International Journal of Database Theory and Application
Vol.9, No.3 (2016), pp.81-86
http://dx.doi.org/10.14257/ijdta.2016.9.3.09

Mining Analysis on User Search Behavior Based on Hadoop

Jie Fang

(Zhejiang Industry Polytechnic College Shaoxing Zhejiang 312000 China)
shaoxingfj@sina.com

Abstract

User webpage search behavior is the hot research topic on information search at present. In view of the deficiency in Cloud parallel computing search, the thesis has proposed a method for mass user webpagesite search behavior based on Hadoop. On the basis of PageRank algorithm, the method aims to improve the algorithm efficiency of the modified PageRank and user webpage search efficiency by increasing user influence index, time vectors and webpage relevancy into the algorithm. The algorithm has been proved to achieve excellent effects by log inquiry and analysis in Youku laboratory, and will offer some guidance for user behavior analysis in Cloud computation.

Keywords*: Hadoop user search; behavior analysis; mass log; PageRank algorithm*

1. Introduction

With the advent of Cloud computation concept, there has been increasing information to be shared and communicated by the internet, and network information is expanding at exponential class rate. Under such circumstance, the search engine under Cloud algorithm has rapidly developed to be an important information acquisition method. currently now, PageRank [1] has been raised by American Stanford University and HITS technology by IBM [2-3]; then, some new usable information has also sprung up in Webpage query mode, which can be reflective of some user search behaviors from another perspective and thus helpful for Cloud computation server to analyze user information search quality and behaviors. Large search engines can achieve hundreds of millions of page views every day and mass files of user query log records. In face of the new characteristics of the mass logs, the traditional data storage and computing method can no longer be able to tackle search engine user behavior analysis. Therefore, on such basis, the thesis made a search for mass webpage information under Hadoop framework, and added user influence index, time vector and webpage relevancy into PageRank algorithm to improve its computing efficiency and user search efficiency. The superiority of the analysis in the thesis has been proved by simulation experiment.

2. Introduction of Hadoop Framework

Hadoop framework is an open-source distributed computational framework and widely used in some large enterprises [4]. It has achieved MapReduce parallel programming model, provided a distributed file system HDFS and basic storage function for Cloud distributed computation. There is a JobTracker in Hadoop mainly used for scheduling and managing other TaskTrackers. JobTracker mainly runs on any of the computers in computer group, while TaskTracker is mainly to implement tasks and must operate at Cloud computing ports. HDFS [5] uses Master/Slave frame and is formed by NameNode management node and several data nodes, mainly functioning to segment the large file into several blocks and separately storage them on different DataNode, among each block

ISSN: 2005-4270 IJDTA

will duplicate the data on different data nodes to make it fault-tolerant. Hive [6] is a basic frame based on Hadoop data warehouse. It offers a host of translation optimization service for from HiveQL to MapReduce and thus makes sure the high efficiency of MapReduce. In practice use, Hive can process TB and even PB data in great efficiency.

3. Work Related to User Search

3.1 The Weights of the Enquiring Indexes

There is a mass of Cloud computation resource. When the search results of the search engine are acquired from Cloud computation, it requires to judging which result is needed and how to identify the relevancy between webpage documents and user's query contents. First, the search engine segments the query keywords, deletes the stopwords unrelated to the webpage theme, and convert webpage documents to unit entries. Then, it calculates the similarity and matching degree between webpage document and user request in keyword weight calculation approach, and at the same time, offers different weights to the keywords after segmentation. In the end, according to the calculated correlation value, submit to the user the visiting contents in big-to-small order.

The thesis has acquired such as result as below, by use of vector space modal and the item sets respectively grouped by the uniterms of word weights in the document and the weights of query sequence key words. Among, $X_{i,j}$ means the weight of a word in the webpage document after being segmented by search engine; while $Y_{i,j}$ refers to the weight of a word in query sequence.

$$f(d_j, t) = \frac{\vec{d_j} \cdot \vec{t}}{|d_j| \times |t|} = \frac{\sum_{i=1}^{t} X_{i,j} \times Y_{i,j}}{\sqrt{\sum_{i=1}^{t} X_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} Y_{i,j}^2}} \quad (1)$$

$$X_{i,j} = \frac{fre_{i,j}}{\max_i fre_{i,j}} \times \log \frac{N}{n_i} \quad (2)$$

$\vec{d_j} = (X_{1,m}, X_{2,m}, \cdots X_{i,m})$, $\vec{t} = (X_{1,n}, X_{2,n}, \cdots X_{i,n})$. In this formula, $fre_{i,j}$ denotes the frequency of the key word in the sequence of i appearing in webpage in the sequence of j; $\max_i fre_{i,j}$ denotes the total frequency; $\log \frac{N}{n_i}$ represents inverse document frequency index; N represents the quantity of the resource on all the web pages; and n_i means the total quantity of the web pages where the key word in the sequence i appears.

