International Journal of Soft Computing 12 (1): 26-30, 2017
ISSN: 1816-9503

Improved Distance Page Rank Algorithm

[1]L. Lakshmi, [1]P. Bhaskara Reddy and [2]C. Shoba Bindhu
[1]MLR Institute of Technology, 43 Hyderabad, India
[2]Jawaharlal Nehru Technological University, Anantapur, Anantapur, India

Abstract: World wide web is full of information as sources of information are increasing every day. Information retrieval process is complicated today due to it depends on number of factors like the number of hyperlinks the page contains, the content of the page and the number of pages that are pointing to the particular page. Owing to massive increase in the size of the world wide web, retrieval of pertinent and worthwhile results for a given query is very difficult task. To provide most appropriate results for a given query, by reducing non relevant results and to decrease the time taken for retrieval, we propose an improved distance page rank algorithm.

Key words: World wide web, hyperlinks, page rank, distance rank, query, unique visit count

INTRODUCTION

The usage of world wide web is increasing every day. Millions of web pages are added to web every day. The web mining is defined as finding out analysis of useful information. Web mining mainly classified into web usage mining, web content mining, web structure mining (Haveliwala, 2003; Brin and Page, 1998). Web usage mining is utilized to look at information identified with the customer end, for example, the profiles of the users of the site, the site utilized for the particular time and period. Web usage mining techniques include pattern analysis, pattern filtering, aggregation and characterization (Wang *et al*., 2008; Pawar and Natani, 2014). Web content mining is used to search, extract useful information and examine data by search engine algorithms. Web content mining techniques includes information extraction, topic tracking, summation, categorization and clustering and information visualization (Zhang *et al*., 2008). Web structure mining is utilized to inspect the structure of a specific site and examine related information (Kumar *et al*., 2016). Ranking, link based classification, link based clustering, link strength and link cardinality are the techniques mainly used with web structure mining (Dubey and Roy, 2011). Working of search engine mainly categorized into three stages that is crawling, indexing and ranking and retrieval (Hemayati *et al*., 2012; Tin *et al*., 2014). Crawling involves acquisition of data about a site; it scans the site and gets a complete list the web page title, pictures, keywords it contains and links to other pages. Now a days, crawlers may read a copy of the complete page and also additional information like page layout and links are on the web page (Haveliwala, 2003). Indexing is the process of reading the data from crawler and keeping the data in a big database (Nagappan and Elango, 2015). All the data stored in vast data centers. Indexing is designed based on the factors like merge factors, storage techniques, index size, lookup speed and maintenance and fault tolerance (Brin and Page, 1998; Kumar *et al*., 2016). Ranking and retrieval is the process when the user enters a search query and the search engine displays the most relevant documents it finds that matches your query and it is the most complicated step (Tyagi and Sharma, 2012). Ranking is reduced to a computation of numeric scores (Huang and Li, 2011). Ranking is based on the factors like relevance feedback, document frequency and inverted document frequency. When the user enters a search query, if the size of the search query is lengthy, it is very difficult for the search engine to match all keywords present in the query (David *et al*., 2012). Every time the user has to refine the query until the user gets the required results (Haveliwala, 2003; Pawar and Nantani, 2014). The most challenging issue in most of the search engines is retrieval of most relevant web pages are arranged from top to bottom on the basis of user preferences and ranking (Kleinberg, 1998). To develop more efficient and effective web we need more efficient algorithms for web searching and crawling (Haveliwala, 2003). All search engines cannot index and crawl the entire web. So, we have to concentrate on most important web pages (Brin and Page, 1998). To index most important web pages we need a better ranking algorithm. Relevancy of web pages is based on the ranking of users. With better ranking algorithm, web search engine serves the user with more relevant web pages (Wang *et al*., 2008; Zhang *et al*., 2008). Most of the ranking algorithms used are having low relevance and precision. Most of the algorithms are based on upscale-get-upscaled problem (Tyagi and Sharma, 2012). In this study, we propose a solution to this problem by

Corresponding Author: L. Lakshmi, MLR Institute of Technology, 43 Hyderabad, India

using improved distance page rank algorithm in which authority update score and hub update score are calculated based on distance between the web pages. We present an algorithm for improved distance rank with this we retrieve more relevant information for the given query within less time.

