Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(5):1420-1425

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Large-scale User Behavior AnalysisBased on Cloud Computing

Dao Jiang[1] and Zhao Yu[2]

[1]*Department of Electronic and Information Engineering, Shunde Polytechnic, Guangdong*
[2] *school of computer science and technology， ZhouKou Normal University,Henan,China*

ABSTRACT

The user behavior analysis is the effective way to enhance the user viscosity, maintain traffic flow among the large Internet network. This paper proposes the idea of user behavior analysis engine. It will combine the user behavior of static analysis with real-time acquisition of Web log. Using the data mining model based on cloud computing technology analyzes the contextual information that is obtained by accessing the page in Real-time, processing and storage at the same time. Experiments show that the system can improve the effect and efficiency of the user behavior analysis.

Keywords:user behavior; cloud computing; behavior analysis engine

---

INTRODUCTION

Analysisof user behavior refers to the site access to basic data, through the study on statistical analysis of the relevant data, found that the laws of the useri to access the website, to allow enterprises to more detailed, clear understanding of user behavior, to find out the existing problem of business website, marketing channels,marketing environment, to help enterprises to obtain high conversion page, the enterprise marketing is more accurate, efficient, improve business conversion rate, so as to enhance the enterprise income[2].

LARGE-SCALE USER BEHAVIOR ANALYSIS ENGINE BASED ON CLOUD COMPUTING

ENGINE ARCHITECTURE

In this study, "user behavior analysis engine" is defined as: according to certain strategy, respectively, to obtain user dynamic behavior and behavior, and summarize, analysis and reasoning, the system user behavior habit and characteristics.

"The scale of user behavior based on cloud computing analysis engine" for user behavior information, the use of cloud technology, storage and analysis of its, efficiently find, mining user behavior, its structure as shown in figure 1. It from the client to obtain real-time dynamic behavior of context information, asynchronous upload to the server to save; trigger server processing module pre-processing, aggregation analysis; access to Web log from the server, filtering, denoising and mining, and according to the point in time restore user history context information; at the same time, the dynamic behavior, treatment history stored in HBase database which:

◆Static behavior analysis: mainly completes the Web log mining, and according to the time point the user to restore the historical context of user behavior, filtering, denoising, fusion operation, save the processed results, a user behavior database.

◆Dynamic behavior analysis: Based on the Markoff model, the dynamic behavior of reasoning, to collect, analysis

of user behavior and characteristics.

◆The dynamic behavior of acquisition and preprocessing: access to information users real-time operation page from the client's behavior, and pre processing, storing the results in HBase database. Including data cleaning, transformation, reduction, delete the useless content, check the information completeness and consistency.

◆User behavior information storage based on HBase: storage of user behavior information from the client and the server, the dynamic and static user behavior data, results and analysis.

◆The polymerization behavior of dynamic user: dynamic user behavior data filtering, integration, excluding those correct but invalid information user behavior but invalid.

RESEARCH ON KEY TECHNOLOGY

THE DYNAMIC BEHAVIOR OF THE USER ACQUISITION AND PREPROCESSING

The dynamic behavior of user refers to the user (including login and not logged in two cases. The user login, registered account user identification by acquiring ID; for users who are not logged in, record visit their website SessionID logo) occurred in accessing the page a moment of behavior, the behavior includes the occurrence time, the page (contains the page title and page URL), related to the operation and behavior subject, real-time capture them and carry on the effective analysis, has an important significance for understanding user behavior characteristics.

From the client gets information including the user dynamic behavior and context information, the context, including: behavior occurrence time; the current user ID or SessionID; the current page title; the address URL of the current page; the current user search conditions; access to the same page number; page retention time; do you want to save the page; printed the page whether or not ;whether add to favorites; copy or cut the page content and so on ; the environmental context information includes: the client machine configuration, current network condition, the server working condition etc..

Because the user behavior data, acquired it, by using MapReduce model in cloud environment, including filtering, eliminate duplication, delete the useless content, check the information completeness and consistency. As the following methods used:

1) Data cleaning: removal of the incomplete data, delete duplicate data, delete access to pictures, delete pages of animation, the user behavior analysis of useless data[8].

