Advances in Computer and Communications, 2023, 4(1), 46-50
https://www.hillpublisher.com/journals/acc/
ISSN Online: 2767-2875

User Characteristics Analysis Based on Web Log Mining

Cui Zhang[1,*], **Xiaofei Li**[2]

[1]School of Electrical and Automation Engineering, Liaoning Institute of Science and Technology, Benxi, Liaoning, China.
[2]MCC Coke Resistance (Dalian) Engineering Technology Co., Ltd, Dalian, Liaoning, China.

How to cite this paper: Cui Zhang, Xiaofei Li. (2023) User Characteristics Analysis Based on Web Log Mining. *Advances in Computer and Communication*, 4(1), 46-50.
DOI: 10.26855/acc.2023.02.007

Received: February 15, 2023
Accepted: March 10, 2023
Published: April 7, 2023

***Corresponding author:** Cui Zhang, School of Electrical and Automation Engineering, Liaoning Institute of Science and Technology, Benxi, Liaoning, China.

Abstract

In order to solve the contradiction between massive network information and limited learning needs, personalized recommendation service becomes the hotspot in research area. User characteristics analysis is the key point in personalized recommendation service. Based on the massive web access log in the web server, this research gradually puts forward the steps of user characteristics analysis which including data pretreatment, user feature extraction and user clustering. This paper focuses on the rules of user recognition, definition of user feature, user feature extraction algorithm and user group clustering. Finally, take access log files of a web server as sample, simulation experiments have been made to prove the thought put forward by this context. Improvement has been taken to the push service repository on the basis of the experimental results, which achieved good practical results. Behavior characteristics and laws of visitors access based on log analysis may lay the foundation of recommender service for resources platform.

Keywords

Web Log Mining, User Characteristics, Characteristics Analysis, Data Mining

1. Introduction

With the rapid development of network technology, the Internet has become an important source of information and knowledge. Information overload interferes with the visitor to the required information selection. Contradiction between vast information and limited need, forced us to seek a fast, accurate ways to find data needed in the ocean of information. Driven by this demand, personalized recommendation services have been emerged as the times require. Personalized service is a kind of targeted service mode, which will collect and collate resources, then classify the resources through various channels. Recommendations will be made based on the research, in order to meet the needs of visitors.

Visitor characteristics analysis is one of key content for the personalized recommendation service [1]. Now the research on visitor characteristics analysis or focus on the algorithm of [2, 3], or focus on a particular application field [4], Such as search behavior analysis [5, 6]. But there is few research which focused on resource personalized service.

The servers will store a lot of web page access parameters in certain log files while providing the web resources service. These access log files recorded massive Internet information. In this study, the web log files of a university teaching resources server are used as the sample data. By constructing a visitor characteristics analysis system, extract the online behavior of related data from the web log files, then mining the data to get the user characteristics of visitor, finally clustering to get some visitors group, which can be the hint of personalized recommendation service. In this study, a sample personalized recommendation service has been built according to the method which will be introduced as follow [7]. A resources server will be taken as a sample. Based on the data mining results of the web log files in the resources server, changes will be made to recommendation system of the resources server, which will provide better resources service for users of the system.

2. Process Design of Visitors Characteristics Analysis

By analyzing the Web access log content composition, the steps of user characteristic analysis can be determined, which goes as follow.

There are four steps in user characteristics analysis process. They are log pretreatment, webpage pretreatment, study characteristics extraction and study group clustering.

Step 1: Extracting and cleaning the web log, identifying all the visitors. Firstly, gathering web log files from server. Because there is a lot of irrelevant information contains in collected log files, so log files cleaning must be done before the user characteristics analysis. The content of the web log file will be organized into structured data files according to the requirement of analysis [13]. Then based on the structured data files identifying the users according to some rules, such as according to different IP address etc.

Step 2: Use crawl tools to gather the webpage content according to the urls recorded in web log files. There is a lot of irrelevant information in the original content which crawl out by tools, such as advertisement, pictures etc. So Before further analysis has been mad, it should be made clean. Then word segmentation should be made, which will be used in user characteristics analysis algorithm.

Step 3: Taking the definition of TF-IDF as reference, release the definition of L-TF-IDF according to the requirement of this research. Then design the user characteristics analysis algorithm according to the L-TF-IDF definition.

Step 4: Classification and clustering the visitors group according to K-Means in Mahout. The results of user clustering will be the important reference of personalized recommendation service.

The user characteristics and user group which got from step 3 and step 4 would provide strong support for the personalized recommendation service. These two results are the core contents of this research.

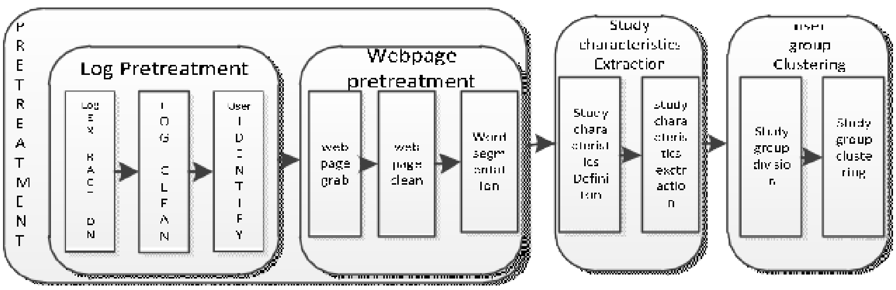

Figure 1. Diagram of the user characteristics analysis process.

