapolis, MN, USA {desikan, srivastava, kumar, ptan} @cs.umn.edu.

[23] A Comparative Analysis of Web Page Ranking Algorithms, Dilip Kumar Sharma et al. / (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 08, 2010, 2670-2676.

[24] J. Srivastava, R. Cooley, M. Deshpande, and P. –N. Tan. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data" (2000), SIGKDD Explorations, Vol. 1, Issue 2, 2000

[25] Alta Vista Search Engine; http:// www. altavista.com

[26] Kaur, M., Singh, C., "A Hybrid Page Rank Algorithm using Content and Link Based Algorithms", Global Journal of Advanced Engineering Technologies (GJAET) Vol 3, Issue-2, 2014, 2277-6370

[27] Kaur, M., Singh, C., "Content Based and Link Based Page Ranking Algorithms: A Survey", International Journal of Advanced and Innovative Research (IJAIR) Vol 3, Issue-4, 2014, 2278-7844