International Journal of Education and Science Research Review P-ISSN 2349-1817, E- ISSN 2348-6457
www.ijesrr.org December- 2017, Volume-4, Issue-6 Email- editor@ijesrr.org

WEB PAGE RANKING BASED ON TEXT CONTENT OF LINKED PAGES

DESHMUKH KAMAJI VITTHAL RAO
Research Scholar
OPJS University
Rajasthan

DR. YOGESH KUMAR
Supervisor
OPJS University
Rajasthan

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 outli nked pages.

KEY WORDS: multimedia, algorithm, image, video

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 Retrieval 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. These search engines use one of the following approaches to organize, search and analyze information on the web:

- Analyze the similarity of the word usage at different link distance from the page of interest and demonstrate that structure of words used by the linked pages enables more efficient indexing and search.
- Search engine selects the terms for indexing a web page by analyzing the frequency of the words (after filtering out common or meaningless words) appearing in the entire or a part of the target web page.
- The structure of the links appearing between pages is considered to identify pages that are often referenced by other pages. Analyzing the density, direction and clustering of links, such method is capable of identifying the pages that are likely to contain valuable information.
- Anchor text of a hyperlink is considered to describe its target page and so target pages can be replaced by their corresponding anchor text.
- Nature of Web search environment is such that the retrieval approaches based on single sources of evidence suffer from weaknesses that can hurt the retrieval performance. For example, content-based Information Retrieval 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 can suffer from incomplete or noisy link topology. This inadequacy of singular Web Information Retrieval approaches make a strong argument for combining multiple sources of evidence as a potentially advantageous retrieval strategy for Web Information Retrieval. Ranking of a web page is highly influenced by the following factors:
  1. In-links to a page and out-links from a page
  2. Text content of a web page
  3. Anchor text of the hyperlinks in the page

NATURE OF WEB PAGES
Web pages of different types are retrieved as a result of search query from the WWW. The nature of information available in these pages varies. There are pages having no forward links and discusses about the relevant topic. There are also pages which are index pages having hyperlinks only without any description

on the search query topic. Sometimes some pages are retrieved which are not relevant to the topic. All the possible kinds of web pages are listed in the table 1 given below.

TABLE 1: LIST THE DIFFERENT CATEGORIES OF WEB PAGES.

Category	Web page discussing query topic	Web page having Forward links related/similar to query topic	Web page having Back links related/similar to query topic
1	Y	Y	Y
2	Y	Y	N
3	Y	N	Y
4	Y	N	N
5	N	Y	Y
6	N	Y	N
7	N	N	Y
8	N	N	N

Let us discuss by an example the different nature of possible information contained in a web page.

- A web page discussing the same topic as listed in search query. For example, retrieved a web page on "text mining" for the search query "text mining".
- A web page containing Forward links on the same topic as listed in search query. For example, retrieved a web page containing forward links to pages discussing the topic "text mining" for the search query "text mining".
- A web page containing Back links on the same topic as listed in search query. For example, retrieved a web page with Back links discussing the topic "text mining" for the search query "mining".

- A web page containing Forward/Back links discussing topic not given in search query. For example, retrieved a web page containing forward links to pages discussing the topic like "Spanning tree protocol" for the search query "Spanning tree" whereas the user is interested in "Spanning tree graph".

PRELIMINARIES

All the link analysis ranking algorithms use either in-links (backward links pointing to a page), out-links (forward links pointed by the page) or both in a web page to score the retrieved web pages. Initially a search engine is used to retrieve a set of web pages relevant to the given search query. This creates a Root set. Then this Root Set is expanded to obtain a larger Base Set of Web pages by adding those pages which are pointing to the pages (backward links) of the original Root Set and the pages which are pointed to by the pages (forward links) of the original Root Set. Next, a hyperlink directed graph $G = (V,E)$ is constructed from the Base set with the web pages defining the nodes $1, \ldots, n$, and the links between the web pages defining the edges in the graph. This graph is described by an $n \times n$ adjacency matrix A, where $a_{ij} = 1$ if there is a link from page i to page j and $a_{ij} = 0$ otherwise. The vector $B(i) = \{j:a_{ji}=1\}$ represents the set of nodes that point to node i (backward links) and the vector $F(i) = \{j:a_{ij}=1\}$ represents the set of nodes that are pointed to by node i (forward links). All the link-based ranking algorithms are based on the idea that a web page serves two purposes: to provide information on a topic, and to provide links to other pages giving information on a topic.

