A Kind of Algorithm For Page Ranking Based on Classified Tree In Search Engine

TIAN Chong
Faculty of Science
North University of China
Shanxi PRC 030051
chinaapplechong@163.com

***Abstract*—Algorithm of page ranking is the core of search engine. This paper proposes a new type of algorithm of page ranking by combining classified tree with static algorithm of page ranking-PageRank, which enables the classified tree to be constructed according to a large number of users' similar searching results, and can obviously reduce the problem of Theme-Drift, caused by using PageRank only, and problem of outdated web pages. It improves the searching efficiency without reducing the searching speed, which provides the users with the abundant expanded information relevant to searching content.**

Keywords-Algorithm of page ranking; PageRank classified tree Architecture of search engin

I. INTRODUCTION

With the rapid development of internet, the number of internet users in the world is increasing very quickly and suddenly. Simultaneously, internet, as one huge information source composed of complicated hyper-texts, expands very rapidly, with 7 million new web pages every day. Therefore, the most concerned issue for the users would be how to collect the useful one from the massive information, and to find their genuine information effectively and quickly.

More and more people rely on the search engine, which plays an important navigate role in the information sea. As for the specific search request, the users always expect to get priority to find the most appropriate web pages, and business users would be inclined to provide their information to the targeted customers prior to their competitors. In the web page aggregation meeting the users' searching demands, those web pages that rank the front of results, narrow the searching range and searching time, and improve the users experience, are the problems to be solved by algorithm of page ranking. No doubt to say that algorithm of page ranking is the core of internet searching study.

There are two main types of page ranking: one is the traditional searching method based on the web page content, however, the huge data of internet would be great challenges to the traditional information searching technology. Meanwhile, web page is different from the general text, the former is semi-structured text, including much structured information; and web page does not exist independently, because the links indicate the interrelation between the web pages. The other type of types of page ranking is based on hyperlinks structure analysis, which distinguishes the internet searching and traditional information searching. And it is applicable to the huge data searching in internet.

The page ranking method based on linking structure utilizes the features of web page aggregation to evaluate significance of the page and linking, which determines the ranks of searching results. Marchiori believes that the authority or quality of the web page is up to the amounts of linking to it[1]. Sergey Brin and Lawrence Page put forward with static algorithm-PageRank[2]. And J. Kleinberg introduces dynamic algorithm of ranking page-HITS[3]. As classical hyperlink algorithms, they are all based on Markov Model's Random Walk. Other linking analysis are all originated from those two algorithms.

II. PAGERANK ALGORITHM

Priciple 1. Based on regression relation of "the web pages linked to high quality one must be also high quality", the significance of the web page is determined. PageRank does not calculate the direct links amounts, while interprets the link from A to B as the vote from A to B.

$$P(url) = \sum_{v \in B(url)} P(v)/N(v) \qquad (1)$$

Here,

url is one web page,

$P(url)$ is the PageRank value of *url*.

$B(url)$ is the page aggregation pointed to *url*, i.e. backward chaining of *u*rl.

$P(v)$ is the PageRank value of web page *v*.

$N(v)$ is the external link numbers of *v*.

The web page linked by many page is surely to be the high quality one, because the backward chaining from the other pages could be regarded as the recommendations to the first one. The backward chaining of web page high evaluation will receive high evaluation, too. At the same time, few links to total amounts of links would receive high evaluation, and the page with a great many links to it would receive low evaluation.

Priciple 2. The users' reviewing conforms to a random surfer model. Suppose the users visit one page in an aggregation at random, with probability rate of d along external links without any backspace browsing. The probability rate of visiting the next page is the value of PageRank of viewed page. Or the new page is viewed at the rate of $1-d$. All this can be expressed as:

978-1-4244-7237-6/$26.00 © 2010 IEEE

$$P(url) = (1-d) + d \sum_{v \in B(url)} P(v)/N(v) \quad (2)$$

The probability of reaching one page in random surfing would be the sum of clicked links in other web pages linking to it. And damping coefficient can be introduced here, which usually equals 0.85. The probability would be reduced because it is impossible for users to click the link unlimitedly. More commonly, they happen to browse to another page stochastically. Damping coefficient d defines the probability of links to be clicked from users. It depends on the times of clicks. The more of value d is, the more probability of links to be clicked is. Therefore, the expression of random surfing to another page is $1-d$. $1-d$ is also the PageRank value of page.

