Ignoring Irrelevant Pages in Weighted PageRank Algorithm using Text Content of the Target Page

Sunil Kumar
M.Tech. Scholar
Shobhit University, Meerut

Niraj Singhal
Associate Professor
Shobhit University, Meerut

ABSTRACT
The web is expanding day-by-day and people generally rely on search engines to explore the web. The web has created many challenges for information retrieval. Degree of quality of the information extracted is one of the major issue to be taken care of, and current information retrieval approaches need to be modified to meet such challenges. While doing query based searching, the search engines return a list of web documents containing both relevant and irrelevant pages and sometimes show the higher ranking to the irrelevant pages as compared to relevant pages. This paper presents a novel approach to ignore irrelevant pages in weighted pagerank algorithm using text content of the targeted pages.

General Terms
Web Page Ranking for information retrieval

Keywords
Page rank, Irrelevant pages, Page content, Links.

1. INTRODUCTION
The web is a collection of document pages and hyperlinks that interconnect them. It is very large in size and heterogeneous in nature. Web is expanding day-by-day and people generally rely on search engine to explore the web. The web has created many new challenges [1] for information retrieval like huge size of web, some pages of the web do not possess the quality of self-descriptiveness, degree of quality of the information extracted and the conclusion of the knowledge from the extracted information and semi-structured nature of web pages (i.e. in the form of lists, tables etc.). According to Google [2] on 25th July 2008 there were one trillion unique URLs on the web. Today the number would be many folds than that.

To manage the rapidly growing size of the web and to retrieve only relevant web pages when given a searched query, current information retrieval approaches need to be modified to meet these challenges. Currently while, doing query based searching, the search engine returns a list of web documents containing both relevant and irrelevant pages and sometimes shows the higher ranking to the irrelevant pages as compared to relevant pages. The search engines use one of the following approaches to organize search and analyze information on the web. In the first approach [3], the search engine selects the terms for indexing a web page by analyzing the frequency of the words appearing in the target web page. The second approach [4, 5, 9, 11] uses the structure of the links appearing between pages to identify pages that are often referenced by other pages. Another method [6, 7, 10] analyzes the content of the pages linked to or from page of interest.

Due to the heterogeneous nature of the web the retrieval approach based on single source of information suffer from some weaknesses that can affect the retrieval performance. For example, content-based information retrieval approach does not consider the link of the page while ranking the page and thus affect the quality of web page, while link based approach [8, 9,11] can suffer from incomplete or noisy link topology. This inadequacy of single source information retrieval approaches gives a strong argument for combining multiple sources of information as potential strategy for information retrieval. In this paper, a novel approach to rank relevant pages higher in the retrieved document set is presented that considers both analysis of links and analysis of text content.

2. RELATED WORK
In this section, we review the various existing ranking algorithms and their limitations, which are then used as a base for ignoring ir-relevant pages in Weighted PageRank algorithm. The web is growing tremendously, providing proper and relevant information of the highest quality to the users based on their search query becomes increasingly difficult. This is due to the reason that some web pages are not self-descriptive and some pages are made only for navigation purpose. Therefore, searching relevant pages through a search engine that makes use of hyperlink information is very difficult. For ranking of web pages, several algorithms have been proposed. Among them are PageRank [14] and Hypertext Induced Topic Selection(HITS) [15,16] algorithms. PageRank ranks pages based on the link structure of the web pages. It measures the importance of the pages by analyzing the links [17, 18].

2.1 The Hits Algorithm-Hubs and Authorities
In this algorithm, there are two forms of web pages called authorities and hubs (see figure1). Authorities are the web pages that are pointed to by many hyperlinks whereas web pages that point to many hyperlinks are called are called hubs [19, 20, 21].