2) Data conversion: the pages print, collection, preservation, download operation, in the acquisition, will be converted into the corresponding data format in the database.

3) Data reduction: the user behavior data in large quantity, to standardize the data quantity, reduce the very necessary, but must maintain the integrity of the data.

STUDY ON THE VECTOR SPACE MODEL OF RETRIEVAL BASED ON USER BEHAVIOR

Vector space model (VSM:Vector Space Model) proposed by Salton[12]and others in twentieth Century 70 years, it is the basic idea of each text and query contains some features independent properties reveal its content, and each feature attributes can be regarded as a dimension vector space, then the text can be expressed as a collection of these attributes, ignoring the complex relationship between paragraphs, sentences and words in the text structure. At the same time, given the feature weight vocabulary certain (weight), anti should vocabulary in the importance and the value of the contents of the file identification, this value is called the indexing vocabulary "significant value (Term Significances)" or "weight", by the lexical statistics calculate the document and to, such as: the feature words appear frequency (Term frequency, TF). Vector of each file is in fact all the document feature through a combination of computing, called "the document feature item vocabulary matrix". And then all of the document vector based on specific computing methods of similarity measure between each other.

Vector space retrieval model can be described as $I=(D,T,Q,F,R)$ Among them: $D=\{d_1,d_2,...d_n\}$ As a collection of text, n text collection number;

$T=\{t_1,t_2,...t_n\}$ Set as a feature, m feature of all. A text m feature indexing can be represented as a vector space $d_i=\{w_{i1},w_{i2},...w_{im}\},i=1,2,...,n$, w_{ij} is characteristic t_j for the text d_i of the weight, if the weight value w_{ij}

is 0, indicating t_j that it is not appeared in d_i

$Q = \{q_1, q_2, ...q_m\}$ for the query set, a query q_r can be represented by vectors $q_r = \{q_{r1}, q_{r2}, ...q_{rm}\}$, q_{rj} is a characteristic to t_j the query q_r weights, if the weight value q_r is 0, indicating that t_j is not appeared in q_r.

Further definition:

Frequency tf_{ij}: t_j is the feature for text d_i appear in the frequency;

Inverse document frequency word idf_i (inverse document frequency): the word in the quantitative distribution of document collection, the calculation $\log(N / n_k + 0.5)$ is usually, where N is the total number of document centralized, n represents a number of documents containing K, called the document frequency of the term.

The normalization factor: in order to reduce the inhibitory effect of high frequency characteristics of individual word on other low-frequency feature words, the standardization of components.

Based on the above three factors to term weighting formula:

$$w_{ik} = \frac{tf_{ik} \log(N / n_k + 0.5)}{\sqrt{\sum_{k=1}^{n} (tf_{ik})^2 \times [\log(N / n_k + 0.5)]^2}} \quad (1).$$

The similarity between the text and the query can be used to measure the distance between two vectors. There are many kinds of calculating method of similarity, commonly used methods of inner product, Dice coefficient, Jaccard coefficient and cosine coefficient, usually uses the cosine coefficient method, namely the cosine of the angle between two vectors to represent the similarity between the text and the query $Sim(d_i, q_j)$, see equation (3.2).

Cosine similarity calculation method is a normalization, the angle between the two vectors of the smaller, the greater the degree of correlation between documents, correspondence $\cos$ is higher. Two vector included angle cosine is equivalent to their standard vector inner product unit length, itreflects the similarity term component two vector of relative distribution.

$$Sim(d_i, q_j) = \cos\theta = \frac{\sum_{k=1}^{n} w_{ik} \times w_{jk}}{\sqrt{(\sum_{k=1}^{n} w_{ik}^2) \times (\sum_{k=1}^{n} w_{jk}^2)}} \quad (2)$$

ALGORITHM DESIGN

Input: user access to key words each time the user query log in;

Output: the similarity with the query keywords vector existing values in the database is not paper and similarity of 0, and according to the similarity value from big to small order;

1)extracted from each Webpage keywords as the feature word, and these feature words andkeywords query every time the user binding, rearrangement and according to a lexicographic order, combined together to form a standard feature set of words;

For example, there are Webpage document set (NDoc1, NDoc2, NDoc3, NDoc4), all the feature words together for (W1, W2, W3, W4, W5, W6), and when the query words (WQ1, WQ2), where wq1= W4, the standard set of words as features (W1, W2, W3, W4, W5, W6, WQ2), Webpage document feature item vocabulary matrix in table 1.

