Study on the Application of Data Mining Technology in Software Knowledge Base

Guo Song, Yin Xiao-hong
NanChang Institute of Science & Technology

Abstract—With the rapid development of normalization, computer database system has related to various aspects of social life, which leads to an explosive growth of industrial data. Faced with massive data, computers with limited storage have to abandon some out-dated data. However, these data include much valuable information. Faced with this difficulty, data mining technology comes into being for application. It is an interdisciplinary subject combining database, artificial intelligence, machinery study and statistics and so on with basic functions of correlation analysis, classification analysis, clustering analysis and exception analysis and so on. Meanwhile, various tools related to data mining is becoming mature. In this paper, origins, classical algorithm and application of data mining technology are introduced briefly. Then software engineering is also introduced with application of data mining technology in the enormous software engineering knowledge base.

Keywords-data mining; software development; software knowledge base; software engineering.

I. INTRODUCTION

With the development of society and scientific technology, computers, communication and Internet technology have permeated into various social fields, which is changing people's way of life. Application of various new technologies in computer field has generated, collected and stored large amount of data by various industries, which needs to be solved urgently. While bringing convenience for social industries, the massive information also brings lots of problems. First, information is too much to be digested. Second, the authenticity of information is difficult to be identified. Third, information safety can't be guaranteed. Fourth, forms of information are inconsistent, which needs to be unified. At present, database system can input, check and account data efficiently, but the potential relations and rules of data can't be detected, so the development trend can't be concluded. Thus, a new slogan is advocated, that is, we should learn to abandon information. Meanwhile, people begin to consider how to discover useful knowledge and promote the usage rate of information. Faced with the challenge of rich data but poor knowledge, data mining and knowledge discovery have emerged and developed vigorously with the new technologies of computer and new theories with strong vitality in various fields such as telecom, bank, biology, fraud and super markets.

Maintenance of software system is the most important stage of the life circle of software, which needs the longest time. In order to achieve a normal operation of software system in the practical complicated computer environment with constant changes, maintainers should correct mistakes and improve system timely and constantly so that software system can be operated normally. Maintenance of software has become the main means to extend the life span of software.

II. CONCEPT OF DATA MINING

Data mining is a process to explore interesting knowledge from mass data stored in the database, data warehouse or other storage base. It is widely believed that data mining is a process to explore potential, unknown, effective, original and useful knowledge from the massive, random, incomplete, noisy and vogue data so as to be understood finally. Sometimes data mining is also called knowledge discovery in database. KKD refers to discover connotative, unknown and useful information easy to be understood. The two concepts are basically the same but with different names in different fields. For example, academicians studying on artificial intelligence prefer to call it KDD, while experts in computer technologies usually call it data mining. There are different kinds of data in natural world. Besides the simple and common data like numbers and characters, there are also many complicated data, which mainly include spacial data mining, multi-media data mining,

timing sequence data mining, text data mining, web data mining as well as stream data mining and so on.

The process of data mining is very complex. For instance, when cleaning and preparing data, you may find that a certain data not applicable from isomeric data source, then that data should be processed in advance. Generally, new data without being mined before should be cleaned at first. See Fig .1.