MATERIALS AND METHODS

Improved distance page rank: To overcome pitfalls of existing page ranking algorithms we propose a good ranking algorithm which calculates the distance between the authoritative pages and hub pages with good authoritative score and hub score and with the user can reach authoritative page from hub page with minimum number of clicks and can reach a hub page from authoritative page with minimum number of clicks.

Definition 1: A good authoritative page for a given query is pointed by many good hub pages:

$$a(i) = \sum_{j:\,(j,i)\in F} h(j)$$

Definition 2: A good hub page for a given query is pointing to many good authoritative page:

$$h(i) = \sum_{q:\,(i,j)\in F} a(j)$$

Definition 3: If hub page i points to authoritative page j then the weight of link between pages hub page i and authoritative page j is equal to log out(i) where out(i) is the out degree of i.

Definition 4: Visitor count of a page VC can be defined as number of unique visitors to the page.

Definition 5: The distance between two hub and authoritative pages is the weight of shortest distancebetween i and j denoted $dist_{ij}$. So, the average distance of authoritative page from i to j with n number of web pages as hubs can defined as:

$$dist_j = \sum_{i=1}^{n} distij/n$$

In this study, we have used URL data set from UCI repository. Here we have considered a portion of web pages as web graph (Fig. 1). First we need to construct the adjacency matrix for the web graph:

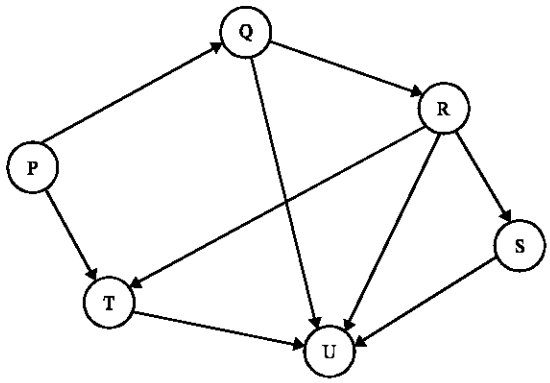

Fig. 1: Graph for improved distance page rank

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The authority score of all node of web graph are calculated by $AS = A^T V$. The default initial hub scores of all nodes of web graph are V = {1, 1, 1, 1, 1, 1}. The authority score of all nodes of web graph are:

$$A.S = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 2 \\ 4 \end{bmatrix}$$

The hub score of all node of web graph are calculated by HS = AU. The default initial authority scores of all nodes of web graph are U = {1, 1, 1, 1, 1, 1}. The hub score of all nodes of web graph are:

$$H.S = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 5 \\ 7 \\ 4 \\ 4 \\ 0 \end{bmatrix}$$

Form the calculation of hub and authority scores for the given web graph the node R is good hub which means

it is pointing good authoritative node and U is a good authoritative node which means it pointed by good hub node. By the consideration of only hub and authority score we will get the problem rich gets richest. To avoid this problem we are considering distance between the nodes of the web graph.

Now are calculating the distance between the hubs and authorities. We calculate the distance between the hubs and authorities based on hub score and authority score. Here we are considering the logarithmic values of hub and authority scores. If we want to calculate the distance between P and U to reach from P to U we have two paths with level 1 through T and Q. To calculate the distance we are considering minimum distance of these two paths. For example in Fig. 1, the logarithmic scores of P, T and Q are log (3), log (4) and log (5). If we choose the path P-T-U then distance is equal to log (3)+log (4). If we choose the path P-Q-U then the distance is equal to log (3)+log (5). Out of these distance the minimum distance is log (3)+log (4). Even through R is a good authoritative hub but we preferred to use shortest distance between pages to move from one web page to another web page. This reduces the problem of upscale-get-up scaled problem.

In addition to the consideration of minimum distance between the web pages we are also considering the unique visited count of the web pages. For example in Fig. 1, x, y are the users who visited the pages T and Q, if x visited the page T three times and Q two times, the user y visited the page T two times, then the visit count of T is two and visit count of Q is one. So, the best path to choose visit count is P-T-U instead of P-Q-U.

To find more relevant web pages for the given query in terms of hyperlinks, we need to consider the less distance between the web pages and more unique visit count of all web pages. With the usage of hubs and authority scores we are performing content search for the given query.

