Research Progress on Software Engineering Data Mining Technology

Deng Fengxian

(Hainan College of Software Technology, Qionghai, Hainan, 571400).

Keywords: Software engineering; Data mining technology; Research

Abstract: At present, with the scale expansion of computer software, only rely on manual for software development, maintenance and other work is more difficult. Data mining technology can accelerate the speed of software development, and can in many databases find valuable data. This paper makes in-depth studies on software engineering data mining technology, and introduces the influence of data mining technology.

Software engineering data mining technology is to use existing technology or new data mining algorithm in massive databases, and is the process of collecting valuable information for software developers through a series of steps, such as selection, analysis, formulation. It is a process of clear grasp and management of software development. Software developers must collect the required data, which is the practice of software development industry. To complete the above work, we must extract the required data information from large amounts of data, and the process of collecting and selecting information is the process of data mining. At present, data mining technology has been widely used in software engineering. This paper introduces the relevant knowledge of data mining technique and its application in software engineering.

I. Relative concepts of data mining technology

The rapid development of computer information technology and network technology provides convenient conditions for users to obtain the required data information. With the help of data mining technology, find the potential data, rules for feedback system software engineering activities in software engineering by classification, clustering, statistical analysis, and improve software product quality, and the effect of software development efficiency.

A. The concept and its classification of data mining

Data mining is to obtain needed valuable data information in huge amounts of data. This process is called "dig" or "catch". Data mining is to test drive analysis way turning to drive analysis data way. To verify this driver, user can assume the existence of information, collect and analyze, then gradually verify the original hypothesis. At present, data storage is of large scale and has certain complexity, validation method alone cannot make full use of all available data mining database. Data driven approach can perform real-time effective screening for huge amounts of data, and identify useful information hidden inside automatically. In data mining process, information collection can help improve their products, so during data collection, use all kinds of software metrics. Data mining technology can mainly be divided into: classification tree technology, clustering technology, artificial neural network, correlation technology, and visual data mining technology, and so on. Software metrics data generally has the characteristics of high coupling, multi-dimensional. Software engineering often use statistical analysis, neural network and regression modeling and other special processing technology in data mining, and in practical application, choose which kind of mining technology has important influence on software engineering practices to achieve ideal goal.

B. Software engineering measures

As software engineering and large scale growing, there is a significantly increased difficulty of obtaining valuable information by developers in this part. Based on this, software developers use traditional methods such as browse documentation, code. Getting data information in software development method has been unable to meet the demand of times development. Software developers, while project developing, slowly implement quantitative processing of the indicators in the process of monitoring and controlling software to ensure that users can clearly understand the product throughout development process. At present, measurement data is gaining more and more attention and concern; software engineering measure must be a reasonable process combining with data collection, analysis, etc. The chart design product has the characteristics of diversity, generally uses static form in description, and the chart will change with time, so it leads to that measurement data is limited in practical application. For example: most charts can clearly react throughout production process and product quality, but cannot be seen as good judgment. Because of its particularity in numerous data, software engineering brings certain restriction and influence for further research on data mining.

II. The operation steps of data mining technology

Generally, data mining technology is mainly divided into the following steps: selecting data, data preprocessing, data mining, and absorption. Data mining process has the characteristic of interaction, and sometimes may have to select data again or improve the pretreatment process. Based on the above situation, we need to design a feedback loop during data mining. Data mining's first priority is to reflect management and target up to mining tasks, and the whole implementation process is mainly divided into the following steps. (1) Assessment of products: product evaluation is software production process. Resource property implements corresponding inspection, and according to the resources' all kinds of attributes, attribute the unknown assignment, and pay attention to quantitative unknown attribute for processing. After evaluation work, forecast the attribute value. (2) Associated attributes: association found can identify associated attributes in certain content. For example, find points out that software development attributes are associated with product attribute. (3) Clustering process: divide different groups of a structure to another with the same subgroup within a collection of the structure, this operation is called clustering process. (4) Data visualization processing: data visualization processing is the description of complex information with visualization methods, and explores visual data of the described content and uses data visualization interaction control for huge amounts of data analysis and review. Software engineering data mining's concrete operation process can satisfy the requirements of general data mining technology or field. Generally speaking, data mining process mainly includes: data preprocessing, mining, results evaluation and data mining process is shown in Figure 1. Data preprocessing is to transfer unprocessed data to one adapting dig out. Pretreatment process involves a variety of sources and format of data. After transforming data format for formatting data, select records and characteristics relating to the current data mining tasks and clean data to achieve the purpose of eliminating noise. Mining operations is to find essential reaction or regularity of information in huge amounts of data. The whole process uses a series of mining algorithm, and mining tasks include frequent sequences, association rules, and anomaly detection, etc. Results evaluation is to show useful information for users and the difficulties in place is that there is difference between computer understanding and expression and human's, and data mining can facilitate people to understand. Results assessment is mainly composed of: mode filter and mode expression. Based on the design of different tasks, data mining algorithms include classification, valuation and prediction, clustering and anomaly detection.

