ATLANTIS PRESS

Advances in Engineering Research, volume 123

2nd International Conference on Materials Science, Machinery and Energy Engineering (MSMEE 2017)

Application of Data Mining Technology in Software Engineering

Jie Ma

Hainan College of Software Technology, Qiong Hai, Hai Nan, 571400

Keywords: Software Engineering, Data Mining, Application Research

Abstract. Data mining represents a change from validation-driven data analysis to discovery-driven data analysis. In the verification-driven approach, the decision maker must assume the presence of important information, collect information and use it to substantiate such assumptions. Due to the size and complexity of today's data storage, this approach does not effectively explore the available data in an organization, and the discovery method can filter out large amounts of data and automatically or semi-automatically discover hidden information. The information collected in data mining is designed to cater to the needs of software organizations that have their own products and processes to improve their goals. So you might use different software metrics when collecting data.

Introduction

Data mining is the process of digging the various data stored in the database. Data mining generally includes three processes: data preprocessing, which mainly includes data collection, cleaning, sampling and data exchange; data mining, in determining the specific mining data, it is necessary according to the corresponding methods such as summary, classification, Association rules and so on to determine the specific mining algorithm; model evaluation and knowledge representation, that is, after determining the algorithm of data mining, according to the results of mining, evaluation of the results of the process. Through the analysis of the data mining process, the main task of data mining is to find hidden in the data in the interesting mode, because the massive data in the user is difficult to quickly find useful information with their own, so this requires data mining technology to seek with the user Demand information closely linked to the information model, in order to meet the user's expectations and needs. The description model and the forecasting model are the main functions of data mining. Data mining technology is based on the database, so the data mining technology is mainly application-oriented, data mining technology application and software engineering time is in the late twentieth century, with the continuous development of computer technology, data mining technology in software engineering applications More and more widely.

Data mining research covers a wide variety of content, but it mainly includes the following aspects: First, the development of dedicated data mining system. A different data mining system can work in different databases, so it is very important to develop a perfect data mining system for different data mining systems. Second, the mining system operation algorithm can run efficiently. The function of data mining is to save time and improve work efficiency. The implementation of the most important functions needs to be built on the very fast running speed, which requires that the algorithm used in the data mining system must be efficient, so that the working time can be been accepted. To ensure that the results of data mining accuracy and effectiveness. The results of data mining must meet the needs of the user, the relevant data according to the rules of reality, irrelevant data suppression display, similar data display. Fourth, visualize the mining results. The resulting data must be legible and need not be processed in any decode mode, and the results of the data mining are presented by visualization. In the database, many of the data are dynamic changes, or exist in the form of interaction, which requires data mining system with multi-level, dynamic search features to the database for different angles of mining. Many databases are connected by Internet technology to develop a data mining method that can extract the required data in different databases.. Network and the existence of hackers, data mining is facing a great risk, so to improve the data mining process security, to prevent disclosure of personal information to ensure reliable and effective data.

Typical Application of Data Mining in Software Engineering

Software vulnerability detection is used to identify errors or vulnerabilities in the software development process in order to correct in a timely manner to ensure software reliability and quality. Application of data mining to software vulnerability detection can be divided into the following five steps, (1) the establishment of software testing projects. From the user's point of view, to determine what aspects of the software need to test and how to test, the establishment of test plans and strategies; (2) the vulnerability database for data collection, data cleansing and data conversion. (3) select the appropriate data mining model for training and (3) select the appropriate data mining model, and select the appropriate data mining model, and then select the appropriate data mining model for training and verification. (4) classify, locate and describe the software vulnerability through the method in the previous steps. (4) The software vulnerability is selected according to the project requirements, and the training method is selected for the method. The method identified in step 3 is applied to the software database to find the unknown loopholes, according to certain rules to classify and describe the loopholes; (5) the knowledge to be applied to the software test project. The results of the front of the excavation into knowledge, save to the database, the software re-test to confirm the loopholes found, the results will be applied to the software development project.

The data warehouse is a subject-oriented, integrated, and stable data collection that supports the decision-making process in the management through "warehouse building". This process mainly includes the following stages: First, the source data phase: This stage is mainly on the historical data, the current data and comprehensive data collection. Second, the source data preprocessing phase: it mainly includes relational databases, software data documents and others. The third is to enter the warehouse management phase: mainly including data warehouse management tools, extraction, conversion, loading, metadata and data modeling tools. Fourth, knowledge base DM analysis tools: classification analysis tools, clustering analysis tools, correlation analysis tools and sequence analysis tools. Five is the visualization of software domain knowledge: reveal the inevitable factors that affect the quality of the software. Five aspects constitute the entire data warehouse management system.