According to the design mentioned above, the flow of the user's characteristics analyzing which based on the web log analysis is shown in Figure 1. There are three components in user characteristics analysis based on web log mining [14]. The first is pretreatment modular, then goes the study characteristics extraction modular, finally is the classification and clustering the visitors group modular.

As shown in Figure 1, there are two sub modular in pretreatment stage. They are log pretreatment and webpage pretreatment. Log pretreatment is made up of log extraction, log clean and user identify. Web pretreatment is made up of web page grab, web page clean and web segmentation.

The study characteristics extraction modular is made up of study characteristics definition modular and study characteristics extraction modular.

The user group clustering modular is made up of study group division and study group clustering.

Then introduce will be made to the data pretreatment modular, user characteristics extraction modular and user group clustering modular in order.

3. User Characteristics Analysis

User characteristics analysis will analyze the WordList field and select several words as the characteristics of the user, which called user characteristics. So if the user characteristics must be mined out, first, the definition of user characteristics must be made clear firstly [8]. Then the algorithm of how to mine the user characteristics will be defined as follow.

3.1 Definition of User Characteristics

The text representation and feature extraction is a basic problem in the field of text mining [9]. Feature item must correctly express the content of text firstly. So the research must distinguish the target text and other texts. The number of feature items must not be too much [10]. In this research, the feature weights will be calculated based on TF-IDF, then screening according to the result of calculation is a effective method.

The formula of TF-IDF is shown as Equation 1.

EQUATION 1 TF-IDF

$$W_{ij} = tf_{ij} \times \log(\frac{N}{df_t}) \tag{1}$$

In Equation L-TF-IDF, W_{ij} means the TF-IDF value of word i in file j. tf_{ij} means the appearance frequency of word i in file j. N is the total number of files. df_t is the number of files which contain the word i. The feature weight will be calculated based on Equation TF-IDF. More higher the frequency of the word appears in one file and more lower it appears in other files [11]. Then the word can distinguish this file with higher degree and the weight of the word is much greater.

TF-IDF can only calculate the feature of one file. But in this research a user usually visited a set of urls, which refer to a lot of files. User characteristics can't be calculated according to TF-IDF. For this reason, this research redefine the TF-IDF, which called L-TF-IDF.

The formula L-TF-IDF is shown as Equation 2.

EQUATION2 L-TF-IDF

$$W_{ij} = \frac{\sum_{j \in u} tf_{i,j}}{\sum_{j \in u} tf_i} \times \log(\frac{N}{df_t}) \tag{2}$$

In Equation L-TF-IDF, W_{ij} means the L-TF-IDF value of word i for user u. $\sum_{j\in u} tf_{i,j}$ means the sum frequency of word i in the visited pages of user u. $\sum_{j\in u}$ tfi means the all word counts of the webpage user u visited. $\frac{\sum_{j \in u} tf_{i,j}}{\sum_{j \in u} tf_i}$ means the frequency weight of word i in all the webpage user u visited.

$\log(N/df_t)$ means the inverse document frequency of word.

$\frac{\sum_{j \in u} tf_{i,j}}{\sum_{j \in u} tf_i}$ plus log (N/df_t) obtain the value of L-TF-IDF, which segment will be done based on the NLP algorithm, which means the user characteristics of word. The setting of equation 2 consider about both the word visited frequency of user and the discrimination of the word for users. The user characteristics calculated through Equation 2 can represent the visitor much better.

3.2 Extraction of User Characteristics

User characteristics extraction algorithm can calculate the weight of each word stored in the wordlist field of each user and make the wordlist sorted from large to small according to the weight. Then select the top ten words as user characteristics, which have greater weight. Because there are many users for web servers, and every user will explore many webpage in this server [15]. Every webpage the user visited will be segmented out a lot of words. The mass calculation make the calculate much difficult. This research design the User characteristics extractio algorithm based on Hadoop platform. The Hadoop platform using distributed system as its basic architecture. The computational task can be calculated by decomposed to huge sub nodes. After all sub nodes completed calculation, the platform will summary all the sub results into a final result. This property is suitable for massive computing during visitors feature extraction process [16]. The input of user characteristics extraction algorithm is the data items which calculated by word segmentation. The structure of input data items are shown as follow.

(*UserId*, *Content*, *WordList*, Browserid, date, time, c-ip, user-name, method, uri, status, bytes, version, UserAgent, Referer)

The fields UserId, Content and WordList will be used by this algorithm.
The output of the algorithm is shown as follow.

