Finding Page Rank using Transition Matrix and Random Vector

Sreekanth Kavuri, Vedavathi Katneni

Abstract: ***Today, the web is growing at a very fast and rapid rate. Also, there is afast growth in using of the internet compared with a past years. Due to the dynamic nature of web, the information on the internet in form of pages is added and removed in no time. The information on the web had become very important and a large amount of information is hidden inside the web. Getting the information, which is in need has become very difficult. Hence mining of the web data deeply in-terms of the content, structure, and usage is necessary. The search engines, in general, give us a list of web pages for user queries. For the users to move on that list comfortably a ranking mechanism is applied. Many of the rank basedmechanisms are based upon content-based or link-based. Analgorithm is proposed to find the rank of the mined web pages is presented in this paper. The proposed algorithm is compared and analysed with existing mining algorithms namely page rank and HITS algorithms. This paper highlights respective strengths, weaknesses, variations, and carefully analyses all the algorithms with an example. The added feature of the algorithm is that the most valuable page of the list, which is given by the search engine, is displayed at the top of the list.***

Keywords: Page Rank, HITS, Random Vector, Graphs, Web mining.

I. INTRODUCTION

Information plays an important role nowadays. Many techniques or algorithms are developed and are utilized in the information industry in-order to extract the desired information. Data mining has become a natural technology in extracting information. According to Han, Kamber and Pei [12] data mining is the concept of retrieving knowledge from a large amount of data. The information which we got after mining is useful. Hence data mining is also regarded asawareness mining from data. The process of mining of data is iterative. Traditional data mining takes data from a structured format like tables, spreadsheets or files. Now with the rapid growth of the web and its documents demanded the web text mining. The volume of web-based information is massive and continues to grow. The coverage of the information is diverse and heterogeneous. Shadab Irfan, Subhajit Ghosh [4] has described about text, data and web mining.

Different authors are presenting the same information on the web using different words. Li et.al. [7] has compared similarity between two vertices and is based on Web Structure Mining.Duhan et.al. [9] Has given his survey regarding different existed page rank algorithms. Nguyen et.al [2] proposed webpage traverse using Page Rank algorithm and sequence predictions.

Keita Sugihara [1] has used complex numbers also to find page rank for a web page. S. Hassena et.al [3] given a scalable ranking to the web pages based on semantic search on web pages. Weighted Page Rank algorithm is the PageRank algorithm extension, described by W. Xing and Ali Ghorbani [11].ShilpaSethi, Ashutssh Dixit [5] has taken the browsing patterns of the user to calculate page rank, which comes under Web Usage Mining. Also JinalV.Patel, Prof. Rimi Gupta [6]has told about algorithms based on time spent by user on browsing and the hyperlinks used by user. Hence integrating information is therefore a challenging task. Nowadays, Web mining has become very popular and interesting.

According to Bing Liu [14], Web Mining is one among the data mining applications. Web mining extracts the data or information on the web, either related to web pages, links between those web pages, logs of the web pages, etc.

World Wide Web, in short web, has become very important in everyone's life. The web is the world's largest source of information, which can be easily searchable. It consists of billions of web documents or pages which are interconnected with hyperlinks. The work on the web is mostly dependent on hyperlinks. Users simply follow these hyperlinks to move from page to page. When there is a query to the search engine on the web then the crawler, indexer and ranking mechanism plays an active role consecutively.

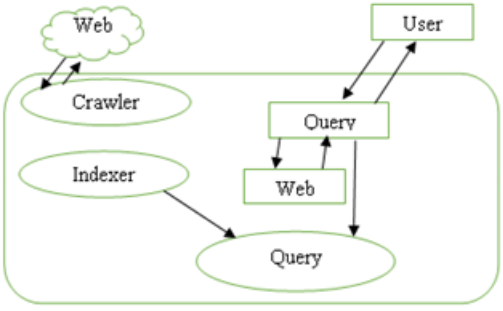

Fig. 1. Search Engine Structure

Fig 1 shows the web search engine's structure. The Crawler will crawl the entire web and gets the web document, and then the documents are given to indexer for indexing. The indexing is done based on keywords on the page. Based on the query of the user the indexed pages URLs will be given a ranking. To assign a rank to the web pages, there are several methods or models. After ranking, the page that gets the highest rank will be shown first and there-on.

