Approaches of Page Ranking Algorithms: Review

Nripendra Narayan Das
Dept. of CSE & IT
ITM University, Gurgaon

Ela Kumar, Ph.D
Dept. of ICT
GBU,Greater Noida

Sheetal
M.Tech. (CSE) 3rd Semester
Dept. of CSE & IT
ITM University, Gurgaon

ABSTRACT
With the use of World Wide Web internet engines like Google will be the tools which in turn allow you in order to understand along with speedily look for solutions. Even so the present engines like Google will not think about user's actual needs. Criteria associated with page position are usually keys associated with search engine optimization. Most engines like Google tend to be position their own listings according to user's dilemma. The particular PageRank formula can be used inside the search engines search engine optimization in order to status listings. With these cardstock examination present page position algorithms, strategies are usually presented plus comparability among these is usually carried out.

Keywords
Page ranking, Search engine.

1. INTRODUCTION
Internet has provided men with single but vast source of knowledge and information. Now a days web search tools (engine) have become one of the indispensable tools for people who do surf on internet [1].But with the growing use of internet, it is expanding rapidly in its content. With rapid growth in information source, there comes a difficulty in managing that information according to user requirements. The problem is like "drowning in data but starving for knowledge".

Search engines are used to find information from the web. Unlike web directories, which are maintained only by human editors, search engines also maintain real time information by running an application algorithm on a web crawler. When user search any web page, a list of web pages is generated as a result referred to as search engines results pages (SERPs). The information may be a specialist in the web pages, images, information and other type of files. The ordering of the web pages in the resultant list is very important. The most important page should come at the top of the list and the less important page should come below it with respect to the searched content by the user. So there is a need of some mechanism that can arrange the pages according to their importance dynamically whenever user searches any content.

Now here comes the concept of ranking the web pages present on the web.

Ranking here means assigning some value to a web page among several web pages of its kind which can decide its importance level. This process is called page ranking. There are two main types of page ranking: based on the web page content and based on the hyperlinks structure analysis. The first type is the traditional one but the huge amount of data of internet would be great challenges to the traditional information searching techniques [4]. The former type involves the hyperlinks which link one web page to other web page. This type utilizes the features of web page aggregation to evaluate significance of the page and linking.

In this paper, different algorithms, techniques and approaches will be presented that are being used to implement page ranking and their comparison.

2. RELATED WORK
2.1 Page Rank Algorithm
This algorithm is used by Google and developed by Surgey Brin and Larry Page. This algorithm is based on the link structure of the web. It divides the page rank of a page evenly among its outgoing links. According to this algorithm the page rank of a page can be given by:

$$PR(u) = (1 - d) + d \sum_{v \in B(u)} PR(v)/N_v$$

Where

PR(u) : page rank of page u, PR(v) : page rank of page v, N(v) : number of outgoing links of page v, B(u) : set of pages that points to u, D : damping factor(the probability of following direct link, usually taken 0.85).

2.2 Weighted PageRank Algorithm
This algorithm is proposed extension to PageRank algorithm by Xing and Ali Ghorbani[2]. This is also a link based algorithm but it does not divide the page rank evenly. It assigns more page rank to more popular pages. It assigns page rank on the basis of incoming and outgoing links to the page. According to this algorithm page rank of a page is given by:

$$WPR(u) = (1 - d) + d \sum_{v \in B(u)} WPR(v) W^{in}_{(v,u)} W^{out}_{(v,u)}$$

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{p \in R(v)} I_p} \quad W^{out}_{(v,u)} = \frac{O_u}{\sum_{p \in R(v)} O_p}$$

Where: I_u and I_p : number of incoming links to page u and p, O_u and O_p : number of outgoing links of page u and p.