This gives rise to two ways of categorizing a web page:

- A web page is defined as an authority node in graph G having nonzero in-degree, if it provides good information about the topic given in search query.

- A web page is defined as hub node in graph G having nonzero out-degree, if it provides links to good authorities on the topic mention in search query.

Let *a* represent the set of authority nodes, *h* denotes the set of hub nodes, $G_a = (a, E_a)$ denotes the undirected authority graph on the set of authorities *a* having an edge between the authorities *i* and *j*, if $B(i) \cap B(j) \neq \emptyset$.

LIMITATIONS OF THE EXISTING LINK ANALYSIS RANKING ALGORITHMS

Following observations have been made about the different link analysis ranking algorithms:

- 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.
- Inappropriate zero weights can be seen in HITS regardless of the output's dependence on or independence of the initial vector.
- 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.
- 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.
- 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.
- 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.
- Threshold algorithms 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.
- 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.

EXPERIMENTAL RESULTS

The proposed method considers the page content of backward link pages, forward link pages and the content 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 discussed 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 content. This is clearly depicted in the results obtained by implementing the proposed algorithm for different queries on the base dataset as shown below.

TABLE-2 : ALGORITHM FOR DIFFERENT QUERIES

Web Page	Category	URL	TITLE
P-5	5	http://www.cs.duke.edu/CGC/workshop97.html	Second CGC Workshop on Computational Geometry
P-10	6	http://jeff.cs.mcgill.ca/cgm.html	Computational Geometry Lab at McGill
P-12	7	http://cs.smith.edu/~orourke/books/discrete.html	Handbook of Discrete and Computational Geometry
P-20	5	http://www.ics.uci.edu/~eppstein/266	Computational Geometry
P-21	5	http://dimacs.rutgers.edu/Volumes/Vol06.html	Volume 6 "Discrete and Computational Geometry: Papers from the DIMACS Special Year", Goodman, Pollack & Steiger, Eds.
P-23	5	http://archives.math.utk.edu/topics/computationalGeom.html	Mathematics Archives - Topics in Mathematics - Computational Geometry
P-27	3	http://www.siam.org/meetings/archives/an97/ms8.htm	MS8 Computational Geometry Approaches to Mesh Generation
P-28	5	http://www.math-inst.hu/staff/geometry.html	Convex and Computational Geometry research group
P-31	5	http://www.sonic.net/~sjl/compgeom.html	Computational Geometry
P-50	5	http://www.risc.uni-linz.ac.at/projects/basic/cgal	Computational Geometry Algorithms Library

TABLE 3: SHOWING THE RANK SCORES OF TEN WEB PAGES OBTAINED BY DIFFERENT LAR ALGORITHMS

Page	Kleinbe	pSALS	SALSA	HubAvg	AThres	HThres	FThresh	BFS	New
P-5	0.001454	0.003289	0.003055	0.000082	0.011785	0.073141	0.001335	243.070312	14.837037
P-10	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	5.000000
P-12	0.001450	0.001771	0.001645	0.000023	0.011436	0.073141	0.001246	241.453125	7.800000
P-20	0.003105	0.001265	0.001175	0.000059	0.028738	0.127205	0.002447	250.835938	15.500000
P-21	0.000216	0.000506	0.000470	0.000004	0.002680	0.011887	0.000267	138.656250	5.500000
P-23	0.003365	0.002530	0.002350	0.000138	0.035451	0.188979	0.003601	245.226562	16.416667
P-27	0.000054	0.000506	0.000470	0.000002	0.000760	0.002674	0.000067	92.156250	6.000000
P-28	0.002133	0.001265	0.001175	0.000044	0.023483	0.119727	0.002345	201.992188	4.000000
P-31	0.003403	0.001771	0.001645	0.000097	0.035314	0.188979	0.003329	243.351562	18.839286
P-50	0.000423	0.000506	0.000470	0.000048	0.007112	0.022403	0.000693	150.828125	7.000000