Revised Manuscript Received on December 13, 2019.

***** Correspondence Author

Sreekanth Kavuri,* IT Mentor, IT Mentor, Rajiv Gandhi University of Knowledge and Technologies, Hyderabad, Telangana. Email: skavuri47@gmail.com

Dr. VedavathiKatneni, Computer Science Department, GIS, GITAM (Deemed to be University), Visakhapatnam, India. Email: vedavathi.katneni@gitam.edu

The parameters that may be considered for mining the web are the content or the activities that are done by the user on the internet in the past or the time spent by the user on a particular webpage.

According to Jaideep Srivastava [13] Web mining is classified into three types based on the information that the user gets. They are Web Content Mining (WCM) where content inside the web pages is mined, Web Structure Mining (WSM) as described by Bing Liu, [14] where mining is done by the content of the document or the hyperlinks between web documents and Web Usage Mining (WUM) where mining is based on the user's profile or his activities which are recorded in the log during his previous visits. Fig2 presents the categories of Web Mining.

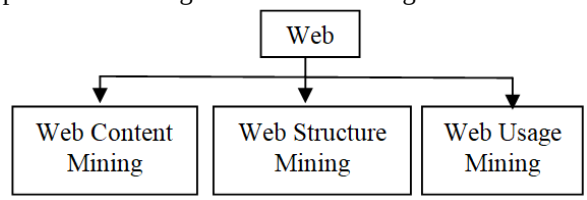

Fig. 2. Categories of Web Mining

The main theme behind all these techniques is to mine the web and get the related information for the user. In this process, when the user browses the information then multiple results are shown in the form of a URL. Various algorithms are used to get that URL in their order of ranking. Section II presents the relevant work done on Page Rank. Section III discusses the proposed algorithm. Section IV gives results of the algorithms discussed and section V gives the conclusion.

II. RELATED WORK

A. *PageRank*

Of all existing Models and algorithms PageRank is one of the techniques used by google search, which was introduced by Page and Brin [8]. This was the first page rank algorithm, which calculates the rank of the pagesbased on web link structure. The rank of a web page is calculated by the number of page backlinks. If the amount of backlinks is more then it implies that the page has more importance. Here PageRank is the measurement for giving ranks to web documents. A page will have high rank if (i) it has more backlinks (ii) swap links with websites which have high page rank value (iii) Adding new pages to the web sites. So PageRank is both qualitative and quantitative. The resultant rank vector is independent of the query posed by the user. PageRank is an iterative method and is iterated until all the web pages converge. Rank for a document is given by the documents which are linked to the same page. In turn, the rank is given by the rank of the pages linked to them. In this way this algorithm iterates. The PageRank of a page 'u' is given by the following equation:

$$PR(u) = (1-d) + d * \left(\frac{PR(T1)}{C(T1)} + \ldots + \frac{PR(Tn)}{C(Tn)}\right) \quad (1)$$

where,

PR(u) is the PageRank of page u.

$PR(T_i)$ is the PageRank of pages $T_i$which link to the page u.

$C(T_i)$ is the number of out links on page T_i

d is a damping factor usually set between 0 and1 as proposed by Page and Brin [8]

Here, assume that page u is linked by the pages T_1 to T_n.

The drawbacks of the Page Rank algorithm can be presented as:

- When pages are linked as loop it is difficult to find PageRank.
- Pages which are not relevant to content of query can gain higher rank due in links.
- PageRank does not take content of web document and query into consideration.
- PageRank algorithm gives more importance to old web documents.
- Results are not query dependent.

B. *Hyperlink Induced Topic Selection (HITS)*

The next method, which is introduced by Klienberg [10] is the Hyperlink Induced Topic search algorithm known as the HITS algorithm. This algorithm is based on web structure mining and web content mining. In this model web pages are divided into two types of pages i.e. Authority pages and Hub pages. Hubs are the web pages with the number of in links to it and Authority pageis the web page with good content. Therefore hubs are based upon web structure mining, having good quality links and authorities are based on web content mining, having good source content. These two pages show mutual relationship because a webpage can be both a hub and authority. HITS algorithm gives ranking to the web documents by their in links and out links. Hubs mostly point the authorities, i.e. hub pages will have more out links and authority pageshaves more in links. This description is presented in Fig 3.

