State of Art Different Web-Page Ranking Approaches

Bhushan Thakare
Asst. Professor Pune

Rohan Rawlani
Student Pune

Sahil Pathak
Student Pune

Dipali Salve
Student Pune

Ritesh Natekar
Student Pune

ABSTRACT

The internet engines like Google will be the tools which in turn allow you in order to understand along with speedily look for solutions. The present search for e.g. Google doesn't think about user needs and wants. Keys are associated with search engine optimization and search engines use it according to the user's dilemma. The particular PageRank formulas 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.

General Terms

Networking, Performance, Efficiency, etc.

Keywords

Page ranking, Search engine optimization, indexing, re-ranking.

1. INTRODUCTION

Internet is a vast source of information. Now a days web search tools (engine) have become one of the indispensable tools for people who do surf on internet. But with the growing use of internet [1], it is expanding rapidly in its content.

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 which they run on 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 ordering of the web pages 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. So there is a need of some mechanism that can arrange the pages according to their importance dynamically whenever user searches any content.

Now we present the techniques available for ranking web pages on the internet.

Ranking is nothing but assigning some value to a webpage among various web pages available on the internet. This 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 The former type involves the hyperlinks which link one web page to other web page.

In this paper we are going to study different algorithms, ranking techniques and methods illustrated.

2. RANKING ALGORITHMS

2.1 Page Rank Algorithm

This algorithm is developed and used by Google and is developed by Google founders Sergey Brin and Larry Page. This algorithm is based on the link structure of the web. It divides the page rank of 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 and is an extension to PageRank algorithmby Xing and Ali Ghorbani[2]. This is also a link based algorithm but it does not divide the page rank evenly but it instead 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:

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

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

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

Where: and : number of incoming links to page u and p, and : number of outgoing links of page u and p.

2.3 HITS Algorithm

It is called Hyper Induced Topic Search which is both link based and content based [3]. 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 applied 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

- For each page p E S, do Steps 3 to 5
- Let T be the set of all pages S points to.
- Let F be the set of all pages that point to S.
- Let S = S + T + some or all of F.
- Delete all links with the same domain name.
- 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.

- Let a page p have a non-negative authority weight and hub weight . Pages with relatively large weights will be classified to be the authorities, similarly hubs with large weights .
- The weights are normalized so the squared sum for each type of weight is 1.
- For a page p, the value of is updated to be the sum of overall pages q linking top.

2.4 Query Independent Algorithms

It is query and content independent algorithm which assigns a value to every document independent of the query [5] and it is also concerned with the static quality of web page. Page ranking is computed 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

```
1)  If node having no out link then equally distribute probability otherwise distribute it according to out links
2)  For all i,j if(counter==0) then
3)  Prob[i][j] =1/N else
4)  If(prob[i][j]==1 then
5)  Prob[i][j]=1.0/counter
6)  Multiply the resulting matrix by 1-m
7)  Add m/N to every entity of the resulting matrix, to obtain probability transition matrix
8)  For all I,j do prob[i][j]=(prob[i][j]*(1-m))+(m/N)
9)  Randomly select a node from 0 to start a walk say s_int.
10) Initialize a random surfer and itr to determine no of iterations required to 0.
11) Try to reach at steady state within 200 iterations, otherwise toggling occurs.
12) Multiply probability transition matrix probability vector to get steady state and check if system is in state or not.
13) Print the rank stored in the 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				

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

1) There is only root when initializing tree of classified tree.

2) In inquiry, multithreading parallel splits all the inquiries from the different users at the same time to key aggregation……… {keyi1, keyi2,.. keyin} i=1,2,...

1) 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 minutes, the users make the second same search), the index database can be multithreading parallel.

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

3) 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.

4) 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;

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

6) Repeat the process from 1 to 6.

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:

1) 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

2)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.

3) 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.

4) 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 presentin 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%**

2.8 EigenRumor Algorithm

This algorithm enables a higher score to be assigned to a blog entry entered by a good blogger but not linked to by any other blogs based on acceptance of the blogger's prior work. In the recent scenario day by day number of blogging sites is increasing, there is a challenge for internet service provider to provide good blogs to the users. Page rank and HITS are very promising in providing the rank value to the blogs but some issues arise, if these two algorithms are applied directly to the blogs.

These issues are:
1. The number of links to a blog entry is generally very small. As the result, the scores of blog entries are calculated by PageRank, for example, are generally too small to permit blog entries to be ranked by importance.

2. Generally, some time is needed to develop a number of in-links and thus have a higher PageRank score. Since blogs are considered to be a communication tool for discussing new topics. It is desirable to assign a higher score to an entry submitted by a blogger who has been received a lot of attention in the past, even if the entry itself has no in links at first.

2.9. Distance Rank Algorithm

A distance rank algorithm is proposed by Ali Mohammad Zareh Bidoki and Nasser Yazdani. This intelligent ranking algorithm based on reinforcement learning algorithm based on novel recursive method. In this algorithm, the distance between pages is considered as a distance factor to compute rank of web pages in search engine. The main goal of this ranking algorithm is computed on the basis of the shortest logarithmic distance between two pages and ranked according to them so that a page with smaller distance to assigned a higher rank. The Advantage of this algorithm is that, being less sensitive, it can find pages faster with high quality and more quickly with the use of distance based solution as compared to other algorithms. If the some algorithms provide quality output then that has some certain limitations. So the limitation for this algorithm is that the crawler should perform a large calculation to calculate the distance vector, if new page is inserted between the two pages. This Distance Rank algorithm adopts the PageRank properties i.e. the rank of each page is computed as the weighted sum of ranks of all incoming pages to that particular page. Then, a page has a high rank value if it has more incoming links on a page.

3. CONCLUSION

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 recommending ranking mechanism to make users navigation easier and faster. Hence we have discussed different page ranking algorithms.

4. 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] 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.

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

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