ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 11, November 2017

A Survey of Page Ranking Algorithms

Y Sravani Devi [1], B.Geetha Kumari [2]

Assistant Professor, Dept. of Computer Science & Engineering, G.Narayanamma Institute of Technology & Science, Shaikpet, Hyderabad, India[1]

Assistant Professor, Dept. of Computer Science & Engineering, G.Narayanamma Institute of Technology & Science, Shaikpet, Hyderabad, India[2]

ABSTRACT: One of the most popular algorithm in processing internet data i.e. webpages is page rank algorithm which is intended to decide the importance of a webpage by assigning a weighting value based on any incoming link to those webpage. However, the large amount of internet data may lead into computational burden in processing those page rank algorithm. Many algorithms exist for web page ranking and these algorithms are based upon one or more parameters such as forward links, backward links, contents and user interaction time. The efficiency of an algorithm may be based upon the parameters that are applied to determine the ranking of the page. In this paper some important page ranking algorithms are discussed.

KEYWORDS: WWW; Web Crawling, Web Mining, Page Ranking Algorithms.

I. INTRODUCTION

In recent days, we are experiencing the rapid growth of internet data i.e. (web) pages. For example, Google currently processes over 20 petabytes of data consists of webpages each day [1]. On the other side, there is a need for analyzing those large (internet) data to obtain any valuable information. One of the famous analysis in recent large data processing is page rank algorithm which is intended for ranking a webpage by assigning a weighting value based on any incoming links to those webpage. Basically, the page rank algorithm is relatively easy to be implemented in a single machine since this algorithm only needs page information (link, id or title), outgoing link information and total amount of the pages. However, processing those large amount of internet data in single machine may leads into scalability problems such as computational time.

In the literature, there have been several research regarding to the page rank algorithm for several purposes. In [2], the authors propose a basic idea of page ranking as an algorithm for measuring the importance of a web page. Those paper is the first paper which introduce the concept of page ranking. In [3], the authors proposed a relation-based page rank algorithm for semantic web search engines. However, as far as we aware of, there is a few paper which explain an implementation of simple page rank algorithm in a distributed system.
Calculation of Page Rank is to be done for all incoming links. Wu and Davison studied the combination of both incoming, outgoing links and a broadcast step to notice link farms. Drost and Scheffer [8] gave a machine learning method to discover link spam. From the beginning, the knowledge of a focused or custom PageRank vector has occurred [8], Haveliwala initially intended to get topical information into PageRank calculation. In this paper, different topics are required to be selected and then apply PageRank to find good pages within topical networks using Automatic seed set selection.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 11, November 2017

II. PAGE RANK

PageRank algorithm is an eminent algorithm that is based on considering the link information for giving all the pages global significance scores. Larry Page originally gave this proposal [4] in which PageRank of a page becomes important if the number of other important web pages is pointing to that page. Consistently, this will be based on a mutual support between pages. Here, the importance of page effects and is effected by the rank of other pages in the web. The PageRank score r (p) of a page p is given as: ----- 1

Later, we can sum up two elements to get the importance of a page p: one element is from the score comes from incoming links to p and the other (static) element is equal for all web pages. ----- 2

A key factor to be considered is, the regular PageRank algorithm allocates an alike static score to each page, this rule is cancelled in a biased PageRank version. PageRank in the form of matrix equation, ----- 3

d is a static score distribution vector of random, nonnegative entries adding to get total as one.

PageRank is the famous algorithm implemented by Google to help define where websites should rank in their search engine index. PageRank can measure the status of a page. PageRank works by calculating the number of inbound hyperlinks to a particular page. The primary hypothesis is that, websites that are more reliable are likely to get more links from other websites. If we consider each link as a vote from another web master, Google applies this data to determine if a site has good contents and promotes it accordingly.

III. RANKING ALGORITHMS

A. PAGE RANKING ALGORITHM

Page Ranking Algorithm [1] is the mostly used algorithm which is given by S. Brin and L. Page (co-founder of Google) and used in Google search engine. In this algorithm a web page propagates their rank value to its outlink WebPages, in other words rank of a webpage depend on the rank of its backlink web pages. So it states that a page is important than its out link pages also important.

