Research and implementation of web transactions clustering analysis

Fengzhang Wang
Hubei University of Technology,Wuhan,430068, China.
Zaozhuang Vocational College of Science & Technology
Zaozhuang, 277599, China;
E-mail:81883899@qq.com

Abstract—This article system, fully discusses the Web transaction clustering analysis of two stages, namely data preprocessing and cluster analysis phase. Data preprocessing phase, including log analysis, data cleaning, the user identification process; at the same time, studied the Web transaction clustering analysis based on k-means algorithm.

Keywords-Web transaction; Clustering analysis; k-means

I. INTRODUCTION

In the current world, with its unique power affects the Internet is the life of mankind. Nowadays, almost every government department, companies, schools and institutes and other units have or are being set up their own web sites. People's work, study and life are getting more and more dependent on the Internet. The number of Web sites on the Internet is growing rapidly. Each site scale is continually expanding. Web brings the extremely irritating browsing information and great convenience.

II. WEB MINING AND WEB TRANSACTION CLUSTERING ANALYSIS

A. Web mining

In the large-scale data source often hides many have a potential value of knowledge. How to find useful knowledge is the research focus on subjects such as artificial intelligence and database. But only rely on the traditional retrieval mechanism and statistical analysis method has far cannot meet the need. Therefore, in recent years saw the emergence of a new knowledge extraction technology, Data Mining, Data Mining,. Data mining aims to extract from the database right, non trivial, unknown, has potential application value, and ultimately can be user understand the model. Its emergence as the automatic and intelligent to vast amounts of data into useful information and knowledge provides an opportunity. Involved in data

mining, such as machine learning, pattern recognition, statistics, database and artificial intelligence, and many other disciplines, cross discipline is the database theory and machine learning.

Data mining is sometimes known as the Database Knowledge Discovery (Knowledge Discovery in Database, the KDD). Term KDD first appeared in 1989. Fayyad defines it as "from the data set to identify valid, novel, potentially useful, and ultimately understandable patterns of non trivial process". Its purpose is to improve the decision-making ability, detect abnormal market pattern, on the basis of past experience predict future trends, etc. These patterns are implicit, previously unknown, is potentially useful information for decision-making. Through data mining, valuable knowledge, rules, or a high level of information can be extracted from the database related data collection, provide the basis for decision-making, so as to enrich the database as a reliable resource, inductive for knowledgeable service.

Web mining is the application of data mining on the Web, which USES data mining techniques drawn from WWW related resources and the behavior of interest and useful model and implicit information, Web technology, data mining, computer linguistics, informatics, and other fields, is a comprehensive technology. The technology can be divided into three types: one is the Web Content Mining, Web Content Mining,), namely of the Web page Content Mining; Second, the Web Structure Mining (Web Structure Mining,), namely the Mining Structure between Web pages; 3 it is to use Web information Mining (Web Usage Mining,), namely the user access to Web access records left by the excavations. This topic - Web transaction clustering analysis, belongs to use Web information mining.

B. Web transaction clustering analysis

Clustering analysis is one of the key technologies of Web mining. In Web use mining, can undertake two clustering: user clustering (or user access transaction clustering) and page clustering. Personalized service user clustering is mainly refers to through the analysis of the WWW server log files for the web user behavior patterns, and its quantity, then use a certain algorithm clustering - that is, to build a similar view of users - process. Such a rule in e-commerce market decision and it will be very useful to provide customers with personalized service. Page clustering cluster is related to mine content page for the Internet search engine and web provider is very useful. This paper focusses on the user clustering and session affairs as the clustering analysis of data objects.

III. THE OVERALL DESIGN OF THE WEB TRANSACTION CLUSTERING ANALYSIS SYSTEM

A. Development goals

This system development goal is to achieve a Web log as data sources, using Web usage mining and Web transaction clustering analysis of related technologies and user's access behavior of algorithm analysis, found that users access the main interest of Web transaction

clustering analysis system, provide the basis for personalized service recommended design.

B. *Mining the data source*

Web service system is a logical structure of multi-level, including customer, agency service layer and Web service layer, etc. The Web access to the data collection can be client, proxy server and a Web server. Client data from the user browsing behavior occurs due to be collected, it involves the privacy of users, at the same time also requires users to cooperate, therefore not suitable for as the data source of this article; Proxy server data collected by commonly used to provide a more user - site information, but in this paper, the present study is a multi-user - single site, so did not choose the proxy server side data as data sources in this paper. After comprehensive consideration, this article chooses from a Web server to collect the server log file as a clustering analysis of data, and using the Beijing institute of electronic science and technology Website (http://www.besti.edu.cn) logs file on the server as test data.