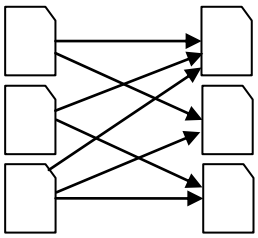


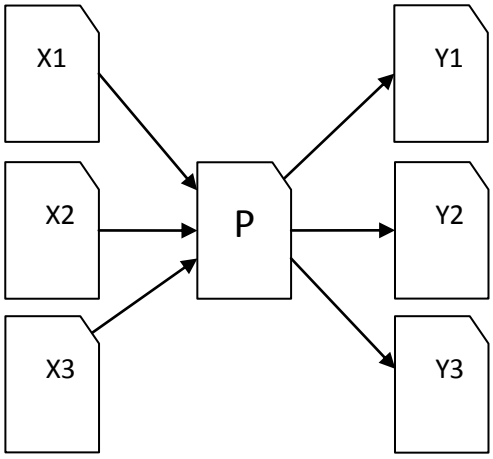
Fig 1: Hubs and authorities

Scores assigned to hubs and authorities are computed in a mutually reinforcing way. An authority pointed to by several highly scored hubs should be a popular authority while a hub that points to several highly scored authorities should be an important hub [19, 20]. The scores of hubs and authorities are calculated as follows [7, 19, 20]:

$$A_p = \sum_{q \in B(p)} H_q \quad (1)$$

$$H_p = \sum_{q \in I(p)} A_q \quad (2)$$

Where Ap and Hp are the authority and hub scores of page 'p', respectively. B(p) and I(p) denote the set of referrer and reference pages of page 'p', respectively. The page's authority score is equal to the sum of hub scores of pages that it points to it [9]. Similarly, the page's hub score is equal to the sum of the authority scores of pages that is links to. Computation of both is shown in figure 2.



$A_p = H_{x1} + H_{x2} + H_{x3}$ $\qquad$ $H_p = A_{y1} + A_{y2} + A_{y3}$

Fig 2: Computation of Hubs and Authority Scores

Some problems in the HITS algorithm are as follows [15]:-

(i) High rank value is given to some popular website that is not highly relevant to the given query.

(ii) Drift to the topic occurs when the hub has multiple topics as equivalent weights are given to all the outlinks.

2.2 PageRank Algorithm

The PageRank algorithm is the most commonly used algorithm for ranking web pages. The PageRank algorithm is based on the citation analysis, states that if a page contains important links toward it then the links of this page towards the other page are also considered important. PageRank takes the backlinks into accounts in deciding the rank score. A page has a high rank if the sum of the ranks of its backlinks is high [14, 18]. Figure 3 illustrate the backlinks (Page 'U' is the backlinks of page 'V' and 'W' while page 'V' and page 'W' are backlinks of page 'X').

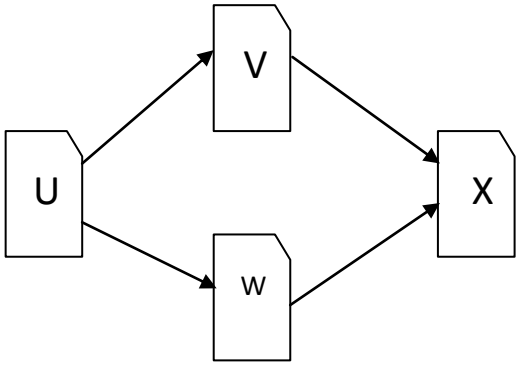


Fig 3: An example of backlinks

A simplified version of PageRank is given as [18]:-

$$PR(n) = \sum_{m \in B(n)} \frac{PR(m)}{N(m)} \quad (3)$$

Where 'n' is the web page for which page rank is dependent on the page rank values for each web page m out of the set B(n) (i.e. the set of pages that point to n). PR (n) and PR (m) are rank scores of page 'n' and 'm' respectively. N(m) represents the number of outgoing links of page 'm'. c is a normalization factor.

In PageRank algorithm, the rank of a page, 'p', is equally divided between its outgoing links, which in turn is used to calculate the ranks of the pages to which page 'p' is pointing. If two or more pages are connected to each other to form a loop and these pages did not refer to but are referred by other webpages outside the loop, they would collect rank but never distribute any rank. This is referred to as rank sink [18].

To solve the rank sink problem, the original PageRank is published [14, 18]:

$$PR(n) = (1 - d) + \sum_{m \in B(n)} \frac{PR(m)}{N(m)} \quad (4)$$

Where 'd' damping factor is the probability at each page the "random surfer" will get bored and request another random page.

