Improvement on PageRank Algorithm Based on User Influence

Yang Wang

Basic Medical College
Shaanxi University of Chinese Medicine
Xianyang, Shaanxi, China

Abstract — With the rapid development of the Internet, web search engines have become the most important Internet tools to retrieve information. The rational PageRank algorithm is mainly based on the relationship between links page sorting, thus, is easy to ignore the professional site, too much old web pages and other shortcomings. To solve these problems, we add user influence factors based on the website authority weighting and time to improve the weight. Our experimental results show that comparing the traditional PageRank algorithm, the improved algorithm can effectively improve the precision and user satisfaction with the sorting results.

Keywords - PageRank algorithm; User influence; Topic relevance; Page sorting

I. INTRODUCTION

With the rapid development of the Internet, all kinds of data and information grow exponentially. Internet has become the most abundant and the most intensive source of information in the world. For so much rich information resources, web search engines have become the most important Internet tools to retrieve information. When retrieving the web information we always hope to be able to get the latest most relevant and authoritative web pages, [1-3] and obtain them at the front of the search result pages. User time can be saved largely, and on the other hand, the burden of network can be reduced in this way. Usually, web search engines work with the several steps. Firstly, query the keywords input by users, secondly, find related pages based on the words. Then according to the specific algorithm, order the relevant web pages and feedback the orderly results to the user. Search engine systems are usually made by the three modules, and they are searching, indexing and retrieval. The task of search the web spiders collect the web pages from the Internet, Starting from the initial page URL to provide hyperlinks for the index part of its original web set. Index part is responsible for sorting the spiders web information and extracting keywords, and then generate the inverted index table inverted index for user input keywords retrieval section provides information database. Retrieve part works in the index table with all the lost keywords in web pages and sort them according to a certain algorithm and return the results to the users [4-6].

II. METHODOLOGY

A. PageRank Algorithm

PageRank algorithm is used to calculate the weight of the web page. The value is Google to retrieve web page sorting is an important basis of PageRank algorithm presented by S.B. and L.R. held in April 1998, the 7th international conference www7. The algorithm is as follows: the basic idea of PageRank web page link to web page B is A recognition of web page B, as A web page for web B cast A vote, A web page, the more votes the authority of the web page value, the greater the values of the authority of the web page according to the above ideas into A formula:

\[ P_i = \sum_{j \neq i} \frac{P_j}{O_j} \]

in which \( O_j \) is the number of the links.

Internet link structure formation is spontaneous and disorderly, [12,13] there may be a precipitation phenomenon, some web pages only chain clusters within the cluster to other web pages, but not outside the chain to cluster web pages, passed in PageRank values stranded in the web page cannot be passed out within clusters. The solution is as follows: set your visitors will randomly select a link to set your visitors browsing is the probability that \( d \) is not through the link, and jump to the other is the probability that a random page \( 1 - d \). The formula is shown as follow:

\[ P_i = d \sum_{j \neq i} \frac{P_j}{O_j} + (1 - d) \]

Because of the link structure of spontaneous and disorder, matrix A is black hole effect, existing in the network link graph into the degree of 0 page, after finite iterations, the link of PageRank value of all the pages are 0. Solution: to have no outside chain outlets, set the link of all the other nodes on the figure is it outside the chain, namely in the adjacency matrix, A whole of 0 will all element value set to 1.
The formula is shown as follow:

\[ A' = \left( A_0 + \frac{t}{N} \right) \]

in which variable \( t \) is used to solve the black hole effect.

Therefore, the formula of PageRank algorithm is described as follow:

\[ P = \left( 1 - d \right) E/n + \left( d \right) A'^T P \]

PageRank algorithm used in the power of simple iterative method to calculate but slow convergence speed, and time consumption is large.

### B. Hadoop and MapReduce

Hadoop is put forward by the Apache software foundation open source framework, including graphs of parallel programming model interfaces and HDFS (Hadoop Distributed File System) two parts graphs is put forward by Google parallel programming model, the model will be huge amounts of data analysis for the Key - the Value of the form, through mapping (Map) and specification (Reduce) process analysis and calculation of the complete data parallel first, will become one piece of data segmentation (Split), according to the defined input format, input into the Map in the Map phase, data input by the user custom Map function processing, mapping and the Key - the Value of the shuffle (shuffle) process automatically sorted output to Reduce end; In the Reduce phase, will have the same Key Value, the Value of the aggregate according to the user defined way.