C. *The working process*

Analysis of the Web log need to pass a series of data preparation and modeling work. As showed in Fig .1, the working process of the system mainly includes data collection, data pretreatment, clustering analysis and visualization of the results of the analysis of four stages. The data preprocessing and includes the following several stages:

Log analysis, the text format and structure of the log file can be transformed into structured data records.

Data cleaning, remove, and clustering analysis have nothing to do with redundant records.

User identification, record from the data of identifying the different users access the Website.

Transaction identification, from the user's access records identifies users of different transactions when accessing a website.

D. *System architecture and development technology*

Fig .2 shows the Web transaction clustering analysis system architecture.

As shown, the Website of the Web server creates a regular text format logs files, in this paper, the log files are called the original log, and the log file as input data. For performance reasons, the system will log parsing and data cleaning combined into one process, the use of Java I/O reads log files, and resolution of the log file format. At the same time with the resolution of the log file, the process to resolve the logging for cleaning, remove useless visit record, and record the filtered through a JDBC connection stored in the MySQL database.

Also for performance reasons, the system incorporating user identification and transaction identification is a work. The process of the database to read the log output data record extraction and data cleaning process, and identify the users and transaction from these records, will identify the results stored in the database again.

Clustering analysis process is delivered from the database transaction sets are identified, and use the cluster analysis algorithm to group similar affairs gathered into a class. System using Java 2 d the clustering results presented to users of the system in a graphical manner.

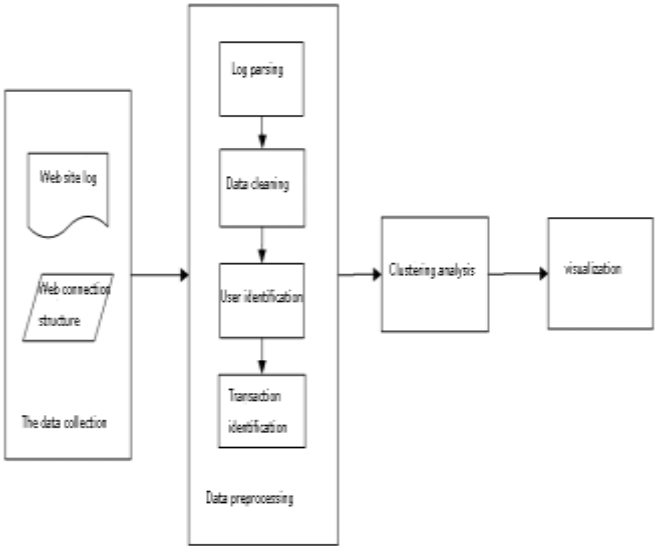

Figure 1. Web transaction clustering analysis work flow of the system

This article USES the JFC Swing implementation system of the user interface.

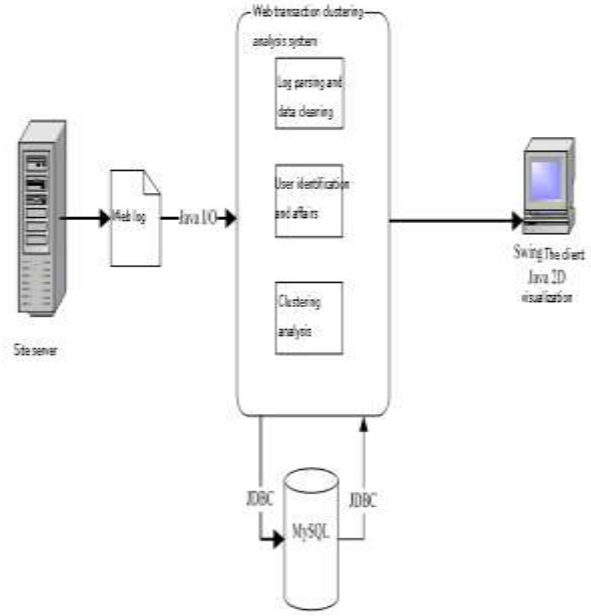

Figure 2. The Web transaction clustering analysis system architecture

IV. DATA PREPROCESSING

Before clustering analysis, the first thing to preprocess the site logs, to convert the original log files or abstract as a suitable for clustering analysis of data. In user clustering analysis, the main work includes data preprocessing log extraction, data cleaning, user identification and transaction identification. Here by Beijing Institute of electronic science and technology Website (http://www.besti.edu.cn) in the server log files as the data set.

