International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Volume 2, Issue 6 (Nov-Dec 2014), PP. 57-60

WEB PAGE RANKING BASED ON TEXT SUBSTANCE OF LINKED PAGES

SWATI KUMARI[1], Mr. ASHOK SHAKYA[2]
[1] Research Scholar, [2] Department of Computer Science,
SAROJ INSTITUTE OF TECHNOLOGY AND MANAGEMENT,
Lucknow, UP, INDIA
katiyardevesh@gmail.com

Abstract: World Wide Web is large sized repository of interlinked hypertext documents accessed via the Internet. Web may contain text, images, video, and other multimedia data. The user navigates through this using hyperlink. Search Engine gives millions of results and applies Web mining techniques to order the results. The sorted order of search results is obtained by applying some special algorithms called—Page ranking algorithms. The algorithm measures the importance of the pages by analyzing the number of inlinked and outlinked pages. Our proposed system is built on an idea that to rank the relevant pages higher in the retrieved document set, an analysis of both page‘s text substance and links information is required. The proposed approach is based on the assumption that the effective weight of a term in a page is computed by adding the weight of a term in the current page and additional weight of the term in the linked pages. In this chapter, we first study the nature of web pages, the various link analysis ranking algorithms and their limitations and then show the comparative analysis of the ranking scores obtained through these approaches with our new suggested ranking approach.

I. INTRODUCTION

To manage the rapidly growing size of World Wide Web and to retrieve only related Web pages when given a search query, current Information Recuperation approaches need to be modified to meet these challenges. Presently, while doing query based searching, the search engines return a list of web pages containing both related and unrelated pages and sometimes showing higher ranking to the unrelated pages as compared to relevant pages. Nature of Web search environment is such that the recuperation approaches based on single sources of evidence suffer from weaknesses that can hurt the recuperation performance. For instance, substance-based Information Recuperation approach does not consider the pages link by the page while ranking the page and hence affect the quality of web documents, while link-based approaches [1] can suffer from incomplete or noisy link topology. This inadequacy of singular Web Information Recuperation approaches make a strong argument for combining multiple sources of evidence as a potentially advantageous recuperation strategy for Web Information Recuperation.

II. LINK ANALYSIS RANKING ALGORITHMS

Different existing LAR algorithms [2], [3] along with their limitations are described in this section.

A. *HITS (Hyperlink Induced Topics Distillation)*

HITS algorithm defines a mutual reinforcing relationship between the authorities and hubs and shows that a good hub points to good authorities and a good authority is pointed to by good hubs. In order to quantify the quality of a page as a hub and an authority, Kleinberg associated with every page a hub and an authority weight. It uses an iterative algorithm for computing the hub and authority weights. Initially all authority and hub weights are set to one. At each iteration, the authority and hub weight of a node is computed. The algorithm iterates until the vectors converges. HITS consider the whole graph, taking into account the structure of the graph around the node to compute its hub and authority scores. The automatic Resource Compilation (ARC) system, described in [5], augments Kleinberg's link-structure analysis by considering the anchor text also, the text which surrounds the hyperlink in the pointing page. ARC computes a distance-2 neighborhood graph and weights edges. The weight of each edge is based on the match between the query terms and the text surrounding the hyperlink in the source page. The reasoning behind this is that many times, the pointing page describes the destination page's substances around the hyperlink, and thus the authority conferred by the links can be better assessed. Bharat and Henzinger [1], studied Kleinberg's approach and have proposed improvements to it. The connectivity analysis has been shown to be useful in identifying high quality pages within a topic specific graph of hyperlinked documents. The essence of their approach is to augment a previous connectivity analysis based algorithm with substance analysis. The results of a user evaluation are reported that show an improvement of precision at 10 documents by at least 45% over pure connectivity analysis. Nomura et al. [8] also attempted to improve HITS algorithm and proposed two types of link-analysis based modification: the projection method and the base-set downsizing method to solve HITS‘s topic drift problem. Topic drift problem occurred when in the base set, a large and densely linked set of unrelated Web pages are included so that the authorities converge into these densely linked irrelevant pages and couldn‘t extract appropriate relevant Web communities. The proposed two methods in [8] take into account the number of links to/from pages included in the root set to extract appropriate web communities.