Some scholars have proposed to the problem using graphs for parallel processing, reduce the waste of time. Mapreduce is a software architecture put forward by Google, and is a kind of deal with huge amounts of data parallel programming model, for large-scale data sets. Usually greater than 1 TB of parallel computing, graphs divided into stages of Map mapping and Reduce reduction process of graphs. First of all, the input file into separate N blocks, each assigned to a Map function of the processor.

Perform the Reduce function processor will each intermediate results produced by the Map function to merge.

The Reduce function processor can also parallel execution.

\[ PR(u) = (1 - \alpha) + \alpha \sum_{i \in B(u)} PR(v) W_{(v,u)}^{in} W_{(v,u)}^{out} \]

in which \( B(u) \) is a set of web page \( u \) all backlinks web page, the chain set \( u \) all of the web page to a web page.

\[ W_{(v,u)}^{in} = 1_u / \sum_{p \in F(v)} I_p \]

\[ W_{(v,u)}^{out} = O_u / \sum_{p \in F(v)} O_p \]

In which \( F(v) \) is a set of web page \( v \) before link to web pages.

Hadoop framework graphs by running Jobtracker service in a Master node and run in a number of Slave nodes Tasktracker services Master Jobtracker node management could be divided into multiple tasks Tasktracker, respectively on the Slave node Jobtracker running management and supervise the completion of Tasktracker, Tasktracker respectively complete the node of the data processing, implementation of large-scale data parallel processing. Hadoop framework is described in Figure 1.

The current research shows that the PageRank algorithm considering still only web link structure, so the there are the same as the traditional PageRank algorithm shortcoming includes:

1) NO subject

Because there is no web content and search topic relevance, so we can not identify search to the results of the web pages are pertinent to the topic search, and cause the search to the part of the web has nothing to do with the theme.

2) Old page

Relative to the old website, new web existence time is short, less link relations, obtain the smaller values of the PR, ranking position Obviously it for new pages and hot topic, is not reasonable.

3) User's interests ignorance

Only according to link relations without considering the user's interest in web pages, is also a Weighted PageRank algorithm design is not reasonable.

### User Influence

Users browse information will just click on their interested web pages, in this paper, on the basis of the PageRank algorithm is optimized when calculating the page PR value, such as single from the clicked pages on the importance of considering the number of the chain, namely the longer survival time for the web, it in live cycle by click on the more the number of times, so unfair to the distribution of the new web page, in order to solve this problem.

Based on the statistics of the web crawl after clicks and last web crawl the statistics of click on the difference between the number to allocate the PR value of the page If the web site for a period of time has been clicks increase the faster So that the web is the greater the PR value because of the web page clicks can be artificially controlled So there are artificially high PR value of a kind of web page of web traffic situation Aimed at the problem in statistics to consider how to reduce cheating importance for web of shadows.

1 Data cleaning
If some of the IP address of the user clicks a day to a fixed web page above a certain threshold S according to clean up

2 Users’ usage

Not completely on the user clicks to calculate scores But correspondingly define user clicks pages within a certain amount of time.

In a certain period of time of clicks could be obtained by the input url through the boost values of high and low degree of importance to set up the web page Through the weight of high and low Set a weight high user clicks Low click-through rates set of weights is small So users when using search engine to search to give full consideration to the user's feedback information.

The user feedback is the user feedback on the search results, divided into explicit feedback and implicit feedback. Feedback is explicitly in the active participation of users, by users to upload link address and the corresponding key words to reflect user. Esteemed implicit feedback is in the absence of explicit user participation, through the analysis of users’ search behavior information to users of the search results were the feedback factor using implicit feedback using user click number.

The formula is shown as follow:

\[
W = \frac{V(v \rightarrow u)}{TL(v)}
\]

in which TL (v) is the user for web v all chain link clicks the sum, F (v) is a set of web v positive link pages.