We used c = 1, 00,000 urls for testing our algorithm because web contains billions of web page as we cannot crawl all web pages at a time, so first we kept 1, 00,000 web pages on queue. Processing of those pages has been completed then we will load next one lakh web pages.

Algorithm for unique visit count of a web page:

Use queue to handle the connections to the web server served by thread.
Use timestamp for every connection and vc
For all connections present in queue, write them to file with their unique identifier.
When the connection is closed, dequeue the connection
When user requests for number of connections,
process the file, based on timestamp and return visit count values to user.
This can be run by using R tool for data analytics

Algorithm for improved distance page rank:

Construct adjacency matrix(A) for the web graph with URLs present on the queue.
Find the authority score and hub scores of the web pages by using $AS = A^T V$
$HS = AU$
Identify good hubs and authority in the web graph using authority and hub scores.
Find the unique visit count of all web pages using the algorithm specified.
Find the distance between the pages
$Dist_j = min_i(log(i)+d_i)$
Here the path from i-j, we consider minimum logarithmic distances in the path from i to j
Calculate the rank of the web page, with good authority score, good hub score, good unique visit count and
with minimum distance to reach web page.
This algorithm implemented by using R tool with URL data set from UCI repository

RESULTS AND DISCUSSION

We used URL data set from UCI repository with two million web pages to evaluate improved distance rank algorithm. Here, we used two schemas: unique visit count of a web page and authority and hub distance rank. In the unique visit count of a web page by using r tool we counted the unique visit count for all web pages. The web pages with highest unique count are known as most important web pages (Table 1).

Table 1: Unique Visit Count of a web page (UVC)

URL of web page	UVC
http://www.sinduscongoias.com.br/index.html	33
http://www.helpbackup.com.br/index.html	2
http://www.pontoprofissional.com.br/index.html	12
http://www.coleyglesias.com/index.html	63
http://88logistics.com/index.html	13
http://www.sandroecicero.com.br/index.html	30
http://www.tehobsledovanie.ru/index.html	72
http://www.mundialpisoseblocos.com.br/index.html	43
http://www.pizzariapontual.com.br/index.html	9
http://www.generalcustom.com.br/index.html	100
http://www.ambsal.it/ck.html	1
http://www.homedesign.com.br/index.html	98
http://sohacogroup.com.vn/index.html	45
http://www.amicidelgiocodelponte.it/index.html	2
http://www.uniaoparaobem.com.br/index.html	46
http://www.hotelmajore.it/ck.html	1
http://v-montazar.com/index.html	62
http://www.eca.edu.au/index.html	79
http://www.speyerseminar.de/index.html	78
http://www.jsjgw.com/index.html	59
http://www.leftoverpets.org/index.html	10
http://www.bsc-md.de/index.html	101
http://mov-designtec.de/index.html	20
http://www.reyderocha.com.co/index.html	74
http://www.passion-cosmetics.ch/index.html	27
http://www.surfgadget.com/index.html	53
http://www.tipulsini.co.il/index.html	37
http://www.missclean.rs/index.html	52
http://www.textilexpres.com/index.html	122
http://www.americanbussales.net/index.html	44