B. *SALSA (Stochastic Approach for Link Structure Analysis)*

Like Kleinberg‘s HITS algorithm, SALSA [6], [7] starts with a similarly constructed Base Set. It performs a two-step random walk on the bipartite hub and authority graph, alternating between the hub and authority sides. The random walk starts from some authority node selected uniformly at random. When at a node on the authority side, an incoming link is uniformly selected at random and moves to a hub node on the hub side. Each hub divides its weight equally among the authorities to which it points and the authority weight of a node is computed by summing up the weight of the hubs that point to it. Similarly, when at a node on the hub side, the algorithm selects one of the outgoing links uniformly at random and moves to an authority. Each authority divides its weight equally among the hubs that point to it and the hub weight of a node is computed by summing up the weight of the authorities that it point to.

C. Salsa (popularity SALSA)

It is a simplified version of SALSA and also performs a two-step random walk on the bipartite hub and authority graph, alternating between the hub and authority sides. But the starting point for the random walk is chosen with probability proportional to the "popularity" of the node, that is, the number of links that point to this node.

D. HThresh (Hub Threshold)

The algorithm is similar to HITS, but to compute the authority weight of the i^{th} page, it does not consider all hubs that point to page *i* but only those whose hub weight is at least the average hub weight over all the hubs that point to page *i*, computed using the current hub weights for the nodes. This overcomes the drawback of pSALSA by assigning lower authority weight to a site which points to large number of poor hubs.

This algorithm is similar to HITS, but to compute the hub weight of the i^{th} page, it does not consider all authorities that are pointed by page *i* but only those authorities which are among the top K authorities, judging by current authority values. Hence, for a site to be a good hub, it must point to some of the best authorities. The algorithm is based on the fact that, in most web searches, a user only visits the top few authorities.

E. BFS (Breadth-First-Search Algorithm)

Breadth-First-search algorithm ranks the nodes according to their reach ability i.e., the number of nodes reachable from each node. The algorithm starts from node *i*, and visits its neighbors in BFS order, alternating between backward and forward steps. Every time one link is moved further from the starting node *i*, the weight factor of the node is updated accordingly. The algorithm stops either when *n* links have been traversed, or when the nodes that can be reached from node *i* are exhausted.

III. LIMITATIONS OF THE EXISTING LINK ANALYSIS RANKING ALGORITHMS

Following observations have been made [2], [4], [6], and [7] about the different link analysis ranking algorithms:

1. Kleinberg algorithm is biased towards tightly-knit communities (TKC) and ranked set of small highly interconnected sites higher than those of large set of interconnected sites which is having hub pointing to a smaller part of the authorities.
2. Inappropriate zero weights can be seen in HITS regardless of the output's dependence on or independence of the initial vector.
3. In multi-topic collections, the principal community of authorities found by the Kleinberg approach tends to pertain to only one of the topics in the collection.
4. For both HITS and SALSA, there are some graphs that give rise to repeated eigenvalues. The output of such graphs is sensitive to the initial vector chosen.
5. pSALSA algorithm place greater importance on the in-degree of a node when determining the authority weight of a node and favors various authorities from different communities. The algorithm is local in nature and the authority weight assigned to a node depends only on the links that point to the node. But counting the in-degree as the authority weight is sometimes imperfect as it sometimes results in pages belonging to unrelated community ranked higher than the pages belonging to related community.
6. Hub-average algorithm also favors nodes with high in-degree. It overcomes the shortcoming of the HITS algorithm of a hub getting a high weight when it points to numerous low-quality authorities. So to achieve a high weight a hub should link good authorities. But the limitation of the algorithm is that a hub is scored low as compared to a hub pointing to equal number of equally good authorities if an additional link of low quality authority is added to it.
7. Threshold algorithms (AThresh, HThresh, and FThresh) eliminate unrelated hubs when computing authorities and hence try to remove the TKC effect as seen in HITS algorithm. The results obtained from threshold algorithms are 80% similar to HITS algorithm.
8. BFS algorithm exhibits best performance among all LAR algorithms. BFS is not sensitive to tightly-knit communities as the weight of a node in the BFS algorithm depends on the number of neighbors that are reachable from that node. It also avoids strong topic drift as seen in HITS algorithm.

IV. PROPOSED METHOD

On the basis of the limitations of existing Link analysis ranking algorithms described in section 4.5, a method is proposed to compute the degree of relevance of all linked pages (both forward links and backward links) to target page based on text substance analysis of the linked pages. The proposed ranking algorithm represents each page as a vector of terms using Vector Space Model technique (VSM). VSM estimates the relevance of each term to the page using the term frequency information to generate weights for all the terms in a document and represents the documents as term frequency weight vectors, so that document *j* is represented by the vector $(wij) = 1 \ldots.. m$

Where, *m* is the total number of unique terms appearing in the document

To calculate the weight of each term in given page, we use Term Frequency (TF) weighting approach. The weight of i^{th} term using TF weighting is

$$w_i = \frac{tf_i}{T}$$

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 current page p

Page is ranked higher if it contains functional links (i.e., links to pages related to similar topic). To differentiate forward links as functional or navigational links, the substance of forward linked pages is considered and if they are related to similar topic then forward links are considered as functional links. The idea behind is that if there are two pages having same number of forward links and let first page contains forward links discussing similar topic and other page contains forward links not related to similar topic as described in target page, then in such case first page should be ranked higher than the second irrespective both have same number of out-links.

