Design and Realization of Data Mining System based on Web

HE Defu[1, a]

[1]Department of Quartermaster, Wuhan Economics College, Wuhan 430035, China

[a] hedefu1023@sina.com

Keywords: Data mining, Web mining, Database.

Abstract. For enterprises, it appears especially important to survive in the increasingly fierce competition in the business and strive for more customers. If they understand what customers need and interested better, and make key decisions that can attract more users to buy their goods, then they will make more profits. As a new business technology, data mining can be becomes the key to obtain competitive information for commercialization. In the usual process, people often leave a lot of records such as in browsing the web, etc. Based on the analysis of extraction, transformation, or processed by other modeling methods, this information seemingly useless in the data mining system can be important data for business decisions. By extracting records user left in the XML file as the basic data source, this paper put forward the general model of data mining. Through clustering, correlation algorithm, this paper studied the common rules of users' surfing the Internet behavior, and figured out the subconscious demand of users.

Data mining

Concept of data mining

Data mining in simple terms is to find "knowledge" of gold bullion in the "data mining". It means to extract implicit, unknown information and knowledge in advance, which is potentially useful and ultimately understandable from incomplete, noisy, fuzzy, random practical application in the data. It involves in a very wide range of interdisciplinary sciences, including machine learning, database, statistics, pattern recognition, data analysis and related technology. Different from traditional data analysis such as query, reporting, online analytical processing, data mining is to mine information and discover knowledge without clear hypothesis.

Process of data mining

Data mining is a complete process, which process mining unknown, effective and practical information from a large database of previously, and use the information to make decisions or enrich knowledge. Data mining can be regarded as knowledge discovery in database (KDD); data mining can also be regarded as just a basic step of the KDD. [1]The data mining process can be roughly interpreted as trilogy: data preparation, data mining, and estimate the interpretation of the results, the process is shown in figure below.

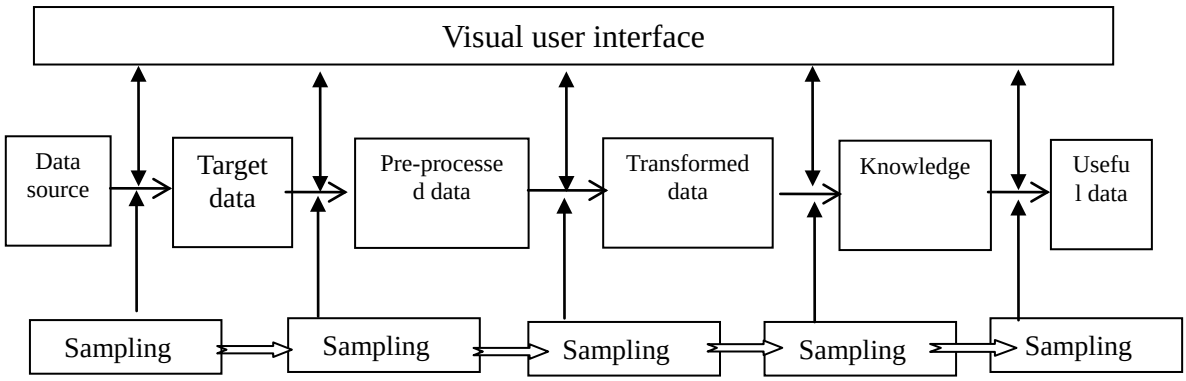

Figure 1 The basic process of data mining

Typical data mining algorithm

The fundamental goal of data mining is found standard available mode in the database mining. This mode can be roughly divided into two types, one is the type of descriptive model, and its main expression is description of the useful information in the data standardization, set up data characteristic; the other is a prediction model, which is based on time series to predict in advance information value. In terms of algorithm, data mining algorithms mainly has five categories, namely, clustering analysis, classification analysis, regression analysis method of data mining, association rules method and time.

Association rules analysis

The main way of the association rules in essence is to retrieve the corresponding relationship between each other from a large amount of data collection. [2]Data business develops with the development of modern society. New data item produced every day is amazing, and various kinds of information hidden in these data items are very valuable, so association rules, mining useful information essential work becomes very important. Especially in business associated with categories of information mining, it can provide very effective decision support for business decisions, thus are attached much importance in business activities.