Tab.1 Webpage document vector space model

	w1	W2	W3	W4	W5	W6	wq2
Q1	0	0	0	0	0	0	1
NDoc1	14	21	33	0	0	0	0
NDoc2	0	11	15	0	0	22	0
NDoc3	8	0	0	14	15	17	0
NDoc4	0	8	9	12	0	15	0

2) t_j is calculated for each term tf_{ij} appears in the Webpage text d_i frequency; calculation of $\log(N/n_k+0.5)$ words formula of inverse document frequency idf_i; then the formula with weight (1) to calculate the weight of each feature words each Webpage of document vectors, forming the Webpage document vector in a vector space;

3)to calculate the similarity of each document vector and the query vector Webpage between the cosine coefficient method, see equation (2). The interception of similarity values greater than 0.2000 articles, and from high to low return results;

Study on purchasing behavior of modified matrix vector association rules
method based on userAssociation rulemining is used to find correlation between the attributes of databases. Association rules is the initial motive of shopping basket analysis problem, the goal is to find the different commodities of association rules mining in transaction database, the relationship between the useful knowledge description data item value. These knowledge characterizes customer buying behavior and mode, use these rules, can effectively guide the scientific arrangement and design business purchase goods shelves. The form of association rule is a rule is, "to buy milk and bread customers, 90% of people bought butter", namely "(milk, bread)→ butter" issue.

Let be the $I=\{i_1,i_2,...,i_m\}$ set of items. A related task data D is a collection of database transactions, where each transaction is a set ofT $T\subseteq I$, so. Each transaction is an identifier, called TID. Let A be a set of transaction, $A\subseteq T$ T contains A if and only if. Association rules are shaped implication $A\Rightarrow B$, such as one $A\subset I, B\subset I$, and $A\cap B=\varnothing$. Rule $A\Rightarrow B$ D in the transaction set, with the support of S, where s is the D transaction contains the percentage of A ∪ B, namely P (A, B). Rule $A\Rightarrow B$ D C has confidence in the transaction set, where C is contained in the D A transaction also includes a percentage of B, namely P (B |A).

$$Support(A\Rightarrow B)=P(A\cup B) \quad (3)$$
$$Confidence(A\Rightarrow B)=P(B\mid A) \quad (4)$$

The support and confidence are two important concept description of association rules, the former for statistical measure of the importance of association rules in the data, said the rules which is used to measure frequency; credible degree of association rules, said the strength of the rules. In general, only the support and confidence of association rules are high may be only the interesting rules, useful. Association rules mining is mainly realized by the two steps:

Step one, according to the minimum support degree to find the database in D all the frequent item sets.
Step two, according to the frequent item sets and minimum confidence generated Association rules.
Task one step is to quickly and efficiently find all frequent itemsets in D, is the central problem of the association rule mining algorithm of association rules mining, is a measure of the standard; step two, relatively easy to achieve, so now all association rules mining algorithm is designed for the first step forward.

APPLICATION EXAMPLES
Using the above research results, relying on the research project, the analysis engine system user behavior based on cloud computing is a design and development.