Hence, the additional weight of i^{th} term in current page *p* due to forward links is computed as:

$$aw_i = \frac{1}{H}\left(\sum_{j=1}^{H} L_{ji}\right), \quad L_{ji} = \frac{tf_{ji}}{T}$$

where, H is the number of pages linked by page p tf_{ji} is the number of times the i^{th} term appears in the j^{th} document
T is the maximum frequency of any term in j^{th} page

Likewise, a page is ranked higher if it is pointed to by pages that are also related to the similar topic. This will remove the problem of assigning high rank score to a page linked by pages showing no topic similarity to target page, but showing links just to improve their ranking as seen in pSALSA LAR algorithm. Based on this concept, higher score can be assigned to a page with few backward links but having functional links in comparison to a page having large number of non-functional navigational backward links.

Hence, the additional weight of i^{th} term in current page *p* due to backward links is computed as:

$$aw'_i = \frac{h'}{H'}$$

where, h' is the total number of pages pointing to page p having tf_{ji} >= average term frequency in page

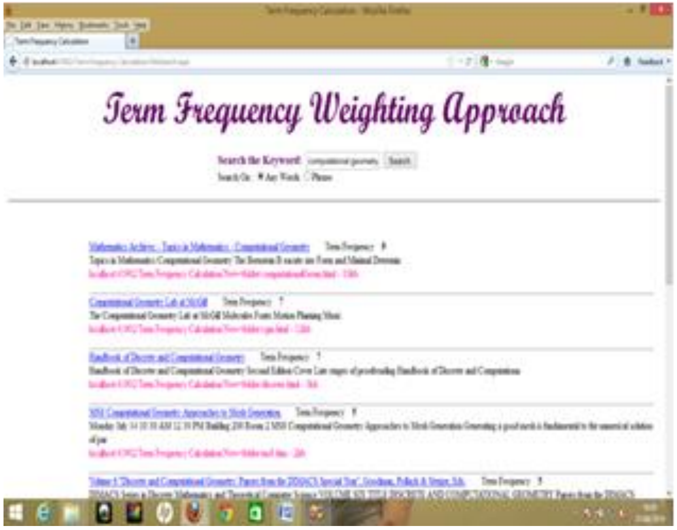
The effective weight of the i^{th} term in page *p* is thus given

Similarly, we can calculate the effective weight of each word in a page and stored it in the inverted_word_document table against the corresponding word with the page information in its posting list [Sergio]. Whenever, a search query is given, for each search query term, inverted_word_document table is searched to retrieve documents list against query term from the table.

We analyze the similarity of the word usage at single level link distance from the page of interest and demonstrate that information about the linked pages enables more efficient indexing and search. The computational overheads to gather the link information of the pages is ignored while computing the total cost to solve the user query as the link information is collected prior to user query input, once the pages are indexed on the local databases by the crawlers. Also the overheads to collect the linked pages are less since we consider the immediate single distance backward links and forward links.

V. EXPERIMENTAL RESULTS

To calculate the weight of each term in given page, we developed a search tool in .Net Environment using C# to find the term frequency. The screenshot of output is shown below:



The proposed method considers the page substance of backward link pages, forward link pages and the substance of the target page to compute the rank score of the target page. The proposed algorithm reduces the limitations of the other link analysis ranking algorithms by differentiating between navigational and functional links. It is also based on the concept that only good hubs are considered in computing the ranking of the target page and only good authorities contribute in computing the final ranking of the target page. A hub is considered good if it points to pages which are related or similar to same topic as described in its own page. Similarly a good authority is the one which is pointed to by pages which are discussing the related/similar topic as given in its own page substance. This is clearly depicted in the results obtained by implementing the proposed algorithm for different queries on the base dataset as shown below.

The ranking of web pages computed by our proposed algorithm is comparable to the ranking score obtained by other LAR algorithms. Slight variations in the ranking of the web pages are due to error in retrieving some of the backward links and forward links of some root web pages. The reasons for this error can be modifications made in the web page or server containing the target page down at the time of searching of the web page.