Table 2: The authority and hub distance rank

SNO	URL of web page	AS	HS	Distance
P1	http://www.sinduscongoias.com.br/index.html	0.423	0.465	0.812
P2	http://www.helpbackup.com.br/index.html	0.001	0.010	0.768
P3	http://www.pontoprofissional.com.br/index.html	0.121	0.152	0.921
P4	http://www.coleyglesias.com/index.html	0.522	0.589	0.145
P5	http://88logistics.com/index.html	0.101	0.123	0.845
P6	http://www.sandroecicero.com.br/index.html	0.254	0.224	0.721
P7	http://www.tehobsledovanie.ru/index.html	0.678	0.654	0.232
P8	http://www.mundialpisoseblocos.com.br/index.html	0.396	0.377	0.782
P9	http://www.pizzariapontual.com.br/index.html	0.12	0.100	0.855
P10	http://www.generalcustom.com.br/index.html	0.981	0.978	0.213
P11	http://www.ambsal.it/ck.html	0.001	0.001	0.986
P12	http://www.homedesign.com.br/index.html	0.978	0.988	0.294
P13	http://sohacogroup.com.vn/index.html	0.354	0.365	0.624
P14	http://www.amicidelgiocodelponte.it/index.html	0.010	0.112	0.981
P15	http://www.uniaoparaobem.com.br/index.html	0.326	0.254	0.372
P16	http://www.hotelmajore.it/ck.html	0.001	0.000	0.954
P17	http://v-montazar.com/index.html	0.569	0.459	0.358
P18	http://www.eca.edu.au/index.html	0.657	0.721	0.126
P19	http://www.speyerseminar.de/index.html	0.689	0.714	0.304
P20	http://www.jsjgw.com/index.html	0.456	0.342	0.452
P21	http://www.leftoverpets.org/index.html	0.100	0.123	0.856
P22	http://www.bsc-md.de/index.html	0.789	0.897	0.115
P23	http://mov-designtec.de/index.html	0.128	0.123	0.821
P24	http://www.reyderocha.com.co/index.html	0.721	0.748	0.164
P25	http://www.passion-cosmetics.ch/index.html	0.869	0.756	0.874
P26	http://www.surfgadget.com/index.html	0.497	0.473	0.399
P27	http://www.tipulsini.co.il/index.html	0.289	0.254	0.256
P28	http://www.missclean.rs/index.html	0.478	0.398	0.245
P29	http://www.textilexpres.com/index.html	0.999	1.000	0.175
P30	http://www.americanbussales.net/index.html	0.356	0.358	0.485

Table 3: Performance of improved distance rank algorithm

Page rank	Distance rank	Improved distance page rank
P29*	P22*	P18*
P10	P18*	P29*
P11	P4*	P22*
P25*	P24*	P10*
P22	P29*	P12*
P24*	P10	P19*
P19*	P7*	P24*
P7*	P28	P7*
P18	P27*	P4*
P19	P12	P20*
P4*	P19*	P26*
P23*	P17*	P28*
P28*	P15	P15*
P21	P26*	P13
P1*	P20	P30*
P8*	P30*	P8
P30*	P13*	P27
P13	P6*	P1*
P15*	P2	P6*
P27	P8*	P25

Precision = 0.6; Precision = 0.7; Precision = 0.80

The authority and hub distance rank, we calculated good authority and hub scores with less distance between them. After calculating authority and hub scores, we are calculating logarithmic of scores to represent them in the scale 0-1. After calculation of HS and AS and distance between the pages, we check for the web pages with minimum distance and with highest visit count of web pages are considered as more relevant web pages for improved distance rank algorithm (Table 2). Top twenty

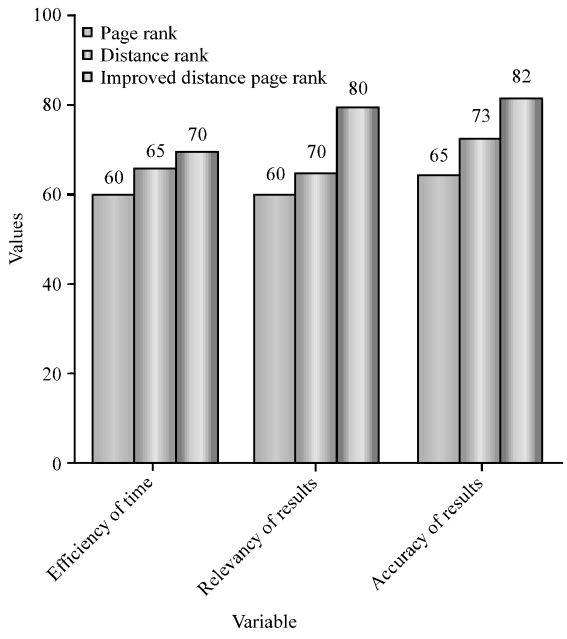

Fig. 2: Graph representing performance of improved distance page rank

web site URLs retrieved by page rank, distance rank and improved distance page rank. We compared the results of our improved page rank algorithm with page rank and distance rank (Table 3) (Fig. 2).

