International Journal of Smart Home
Vol. 10, No. 5 (2016), pp. 297-304
http://dx.doi.org/10.14257/ijsh.2016.10.5.27

Analysis on the Digging of Social Network Based on User Search Behavior

Junyi Du[1,2], Zhiyong Zhang[1] and Changwei Zhao[1]

[1]*Information Engineering College, Henan University of Science and Technology, Luoyang 471023,China*
[2]*College of Mathematics Science, Luoyang Normal College, Luoyang, 471022,China*
hndujunyi@sina.com

Abstract

Traditional social network has a glittering array of demerits such as low speed as well as low efficiency. This paper comes up with social network behavior search algorithm based on Hadoop Cloud Computing, which mainly adds impact factors, time arrow and page correlation factors into digging factors so as to improve the performance of digging computing and the search efficiency. The experiment proves that the computing has good effect and has instructive significance for user analysis of Cloud Computing.

Keywords*: Hadoop; User Search; Social Network*

1. Introduction

At present, sky rocketing information appears in the cloud computing platform and network information grows exponentially. As a result, quick research based on cloud computing becomes an important means for people to acquire information. Besides, Stanford University of the United States has put forward PageRank [1] and at the same time IBM comes up with the idea of HITS technology [2-3]. What's more, a dazzling array of available information has been produced based on Web query pattern, which can reflect some search behaviors of users from the other side and is able to analyze the search quality as well as user behavior. Moreover, the visitor number for large scale search engine can reach 100 million and the log file objects are massive. Based on the characteristics of massive files, it shall be hard for traditional data storage and computing methods to be adapted to user behavior analysis in terms of search engine. Therefore, concerning this problem, this paper comes up with the idea of network behavior search algorithm and proves that it will lead to good effect and shall provide guiding significance for the user search analysis of cloud computing.

2. Hadoop Framework

Hadoop structure is an open distributed computing framework which mainly includes master node and compute node [4]. Hadhoop has a master node (JobTracker), which is mainly used to regulate and manage other compute nodes (TASKtRACKER). JobTracker shall assign the Map task and Reduce task in the cloud computing model to free TraskTracker and then carry out monitoring. Once a TaskTracker breaks down and cannot bear the responsibility JobTracker shall send the task of this machine to another free TaskTracker. Therefore, Hadoop is a kind of reliable distributed computing framework.

ISSN: 1975-4094 IJSH

3. Map/Reduce Model

In cloud computing, resources and distribution of tasks are not correspondent to each other. First of all, tasks are reflected on corresponding resources, and then corresponding physical equipment through resources. Currently, this reflection method mainly adopts the Map/Reduce model of cloud computing as shown in Figure 1:

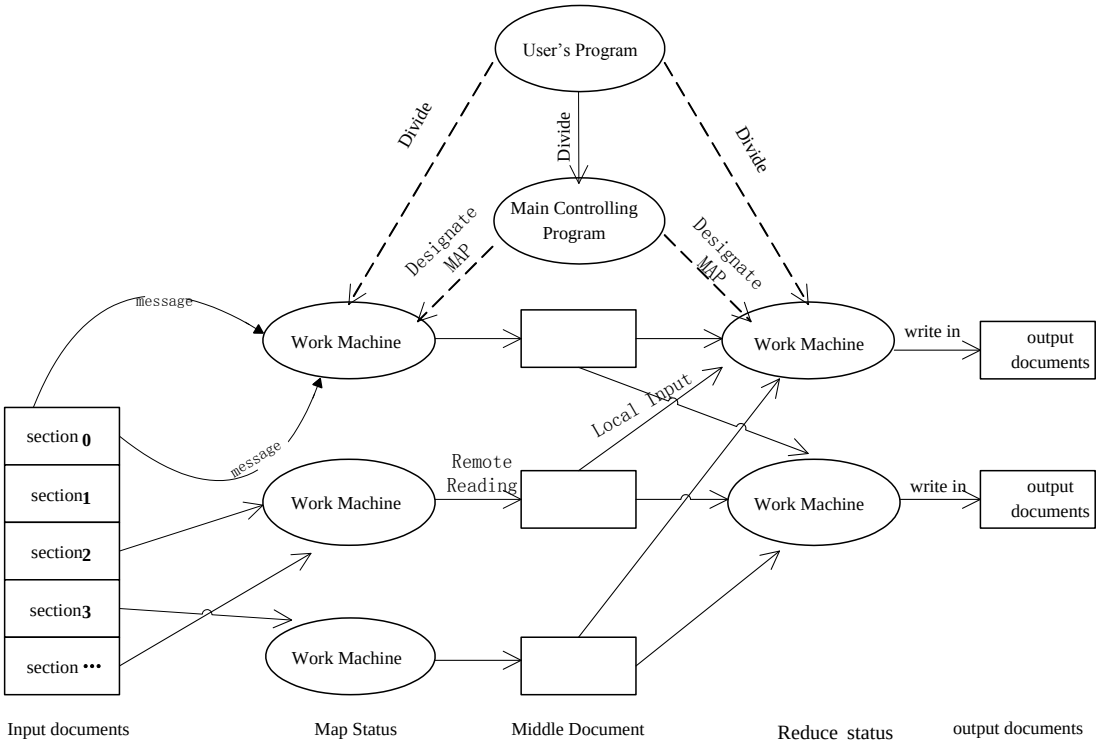

Figure 1 Map/Reduce Model

From the above model, the cloud computing model can be found in the need to user's tasks is divided into several segments in the system, these fragments represent the cloud customer service, and these fragments requires work converting machine can be a physical device output file. Due to the limitations of cloud computing resources and structural heterogeneity characteristics of recurrent resource competition and lead to irrational distribution of resources, how can make a reasonable configuration of resource allocation in the model is the core of the allocation of resources.

4. Relevant Work of User Search

4.1. Weight of Various Search Index

The resource information of social network based on cloud computing is vast. To begin with, search engine will carry out text segmentation based on the key words submitted by users and then delete the irrelevant words. Then, with search engine, we can analyze the weight of searching words. Based on this, the computing value can be adopted to highlight the key words.

This paper takes the word weight as the search uniterm and the term set as well as the weight of key words can be taken as the integration of uniterm. Suppose $X_{i,j}$ is the divided weight after searching and $Y_{i,j}$ is the weight of a certain word in the query

sequence and $z_{i,j}$ is the weight of sequence segmentation. Therefore, we can have the following conclusion:

$$f(d_j,t)=\frac{\sum_{i=1}^{t} X_{i,j}\times Y_{i,j}\times Z_{i,j}}{\sqrt{\sum_{i=1}^{t} X_{i,j}^{2}}\times\sqrt{\sum_{i=1}^{t} Y_{i,j}^{2}}\times\sqrt{\sum_{i=1}^{t} Z_{i,j}^{2}}} \tag{1}$$

$$X_{i,j}=Fre_{i,j}\times\log\frac{N}{n_i} \tag{2}$$

$$Y_{i,j}=\frac{n_{i,j}}{\sum_{i=1}^{N} n_{i,j}} \tag{3}$$

$$Z_{i,j}=\frac{\{N_{i,j}\}}{\sum_{i=1}^{N} N_{i,j}} \tag{4}$$

Among which, $\overrightarrow{d_j}=(X_{1,m},X_{2,m},\cdots X_{i,m})$, $\vec{t}=(X_{1,n},X_{2,n},\cdots X_{i,n})$. $Fre_{i,j}$ refers to the frequency of I key word in j web page. What's more, $\log\frac{N}{n_i}$ refers to the frequency index of reverse document, N refers to the number of resources and n_i is the total number of web page showing for i key words.

4.2. Model Factors to Be Considered

Based on the PageRank algorithm, this paper firstly analyzes the user searching behavior from the perspective of searching frequency as well as preference. Then, the weight shall be taken into consideration, and main considerations show as follows:

(1) User searching behavior: Concerning the searching behavior q, and clicking volume C, users may ignore the return result URL and therefore the searching behavior is greatly affected by it. This paper adopts the following formula to make up for this defect

$$U_q=\sum_{i=1}^{n} c(A,q)*click(A,q) \tag{5}$$

In the formula, $c(A,q)$ is the balance factor of webpage A and $click(A,q)$ is the clicking volume. The higher value of U_q, the page is more popular otherwise the page shall be listed in the bottom.