The analysis engine platform running on the Ubunt12.10 user behavior, the software mainly includes: jdk-1.7.0_11, Jena-2.6.4, Myeclipse-8.0, Hive-0.10.0, HBase-0.94.4, Hadoop-1.0.4, Tomcat-6.0, Jquery-1.6, Spring-3.0, Struts2-2.2.1, the browser is above IE8.0. The context aware behavior analysis as an example, system operation is as follows:

The system calculates the Markoff matrix intermediate important, as shown in figure 2.

chengli	www...	www...	www...	www...	www...	www...	www...	www...	www...	www...	www...
www...	0.196	0.176	0.078	0.176	0.196	0.0	0.176	0.0	0.0	0.0	0.0
www...	0.0	0.0	0.0	0.0	0.0	0.176	0.0	0.0	0.274	0.274	0.274
www...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0

lucky	www...	www...	www...	www...	www...	www...	www...	www...	www...	www...	www...	www...	www...
www...	0.258	0.258	0.0	0.0	0.0	0.241	0.0	0.0	0.241	0.0	0.0	0.0	0.0
www...	0.0	0.0	0.340	0.0	0.0	0.0	0.0	0.340	0.0	0.0	0.0	0.0	0.318
www...	0.0	0.0	0.0	0.378	0.378	0.0	0.0	0.243	0.0	0.0	0.0	0.0	0.0
www...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
www...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.5	0.0	0.0
www...	0.0	0.0	0.0	0.0	0.0	0.0	0.517	0.0	0.0	0.0	0.0	0.482	0.0

Fig.2 Mark off matrix

Select a user's current URL, system using Markov model and collaborative filtering, given URL probably next, as shown in Figure 3, figure 4.

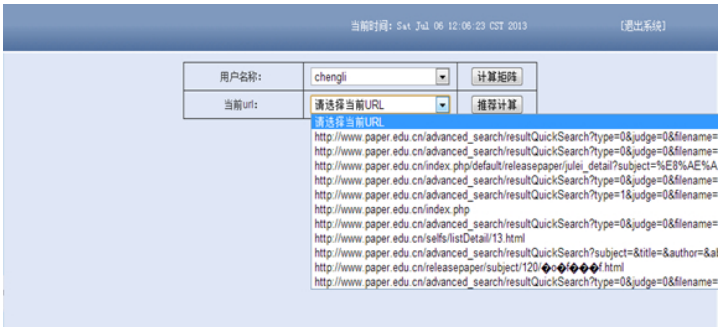

Fig.3 to select the current URL

Fig.4 Mark off results

CONCLUSION

User behavior analysis engine through the WEB log, the user dynamic behavior and the context information, multi-channel, all-round, three-dimensional access to large-scale user behavior information increasing, the use of cloud environment MapReduce parallel computing model, HBase cloud storage capacity, and uses the relevant data mining algorithms, static analysis, dynamic monitoring of user behavior characteristics and synthesis reasoning, behavior of rich information content, comprehensive, complete, and can efficiently for the user to push their interested information and provide the basis for the site structure adjustment.

Because of the time and energy constraints, this work is currently only integrates three cloud platform of the data mining model, looking for more scene data mining model, and transformation, in the cloud platform integrated, make the system more universal, universal, is the next step of work to do.

REFERENCE

[1] AbrahamSiberschatz, Peter B. Galvin,Greg Gagne. Operating System Concepts(8thed.). John Wiley & Sons (Asia). **2010**. Page: 194.
[2] Ian Foster, Yong Zhao, IoanRaicu,etal.Cloud Computing and Grid Computing 360-Degree Compared. Grid

Computing Environments Workshop, GCE '08,**2008**.
[3] Y. Yang, K. Liu, J. J Chen, et al. An algorithm in SwinDeW-C for scheduling transaction-intensive cost-constrained cloud workflows. In: Proceedings of the IEEE 4thInternational Conference on eScience, IEEE Press. Dec **2008**.Pages:374-375.
[4] Luqun Li. An Optimistic Differentiated Service Job Scheduling System for Cloud Computing Service Users and Providers. In: Proceedings of the 3rdInternational Conference on Multimedia and Ubiquitous Engineering, **2009**.
[5] QuXilong, HaoZhongxiao,BaiLinfeng. *International Journal of Advancements in Computing Technology*, vol. 3, no. 10. **2011**. Pages: 99-106.