3.2 The Establishment of Search Model

The thesis has established a search model under Cloud conditions. The whole model is mainly based on simulating a small webpage, where the webpage link relations are built on the basis of internet structure. The user uses formal search terms and involves no falsification and other cheating behaviors, therefore, the webpage search result is based on the user's degree of interest, and the data in search log are accurate.

Definition1: assume the number of webpage is N, use $A_1, A_2 \cdots A_n$ to represent web pages in the number of n, the matrix $A = a_{ij}$ can be used to describe the connected relation between the pages.

Use diagraph $G = (V, E)$ to express the connected relation between the above web pages; assume $V = \{A_1, A_2, \cdots A_n\}$ to represent the group of the node, and $E = \{A_i \rightarrow A_j\}$ the connective relation between the web pages.

Definition 2: for the webpage x, in case the web page is clicked once with the time of

$[0,t]$, it can be assumed 1, that's to say $click(x)=1$, or else, it shall be assumed 0.

Definition 3: if there are several search results for the search behavior q within the time of $[0,t]$, represent these web pages as $x_1, x_2, \cdots x_n$. Assume the web page x is submitted m times in a day, the click rate for the webpage x are be acquired by the formula $U_p = \sum_{i=1}^{m} click(x,q)$, among U_p means user query relevancy vector quantity, $click(x,q)$ means the frequency of webpage x is clicked within the time $[0,t]$ in the sequence of i.

3.3 The Factor Requiring the Consideration for the Model

The model in the thesis mainly aims to analyze webpage click rate and user behavior based on PageRank algorithm and user's visiting frequency and preference for the webpage. At the same time, it takes into consideration the ratio of user's behavior to the webpage, calculate the comprehensive weights by setting the order and gives user search result. But the following elements shall be offered in the process:

(1) User behavior influence index. Within a section of time, for the search behavior q, user's click rate C represents that the clicking possibility is largely influenced by the position of the results in the return webpage, due to the neglect for the URL information in return results while clicking. Therefore, there exists a great relevancy between the webpage and the search. If the information lies at the bottom of the webpage list, it is less likely to be found and clicked by the user. Considering this, formula (3) can be used to make up for the deficiency.

$$U_q = \sum_{i=1}^{n} c(pos(A,q)) * click(A,q) \quad (3)$$

(2) The time for user's consideration. When user discovers related or similar contents in the process of search, the user will spend some time browsing them. However, the length of the time does not represent their degree of satisfaction with the search results, because they may copy or paste the information on the page. Therefore, any length of the browsing time for the search result group decides how the user satisfies with the result. On such basis, formula (4) can be used to describe the weight of the user's browsing time.

Among, t_i represents the time that user spends browsing the webpage A aimed for the query word set q.

$$Time(A,q) = \frac{t_i}{\sum_{i=1}^{n} t_i} \quad (4)$$

(3) The relevancy of the web pages

In the actual user search results, there may be a strong similarity between webpage i and webpage j in terms of contents, but they may be arranged in an order by the search engine. The result of the first query may be the final result, so the equilibrium factor cannot make sure for the webpage listed at the back. Therefore, assume there appear N iterations within the time of $[0,t]$, the clicked web pages form a webpage matrix $C_{N\times N}$, among, $C_{i,j}$ means the times of the webpage i and webpage j to be clicked. If $C_{i,j}$ and $C_{j,k}$ are both above 0, there suggests a relation between web pages i, j, and k as below:

$$K(A,T_i) = K(ID_A, ID_{T_i}) \quad (5)$$

4. The Modified User Webpage Search Pagerank Algorithm

PageRank algorithm is the method for Google to label the grade and importance of the web pages and the sole criterion to judge the quality of a web page. It uplifts the ranks of the web pages with higher "grade/importance in the search results, and thus improve the

relevancy and quality of the search results. While as the users are visiting the web pages, the web pages can be selected by the similarity of the links under the same theme. Assume web page X has 5 links directing to the web pages A, B, C, D and E5, the theme similarity will be respectively 0.15, 0.24, 0.32, 0.42 and 0.65. Therefore, while choosing website links, there is the biggest probability to choose webpage E. PageRank needs to consider the elements in Section 2.3 in the modified webpage algorithm, not only limited to the direct links between the web pages, and also including the elements of implicit brief introduction. On such basis, to modify the traditional PageRank formula, PR calculation for webpage X can be expressed as below:

$$PR(X) = \frac{1-d}{N} + d * \sum_{(X,T_i)\in E} \left(\frac{PR(T_i)}{\sum_{k=1}^{M} click(T_i, X)}\right) * \tag{6}$$

$$(\delta_1 f(X,T_i) + \delta_2 T(X,q) + \delta_3 K(X,T_i))$$

In the formula, δ_1, δ_2, δ_3 respectively represent user influence index, time vector quantity and website relevancy, the formulas $\delta_1 + \delta_2 + \delta_3 = 1$ and $d * (\delta_1 f(X,T_i) + \delta_2 T(X,q) + \delta_3 K(X,T_i)) \le 1$ can better guarantee algorithm convergence. E means the total quantity of the web pages in the internet, d means damping factor, $click(T_i, X)$ refers to the times of web pages T_i and X being clicked at the same time. The bigger figure of $click(T_i, X)$ suggests a greater relevancy of the two web pages. The algorithm has fully considered user influence index, time vector quantity and website relevancy in computing the weights of the web pages.

Algorithm flow

Step 1: user conducts the object web search to get certain number of web pages;

Step 2: based on PageRank algorithm, introduce and analyze user influence index, time vector quantity and website relevancy in turns;

Step 3: analyze the web pages from user influence index;

Step 4: analyze the time taken by the web pages from time vector quantity;

Step 5: select the analysis on web pages relevancy;

Step 6: submit the results of step 3 and 5 to PageRank unit and compute the result;

Step 7: report the result to the user.

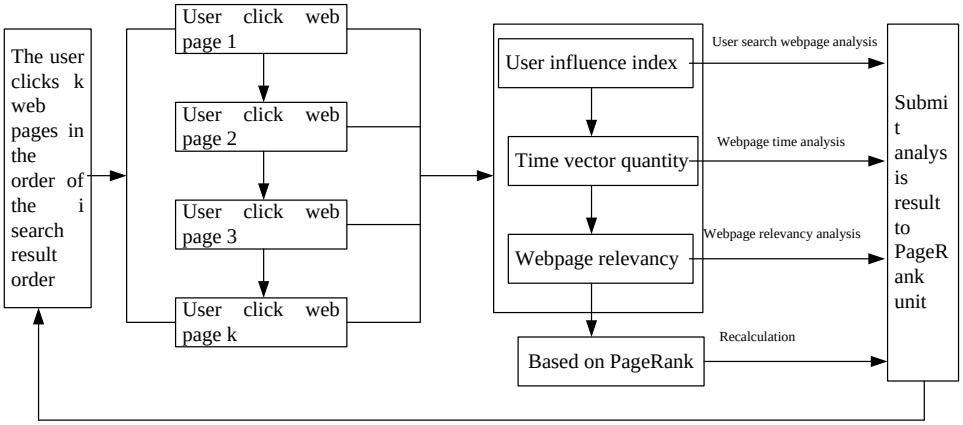

Figure 2. Algorithm Flow

5. Experiment Result Analysis

The thesis uses 5 PCs (respectively labeled from PC1 to PC5) to build a Hadoop distributed computation platform, among PC1 operates Jobtracker as the Master, and the ret four operate Tasktarcker. The configurations of the PCs are as below: CPU: Inter Core2.2Ghz,4GDDR3,500G hardware; software environment: Ubuntu12,Hadoop 0.20.3,OpenSSH。

5.1. Hot List Analysis

By data collection and analysis from Youku, the thesis has acquired that the visit volume of the top 1000 web pages account for 75.36% of the total, suggesting that the search engine are dealing with many repeat requests every day. As shown in Figure3.