3) Our Approach

Improved web page sorting algorithms, we think the web link relations (including chain and chain), the more especially chain into the relationship, the more represents the authority of the web page; Web content related to the themes, the higher link clicks, link to web page should obtain the higher PR value, ranking position more At the same time, the new web page contains information about the new because in the network time is shorter, link relations and clicks were small, add time correlation factor algorithm should make some compensation as a result, according to the Topic - Sensitive thoughts of PageRank algorithm, namely the user browsing the web is affected by the current query subject and content, based on the web page chain into the chain on the basis of the structure, add the query and web content relevance factor at the same time considering the link clicks Users browse the time factor to extract the user interest factor, web page there is time and fusion, the improved algorithm can effectively improve web page sorting effect, more can meet the demand of user queries.

Link after classification, the pretreatment of the output as shown in table 1 Among them: R is expressed as \{PageRank, flag, Num\}, PageRank namely PageRank value of the current node; Flag is used to mark the source of the link: flag = 1 for the same site, flag = 0 means different site; Chain of the nodes of the current node.

<table>
<thead>
<tr>
<th>Output</th>
<th>K</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Map1</td>
<td>S1</td>
<td>R:1;R:3</td>
</tr>
<tr>
<td>Map2</td>
<td>S2</td>
<td>R:6;R:8</td>
</tr>
<tr>
<td>Map3</td>
<td>S3</td>
<td>R:5;R:7</td>
</tr>
<tr>
<td>Map4</td>
<td>S4</td>
<td>R:9;R:8</td>
</tr>
<tr>
<td>Map5</td>
<td>S5</td>
<td>R:12;R:11</td>
</tr>
<tr>
<td>Map6</td>
<td>S6</td>
<td>R:7;R:8</td>
</tr>
</tbody>
</table>

Each station only need two maps of to perform the PageRank calculation, so the figure, a total of six Map is needed to complete the follow-up processing Because calls the Map need to drive load constants of initialization, etc, to spend a lot of resources In the PageRank algorithm of iterative process, decrease the number of calls for the Map process can Reduce the consumption of system resource literature points out that put a large amount of complicated data calculation in the end, only do simple statistical work, Reduce end can improve calculation efficiency system according to this thought, on the Map to separate each Key - Value pairs, with r said the calculated PageRank, r’, said the previous calculated PageRank, c said node chain nodes, f flag values

In the process of mass data processing, data compression is particularly important. This paper works in three levels: data preprocessing layer calculation layer to compress the data.

1) Data layer

The Hadoop - LZO algorithm used in the Hadoop LZO compression algorithm can reduce the size of the storage of data and data reading and writing time Compressed data storage in the HDFS, cluster can hold more data not only that, because the graphs work usually on the I/O bottlenecks, storing compressed data means fewer I/O operations, job runs more efficient Therefore, this article choose open source Hadoop - LZO compression algorithm as the web hyperlink compression scheme was carried out on the experimental data compression Hadoop - LZO compressed data can be partitioned In parallel processing, has a good compression performance and compatibility of graphs.

2) Preprocessing layer

Using Num PR method is usually a web link consists of dozens or even hundreds of characters in the URL of the said, using the web page link to represent the node will take up a lot of memory, increase the quantities of data between the Map and Reduce, Reduce the I/O speed and computing speed the NumPR methods with reference to the literature in

---

**Figure 2.** Data Flow based on Web Search Engine
YANG WANG: IMPROVEMENT ON PAGERANK ALGORITHM BASED ON USER INFLUENCE

DOI 10.2013/IJSSST.a.17.07.13 13.4  ISSN: 1473-804x online, 1473-8031 print

this paper, with a globally unique number to replace the URL that is used to represent a web page. Numbers and the URL mapping relationship is stored locally in the hash table. Until the PageRank calculation is complete, then id number back into the URL of a web page.

3) Computer layer

Add combiner process calculation layer compression or compress the data in the Map end, reduces the data volume, reduce the network overhead, improve the operation efficiency of the algorithm and system.

III. RESULTS AND DISCUSSION

A. Experimental Environment

Our paper uses the 5 PC set up Hadoop distributed computing platform, No1-No5 respectively among them: Master PC1, running tracker; No1 as a Slave, run Tasktracker per PC specific hardware environment: Intel Core 2 Duo CPU 2.20 GHz; 2 GB of memory; 300 GB hard drive; 100 Mb/s net mouth software environment: Ubuntu 13; JDK1.6.0 27; Hadoop 0.20.3.