2.3 Weighted PageRank Algorithm

Weighted PageRank Algorithm [22] is an extension of the PageRank. This algorithm assigns a larger rank values to the more important pages rather than dividing the rank value of a page evenly among its outgoing links. Each outlink gets a value proportional to its importance or popularity (number of inlinks and outlinks). The importance is determined in terms of weight values of inlinks and outlinks denoted as $W^{in}_{(m,n)}$ and $W^{out}_{(m,n)}$, respectively.

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

$$W^{in}_{(m,n)} = \frac{I_n}{\sum_{p \in R(m)} I_p} \quad (5)$$

Where I_n and I_p represent the number of inlinks of page 'n' and page 'p' respectively. R(m) denotes the reference page list of page m.

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

$$W^{out}_{(m,n)} = \frac{O_n}{\sum_{p \in R(m)} O_p} \quad (6)$$

Where O_n and O_p represent the number of outlinks of page 'n' and page 'p' respectively. R(m) denotes the reference page list of page 'm'.

Considering the importance of pages, the formula as proposed by Wenpu et al [22] is as follows:-

$$WPR(n) = (1 - d) + d \sum_{m \in B(n)} PR(m)\, W^{in}_{(m.n)}\, W^{out}_{(m,n)} \quad (7)$$

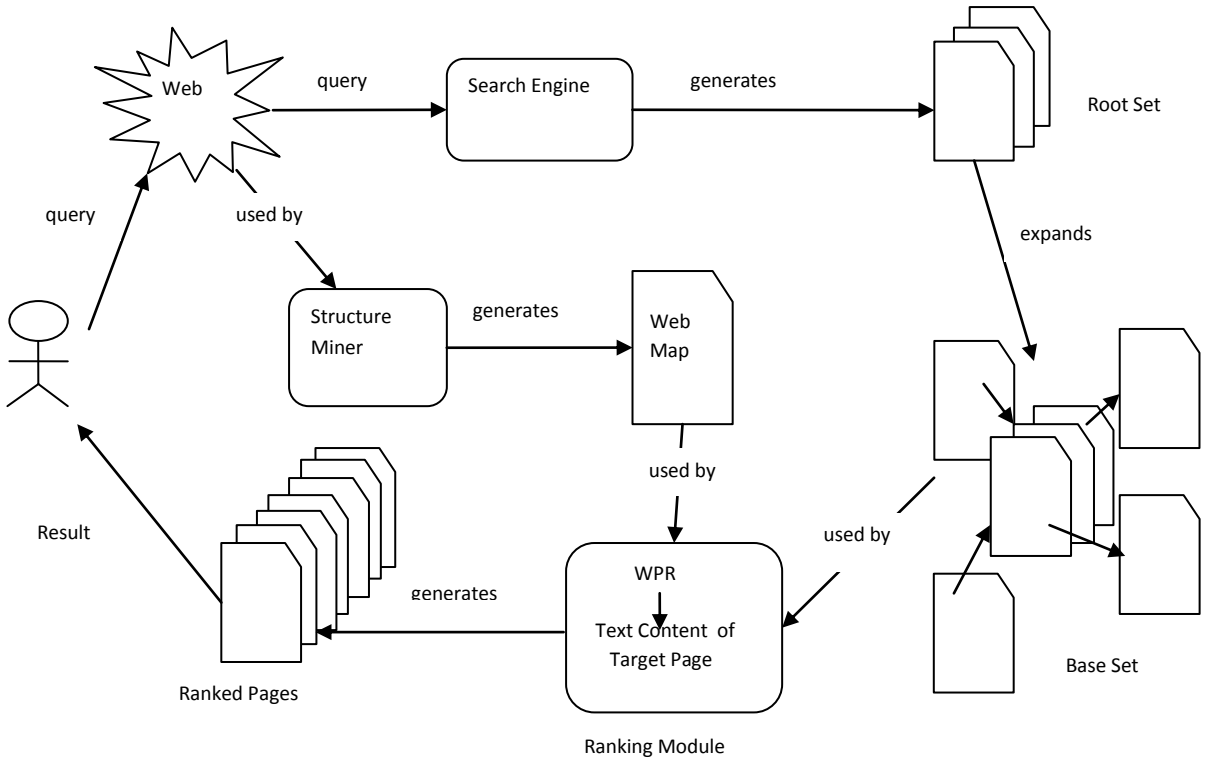


Fig 4: The architectural components of the system

3. PROPOSED WORK

Though the Weighted PageRank algorithm (see Eq. 7) provides important information about a given query by using the structure of the website, but it suffer from some weaknesses that affect the performance. For example, some pages irrelevant to a given query are included in the result as well. This is because Weigted PageRank algorithm uses the structure of the website and some irrelevant pages get the highest rank because of its many existing inlinks and outlinks. So to reduce the noise resultant from irrelevant pages, it is needed to keep relevant pages higher in the retrieved document set. The architecture of the proposed approach is shown in figure 4.

All the link analysis algorithms [8] use the inlinks and outlinks of web page to score the target web pages. Initially a search engine returns a set of web pages relevant to the given search query. This set of web pages is called the root set. Then this root set is expanded to obtain a base set of pages that directly point to or are pointed to by the pages in the root set.