From the analysis of PageRank Priciples, PageRank value is only related to the situation of internet, but not to the searching. If the detailed information of complete web pages can be acquired, every page's PageRank can be calculated theoretically. This value is the important quantities reference value to judge the importance of the page.

Furthermore, the PageRank is off-line calculation, and the page aggregation is acquired by the matched key words when users are searching to get recommendation results, so it responds at a high speed. And the success of Google proves that this algorithm is effective and reasonable.

Still, because of the author considers the linking architecture only, there are some shortcomings of this algorithm:

1) Ignore the relative importance while only treat the external links equally.

2) Similarity of contents cannot be guaranteed, which can easily lead to the problem of theme-shifting, i.e. the result may not meet the needs of users, even if there is no meaning for the users in the important page.

3) PageRank Algorithm stresses on old web pages.

Therefore, a new ranking algorithm of search engine based on classified tree and PageRank is proposed in the following.

III. RANKING ALGORITHM OF SEARCH ENGINE BASE ON CLASSIFIED TREE AND PAGERANK

A. The Structure of Search Engine

Search Engine is usually composed of crawler, indexer, searcher and inquiry, shown as Figure 1. For the great amount of data in web page, the particular rule of index has to be established to improve the search efficiency. The index is one of the core technologies of search engine, which directly influences the quality of result. So far, the most popular and effective index method is inverted file, i.e. the file with word splitter firstly forms the index data, and then these data will be inverted.

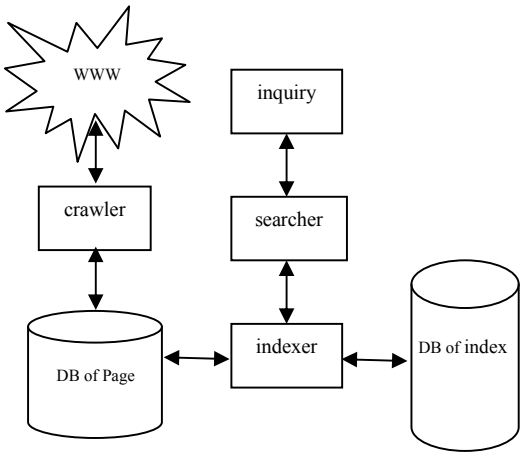

Figure 1 The Structure of Search Engine

Between the crawler and indexer, one parallel layer is introduced, that is, classified tree, shown as elliptical area within the dotted line in Figure 2. In this layer, multiple classified trees form the classified forest.

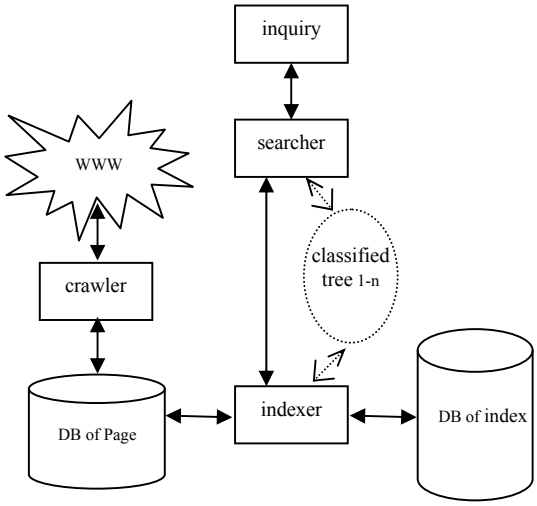

Figure 2 The Structure of Search Engine with classified tree

B. Classified Tree

The template is used to format your paper and style the text. The data structure of classified tree adapts the structure of B. The branches of the tree are relatively more, while the hierarchy of the tree is low. It is required that the relationship has to be established between the leaves of the tree and key words in the inverted file. The visiting between the two sides are doubleaction. 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.