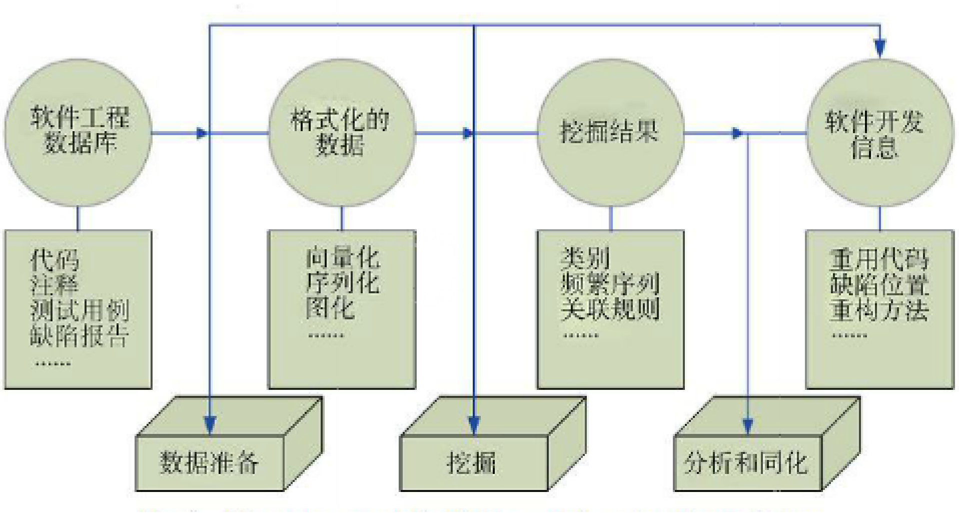

Figure 1 Flow chart of data mining

III. Challenges confronting software engineering data mining

Software engineering data mining process and traditional data mining have certain similarities, and this is a special form of data mining technology, which mainly includes pretreatment, data mining, evaluation results. Compared to traditional data mining, software engineering data mining has its own particularity, and its performance is as follows:

A. Bigger data complexity

Software engineering data has not only software report and version information related to structured data, but also contains a lot of code, comments, and kind of unstructured data. These two kinds of different data structures are unable to use the same algorithm for operations. These two kinds of data information have very important link, which causes increased data complexity of the entire engineering significantly.

B. Particular analysis results

The results traditional data mining get are generally through a variety of forms, such as reports, text, etc., software engineering data mining is not only to provide users with corresponding results, also gives software development personnel concrete examples in detail, and provides the required information for its design. So, software engineering excavation will submit the corresponding method for the result of new type of data.

C. Unified evaluation result

Traditional data mining technology has formed a relatively mature evaluation index, but from the point of view of software engineering data mining, software developers require information which has the characteristics of complexity, embodiment, and the corresponding representation method is also diversified. They can't be compared before expanding, so it is difficult to draw a relatively unified evaluation results. Therefore, software engineering data mining difficulty is in data preprocessing and expression of mining results, in-depth analysis of the problems existing in software development process and effectively solves is particularly important.

IV. Data mining technology in software development process

In recent years, data mining technology is widely used in software engineering, and applying data mining in software engineering can improve the maintenance efficiency of software system, which also to a certain extent, increases system stability.

A. Data mining in programming

Programming, as an important content of developing software, writes the code. Developers need to fully understand the structure and function of programming code, according to their own understanding, select valuable information in the database. Usually programming needs information of following parts: (1) software developers find needed code structure, similar function, and patterns which can be reused, such as data structures, object, method and so on in the existing code library; (2) developers can find static rules for reusing some patterns in the database, for example: class method, inheritance relationships, and so on; (3) developers have a thorough analysis understanding of reuse pattern rules, such as: API call order.