HITS algorithm initially identifies hubs and authorities based upon the content and then calculates the weights of hubs and authorities. This algorithm represents the web pages as a graph (v,e), where v, vertices represents the web pages and e denotes links in between those web pages. HITS algorithm has two stages: (i) sampling stagedefines appropriate query based pages. (ii) Iterative step: Taking the output of the sampling step, iterative step identifies hubs and authorities and calculates weights as following equations 2 and 3.

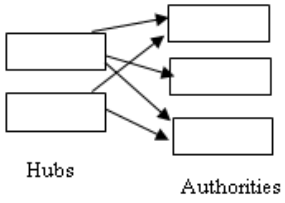

Fig. 3. Example of Hubs and Authorities

Let x denotes the authoritative scores of all pages and y denotes hub scores of all pages. Simply to say x and y are vectors of respective scores and A is the link matrix of in links of the pages. $A^T$is the transpose of link matrix.

$$x = A^T y \quad (2)$$

$$y = Ax \quad (3)$$

Initially hub score vector y is taken as an identity vector. After that authoritative score is calculated using equations (2). Then again hub score is calculated using equation (3). Then nodes having highest hub score than authoritative score are treated as hubs and vice-versa.

The Drawbacks of HITS algorithm are*:*

- Mutual relationships between hub and authoritycanresult in different weights.
- May not give correct results as the algorithm may lead to different weights.
- It is not so easy to identify a page is hub and authority.
- This is not efficient algorithm due to construction of graph.

Due to the above drawbacks only HITS algorithm is not chosen by Google for searching. So PageRank is chosen because of its efficiency.

III. PROPOSED MODEL

When the user wants some information from the web, he asks a query to the search engine and subsequently the query gives some results. In the result we will have number of web pages and hyperlinks between them. These resultant web pages has to be ranked. So for this the web pages and the links are considered as a graph and ranked considering the following procedure.

1. Take agraph (G) with vertices (V) as web pages and edges (E) as hyperlinks in between them.
2. Form a matrix with all in links to each vertex, let this be a link matrix (A).
3. Form a transition matrix (M), where

$$Mij = \frac{Aij}{\Sigma j} \quad (4)$$

4. Take a vector v and start a random surfer with a vertex.
5. Multiply v with M and get v1
6. Repeat step 5, number of node times (minimum number of times).
7. If we have repeated values continue with step 5 until we get unique values in the vector.
8. vV will be the ranks of the respected pages.

Table- I: Terms description of the proposed algorithm

Notations Used	Description
G	Graph with web pages as vertices and hyperlinks as edges between them.
A	In links matrix
M	Stochastic Matrix
V	Random vector
v_V	Vector after multiplying vector and matrix number of node times

IV. RESULT ANALYSIS

Consider a graph G (Fig 4) that has 4 vertices and 6 edges, i.e. we are resulted with 4 web pages from the query we put to the search engine and let us consider the 4 web pages has 6 hyperlinks in between them. Now applying the proposed algorithm, PageRank and HITS algorithm on the Graph G is presented here:

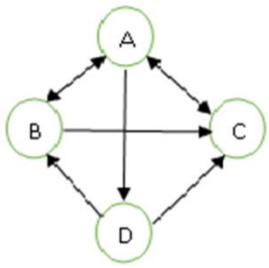

Fig. 4. Example Graph G

For the proposed algorithm, initially form a link matrix A from the above graph in Fig 4 with the in links for each node. Let's say node A have in links from B and C, the link matrix will have [0 1 1 0] in the first row, as node A is considered first. Then node B. node C and node D is considered consecutively.

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

The link matrix A contains the information of all the in links of a particular node. After forming link matrix A form another Matrix M calledTransition Matrix with the help of equation (4), which is step 3 of the proposed algorithm.

$$M = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

Let us take a random unit vector v.

$$v = [1 \quad 0 \quad 0 \quad 0] \quad (5)$$

Apply step 5 in the above proposed algorithm, i.e. multiplying random vector v with transition matrix M.

$$v1 = v * M \quad (6)$$

Therefore,

$$v1 = [1\,0\,0\,0] \times \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

$$v1 = [0 \ \frac{1}{2} \frac{1}{2} \ 0]$$

Repeat the same step node number of times.