(2) User thinking time: While carrying out searching behavior, if users find out that there are similarities between searching behavior they shall stay on for a while which yet is not related to their satisfaction. Therefore, the formula (6) is adopted to describe the weight of searching time.

$$Time(A,q)=\frac{t_i}{\sum_{i=1}^{n} t_i} \tag{6}$$

In the formula, t_i refers to the time users spend to browse website A while inquiring the word collection q

(3) Correlation among pages

In the course of searching in cloud computing, there is correlation between page i and j yet the weight maybe greatly different. Therefore, the balance factor should be relied on to compensate for the pages listed behind. Suppose N iteration is carried out during certain time $[0,t]$ and the pages clicked by users constructs matrix $C_{N\times N}$, among which, $C_{i,j}$ refers to the number of page i and j being clicked. If $C_{i,j}$ as well as $C_{j,k}$ is greater than 0, then we can say that i, j, k have relations and we can conclude the following:

$$K(A,T_i) = \lambda(ID_A, ID_{T_i}) \tag{7}$$

In the formula, $K(A,T_i)$ refers to the correlation between A and T_i, $\lambda(ID_A, ID_{T_i})$ refers to the correlation value found out based on the ID of two pages.

4. Analysis on the Advanced User Searching Algorithm

Page Rank is an important method to identify the page level/importance which is able to
promote the ranking of the pages so as to improve the searching correlation and quality. However, in the course of searching, a dazzling array of users will choose the page based on the similarity with the topic. As a result, there should be imbalance. So, not only the direct link relation is needed the implicit introduction factors are also needed. So, advancement can be carried out on traditional PageRank and the PR calculation for page X is shown as follows:

$$PR(X) = \sum_{(X,T_i)\in E} \left(\frac{PR(T_i)*(\delta_1 f(X,T_i) + \delta_2 T(X,q) + \delta_3 K(X,T_i))}{\sum_{k=1}^{M} click(T_i, X)}\right) \tag{8}$$

In the formula, parameters δ_1, δ_2, δ_3 refer to influence factors, time arrow and page relevance and $\delta_1 + \delta_2 + \delta_3 = 1$. E refers to the total of web pages, d refers to damping factor and $click(T_i, X)$ refers to the number of being clicked of page T_i and X and the higher number of being clicked, they are more related. To conclude, this formula takes user behavior, time arrow and page relevance into full consideration.

4.1 The Setting of δ_1, δ_2, δ_3 Factors

Influence factors, to some extent can be regarded as the key affecting the efficiency of algorithm. This paper adopts data sample to analyze the data elements, and conducts group calculation on the data of searching behavior finding out that influence factors have little effect on the interference factor. The performance feedback is shown in Table 1.

Table 1. Parameter Record

Record Number	δ_1	δ_2	δ_3
500	0.16878	0.31731	0.51391
1000	0.25492	0.36147	0.37461
10000	0.22712	0.42723	0.34565
20000	0.41291	0.21218	0.37491

From the above table, we can see that the value of δ_1 is between [0.16878, 0.41291] and the value of δ_2 is between [0.21218, 0.42723] while the value of δ_3 is between

[0.34565, 0.51391]. That is to say, the order singly carried out by reliability function cannot fully satisfy customers and the final order will affect the customer influence factors, time arrow and page relevance.

4.2. Real-Time Feedback Details

In the course of user searching based on cloud computing we firstly obtain a result set via search engine and then users can click the targeted pages after receiving the query result so as to acquire its ID number. What's more, based on relevant implicit values, the result set will be compared with the implicit degree of correlation and the page is closely attached can be returned to users as the new searching result, showing in Figure 2:

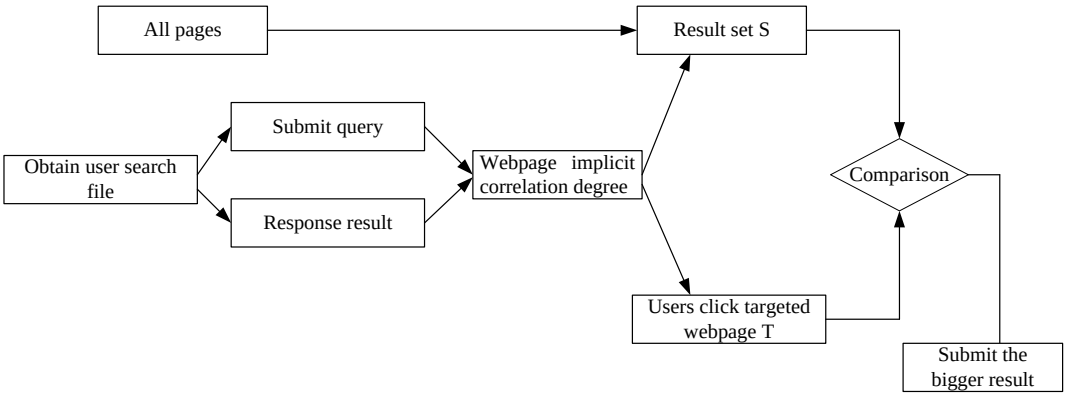

Figure 2. Real-Time Feedback Detail Comparison

4.3. Algorithm Flow

Step 1: Users search based on targets and obtain certain pages;

Step 2: Based on the basic PageRank computing, we introduce the user influence factors, time arrow, page relevance to conduct analysis.

Step 3: Analyze the webpage concerning the user influence factors.

Step 4: Carry out analyze based on the time needed based on time arrow.

Step 5: Choose based on the analysis on the page relevance.

Step 6: Submit the result of step 3 and step 5 to PageRank and calculate the result.

Step 7: Feedback the results to users.

5. Massive Data Processing Model Design Based on Hadoop

This experiment adopts 5 PC to construct the Hadoop distributed computing platform and one of the PC is taken as the server, which is mainly responsible as the main node and the other four is TaskTracker.

5.1 Data Collecting

The data used in the experiment is mainly collected from 58 [7] and this paper collects 10 million items in one week and the query setting is shown in Table 2

Table 2. Query Recording Format

Field Name	Instruction
Query	Query content
URL-Rank	URL Ranking
SessionID	User Cookie information

5.2. De-Duplication

There is a glittering array of duplicate records in iqiyi which ignited based on the different order of searching in 58 for example Zhang Lei male and Male Zhanglei and such demonstration approach will lead to one searching result and the duplicated one should be deleted. This paper adopts Map/Reduce to delete pseudo-code.

```
Map(String No,Stirng Content)
{ String Str[]=”lineContent.spli()”;
   Collect(id,term);// Collecting all data
  }
Reduece(String id,Tree terms)
{ While each<=terms
     {// Duplicate Weedout
     }
   Collect(id,new Terms);
}
```

5.3. User Data Analysis

The Hadoop framework is able to analyze and dig data from multiple perspectives including hot searching words, and single click frequency. Users of Youku website mainly search based on the hot words and these search behaviors can be analyzed. What’s more, the size of data set can be stored and calculated. The pseudo-code of hot topics is shown as follows:

```
Map(String No,Stirng Content)
  { String Str[]=”lineContent.spli()”;
     Collect(id,term);// Collect all data
     While each<=terms
        { collect(term,reduce)// reduce Send the data to reduce
        }
   }
  Reduece(String query,Tree values)
  {    int num=0;// Ser counter
       While each<=terms
       {num=num+values// Accumulative page view
       }
       Collect(query,num);
}
```

5.4. Experimental Result Analysis

This paper compares literature [9-10] and the data resource comes from Heritrix[11] and the webpage page view is distributed based on 100 thousand, 30 thousand, 50 thousand and 1 million with 5 clusters, including 100 thousand, 20 thousand, 30 thousand and 40 thousand. The comparison based on different data volume, nodes and accuracy is shown in Figure 3.