2.3 HITS Algorithm
It is called Hyper Induced Topic Search [3]. It is both link based and content based. It considers two types of pages authorities and hub. The former one is a set of pages that are popular and relevant to the query and the later one contains links to useful sites including link to authorities. This algorithm works in two steps:

1. Sampling Step: In this step the page relevant to the user query are collected and a sub graph of the pages is formed. From this sub graph a root set R is taken and algorithm is applies on this root set for expanding it into a base set S by using the algorithm:
Input: Root set R; Output: Base set S
Let S =R
   a. For each page p E S, do Steps 3 to 5

b. Let T be the set of all pages S points to.
c. Let F be the set of all pages that point to S.
d. Let S = S + T + some or all of F.
e. Delete all links with the same domain name.
f. Return S

2. Iterative Step: Using the output of sampling step that is the base set hub and authorities are identified using the algorithm:
Input: Base set S, Output: A set of hubs and a set of authorities.

a. Let a page p have a non-negative authority weight $x_p$and hub weight y_p. Pages with relatively large weights x_p will be classified to be the authorities, similarly hubs with large weights y_p.
b. The weights are normalized so the squared sum for each type of weight is 1.
c. For a page p, the value of x_p is updated to be the sum of y_q overall pages q linking top.
d. The value of y_p is updated to be the sum of x_q over all pages q linked to byp.
e. Continue with step 2 unless a termination condition has been reached.
f. Output the set of pages with the largest x_p weights i.e. authorities, and those with the largest $y_p$weights i.e. hubs.

2.4 Query Independent Algorithm

It is query and content independent algorithm that assigns a value to every document independent of the query [5]. It is concerned with the static quality of web page. It computes page rank using web graph.

In this algorithm N is the number of documents in the collection, m represents the probability that the random surfer will get bored and restarts from some another random document, "prob" represents the probability transition matrix which is a N*N matrix considering total N pages, "adj" is adjacency matrix and x is probability vector all entities of which are in the interval [0,1]. The algorithm is:

Create a Web Graph
Initialize the probability transition matrix for all I. j€1 to N

Compute a no of out links from a particular node say counter

a. If node having no out link then equally distribute probability otherwise distribute it according to out links
b. For all i,j if(counter==0) then
c. Prob[i][j] =1/N else
d. If(prob[i][j]==1) then
e. Prob[i][j]=1.0/counter
f. Multiply the resulting matrix by 1-m
g. Add m/N to every entity of the resulting matrix, to obtain probability transition matrix
h. For all I,j do prob[i][j]=(prob[i][j]*(1-m))+(m/N)
i. Randomly select a node from 0 to start a walk say s_int.
j. Initialize a random surfer and itr to determine no of iterations required to 0.
k. Try to reach at steady state within 200 iterations otherwise toggling occurs.
l. Multiply probability transition matrix probability vector to get steady state.
m. Check either system enter in steady state or not.
n. Print the ranks stored in probability vector and exit.

2.5 Algorithm based on classified tree in search engine

Classified tree: The data structure of classified tree adapts the structure of B[4]. The branches of tree are relatively more than the height. It is required that the relationship has to be established between the leaves of the tree and keywords in the inverted file. The visiting between the two sides are double action. The amount of the trees has no restriction to form the forest, show as Figure 3. key1, key2,....., keyi,..... ,keyn is the value of node respectively (key words aggregation). The arrow is the related page aggregation of the node.

The format of the items in the inverted file is as follows:
Keyi→ {[Pid1,ni1(hit1,hit2,…,hitni1)]
[Pidn,nin(hit1,hit2,…,hitnin)]}

Table 1. INVERTED INDEX

Keyi	**Pid**	**Other**	**$Tree_i$**	**$Queue_i$**
Key1				
Key2				
Key3				
Key n				

Here : Keyi :the serial number of keywords, P_{id} : the serial number of page file, n_{in} : the no of times as keywords in the file, hit_i : the place where key words appear in the file, $Tree_i$: the place where keywords appear in the classified tree, $Queue_i$: user queue

The following would be the description of ranking algorithm process with classified tree.

a. There is only root when initializing tree0 of classified tree;
b. In inquiry, multithreading parallel splits all the inquiries from the different users at the same time to key aggregation.