Classification analysis

Similar with association rules, classification analysis also have a wide application in today's business activities because of its more focus on the analysis of the data, therefore have considerable reliability and scientific. Data analysis establishes a reasonable classification model based on classification analysis and training, and then based on the effect of classification model, to difference and classification of the data to achieve the ultimate goal. The nature of the classification and regression are similar, both of which can forecast data, the difference is, compared with regression method, the output of the classification results relative to some scattered, not like regression output data continuity. Now data classification algorithm in the level are various, more than the C4.5 decision tree algorithm and ID3 algorithm. In addition, the two kinds of calculation methods the KNN classification and adjacent directional transmission classification algorithm are also included.

Logistic regression algorithm analysis

Generally speaking, regression analysis method to predict the way of data in advance based on a certain characteristics of data. In practice, for the simple data prediction often uses linear regression techniques to complete, and complicated technical method such as logistic regression method, decision tree regression and logistic regression model is to meet a situation more complex environment.

To calculate logistic regression algorithm, we first set a variable Y, and the critical value of 1 and 0 respectively. The Y's argument for the $X_1, X_2, ..., X_N$. Regression algorithm is expressed by the following formula.

$$P = \frac{\exp(\beta 0 + \beta 1X1 + \beta 2X2 + \cdots \beta mXm)}{1 + \exp(\beta 0 + \beta 1X1 + \beta 2X2 + \cdots \beta mXm)}$$

Then we make the following change:

$$\ln\left(\frac{P}{1-P}\right) = \ln\left(\frac{\frac{\exp(\beta 0 + \beta 1X1 + \beta 2X2 + \cdots \beta m1Xm)}{1 + \exp(\beta 0 + \beta 1X1 + \beta 2X2 + \cdots \beta mXm)}}{1 - \frac{\exp(\beta 0 + \beta 1X1 + \beta 2X2 + \cdots \beta mXm)}{1 + \exp(\beta 0 + \beta 1X1 + \beta 2X2 + \cdots \beta mXm)}}\right)$$

$$= \ln[\exp(\beta 0 + \beta 1X1 + \beta 2X2 + \cdots \beta mXm)]$$

$$= \beta 0 + \beta 1X1 + \beta 2X2 + \cdots \beta mXm$$

When all various influence factors are of 0

$$\ln\left(\frac{P}{1-P}\right) = \beta 0 + \beta 1 X 1 + \beta 2 X 2 + \cdots \beta m X m$$

$$= \beta 0 + \beta 1 * 0 + \beta 2 * 0 + \cdots \beta m * 0$$

$$= \beta 0$$

By the formula above, you can see that constant term beta is the natural value ratio of individual value and probability of zero when the various influence factors of 0.

Web mining

Definition of web mining

Web mining first appeared in 1996, its name is not the only one. Web knowledge discovery, Web information mining, Web information mining and Web mining is the same one.

Web mining is a comprehensive technology, the use of different mining algorithm, hiding in Web data mining model of data in the data source. [3]The Web mining prospect is good, will inevitably puts glorious greatly in the future.

Characteristics of web mining

Web data mining and traditional data mining is different; I summarized the following features:

(1) Complexity. The data on the web involves industry is multifarious. Under the background of rapid development of Internet, web data also develops by leaps and bounds. This makes when we need a quick and good mining method when mine a large amount of data mining in order to meet the mining efficiency.

(2) Dynamic. We all know that the Internet information is updated in real time. Every day there are all kinds of new information is presented in front of us, about the international news, gossip entertainment, hot spots of the people's livelihood, walking in the street also is not difficult to find that people use mobile phone to brush the Weibo, release the anecdotes. The information of rapid change is the dynamic characteristics of the web, so we need to adopt a technology to store the data.

(3) Diversity. After filtering processing, web data will present a different form. So we have to slightly change the method of data mining to web data mining effectively do better with the data of one form or another.

(4) Heterogeneity. Web data heterogeneous mainly for heterogeneous database, as each Web site have their own data sources, the differences between these data sources contributed to the diversity of Web data mining.