CONCLUSION

Retrieval of pertinent results for a given query is most complicated task, it depends on so many factors. In this study we proposed a new efficient ranking algorithm for web pages called improved distance rank algorithm. This algorithm is based on unique visit count of web pages and distance between hubs and authorities of web pages. In this algorithm unique visit count gives us information about the web pages which are visited more number of times so we will get more relevant web pages in addition to this we are also considering the distances between hubs and authorities, if less distance between web pages then they can be reached in less amount of time. Both these factors, web pages with more unique visit count and less distance between them are considered are more relevant web pages, can be retrieved by improved page rank algorithm.

RECOMMENDATIONS

In this study, we developed Improved Distance Rank algorithm to retrieve more relevant web pages, by considering unique visit count, hub scores, authority score and distance between the web pages. In addition to this we can also consider reduction of hidden fraudulent web pages. Most of the page rank algorithms use static navigation. Use of dynamic navigation to reduce number of non relevant pages as future work.

REFERENCES

Brin, S. and L. Page, 1998. The anatomy of a large-scale hypertextual web search engine. Proceedings of the 7th International World Wide Web Conference, April 14-18, 1998, Brisbane, Australia, Elsevier Science, pp: 107-117.

David, F., G. Ryan and A. Rossi, 2012. A Dynamical System for PageRank with Time-Dependent Teleportation. Vol. 20, Purdue University Press, West Lafayette, Indiana, pp: 126-137.

Dubey, H. and B.N. Roy, 2011. An improved page rank algorithm based on optimized normalization technique. Intl. J. Comput. Sci. Inf. Tech., 2: 2183-2188.

Haveliwala, T.H., 2003. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. IEEE Trans. Knowledge Data Eng., 15: 784-796.

Hemayati, R.T., W. Meng and C. Yu, 2012. Categorizing search results using Wordnet and Wikipedia. Proceedings of the International Conference on Web-Age Information Management, August 18-20, 2012, Springer, Berlin, Germany, ISBN: 978-3-642-32280-8, pp: 185-197.

Huang, W. and B. Li, 2011. An improved method for the computation of PageRank. Proceedings of the IEEE International Conference on Mechatronic Science, Electric Engineering and Computer (MEC), August 19-22, 2011, IEEE, Guangzhou, China, ISBN: 978-1-61284-719-1, pp: 2191-2194.

Kleinberg, J., 1998. Authoritative source in a hyperlinked environment. Proceedings of the ACM-SIAM Symposium on Discrete Algorithm, April 03, 1998, ACM Press, New York, USA., pp: 668-677.

Kumar, P.R., A.G.K. Leng, A.K. Singh and A. Mohan, 2016. Efficient methodologies to determine the relevancy of hanging pages using stability analysis. Cybern. Syst., 47: 376-391.

Nagappan, V.K. and P. Elango, 2015. Agent based weighted page ranking algorithm for web content information retrieval. Proceedings of the IEEE International Conference on Computing and Communications Technologies (ICCCT), Feburary 26-27, 2015, IEEE, Coimbatore, India, ISBN: 978-1-4799-7624-9, pp: 31-36.

Pawar, S.G. and P. Natani, 2014. Effective utilization of page ranking and HITS in significant information retrieval. Proceedings of the IEEE International Conference on Convergence of Technology (I2CT), April 6-8, 2014, IEEE, Pune, India, ISBN: 978-1-4799-3760-8, pp: 1-6.

Tin, P., T. Toriu, T.T. Zin and H. Hama, 2014. A cluster based ranking framework for multi-typed information networks. Proceedings of the 10th IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), August 27-29, 2014, IEEE, Osaka, Japan, ISBN: 978-1-4799-5391-2, pp: 415-418.

Tyagi, N. and S. Sharma, 2012. Weighted page rank algorithm based on number of visits of links of web page. Intl. J. Soft Comput. Eng., 2: 441-446.

Wang, X., T. Tao, J.T. Sun, A. Shakery and C. Zhai, 2008. Dirichletrank: Solving the zero-one gap problem of Pagerank. ACM. Trans. Inf. Syst., 26: 1-10.

Zhang, Y., L.B. Xiao and B. Fan, 2008. The research about web page ranking based on the a-pagerank and the extended VSM. Proceedings of the 5th IEEE International Conference on Fuzzy Systems and Knowledge Discovery FSKD'08, October 18-20, 2008, IEEE, Lanzhou, China, ISBN: 978-0-7695-3305-6, pp: 223-227.