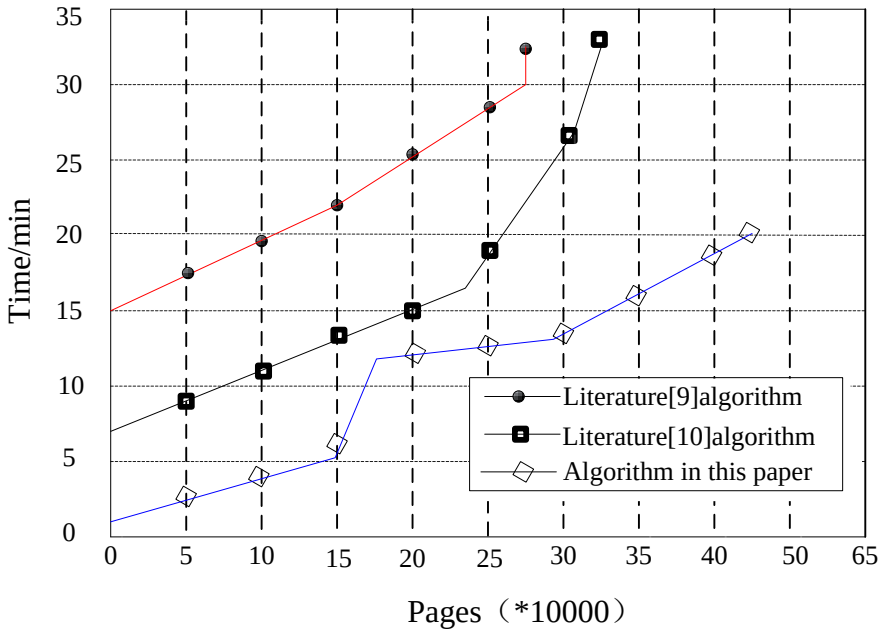

图 3. The Comparison of Three Algorithms Based on Different Page View

6. Conclusion

To conclude, the user search behavior based on Hadoop is able to help users acquire information via query log and data digging technique, which can be used in massive file processing. In addition, the analysis on 58 dataing data as well as the Hadoop distributed computing framework is beneficial to effectively make up for the demerits of computing model based on cloud computing, which is also of guiding and practical significance.

References

[1] L. Page, S. Brin and R. Motwani, "The Pagerank Citation Ranking;Bringing Order to the Web", [R].Techical Report,Standford Digital Library Technologies Project, **(2011)**.
[2] J. M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", [J].Journal of the ACM, vol. 46, no. 5, **(2012)**, pp. 604-632.
[3] S. Chakrabarti, B. Dom and P. Raghavan, "Automatic Resource List Compilation by Analyzing Hyperlinked Resource List Compilation by Analyzing Hyperlink Structure and Assocaitaed Text[EB/OL]., http://citeseer.ist.psu.edu/chakrabarti98automatic.htm,
[4] PoweredBy-HadoopWiki[EB/OL].http://wiki,apache.org/hadoop/PoweredBy,
[5] D Borthakur, "HDFS Architecture" [EB/OL]. http://hadoop.apache.org/common/docs/current/hdfs_design, **(2013)** November 17.
[6] G.-J. Mao and L.J. Duan, "The principle and algorithm of data mining", Beijingtsinghua university press, **(2009)**.
[7] http://labs.58.com
[8] L. Jian, L. Yi-qun and M. S.-Ping, "Analysis into the Relationship Between Search Engine User Behavior and User Satisfaction Evaluation[J]. f Chinese Information Processing, **(2014)**, vol. 28, no. 1, pp. 73-79.
[9] C.Chen, Zhan Yin-wei and Li Ying, "Page Rank parallel algorithm based on Journal of Computer Applications", vol. 35, no. 1, **(2015)**, pp. 48-52.
[10] C.Shan-Shan and W. Chong, "Improved PageRank Algorithm Based on Links and User Feedback", [J],Computer Science, vol. 41, no. 12, **(2014)**, pp. 179-182

Fund Project:

The work was sponsored by National Natural Science Foundation of China Grant No.61370220, Program for Innovative Research Team (in Science and Technology) in University of Henan Province Grant No.15IRTSTHN010，Program for Henan Province Science and Technology Grant No.142102210425，Key Program for Basic Research of The Education Department of Henan Province Grant No.13A520240 and No.14A520048，Training Foundation for Scientific Innovation Ability of Henan University of Science and Technology Grand No.2013ZCX022.

Authors

Junyi Du（1983.05-），male，master. He research interests are mainly focused on cloud computing; multimedia.

Zhiyong Zhang（1975-），male，doctor, Professor. He research interests are mainly focused on Digital rights management, trusted computing and access control, multimedia networking.

Changwei Zhao(1971-), male，doctor, Associate Professor. He research interests are mainly focused on Digital rights management, trusted computing and access control, multimedia networking.