SK_i…… {keyi1,keyi2,...keyin} i=1,2,...
After collecting the key words from m users, the author merges the repeated ones, calculate

$$\bigcup_{i=1}^{m} SK_i$$

i. According to the key words, the classified tree in the classified forest will be searched by multithreading parallel to get the corresponding page aggregation. If the result is blank or cannot meet the user's demands (Specifically ,within a period of time, like 3 mins, the users make the second same search), the index database can be multithreading parallel by indexer to get the Page as aggregation on each keyi;

ii. All the users can be searched according to the keywords in the inverted file within a period of time.

iii. Similarity of users' searching in users' aggregation can be calculated based on VSM, i.e. the key words aggregation Ski can be taken out to be calculated[5];

iv. Find out the users with over 0.87 similarity, and put together the page aggregation with corresponding key

words, eliminate the repeated page, and make these key words as one node;

v. Multithreading parallel in all classified trees and if the new generated node value (key words aggregation) is the Sub class of node of classified tree I (Condition i), then along the branch down, find the node with no more than 2 of keywords, the searched page aggregation can be combined to this node's page aggregation, then all information feedback to users in the turn from maximum to minimum;

vi. If the difference between the node meeting the Condition1 and its own key words amount is over 2, then this node can be split to 2 nodes. One is the new generated, and the other is the left one. They are treated as two off springs of original generated node. Then the new generated page aggregation is sent to users as feedback;

vii. If there is no node satisfying 8), then this time searched node will be regarded as new classified tree's root;

viii. repeat the process from i to viii.

2.6 Page Content Rank Algorithm

In this ranking is based on the content of the page [6]. The terms used in the page determines the page importance. Importance is calculated on the basis of user query. The frequency of a term in a page is used to rank the page.

PCR works in four steps:

I. **Term extraction:** An HTML parser extracts terms from each page in Rq. An inverted list is built in this step and used in step 4

II. **Parameter Calculation:** Statistical parameters like term frequency and occurrences position, as well as linguistic parameters such as frequency of words in natural language are calculated and synonym classes are identified.

III. **Term Classification:** Based on parameter calculation in step 2, the importance of each term is determined. A neural network is used as a classifier. Each parameter corresponds to excitation of one neuron in the input level and importance of a term is given by excitation of the output neuron in the time of termination of propagation.

IV. **Relevance Calculation:** Page relevance scores are determined on the basis of importance of terms in the page, which have been calculated in step 3. The new score of a page P is equal to the average importance of terms in P.

2.7 Ranking Web Pages Using Machine Learning

A suitably trained machine learning method called Graph Neural Network(GNN) can produce a generic model to encompass different types of the numerical page ranking methods[7].GNN is a new class of neural network based algorithms capable of processing general type of input in terms of graphs in supervised manner. It computes the graph's output based on information present in node and links. Each node is a MLP (multi layer perceptron). Once trained the GNN can be used to compute unknown outputs to any given input.

Algorithms that are already being implemented through machine learning are:

A. PageRank
B. Adaptive PageRank
C. Trust Rank
D. HITS
E. OPIC

There relevant performance is shown in the table:

Ranking Scheme	**Performance**
PageRank	99.27%
Adaptive PageRank	95.05%
Trust Rank(random seed)	55.63%
Trust Rank(balanced seed)	86.57%
HITS	42.86%
OPIC	88.36%