(UserId, Content, WordList, Browserid, date, time, c-ip, username, method, uri, status, bytes, version, UserAgent, Referer, *FeatureList*)

The field FeatureList is the user characteristics extracted by this algorithm.
The algorithm contains three steps, each step can be done by a MapReduce. The process of the algorithm is shown as follow.

Step 1: Calculate the word's weight of current log files group which refer to a user and calculate the file count number which contain this word.

Step 2: Calculate the L-TF-IDF weight of each word.

Step 3: extract user characteristics. Because the L-TF-IDF weight have been calculated in step 2. So this step just to make the wordlist ordered by weight of L-TF-IDF form large to small and choose the top ten as user characteristics.

4. User Group Clustering

User group clustering is a key step of this research. Similar users will be gathered to a user group according to user characteristics. Then the resources can be recommended according to user group.

This research cluster user group based on the K-Means algorithm in Mahout. K-Means is a algorithm based on division. The thought of K-Means is: First select several initial center of clustering. Then calculate the similarity of every user to this center and distribute similar users to the center[11]. Repeated until the standard measurement function begins to convergence. The feature of K-Means is : The cluster itself as compact as possible, and as much as possible to separate clusters.

The first step of clustering is chosen the initial centers of clustering, according to the thought of K-Means. This research select several initial center according to the content of resources, which including Philosophy, literature, science, engineering, law, medicine, education, history, economics, management science.

Sample webpage have been selected from every type. These webpage and the user characteristics will be input to the K-Means. The output of K-Means is shown as follow form.

(UserId, *Feature*, *Type*)

Combined all the data to user group form, which is shown as follow.

(UserId, *FeatureList*, *TypeList*)

The field FeatureList is user characteristics. The field TypeList is the kind of user groups which current user belongs to.

5. Conclusion

With the continuous promotion of e-learning, more and more people are accustomed to online learning resources. How to push more appropriate resources to user actively become the focus of research. Research in this area relates to data mining, artificial intelligence, educational psychology, cognitive science and other aspects of knowledge. Internet behavior analysis is one of the solutions [12]. However, the Internet behavior data are hard to get, which is a bottleneck for this field to development. Web log files record the whole process of interaction between the Web site and visitors. Behavior characteristics and laws of visitors access based on log analysis may lay the foundation of recommender service for resources platform.

This research takes web log files as research object. From the scale of data mining to analyze the web user's behavior, design a user characteristics analysis system and take simulation experiment to this system. Based on the result of the analysis improvement have been made to resources recommender service, which result in good effects. There is still some deviation while data analyzing. So a lot of deeper work need to be done while internet behavior data analyzing with comprehensive knowledge in many fields.

References

[1] Zheng Liangliang. Research and analysis of access logs based on learner characteristics. Donghua University. 2014.

[2] Qiu Di. The study of learner information need identification based on Web log mining. Huazhong Normal University. 2012.

[3] Zhan Shengjun. Research on learning algorithm behavior log analysis based search engine. Hubei University of Technology. 2011.

[4] Ding Yingfang. Recommendation prototype system research and Implementation Based on Web log mining. Nanjing Agricultural University. 2009.

[5] Zhang Jingzeng, Meng Xiaofeng. Research on mobile Web search. Journal of software, 2012, 23 (1):46-64.

[6] XIANG B, JIANG D, PEI J, et al. Context-aware ranking in Web search/ /Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGR′ 10. ACM, New York, NY, USA, 2010:451-458.

[7] Ou Jianwen, Dong Shoubing, Caibin. Method for extracting the template Webpage topic information. Journal of Tsinghua University: Natural Science Edition, 2005, 45 (S1): 174321747.

[8] CAO Yujuan, N IU Zhen2dong, DA ILiu2ling, et al. Extraction of informative blocks from Web pages / /Proc of International Conference on Advanced Language Processing and Web Information Technology. Washington DC: IEEE Computer Society, 2008: 5442549.

[9] Wang Zhanyi. Study on several problems in Web text mining. Beijing University of Posts and Telecommunications, 2012.

[10] Text analysis - Keyword extraction classifier. Http://www.doc88.com/p- 68540434620- 7.html.

[11] Huang He, I luang Hai. Wang Rujing. FCA-Fiascd Web User Profile Mining for Topics of Interest. Proceedings of ihe 2007 \\:A:Al IiUcrnational Conicrencc on Integration Technology. Shenzhen, China. 2007:20-24.

[12] Sofia Stamou, Alexandros Ntouias. Search personalization through query and page topical analysis. User Model User-Adap Iner. 2009, 19:5-33.

[13] Hochul Jeon, Taehwan Kim, Joongmin Choi. Adaptive User Profiling for Personalized Information Retrieval. Third 2008 International Conference on Convergence and Hybrid Information Technology. 2008:836-841.

[14] Cuncun Wei, Chongben Huang, Hengsong Tan. A Personalized Model for Oritology-driven User Profiles Mining. IEEE. 2009:484-487.

[15] Nasraoui O. WEB data mining - exploring hyperlinks, contents, and usage data. SIGKDD Explorations, 2008, 10(2): 23-25.