VI. DISCUSSION

The results of the “abortion” query as shown in table 4.8 shows zero rank score for many web pages as computed by different LAR algorithms. Web page P-81 has zero ranking score in many LAR algorithms (Kleinberg, HubAvg, AThresh, FThresh) or very small rank score in others (pSALSA, SALSA, HThresh, PageRank) but is assigned highest rank score by our ranking algorithm. Page P-81 belongs to category 2 having three backward links and fourteen forward links. But since backward links of P-81 page are few and also all are not related to target topic so shows zero or very small ranking score in many LAR algorithms whereas our ranking algorithm equally considers all the three parameters (page substance, backward links, forward links) for computing a page rank score and hence compute non-zero ranking score for P-81 since the substance of P-81 is related to the target topic. Similar is the case with P-56, P-135 (belonging to category 2 having zero and one backward link respectively), P-74 (belonging to category 6 with neither page substance nor backward link related to given topic). All are having zero or low ranking score in all LAR algorithms and non-zero or high ranking score in our ranking algorithm. P-119 is scored high in all LAR algorithms and low in our ranking algorithm since as compared to others web pages it‘s neither page substance nor backward links are related to target topic. It has only two forward links referring to target topic and hence scored low by our algorithm. The ranking score results obtained by our ranking algorithm shows maximum similarity with the ranking scores obtained by other algorithms as shown in table 4.3 and 4.6 and also our algorithm shows better ranking results by assigning non-zero ranking values to these web pages.

VII. CONCLUSION

In this paper, a method is proposed for learning web structure to classify web documents and demonstrate the usefulness of considering the text substance information of backward links and forward hyperlinks for page ranking. It is shown that utilizing only extended anchor text or just considering the words and phrases in the target pages (full-text) does not yield very accurate results. On the bases of results obtained by analyzing the similarity of the word usage at single level link distance from the page of interest, it is shown that the substance of words in the link pages (Forward and back links) enables more efficient indexing and searching. The new proposed method efficiently reduces the limitations of the already existing Link Analysis ranking algorithms described in the chapter and the results obtained by the proposed method are not biased towards in-degree or out-degree of the target page. Also navigational, functional and noisy links are identified based on similarity between the terms of the link pages with target page. The rank scores computed through given new method showed non-zero values (in case either target page itself or its link pages contains information on given search query) and hence help to rank the web pages more accurately.

REFERENCE

[1] Bharat K., and Henzinger M.R., “Improved Algorithms for Topic Distillation in a Hyperlinked Environment”, Proc. 21st International ACM SIGIR conference on Research and Development in IR, pp. 104-111, 1998.

[2] Borodin A., Roberts G.O., Rosenthal J.S., Tsaparas P., “Finding Authorities and Hubs from Link Structures on the World Wide Web”, Proc. 10th WWW Conference, Hong Kong, pp. 415-429, 2001.

[3] Borodin A., Roberts G.O., Rosenthal J.S., Tsaparas P., “Link Analysis Ranking: Algorithms, Theory, and Experiments”, ACM Transactions on Internet Technology, vol. 5, no. 1, pp. 231-297, 2005.

[4] Bratley P., and Choueka Y., “Processing truncated terms in document retrieval systems”, Information Processing and Management, vol. 18, no. 5, pp. 257-266, 1982.

[5] Chakrabarati S., Dom B., Raghavan P., Rajagopalan S., Gibson D., Kleinberg J.M., “Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text”, Proc. 7th International WWW conference, pp. 65-74, 1998.

[6] Farahat A., Lofaro T., Miller J.C., Rae G., Ward L.A., “Authority Rankings from HITS, PageRank, and SALSA: Existence, Uniqueness, and Effect of Initialization”, SIAM J. Science Computing, vol. 27, no. 4, pp. 1181-1201, 2006.

[7] Lempel R., and Moran S., “The stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect”, Proc. 9th International World Wide Web Conference, Amsterdam, Netherlands, pp. 387-401, 2000.

[8] Nomura S., Oyama S., Hayamizu T., Ishida T., “Analysis and Improvement of HITS Algorithm for Detecting Web Communities”, Proc. Symposium on Applications and the Internet (SAINT'02), pp.132-140, 2002.

[9] “Data Mining”, http://en.wikipedia.org/ wiki/Data mining, 28 May 2011.

[10] Dalal M.K., and Zaveri M.A., “Heuristics Based Automatic Text Summarization of Unstructured Text”, Proc. International Conference and Workshop on Emerging Trends in Technology (ICWET 2011), pp. 690-693, 2011.