DISCUSSION

For query “Computational Geometry”, Page P-10 is assigned zero ranking score by all the LAR algorithms as listed in table 3, because page P-10 belongs to category 6 showing only two functional forward links and hence is poor hub. The page itself doesn‘t contain text related to the topic “Computational Geometry” but contain only forward links related to the target topic. Since all the LAR algorithms discussed above are influenced by the in- degree of a web page while computing the ranking score and hence assigned zero rank score to P-10. While our ranking algorithm consider all the three factors to compute the ranking score of a web page and hence assigned non-zero small rank score to P-10 as it has forward links which are linking to pages related to target topic. Only PageRank algorithm shows zero ranking score to pages P-21, P-27 and P-31 while other ranking algorithms (Alexa Rank, AltaVista results, and AllTheWeb) computes non-zero rank score for these pages which are similar to the results obtained by our algorithm. 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 content, backward links, forward links) for computing a page rank score and hence compute non-zero ranking score for P-81 since the content 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 content 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 content nor backward links are related to target topic.

CONCLUSION

In this paper, a method is proposed for learning web structure to classify web documents and demonstrate the usefulness of considering the text content 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 content 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 discussed in the paper 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.

REFERENCES

1. **[Aptoula, 2009]** E. Aptoula and S. Lefèvre, “Morphological Description of Color Images for Content-Based Image Retrieval,” in *IEEE Transactions on Image Processing*, Vol. 18, No. 11, November 2009, pp. 2505-2517.
2. **[Ren, 2008]** X. Ren, C. C. Fowlkes and J. Malik, “Learning Probabilistic Models for Contour Completion in Natural Images,” in *International Journal of Computer Vision*, 2008, 77, pp. 47–63.
3. **[Tao, 2008]** D. Tao, X. Tang and X. Li, “Which Components are Important for Interactive Image Searching?,” in *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 18, No. 1, January 2008, pp. 3-11.
4. **[Vincent, 2014]** L. Vincent and P. Soille, "Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13, No. 6, June 1991, pp. 583-598.
5. **[Xu, 2010]** J. Xu, A. Janowczyk, S. Chandran and A. Madabhushi, “A Weighted Mean Shift, Normalized Cuts Initialized Color Gradient Based Geodesic Active Contour Model: Applications to Histopathology Image Segmentation,” in *Proceedings of SPIE Medical Imaging 2010: Image Processing*, Vol. 7623, 76230Y.
6. **[Yang, 2002]** M. Yang, D. J. Kriegman and N. Ahuja, “Detecting Faces in Images: A Survey,” *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 1, January 2002.
7. **[Carson, 2010]** C. Carson, S. Belongie, H. Greenspan and J. Malik, “Blobworld: image segmentation using E-M and its application to image querying,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, 24, pp. 1026–1038.
8. **[Fukuda, 2011]** K. Fukuda, T. Takiguchi and Y. Ariki, “Graph Cuts by Using Local Texture Features of Wavelet Coefficient for Image Segmentation,” in *IEEE International Conference on Multimedia and Expo*, 2008, pp. 881 – 884.
9. **[Kass, 2013]** M. Kass, A. Witkin and D. Terzopoulos, “Snakes: Active Contour Models,” in *International Journal of Computer Vision*, 1988, pp. 321-331.
10. **[Medina-Carnicer, 2010]** R. Medina-Carnicer, A. Carmona-Poyato, R. Muñoz-Salinas and F. J. Madrid-Cuevas, “Determining Hysteresis Thresholds for Edge Detection by Combining the Advantages and Disadvantages of Thresholding Methods,” in *IEEE Transactions on Image Processing*, Vol. 19, No. 1, January 2010, pp. 165-173.