B. Experimental Design

To test the efficiency of the improved algorithm, this paper will be under the single PageRank serial algorithm in the literature PageRank parallel algorithm comparing with LCPR algorithm experimental data derived from the open source tools Heritrix - 1.14.4 crawl web pages, the crawler seed set as sina Sohu netease portal, etc. After many crawl, crawl web pages according to the quantity can be divided into 4 groups: 103.0509 billion from the pretreatment of data is loaded into the final PageRank convergence, the time required to as a standard experimental comparison. All of the above algorithms require certain pretreatment time, about 1~2 iteration time. Due to large the whole algorithm iterations, an average of around 50 rounds, so pretreatment time temporarily not consider experiment made four contrast, total number of web links from 100000 to 900000, divided into four groups of data to PageRank for the overall variation is less than 0.001 as a condition of convergence, take 3 times of the average of the experiment as the result of the experiment.

It is easy to see: when the data volume is small, two parallel algorithm running time With the increase of the amount of data, SerialPR (serial PageRank algorithm) running time increased significantly, almost exponentially, when article links 900000 serial algorithm with multiple iterations, excessive memory consumption, CPU resources nervous, increase run time, so that the JVM memory, abnormal program exit; However along with the increase of the amount of data parallel algorithm execution time rising steadily, by contrast, LCPR algorithm due to launch the Map and reduce the Map number is less, the middle of the I/O transmission data through efficient compression, efficiency is relatively high, compared with short operation time, this article will PageRank and LCPR parallel algorithm with good performance are deployed in the number of different nodes on a cluster, test complete the entire algorithm iterative time, with the increase of number of nodes, two parallel algorithm running time are in decline, this is because with the increase of nodes, each node of the full use of the CPU and memory resources, increase the cluster a number of Map data processing, system concurrency enhancement LCPR and PageRank algorithm of time difference in gradually increased, in different cluster, LCPR algorithm has higher efficiency.

![Figure 3. Time consuming of the improved PageRank](image)

![Figure 4. The precision of the improved PageRank](image)

Obviously, under the different web page number, serial execution of PageRank algorithm and parallel execution results were similar, namely web page sorting accuracy. The reason is that these two kinds of implementation process of collation unchanged, cause the page rank is changeless. Because the query is random, so the query result accuracy not completely consistent, on the other hand, a little number of pages, LCPR accuracy is higher than PageRank algorithm, with the page number, LCPR algorithm and PageRank algorithm execution results tend to be more consistent. This is because when the page number is small, LCPR eliminates the intentionally brought about by the mutual reference between web pages in the same web site ranking is unfair, as the page number, this kind of injustice, gradually reduce the influence of the rankings, LCPR algorithm's advantage is no longer obvious.

IV. CONCLUSIONS

Search improved accuracy is the key to the user intention judgment, if in a particular application environment can obtain various background information of the user, user can...
use this information to more accurately determine the user's information needs, and then more accurate to provide the required information. Algorithm in this paper the application of the scene is a large enterprise's internal network, all users are real name, and user information are available as needed. Through the analysis of user attributes, concluded that the user's information demand, and then search the data according to the degree of interest to the user, the user sent to users, the most needed information priority with tolerable time loss greatly improve the accuracy of search, actively explore is to improve the accuracy of information search.

In this paper, we firstly analyzes the PageRank algorithm based on link analysis algorithm and introduces an improved algorithm in order to improve web search precision and meet user requirements well. In the Hadoop distributed platform, research on PageRank algorithm is optimized to PageRank algorithm is improved, and put forward a classification of LCP algorithm based on link, and according to the characteristics of the graphs framework to divide and rule, we design a parallel algorithm of LCP and layer 3 compression methods. Experimental results show that LCPR parallel algorithm in this paper can well adapt to the graphs framework, reducing the consumption of CPU resources, reduce the amount of I/O data transmission and network overhead, save system resources, computing resources and network resources, and a little number of pages, enhance the accuracy in page rank HDFS storage due to PageRank algorithm iterations is more, a large number of intermediate data, how to in the case of huge amounts of data to reduce system to store data, the middle is a main work of the next step.

REFERENCES