$$v2 = v1 \times M = [\frac{5}{12} \frac{1}{6} \ 0 \ \frac{5}{12}]$$

$$v3 = v2 \times M = [\frac{1}{2} \frac{5}{24} \frac{5}{24} \frac{1}{12}]$$

$$v4 = v3 \times M = [\frac{37}{144} \frac{23}{72} \frac{1}{4} \frac{25}{144}]$$

As we have 4 nodes in the graph G (Fig 4) we have vector v4. i.e.

$$v4 = [0.2569 \quad 0.3194 \quad 0.25 \quad 0.1736]$$

Repeating step 5 number of node times, a vector with page ranks of all the pages will be obtained. From this vector the rank of page B is 0.3194, the rank of web page A is 0.2569, the rank of page C is 0.25and the rank of page D is 0.1736. Therefore the pages in the order of highest rank to the least rank are respectively as vertex B, vertex A, vertex C and then vertex D. If we have repeated values in the last vector then repeat the same step until we get unique values.

The obtained results from the above proposed algorithm are compared with Page Rank and HITS algorithms.

So according to equation (1) the PageRank (PR) of all the 4 nodes in Fig 4 can be expressed as below:

$$PR(A) = (1 - 0.85) + 0.85\left(\frac{P(B)}{2} + \frac{P(C)}{1}\right) \quad (7)$$

$$PR(B) = (1 - 0.85) + 0.85\left(\frac{P(A)}{3} + \frac{P(D)}{2}\right) \quad (8)$$

$$PR(C) = (1 - 0.85) + 0.85\left(\frac{P(A)}{3} + \frac{P(B)}{2} + \frac{P(D)}{2}\right) \quad (9)$$

$$PR(D) = (1 - 0.85) + 0.85\left(\frac{PR(A)}{3}\right) \quad (10)$$

Here let's consider damping factor, d as 0.85, since damping factor is a probability of clicking the links. Many theories had taken different values, but Page and Brin [2] assumed as 0.85. Initially page ranks of all the pages are taken as 1. After that for approximation of results number of iterations are made. As they are billions of pages in the web Page and Brin [2] suggested 100 iterations for correct approximation. But they are only 4 pages in Fig 4 only 20 iterations had made. Table II gives the results of iterations.

Table- III: Results of PageRank Iterations

Iteration	PR(A)	PR(B)	PR(C)	PR(D)
0	1	1	1	1
1	1.4250	0.8583	1.2833	0.4333
2	1.6056	0.7379	1.1027	0.5538
3	1.4009	0.8403	1.1539	0.6049
4	1.4879	0.8040	1.1611	0.5469
5	1.4786	0.8040	1.1457	0.5715
6	1.4655	0.8118	1.1535	0.5689
7	1.4755	0.8070	1.1520	0.5652
8	1.4722	0.8082	1.1524	0.5681
9	1.4730	0.8085	1.1521	0.5671
10	1.4729	0.8083	1.1520	0.5673
11	1.4727	0.8084	1.1520	0.5673
12	1.4728	0.8083	1.1519	0.5672
13	1.4726	0.8083	1.1519	0.5673
14	1.4726	0.8083	1.1519	0.5672
15	1.4726	0.8083	1.1518	0.5672
16	1.4726	0.8083	1.1518	0.5672
17	1.4726	0.8083	1.1518	0.5672
18	1.4726	0.8083	1.1518	0.5672
19	1.4726	0.8083	1.1518	0.5672
20	1.4726	0.8083	1.1518	0.5672

After 20 iterations the correct approximation of page ranks will emerge, where page A has highest page rank, then page C will come, and page A and page B consecutively.

For HITS algorithm the graph in Fig 4should be converted to Hubs and Authorities graph first. The graph with the combination of hubs and authorities is presented in Fig 5:

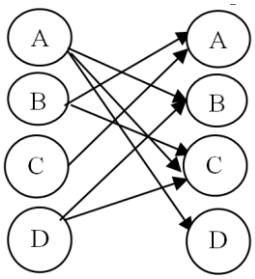

Fig. 5. Hub and Authority Graph

Form a link matrix A from the Fig 4.