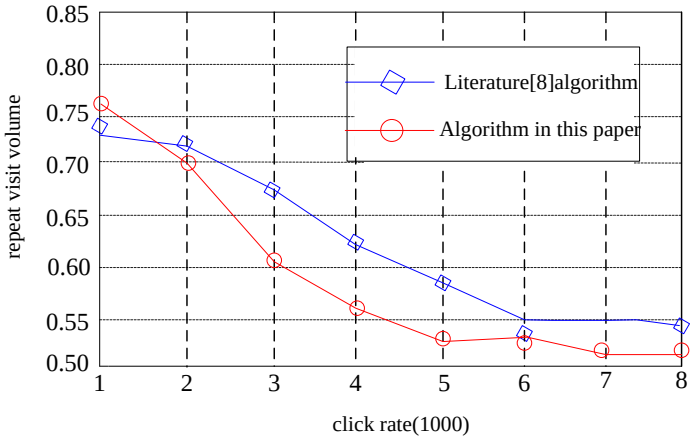

Figure 3. Hot List Analysis

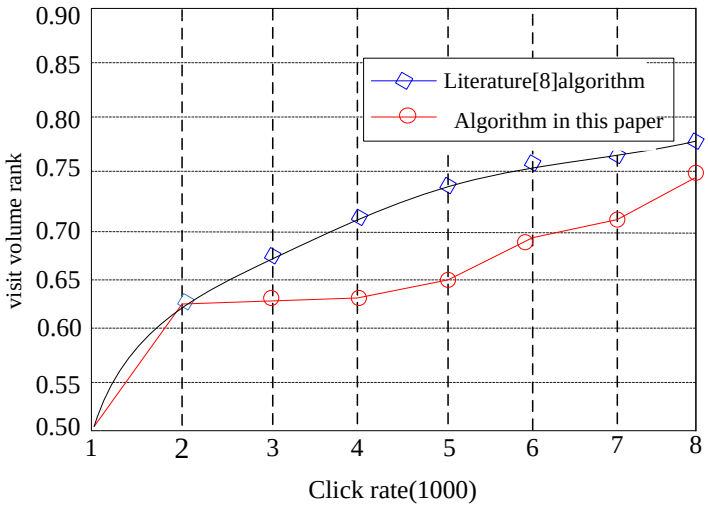

Figure 4. User Click Rate and URL

5.2. User Click Rate and URL

From the statistical result of the data group in this experience, as shown in Figure 4, it can be noticed that the click rate of the top 1000 URL can account for 72.46% of the total, not a big gap from 75% acquired in reference [8], suggestive of the effectiveness and correctness of platform algorithm in the thesis.

5.3. Distributed Platform Efficiency Analysis

The thesis made a test and analysis on the efficiency of data query theme list on the platform based on three different search data scales according to the search logs—basic sample data (1Mb), the data for the whole day (5Gb) and the whole week (40Gb).

Table 1. Data Process Duration

	1 node（s）	2 nodes（s）	3 nodes（s）	4 nodes（ss）
Basic sample data	12.829	17.417	21.821	25.813
Data for the whole day	57.327	52.216	45.741	36.816
Data for the whole week	180.716	162.721	142.715	99.218

From Table 1, it can be discovered that the platform spends more time dealing with smaller scale of data, and obviously less time for the data at a larger scale. That suggests that the modified PageTank algorithm under Hadoop platform is more suitable to deal with large-scale data.

6. Conclusion

User search behavior analysis based on Hadoop platform can improve information acquisition through query log and data mining technology which can be applied into mass file process. The analysis on the data in Youku data base and the application of Hadoop distributed computational framework into search engine analysis can effectively solve the deficiency of Cloud parallel computational model. Judging from this perspective, the research and analysis in the thesis has an excellent guidance and practice significance.

References

[1] L. Page, S. Brin and R. Motwani, “The Pagerank Citation Ranking; Bringing Order to the Web”, Technical Report, Standford Digital Library Technologies Project, **(2011)**.
[2] J. M. Kleinberg, “Authoritative Sources in a Hyperlinked Environment”, 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 Associated Text [EB/OL]”, http://citeseer.ist.psu.edu/chakrabarti98automatic.htm, **(2013)**.
[4] Powered By-Hadoop Wiki[EB/OL].[2013-11-17].http://wiki,apache.org/hadoop/PoweredBy.
[5] D. Borthakur, “HDFS Architecture [EB/OL]”, http://hadoop.apache.org/common/docs/current/hdfs_design, **(2012)**.
[6] M. G. Jun and D. L. Juan, “The principle and algorithm of data mining”, Beijing Tsinghua University press, **(2009)**.
[7] Youku laboratory [EB/OL].[2009-11-17].http://labs.youku.com
[8] L. Jian, L. Y. Qun and M. S. Ping, “Analysis into the Relationship Between Search Engine User Behavior and User Satisfaction Evaluation”, Journal of Chinese Information Processing, vol. 28, no. 1, **(2014)**, pp. 73-79.
[9] C. Chen, Z. Y. Wei and L. Ying, “Page Rank parallel algorithm”, based on Journal of Computer Applications, vol. 35, no. 1, **(2015)**, pp. 48-52.
[10] C. S. Shan and W. Chong, “Improved PageRank Algorithm Based on Links and User Feedback”, Computer Science, vol. 41, no. 12, **(2014)**, pp. 179-182.

Author

Jie Fang, (1981.11-), male, master, lecturer, research direction: information security, cloud computing and Data mining.