A. *Log parsing*

Web site logs generally in the form of text stored in files, and a line of text corresponding to a visit record. No matter what kind of analysis was carried out on the Web log, the first step must be parsed Web logs format, there can be no structure of the text log into a structured database records.

In the actual Web log record does not contain all of the fields. In Beijing institute of electronic science and technology of Web server, for example, the log format for (user IP, server name, user name, access time, request method, request resources, transport protocol, state, transmission of bytes, sources of request, the user agent). Such as a log record below:

59.63.91.22 www.besti.edu.cn - [14 / Oct / 2007:08:13:48 + 0800] "GET/upload/PIC/besti_20040429200438_055065_ph. HTTP / 1.1 JPG" 200 21614 "http://www.besti.edu.cn/page.php?sid=3&ssid=20&did=20" "Mozilla / 4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"

It says the requesting host IP as 59.63.91.22, access to the server name is www.besti.edu.cn, without login user name; Access time is on October 14, 2007, at 8, 13 minutes and 48 seconds + 0800, says www.besti.edu.cn server in east eight time zones; The type of access is GET, the requested resource is/upload/PIC/besti_20040429200438_055065_ph. JPG, and access protocol is HTTP 1.1; 200 indicates the server response the request is successful; The number of bytes of the request and response in the process of transmission is 21614; The user is through http://www.besti.edu.cn/page.php? Sid = 3 & ssid = 20 & did = 20 pages link to the resource; The user's browser types for Mozilla / 4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1).

B. *Data cleaning*

The original log file contains a large number has nothing to do with user clustering analysis records, also known as noise. The first step in cluster analysis, the noise is supposed to be removed, leaving only clustering analysis tasks related data items. The main work in data cleaning is:

1)Delete the log file to JPG, GIF, doc and mp3 players and other images, audio, video, auxiliary resource request record, because this article focuses on the user access to web pages. Different Web sites or the purpose of a Web log mining tasks do work here have a bigger difference, such as image as the main content of the website may keep users request record of image resources.

2)Follow up on all of your request status for 200 ~ 299.

3)Keeping the records access type for the GET request.

4)Delete access records search engine. There are just for two ways to identify search engine request record. One is checking the user agent in the field for Google, Yahoo and basic search engine. This method is simple to implement, but along with the increase in Internet search engines and updating, clustering analysis program also needs to be constantly maintained, and maintenance personnel can also be difficult to fully understand all of the network search engine. Another method is to identify the same user in a short time left by the large number of access records, because it is impossible for human users in a very brief period of time a large number of requests for servers. User identification is key to the second method. Can to IP or login user name to divide the users. But by IP users have their limitations, because if a user USES a proxy server to access the website, when multiple human users use the same proxy server, may also appear the same IP in a short time, the phenomenon of a large number of requests.

This article USES the first method of data cleaning, or delete a user agent field contains the Yahoo, Google, date, MSN and DigExt access records.

Tell from the strict logic, data cleaning is the log phase completed work. However, from a technical perspective, if you will resolve all the logging stored into the database, then to update a record in the database, so will consume large amounts of database connection, poor performance caused by worrying. Is the best way to parse out a log record, will determine whether the record redundancy. If the record is redundant, so doesn't will write to the database. In this trial, it makes log parsing and data cleaning process from 50 minutes to shorten time to 4 minutes, which means to shorten the nearly 10 times.

But, if use this log parsing and data cleaning the practice of merger, then the data cleaning logic is not to rely on the relationship between the record and, that is why this article is based on the user's browser type, not according to user's access records left by the number in a short time to identify the cause of the search engine.

C. *User identification*

This topic is for website users are anonymous, without access to reliable user information; At the same time, considering the topological structure is complicated, many websites according to the relationship between the topology identification pages and pages, can reduce the performance of clustering analysis system, so the structure of the file is not recommended site. Only the Web log file is used to determine the user, can make clustering analysis system on the user identification is more general, but also has certain difficulty. In general, can use the following rules to identify users:

1)The IP address of the difference represent different users;

2)When the same IP address, default different operating system or browser types represent different users;

3)In the same IP address, user uses the operating system and browser type is also the same situation, if the user requests a page cannot arrive from a visit to a page, you can conclude that it is a new user. In particular, to view logging the request source (Referrer) field, if a request page request source is not on the same IP, operating system and browser type has identified a user has requested URL in the sequence, as the request from a new user; If a request page request source appears on the same IP, operating system and browser type has identified a user has requested URL in the sequence, will this request page in the most close to the user request on time sequence of page URL.