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

Transpose of the link matrix A is $A^T = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$

Let us take Hub score vector as, $y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

Calculate authoritative score vector as in equation (2) i.e.

$$x = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 3 \\ 1 \end{bmatrix}$$

Now update hub score vector as in equation (3) as actual authoritative score is obtained and the vector initially taken is an assumption i.e.

$$y = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 2 \\ 2 \\ 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 5 \\ 2 \\ 5 \end{bmatrix}$$

After the convergence step the hub and authoritative vectors are:

$x = [2 \quad 2 \quad 3 \quad 1]$

$y = [6 \quad 5 \quad 2 \quad 5]$

As per hub score page C is having highest rank and then there are same values for page A and B, and at last we have page D. To get unique values instead of same values, there is second iterative step. But there is no guarantee for getting unique, which is the main drawback of HITS algorithm. According to authoritative score page A has highest rank.

V. CONCLUSION

With the continues growth of the web-based applications, especially electronic commerce, there is considerable interest in analyzing web usage data in order to better understand web usage and apply the knowledge to serve better to the users.

This has resulted in several open issues in the field of Web Usage Mining. Due to the introduction of stricter laws, protection for privacy is the major challenge in many practical applications. So for this web mining has become a strong tool for mining of web information. The proposed model in this paper has mined the web information with help of an algorithm which helps the user to get the web information easily.

And the proposed model is compared with the existed page ranked algorithms PageRank and HITS. But the proposed model has been restricted to graph without self-nodes, dangling nodes and null nodes. In future the algorithm will be developed with all the specials cases mentioned.

REFERENCES

1. Keita Sugihara, "Using Complex Numbers in website ranking calculations A non-ad Hoc Alternative to Google'aPangeRank", in *Journal of Software*, February 2019.
2. Nguyen Thon Da, Tan Hanh, Pham Hoang Duy, "Improving web page access predictions based on sequence prediction and PageRank algorithm", in *Interdisciplinary of Information, Knowledge, and Manegemnt*", January 2019.
3. S. Haseena, R. Latha, A.M .Bharani, "Scalable Ranking based Web Pages and Semantic Search", in *International Journal of management, Technology and Engineering*", September 2018.
4. Shadab Irfan, Subhajit Ghosh, "Web Mining for Information Retrivel" in *International Journal of Engineering Science and Computing*, 2018
5. ShilpaSeti, Ashutosh Dixit, "*A Novel Page Ranking Mechanism Based on User Browsing Patterns*" in Springer, June 2018
6. JinalV.Patel, Prof. Rimi Gupta, "Survey on Different Page Ranking Algorithms Based on Links and Time" in *International Journal of Innovative Research in Computer and Communication Engineering*, January 2017
7. Li, C., Han, J., He, G., Jin, X., Sun, Y., Yu, Y., Wu, T., 2010. 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.
8. 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, 2009
9. Duhan, N., Sharma, A.K., Bhatia, K.K,"Page Ranking Algorithms: A Survey". Proceedings of the *IEEE International Conference on Advance Computing. Centric Systems and Applications).* Springer-VerlagNewYork, Inc., Secaucus, NJ, USA. 2009.
10. 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
11. W. Xing and Ali Ghorbani, "Weighted PageRank Algorithm", *Conference IEEE*, 2004.
12. J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques: concepts and techniques*. Elsevier, 2011.
13. Jaideep Srivastava, PrasannaDesikan, Vipin Kumar, "*Web Mining - Concepts, Applications & Research Directions*"
14. Bing Liu., 2006. *Web Data Mining: Exploring Hyperlinks*, Contents, and Usage Data

AUTHORS PROFILE

Sreekanth Kavuri is presently a Research Scholar in the Computer Science Department, GIS, GITAM (Deemed to be University), Visakhapatnam. Completed his Post Graduation from JNTU Hyderabad. Completed his pre-Ph.D. from GITAM, Visakhapatnam. Presently working as IT Mentor in Rajiv Gandhi University of Knowledge and Technologies. Taught many subjects in computer science and engineering department and Information Technology department.

Dr. VedavathiKatneniis the Head of the Department, Computer Scicence, GIS, GITAM (Deemed to be University), Visakhapatnam. Had done her Ph.D. from Andhra University, Visakhapatnam. She is a life member of Indian Science Congress Association and Indian Society of Technical Education. Presently in GITAM (Deemed to be University), she was awarded with many awards, from women empowerment cell in GITAM (Deemed to be University) for her contribution towards women empowerment and initiatives taken towards empowering women