B. HITS (HYPERTEXT INDUCED TOPIC SELECTION)

HITS compute the web page rank dependent on the query and works online. It uses two types of webpages, hub pages which points one or many web pages. And authority pages which contain valuable information and have inlinks from one or many hub pages. A page is a good hub page if contains links to good authority webpages and webpage is good authority if points from good hub web pages. This algorithm starts with base set which contains some pages that are related to given query. In the next step for each web page present in the base set find all inlinks and outlinks pages from the current webpage and add these new retrieved webpages in the base set. Then iterate previous step for all newly added pages until there no new pages remains in the base set. This algorithm is used in IBM's CLEVER search engine. HITS uses two type of score for every web page hub and authority.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 11, November 2017

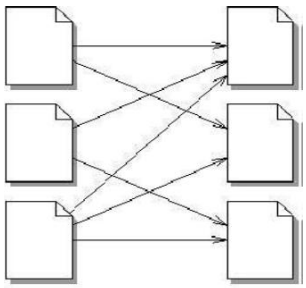

Figure. Hubs and authorities web pages

C. TOPIC-SENSITIVE PAGERANK

It is a extend version of PageRank algorithm, which computes rank of a webpage based on multiple rank vectors. This algorithm [2] overcome the problem of PageRank algorithm, i.e. pages of more inlinks getting high rank for queries that are not related. Webpage ranking value is based on 16 vectors, selected from Open Directory Project (ODP). In this algorithm the linear combination of rank vectors are used to compute the rank value in place of a single PageRank vector. Computation of the PageRank vectors for category cj, the random surfer surf the link or jump to a page randomly. This has the effect of taking the PageRank to corresponding topic.

D. QUERY-SENSITIVE SELF-ADAPTABLE RANKING ALGORITHM

This algorithm [3] reorder the result list of webpages by using query sensitive value and satisfying global as well as local authority. Query sensitive value is calculated based on voting procedure, which is as follows: Firstly, select the top web pages (n) from result list form a voting set VSet,. For a web page docis, give a vote to page docit in following situations. a) If there is a link from docis to docit b) if docit's title, is present in the content of docis c) If docis's information is referred from the content of docit. Only one vote is counted if minimum two conditions are matched from above three conditions. If docis and dociare from the same website, then deal them as follows: 1) Votes between these two pages ignored. 2) Only take unilateral vote 3) Provide limit valid votes for the same domain. Finally count the votes, and page of more votes more sensitiveness it is. Then for two pages doc1, doc2, with their query sensitiveness scores QSl and QS2, and their respected PageRank values PR1 and PR2. The three strategies for ordering the pages are as follows:

1) If (QSl>QS2)||(QSl==QS2&&PR1>PR2) then docl is ordered first than doc2.
2) If (PRI>PR2)||(PR1==PR2 &&QSl>QS2 then docl is ordered first than doc2.

E. TOPIC DISTILLATION WITH QUERY-DEPENDENT LINK CONNECTIONSAND PAGE CHARACTERISTICS

Topic Distillation is aapproach to find key resources about a topic form the web page. The topic distillation approach broadly categorized into two classes: evidence based on link structure and evidence based on web page characteristics [4]. Link based evidences are number query dependent and independent inlinks, outlinks and anchor text. And web page characteristics evidences are anchor density, URL depth and URL length. Anchor density is number of links of a page normalized by the length of the page, URL depth is the number of "/" in the URL, it gives the importance of the web page and URL length is the number of characters in the URL of the page.First retrieve the result list of the web pages according to ranking score of web pages calculated by applying the Okapi BM25 [Robertson et al. 1994] ranking function. The initial search results are reordered based on a combination of external evidence and the original similarity score of a page. The most popular method is a weighted linear combination.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 11, November 2017