3. A RELATIVE COMPARISION

	PageRank	**Weighted PageRank**	**HITS**	**Query Independent Algorithm**	**Classified Tree Based**	**Page Content Rank Algorithm**	**Ranking Using Machine Learning**
Description	Divides page rank equally among outgoing pages	Unequal distribution of page rank among outgoing pages based on popularity	Results in highly relevant and important pages	Assigns value to each web page independent of the user query	Utilize a tree which has more breadth than height	Return relevant documents related to the query	GNN is made to learn PageRank procedure to obtain unknown outputs
Based On	Link structure of web	Link structure of web	Link structure and content	Trust level and link structure	Tree structure	Content	Neural networks
Input Parameters	Back links	Back and forward links	Back link, forward link, content	Graph of links	Tree of links	Content of page	Similar inputs for which the GNN is trained
Advantages	Simple, easy to understand, takes O(logn) time	Less complexity than PageRank i.e.<O(logn)	Gives importance to both structure and content, complexity is< O(logn)	Prioritize the documents on the web independent of the query	Enables multi-threading parallel, improves efficiency	Gives related results only with complexity O(m)*	Less overhead for the implementer
Disadvantages	Ignores relative importance, theme shift problem, stresses on old pages	Do not give importance to relevancy	Less efficiency, topic drift problem	Do not consider query during ranking	Modification in search engine required	Ignores importance of a page	Hard to translate ranking mechanism to GNN

4. CONCLUSION AND FUTURE WORK

Whenever user searches for a query, search engine provides a large number of pages as a result. The user wants to go through only some of these pages which are important to him in spite of navigating all of them. It is the responsibility of the search engine ranking mechanism to make user's navigation easier and faster. For this purpose different page ranking algorithms are used by search engines. The order of the resulting pages depends upon the technique of the page ranking algorithm. Search engine may use different algorithms depending upon the user needs.

As a future guidance, an algorithm which can combine page link structure and page content aspects by removing shortcomings of present algorithms should be developed so that the quality of search results can be improved.

5. REFERENCES

[1]. Yongbin Qin, Daoyun Xu,"A Balanced Rank Algorithm Based On PageRank and Page Belief Recommendation".

[2]. Wenpu Xing ,Ali Ghobrani,"Weighted PageRank Algorithm".IEEE, 2004.

[3]. C.Ding, X. He, P. Husbands, H. Zha, H. Simon, "Link Analysis: Hubs And Authorities On The Web" ,2001.

[4]. TIANG Chong, "A Kind Of Algorithm For Page Ranking Based On Classified Tree In Search Engine ".IEEE.

[5]. Harmunish Taneja and Richa Gupta." Web Information Retrieval Using Query Independent Page Rank Algorithm". IEEE,978-0-7695-4058-0,2010.

[6]. Jaroslav Pokorny, Jozef Smizansky,"Page Content Rank: An Approach to Web Content Mining."

[7]. Sweahh Liang Yong,Markus Hagenbuchner and Ah Chung Tsoi."Ranking Web Pages Using Machine Learning Approach". IEEE,978-0-7695-3496-1,2008.

[8]. Shen Jie, Shen Wen, Fan Xin," Recommending Expert in Q&A Communities by Weighted HITS Algorithm".IEEE,978-0-7695-3600-2, 2009.

[9]. Maurizio Lamberti and Claudio Demartini,"A Relation-Based Page Rank Algorithm For Semantic Web Search Engine".IEEE,1041-4347,2009.

[10]. Chung-Hung,"A Branch And Bound Clustering Algorithm" .IEEE,0018-9472,1995.

[11]. Dung B. Le and Sunita Prasad," TS-LocalRank: A Topic Similarity Local Ranking Algorithm For Re-ranking Web Search Results".IEEE,978-1-4244-5139-5.

[12]. Xia Feifei and Zhang Guangnian," Design and Implementation of a Java Based Search Engine Algorithm Analysis System". IEEE, 978-4244-3521-0.

[13]. Weiguang Xu, YafeiZhang, Jianjiang Lu and Zhenghui Xie, " A Framework Of Web Image Search Engine".IEEE,978-0-7695-3615-6.

[14]. Naresh Barsagade," Web Usages Mining And Pattern Discovery: A Survey Paper", CSE 8331, Dec.8,2003.

[15]. R.Cooley,B.Mobasher and J.Srivastava," Web Mining: Information And Pattern Discovery On The World Wide Web".IEEE,2007.