After that a hyperlink directed graph (see figure 5) G = (V, E) is constructed from the base set with the web pages defining the set of nodes V and the links between web pages defining the set of edges 'E' in the graph. This graph 'G' can be described by an n x n adjacency matrix A, where aij = 1, if there is a link from page i to page j and aij = 0 otherwise. The set of nodes that point to node i (backward links) are represented by the vector B(i) = { j: aji = 1} and the set of nodes that are pointed to by node i (Forward links) are represented by the vector F(i) = { j: a_{ij} = 1}.

It proposes an idea to rank the pages based not only on the link structure but also by analysis the content of the target pages. The proposed method use Vector Space Model (VSM) technique to represent each page as a vector of terms. VSM computes the relevance of each term to the page [23] using the term frequency information to generate weights for all terms in a document and represents the document as term frequency weight vectors, so that document j is represented by the vector

$$W_{ij} = 1.....k$$

where, k is the total number of unique terms appearing in the page.

The weight of a term can be computed using different methods. In the proposed method Term Frequency (TF) weighting approach [23] is used to compute the weight of the term in each page. The weight of an i^{th} term using TF weighting is:-

$$W_i = \frac{tf_i}{T} \qquad (8)$$

where tf_i is the number of times the i^{th} term appears in the document. 'T' is the maximum frequency of any term in the document. In the same way, one can calculate the effective weight of each term in a document and store it in the inverted word document table [24] against the corresponding word with the document information in its posting list.

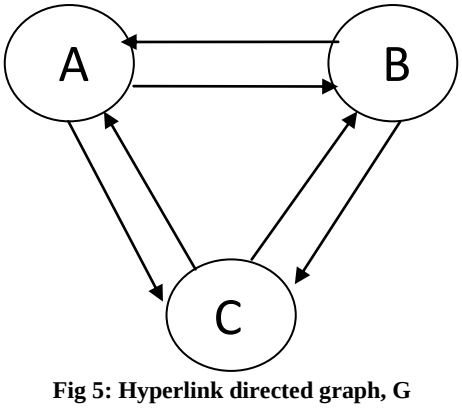


Fig 5: Hyperlink directed graph, G

4. CONCLUSION AND FUTURE SCOPE

This paper describes ranking scores of pages computed through different link analysis ranking algorithms and proposed a novel approach to rank relevant pages that considers both analysis of links and analysis of text content. The new proposed method reduces one of the limitations (i.e. including ir-relevant pages in the retrieved set) of Weighted PageRank Algorithm while computing the rank of the retrieved web pages. The work will extended by applying the text content information of the backward and forward hyperlinks for page ranking.

5. REFERENCES

[1] M. G. da Gomes Jr. and Z.Gong, "Web Structure Mining: AnIntroduction", Proceedings of the IEEE International Conference on Information Acquisition, Hong Kong and Macau, China, pp. 590-595,2005.