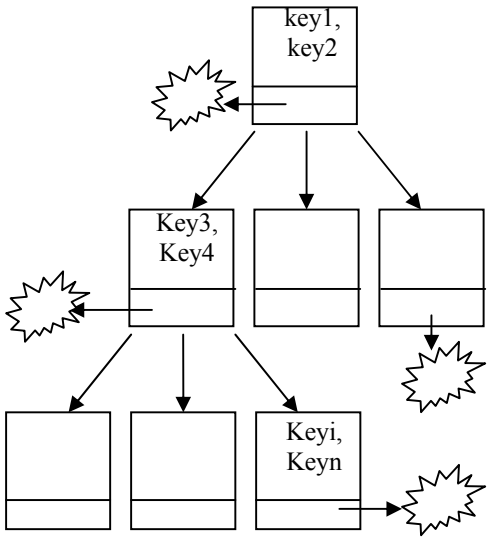

Figure 3 The classified tree

The format of the items in the inverted file is as follows [4]:

Keyi→{[Pid1,ni1(hit1,hit2,…,hitni1)]
,…,
[Pidn,nin(hit1,hit2,…,hitnin)]}

TABLE I. INVERTED INDEX

Keyi	Pid	Other	Treei	Queuei
Keyi				
key1				
Key2				
keyn				

Here,

Keyi The serial number of key words;

Pid The serial number of page file;

nin the number of times as key words in the file;

hiti the place where key words appear in the file;

Treei the place where key words appear in the classified tree;

Queuei User queue.

C. Ranking Algorithm after Introducing the Classified Tree

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

1) There is only root when initializing tree0 of classified tree;

2) In inquiry, multithreading parallel splits all the inquiries from the different users at the same time to key words aggregation:

$$SKi = \{keyi1, keyi2, \cdots keyin\} \quad i = 1,2,\cdots;$$

3) After collecting the key words from m users, the author merges the repeated ones,calculate $\bigcup_{i=1}^{m} SKi$;

4) 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;

5) All the users can be searched according to the key words in the inverted file within a period of time.

6) 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];

7) 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;

8) Multithreading parallel in all classified trees, and if the new generated node value (key words aggregation) is the subclass of node of classified tree I (Condition 1), then along the branch down, find the node with no more than 2 of key words, 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;

9) 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 offsprings of original generated node. Then the new generated page aggregation is sent to users as feedback;

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

11) repeat the process from 1) to 9).

IV. CONCLUSIOIN

From the ranking algorithm based on the classified tree, the process of construction of tree is actually using the users' large searching information to practice repeatedly to get intensive results of page aggregation. By using key words, the result from different users can be expanded relatively to improve the recall of users.

The classified tree is dynamically changing, in accordance with the changing of users' searching. The latest result can rely on the users' behavior to complete the corresponding nodes' page aggregation. Therefore, it solves the problem of theme-shifting or other disadvantages to new page from different method in the PageRank algorithm. Furthermore, the similar results from large number of users'

searching expand the specific user's relative theme, which enrich the user's searching result.

Secondly, in the page aggregation, the author ranks the turn according to the page PageRank value, therefore, the user can get priority to reach the important page in the theme relative aggregation.

Finally, the architecture of classified tree is finished in the server of search engine, therefore, there is no influence on the users' searching speed. And there are many classified trees, which enables the multithreading parallel in the classified forest, in a way to further improve the efficiency.

REFERENCES

[1] Marchiori,M.1997.The quest for correct information on Web: Hyper search engines. In Proceeding of the 6th International World Wide Web Conference.

[2] L.Page , S.Brin ,R.Motwani,and T.Winograd , The PageRank CitationRanking : Bringing Order to the WEB ,January 1998.

[3] J.Kleinberg.Authoritative sources in a hyperlinked envir on ment. Proc. 9th ACM-SIAM Symposium on Discrete Algor- ithms, 1998. Extended version in Journal of the ACM 46(1999).,1-33

[4] http://lucene.apache.org/java/docs/

[5] Xianchao Zhang, Xinxin Fan, Xinyue Liu and Hongyu , A Ranking Algorithm via Changing Markov Probability Matrix Based on Distribution Factor. Fifth International Confe- rence on Fuzzy Systems and Knowledge Discovery, IEEE 2008, 3-7