Web mining system structure design

Database is closely related and data mining. From the definition of knowledge, they are with a lot of weight and degree, mainly reflected in the discovery (KDD). According to the current scholars' general cognitive, data mining and knowledge discovery are put in the same position. Has said it will be the field of artificial intelligence for KDD, is called the data mining in the field of database, data mining mainly by machine learning algorithms, artificial intelligence, fuzzy logic, artificial neural network algorithm and fractal geometry algorithm are discussed in theory for the premise, so as to achieve more efficient algorithms used in huge data to achieve the objective of the discovered knowledge.

Overall structure of the data mining system

The purpose of data mining to find knowledge from large database, which can be roughly divided into four steps:

(l) Data preprocessing

Amount of data in the database is very large and complex, therefore, to improve the efficiency of mining, realize a mining purposes to the pretreatment of the data is necessary, which mainly include the data sorting, integration, filtering, etc. [4]In terms of data sorting, mainly including delete the redundant data, merge data items, check the correct data and information, and so on. After that, we should integrate and select data set come from different sources, and the filter data, select

types meet the requirements of search data. Finally is the translation work, selected data to adjust response into a more suitable format for the next step of work.

(2) Data mining

This step is the main steps of data mining, which is realized according to the special algorithm the system default, it directly process the pretreatment of information, based on the pattern of its hidden knowledge and information, and the characteristic, interconnected and grouping classes related to the analysis of the deviation degree.

(3) Specific comments on model

We explain the model of evaluation is evaluation analysis of the results of data mining; the purpose is to test the results of data mining. Generally speaking, a model evaluation is based on certain values and the corresponding reference material. The process usually need to interact with the data mining module, refer to the corresponding degree of interest, the search results directly positioning on the point of interest, implementation with interest as refer to value filter criteria for data discovery.

(4) Expression of search the knowledge

The definition of knowledge expression is to express the mining results in necessary form, which can be a textual, form and visual graphics, and other expressions. According to the above account, we can use the form of the standard data mining system for expression.

Function model of web mining

Data mining essentially is characterized by discovery process for knowledge. Data statistics and analysis method are mainly used in this algorithm for effective search of information of value and significance. But due to the volume of a database is too big and the complex data volume, the traditional data mining on the search pattern is wet behind the ears. Therefore, compared with the traditional search, data mining system have different implementation styles, namely data mining model.

The characteristics of the data mining model, is that it can search according to the different requirements, make specific data mining model. Namely, it is of flexibility and adjustment. Generally speaking, a more suitable for the data mining model can be made according to the specific needs of customers to solve the problem of the customer, at the same time, combining the model itself and the actual work.

System performance evaluation methods

If you want to evaluate the data mining system, there are two very important indicators, the accuracy rate and recall rate. Among them, the precision accuracy is referred to as precession; At the same time, the recall rate is another way of saying.

Precision:

Data mining system will have a system when a certain interest category, to correctly determine the number of users, the actual number and the sum of the percentage is caused by system accuracy. Generally, the accuracy is inversely proportional to the error rate, namely, when the system is to determine the user interest in the error rate decreased, it represents a high accuracy. For the accuracy of the formula is illustrated below:

$$P_j = \frac{TP_j}{PP_j + FP_j}$$

Recall:

Recall rate in the system design is mainly embodied in the specific interest category, expressed as a percentage in the form of a number of users interested in correct judgment and comparison among all categories of the number of users. The recall rate is high, means in this category, the user's judgment rate is higher. The recall rate calculation can be expressed in the following formula:

$$P_j = \frac{TP_j}{TP_j + FN_j}$$

Summary

This paper studies the data mining system. After a long design, this web data mining system finally achieved the system. And through the recall ratio and the recall rate test, the reliability and stability of the system performance is verified. Some key data was obtained by algorithms, which basically provide good support for enterprise in commercial operation decision.

References

[1] Maha El Choubassi, Oscar Nestares, Yi Wu, et al. An Augmented Reality Tourist Guide on Your Mobile Devices, Lecture Notes in Computer Science 5916, pp. 588-602, 2010.

[2] A. J. Davison, N. D. Molton, I. Reid, and O. Stasse.MonoSLAM: Real-time single camera SLAM. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 29(6):1052-1067, 2007. 1.

[3] Carmen C, Bruce L. A basic primer on data mining. Information systems management, 2002, 12 (4): 34-38

[4] Malerba D, Esposito F, Lisi F A. Apices a mining spatial association rules in census data. Research in Offal Statistics, 2002, 5(l): 19-44