[2] Google Official Blog, http://googleblog.blogspot.com/2008/07.

[3] Justin Zobeland Alistair Moffat, "Inverted Files for Text Search Engines", ACM Computing Surveys, 38 (2), pp. 1-56,2006.

[4] Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, and Panayiotis Tsaparas, "Finding Authorities and Hubs from link structures on the World Wide Web", Proceedings of the 10th WWW Conference, Hong Kong, pp. 415-429, 2001.

[5] David Gibson, Jon Kleinberg, and Prabhakar Raghavan, "Inferring Web Communities from Link Topology", Proceedings of the 9th Conference on Hypertext and Hypermedia, Pittsburgh, Pennsylvania, pp. 225-234, June 1998. Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[6] Kiduk Yang, "Combining text-and link-based retrieval methods for Web IR", in Proceedings of 10th Text REtrieval Conference, pp. 609—618, 2001.

[7] R. Kosala, and H. Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations,Newsletter of the ACM Special Interest Group on Knowledge Discovery and DataMining,Vol. 2, No. 1, pp 1-15, 2000

[8] Boleslaw K. Szymanski, and Ming-shu Chung, "A method for Indexing Web Pages Using Web Bots", in Proceedings of the International Conference on Info-Tech Info-Net ICII'2001, Beijing, China, IEEE CS Press, pp.1-6,2001.

[9] Monika R. Henzinger, and Krishna Bharat, "Improved algorithms for topic distillation in a hyperlinked environment", in Proceedings of the 21st International ACM SIGIR conference on Research and Development in IR, pp. 104-111,1998.

[10] Soumen Chakrabarati, Byron Dom, David Gibson, Jon M. Kleinberg, Prabhakar Raghavan, and Sridhar Rajagopalan, "Automatic resource list compilation by analyzing hyperlink structure and associated text", in Proceedings of the 7th International WWW conference, 30(1-7), pp. 65-74, 1998.

[11] Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, and Panayiotis Tsaparas,"Link analysis ranking: algorithms, theory, and experiments", in ACM Trans. Inter. Tech., 5(1), pp. 231-297, 2005.

[12] Neelam Duhan, A. K. Sharma and Komal Kumar Bhatia, "PageRanking Algorithms: A Survey",proceedings of the IEEEInternational Advanced Computing Conference (IACC), pp 1530-1537, 2009.

[13] C. Ridings and M. Shishigin,"Pagerank uncovered", Technical report, 2002.

[14] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan,A. Tomkins, D. Gibson, and J. Kleinberg,"Miningthe Web's link structure",Computer, 32(8), pp.60–67, 1999.

[15] S. Pal, V. Talwar and P. Mitra, "Web mining in soft computing framework: Relevance, state of the art and future directions",IEEE Trans. Neural Networks, 13(5), pp.1163–1177,2002

[16] S. Brin and L. Page," The anatomy of a large-scale hypertextual Web search engine", Computer Networks and ISDN Systems,30(1998), pp.107–117,1998.

[17] L. Page, S. Brin, R. Motwani and T. Winograd," The page rank citation ranking: Bringing order to the web", Technical report, Stanford Digital Libraries SIDL-WP-1999-0120,1999.

[18] C. Ding, X. He, P. Husbands, H. Zha and H. Simon," Link analysis: Hubs and authorities on the world", Technical report:47847, 2001.

[20] J. Wang, Z. Chen, L. Tao, W. Ma and W. Liu," Ranking user's relevance to a topic through link analysis on web logs", WIDM, pp. 49–54, 2002.

[21] W. Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research, IEEE, 2004.

[22] P. C. Saxena, J. P. Gupta and Namita Gupta, "Web Page Ranking Based on Text Content of Linked Pages" International Journal of Computer Theory and Engineering, Vol. 2, No.1, February, 2010.

[23] Prem Chand Saxena, and Namita Gupta," Quick Text Retrieval Algorithm Supporting Synonyms Based on Fuzzy Logic", Computing Multimedia and Intelligent Techniques, 2(1), pp. 7-24, 2006.