IV. COMPARATIVE ANALYSIS

Algorithms	Methodology	Mining Technique	I/P Parameters	Relevancy	Complexity	Quality of Result	Strength	Limitation
Page Ranking Algorithm [4]	Computes scores at indexing time	Web Structure Mining	Backlinks	Less (this algo. rank the pages on the indexing time)	O(log N)	Medium	Easy Computation of Rank	Rank Computation at indexing time
HITS Algorithm [5]	It computes the hubs and authority of the relevant pages. It relevant as well as important page as the result.	Web Content and Structure Mining	Content, Back and Forward links	More (this algo. Uses the hyperlinks and content of the page so it will give good results)	<O(log N)	Less than PageRank	Working online depend on user query content	Topic Drift
Weighted Page Ranking Algorithm [7]	Weight of web page is calculated on the basis of input and outgoing links and on the basis of weight the importance of page is decided	Web Structure Mining	Backlinks, forward links	Less as ranking is based on the calculation of weight of the web page at the time of indexing.	<O(log N)	Higher than PageRank	No. of inlinks of sibling web pages of current web page give the popularity of current page	It only ranks the web pages on the basis of popularity which may be distinguished easily.
Topic-Sensitive PageRank [8]	PageRank Based on multiple page rank vectors	Web Content and Structure Mining	Inlinks, content of source page	More than WPR as the multiple rank vector is used	>O(log N)	Higher than PageRank	It avoid the Problem of heavily linked pages getting highly ranked	Satiy only local (16 topics) authority not global authority

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 6, Issue 11, November 2017

V. CONCLUSION

In this paper a survey on page ranking in webpage is described, in which various ranking are discussed. Herewith comparison among these algorithms is also described using different parameters. Along with this description, classification of webpage ranking algorithms based on web mining techniques and query dependency is also discussed in this paper. In the future work in the field of webpage ranking, different new parameters along with parameters related to hyperlink analysis such meta HTML tag of a webpage are important which can be useful to improve the ranking of a webpage. Now a days dynamic webpages are mostly used so in crawling these WebPages, HTML scripting languages are also important which will affect the webpage ranking.

REFERENCES

[1] S. Brin and L. Page, "The Anatomy of a Large Scale Hypertextual Web Search Engine," Computer Network and ISDN Systems, vol. 30, no. 1-7, pp. 107-117, 1998. A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting.", IEEE Transactions on Image Processing, vol. 13, no.9, pp. 1200–1212, 2004.

[2] T. H. Haveliwala, "Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search," IEEE transactions on knowledge and data engineering, vol. 15, no. 4, 2003. W. Tao, W. Zuo, "Query-sensitive self-adaptable web page ranking Algorithm," 2nd International Conference on Machine Learning and Cybernetics, 2003. Eftychios A. Pnevmatikakis, Petros Maragos "An Inpainting System For Automatic Image Structure-Texture Restoration With Text Removal", IEEE trans. 978-1-4244-1764, 2008

[3] S.Bhuvaneswari, T.S.Subashini, "Automatic Detection and Inpainting of Text Images", International Journal of Computer Applications (0975 – 8887) Volume 61– No.7, 2013

[4] M. Wu, F. Scholer and A. Turpin, "Topic Distillation with Query Dependent Link Connections and Page Characteristics," ACM Transactions on the Web, vol. 5, no. 2, pp. 6:1-6:25, 2011.Xiaoqing Liu and Jagath Samarabandu, "Multiscale Edge-Based Text Extraction From Complex Images", IEEE Trans., 1424403677, 2006

[5] Baoning Wu and Brian D. Davison (2005), Identifying Link Farm Spam Pages, Proceedings of the 14th International World Wide Web Conference,Chiba, Japan. Mr. Rajesh H. Davda1, Mr. Noor Mohammed, " Text Detection, Removal and Region Filling Using Image Inpainting", International Journal of Futuristic Science Engineering and Technology, vol. 1 Issue 2, ISSN 2320 – 4486, 2013
S. Landset, T.M. Khoshgoftaar, A.N. Richter, T. Hasanin, "A survey of open source tools for machine learning with big data in the Hadoop ecosystem," Journal of Big Data, Vol. 2, No. 24, Nov. 2015.

[6] V.K.Nagappan and P. Elango, "Agent Based Weighted Page Ranking Algorithm for Web Content Information Retrieval", In Proceedings of 2015 International Conference on Computing and Communications Technologies (ICCCT'15),IEEE,2015.