B. Data mining of developing open source software

Open source software refers to source code development software, which is a free provided service for customer. For it is free, the management and control of open source software is more difficult. At this point, data mining technology can improve the quality of open source software. For example, data mining system designed at the University of Oxford, users can real-time track and manage system, so to a certain extent, it improves use efficiency of open source software.

C. Data mining used in program code

Program code refers to clone code, which is a reusable code using copy and paste operations. Data mining used in cloning code is tested early. Cloning code testing mainly adopts the following forms: text comparison method, method based on measurement, latent semantic indexing, etc. But within clone code, using data mining is not enough mature, mainly because data mining must take consideration of semantic mining. At the same time, digging to the crosscutting concerns, crosscutting concerns use more mining methods, for example: in the process of code text analysis, according to different characteristics, divide into analysis based on text and type, based on clustering analysis, formal concept analysis, etc. Call relationship analysis can use fan analytical techniques or measurement method based on coupling and Page Rank for data mining.

D. Data mining technology using for software fault detection

Data mining technology can mine to program code and interactive modes according to the procedure of enforce discipline mining, thus accurate position and detect software failure. Mining technology in the program is reverse modeling for information, so as to strengthen understanding and corresponding maintenance program. At present, commonly used way of mining mainly includes mining based on rules and automation. Mining based on rule is to find the corresponding rules and to express the temporal logic based on program behavior. Automatic mining is a more mature API rule mining method. Traditional positioning software failure uses program slicing, which is more complex, prone to fault and location is not allowed. As traditional positioning method is gradually improved, program spectrum abstract describe trajectories is more successfully used, and compare the running of software with fault running state, according to the difference of the two, judge fault sources, and these new data mining technologies can effectively improve the accuracy and efficiency of software fault detection.

E. Data mining technology in software management

In software project management, data mining technology is used mainly in organizational relationships and version control information. Software project management is a complex project, and the key lies in reasonable personnel organization relationship mining are coordination and allocation of human resources, such as: a project may have hundreds of thousands of people involved in this process, which involves various personnel interactions, which are implemented via E-mail, documents and other interactions. It is prone to order situation in the process, and data mining technology can differentiate the organization relationship between staff, and it is convenient for project management. Version control can record the change of the whole file content in detail, making it easy for users to view the revised version, and apply data mining technology to the late version control information can reduce system maintenance cost. Data mining can provide warning role for daily maintenance software system in time. Some mining data can clearly looking for errors existing in system repair process, according to mistakes records, facilitate software designers to avoid the common mistakes in time, and promote the restoration and management level of software project.

Conclusion

In a word, data mining technology is widely used in code analysis, software fault detection, and software project management, etc., and it can effectively improve the management of software engineering and control ability. The current research of data mining technology is not mature enough, software engineering data mining technology project research must be constantly strengthened to promote better development and management software.

References

[1] Zhang Zhihong. The research and implementation of visual data mining technology [J]. Communication World, 2013, (18) : 28 and 29, 30.

[2] Wang Hao. The standardization of computer software development [J]. Computer CD Software and Applications, 2012 (18) : 206-208.

[3] Lei Lei. Review on data mining technology application in software engineering [J]. Journal of Electronic Testing, 2014, (2) : 128-129.

[4] Li Dapeng. Computer software development language study [J]. Computer CD Software and Applications, 2012 (6) : 196-195.

[5] Yu Shusi, Zhou Shuigeng, Guan Jihong. Software engineering data mining research progress [J]. Journal of Computer Science and Exploration, 2012 (01) : 1-31.

[6] Ding Yue, Zhang Yang, Li Zhanhuai. The research and development of data mining technique [J]. Journal of Computer Applications, 2012, 32 (1) : 182-190.

[7] Han Lei, Zheng Ling. The application of data mining technology in analysis of energy efficiency [J]. Computer CD Software and Applications, 2012, (23) : 153.

[8] Xia Xuefei, Teng Da, Wei Rongkai. Factors influencing the quality of software in computer software development [J]. Journal of Electronic Technology and Software Engineering, 2013 (23) : 89.