V. WEB TRANSACTION CLUSTERING ANALYSIS BASED ON K-MEANS ALGORITHM

There are numerous kinds of clustering analysis in the field of data mining algorithm. The main clustering algorithm can be divided into: dividing method, hierarchical method, the method based on density and grid based method, and the method based on the model. In many algorithms, belongs to a classification method of k-means algorithm is the most understated, is one of the most widely used clustering analysis algorithm.

A. *K-means algorithm*

K-means algorithm (also called K centers or c-means algorithm) to receive an input parameter K and n data objects, and could be divided into K clusters that a data object, the clusters of data objects within the similarity as high as possible, and the data object similarity between clusters as low as possible. Similarity between clusters and cluster depends on the center of the cluster, namely the mean of all the data objects.

The working process of the k-means algorithm is as follows: First, the data set randomly choose K data object, in the K data object represents K initial cluster centers (or average). Then, according to the data object's distance to the center of the k clusters, all the data objects are assigned to the nearest clusters. At this point, to calculate the average k clusters. The process is constantly iteration, until an index functions convergence. Generally can use (yi - (mxi + b)) index function, namely

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2$$

,

Where E is a data set and all the data objects (yi - (mxi + b)); p is a given data object in space, the corresponding points; m_i is C_i a cluster of average (p and m_i are multidimensional). In other words, for each data object to its square sum of the distance of the cluster center. The index function makes the cluster of convergence as compact as possible, within the cluster and the cluster between isolated as much as possible.

B. *The distance between the transaction*

Each data object has r a property, you can use a single point in the r r d $(x_1, x_2, \text{L}\ , x_r)$ To represent a data object o , where x_i is the first o i data object attribute values. The distance between the two data objects for the two points in # d Euclidean distance in space. That set the two data objects to $o_1(x_1, x_2, \text{L}\ , x_r)$ and $o_2(y_1, y_2, \text{L}\ , y_r)$, there are

$$d(o_1, o_2) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \text{L}\ + (x_r - y_r)^2}$$

In Web transaction clustering analysis, a transaction is a data object. The dimensions of the data object r is the number of sites URL address, the value of the object in the first i d is the number of the transaction to access the first i a URL address. For example, suppose that sites with A total of four URL address A, B, C, D, respectively corresponding to the four dimensions of 1, 2, 3, 4 D, the transaction $\vec{\alpha} = \langle A, B, A, D, A \rangle$ point is corresponding to the four dimensional space by,1,0,1 (3), the transaction $\vec{\beta} = \langle B, A, C \rangle$ corresponding points for,1,1,0 (1). As a result, the distance between the transaction $\vec{\beta}$. and transaction $\vec{\alpha}$ is

$$d(\vec{\alpha}, \vec{\beta}) = \sqrt{(3-1)^2 + (1-1)^2 + (0-1)^2 + (1-0)^2} = \sqrt{6} \approx 2.45$$

CONCLUSION

Web transaction clustering analysis is a important content in the field of Web mining, through clustering analysis was carried out on the Web user transaction, can get the attention focus in the user group of the site, users browsing behavior rule of patterns such as knowledge, these knowledge for in the Web personalized recommendation service, improving the structure of the links between pages, and improve the performance of the entire Web system, to carry out the electronic commerce is of great significance in such aspects as intelligent applications.

REFERENCES

[1] Yong Miao. Research and implementation of the anonymous user browsing paths mining. Master degree theses of master of nanjing university of science and technology, 2006.

[2] Huiying Zhang, Linnan Jiao. User access pattern clustering analysis in the application of the web page recommendation. Computer engineering. In August, 2006, 32 (15). 64-66.

[3] Jiawei Han, Micheline Kamber. Data Mining, : Concepts and Techniques (Second Edition). China Machine Press, 2007.

[4] Ying Pan,Jingzhang Liang. Campus network user behavior founded on K - means algorithm clustering analysis. Computer technology and automation, in March, 2007, 26 (1). 66-69.

[5] Yan Guo. Network logs mining and utilization of user interest. Ph.D. Thesis, Institute of computing science and technology of Chinese academy of Sciences in 2004.

[6] [Haiquan Yang. Automobile fault diagnosis and testing technology [M]. People's traffic press, 2004.

[7] Guojun Wang. Automatic control theory development review [J]. Microcomputer and application, 2006.

[8] Lin Sun. Intelligent system with auto intelligent technology [J]. Journal of automotive research and development, 2007.

[9] Qing Nie. Automobile structure [M]. Zhejiang: Labor publishing house, 1990.

[10] Yaoyi Qian. Automobile electronic control system [M]. Beijing: mechanical industry publishing house, 1999.

[11] Zhongguo Liu. Automobile structure [M]. Beijing: mechanical industry publishing house, 2000.