Data warehouse modeling will be the main data together to establish a reasonable data resource library. This information includes customer requirements information, customer evaluation information, software system information, feasibility study report. First of all, the need for customers to the information needed to sort out, so as to do a good job of system functions, interfaces, data and other aspects of the determination. Second, the customer evaluation information is mainly application software testing, (including dynamic testing, including static testing, formal testing) to obtain the appropriate assessment information, with the customer's information needed to aggregate, weigh the quality of the software in the application , From which to find possible errors, and to modify them. Third, the software system information is the basis of software applications, which mainly include the size of the system, scope, the overall requirements, and the required support environment. Fourth, the feasibility report mainly refers to whether it has the feasibility of running, technology, economy, law, use can achieve the desired purpose.

Through a number of similar needs to be divided into a group of customers into a cluster, so that the customer's information more easily be understood by developers in order to provide a higher level of service, and satisfactory service. Some customers clustered into a group, you can specifically for their requirements to develop a special function of the software, through cluster analysis, the customer's software applications can be a screening of the observation, so that the software to achieve a good The use of the effect.

And clustering is completely different, classification analysis is marked by the characteristics of the data classification. Classification is to record the performance to facilitate the description of this type of data has the characteristics. Classification analysis is mainly used in decision tree, neural network and radial basis function software. The results of the classification analysis allow us to be more targeted in the design process of the database, through the software to deal with the customer attributes, for different customers to provide different services or protection.

Sequence analysis is a completely independent analysis algorithm, which is different from the

above two algorithms. This algorithm is mainly based on the data sequence or event detection. Because different customers require that the functionality provided by the same software is generally different, software analysts can classify customers according to their functional patterns. For example, when a customer uses a specific function of a software, the search function will be based on the needs of computer users to prompt, whether to follow the computer algorithm to analyze the needs of the next search operation.

Classical Methods in Software Engineering Data Mining

Classification is the operation of predicting the classification of labels (or discrete values). It is usually necessary to create a model, describe a predetermined set of data or set of concepts, and then use the model for classification. Several commonly used classification methods include deterministic tree method, Bayesian classification, neural network classification, K-nearest classification, support vector machine and so on. The decision tree method is based on the greedy algorithm, through the top-down recursive way to construct the decision tree, the leaf node corresponds to a category label, that is, the final classification result. Commonly used decision tree method has K-nearest classification method, support vector machine and so on. K-Nearest Neighbor (K-Nearest Neighbor), the basic idea is that if a sample in the feature space K most similar (that is, the most close to the feature space) of the majority of the sample belongs to a category The sample also belongs to this category. The algorithm is more suitable for the automatic classification of large sample size, and those with smaller sample size are more likely to use this algorithm.

An association rule refers to an interesting association or associated relationship between items in a large number of data items. The association rule has the following two important attributes: (1) support degree: (PA ∪ B), that is, the probability that the two items of A and B appear simultaneously in transaction set D; (2) confidence degree: P (B|A), that is, in the transaction set D of A, the probability that B also appears. While the rules that satisfy both the minimum support and the minimum confidence are called strong rules. Given a transaction set D, mining association rules is actually generated support and credibility were greater than the user given the minimum support and minimum credibility of the association rules.

Clustering is to classify data objects into multiple classes or clusters, with high similarity between objects in the same cluster, and different objects in different clusters. Clustering and classification is different from the class to be classified is unknown, belonging to a non-guided learning method. Clustering analysis is mainly used in the preprocessing of other algorithms, as a separate tool to obtain data distribution, as well as isolated point mining, indicating fraud and so on.

Conclusion

In summary, the application of data mining technology in software engineering has strong practical value, and the research of data mining technology is of great significance to promote the development of software engineering project. Therefore, the relevant staff should increase the The application of data mining technology in software engineering is more mature and reliable, so that it can also get good application effect in other fields.

References

[1] Miao Yu. Development of computer data mining technology and its application [J]. Urban Construction Theory Research (Electronic Edition), 2016 (22)

[2] Meng Qiang, Li Haichen. Web data mining technology and application research [J] .Computer and Information Technology, 2017 (01)

[3] Zhao Yufei. Data mining technology in the application of information management [J]. China Management Information, 2017 (04)

[4] Zhang Hua. Application of data mining technology in optimal operation of thermal power plant

[J]. Journal of Urban Construction Theory (Electronic Edition), 2016 (27)

[5] Na Guangyi. Data warehouse and data mining technology in the application of health information [J]. Shandong Industrial Technology, 2017 (05)

[6] Wang Zhaofei, Li Lin. Application of data mining based on large-scale database [J]. Digital Technology and Application, 2017 (01)