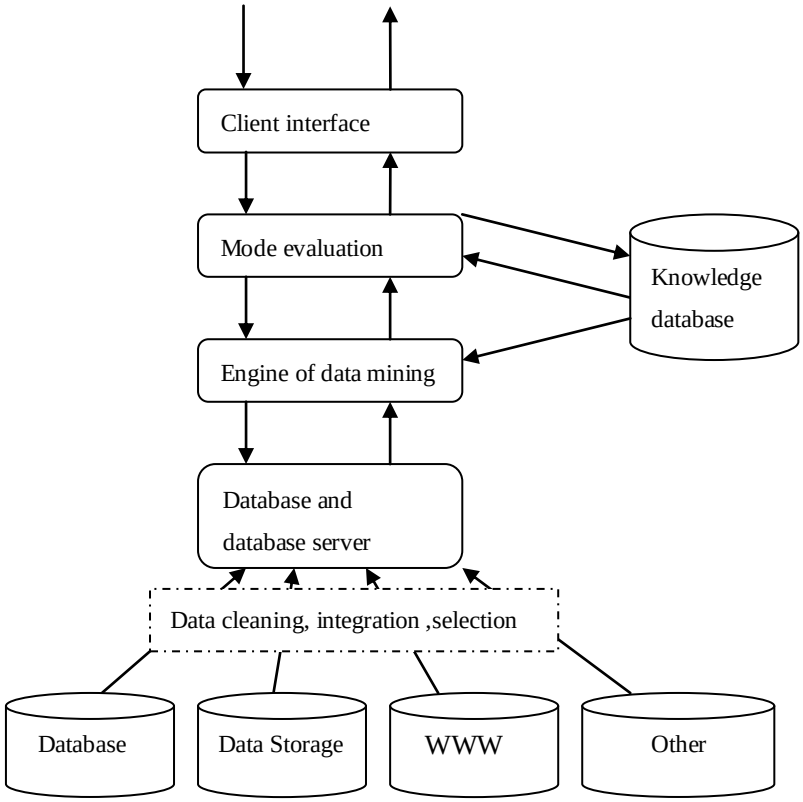

Figure 1. Common process of data mining

III. LIFE SPAN OF SOFTWARE ENGINEERING

In general, life span of software refers to the process of raising, realization, usage, maintenance and disuse of software products. That is to say, the whole period from the emergence of the concept to the end of its use is called the life span of software. Generally it contains feasible study and demand analysis, design, realization, test, commissioning as well as maintenance and so on. These activities can be repeated with replacement during execution. The life span of software has been divided into three stages as Fig .2, which are definition of software, development of software, and operation and maintenance of software.

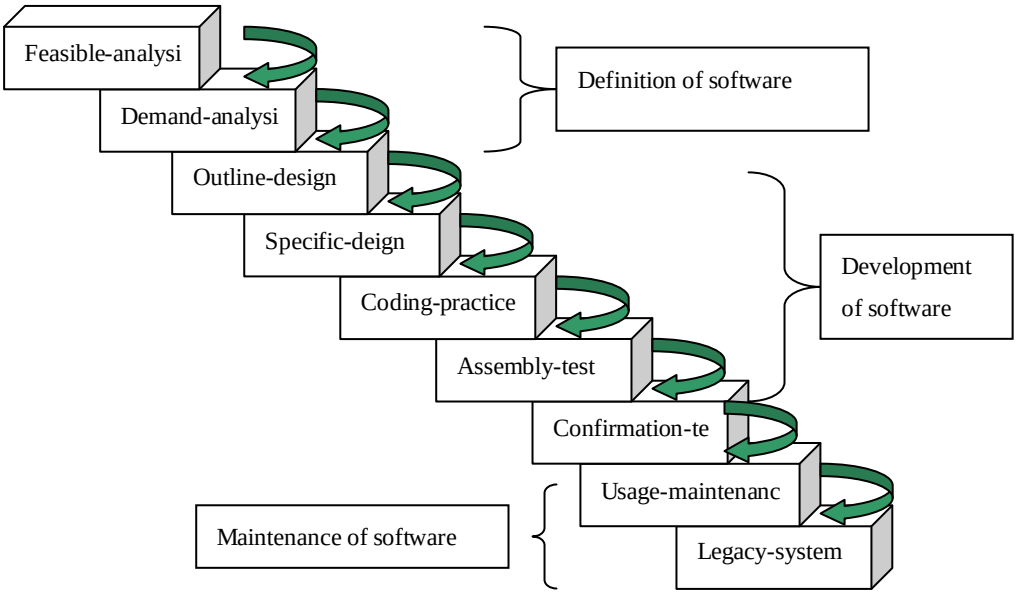

Figure 2. Life span of software system

IV. APPLICATION OF KDD IN SOFTWARE DEVELOPMENT

The system and structure of software system describes us a system structure with highly abstract levels but without much related description of various functions and models. In contrast, they are abstracted to a higher level such as subsystem. The interaction among subsystems should be recorded in the software design document since a sound design document can provide interactive understanding for the whole system structure and subsystems. However, it is a pity that few software designs have been recorded as document. Therefore, developers should try to understand the structure of software system from the perspective of source code.

Generally, software engineers understand the structure of software system from three steps, which are consumption, comparison and investigation (see Fig .3). The developers constantly repeat the three steps until they understand the whole software system. First, the developers will propose a conceptual structure based on their assumption and intuition of current software system, and they define the main components of software system as well as the interaction among different components. Then, developers will compare the conceptual structure with practical source code. Finally, developers will compare the investigation results with update of new knowledge from source code as well as their understanding of the system conceptual structure. This process will be executed repeatedly until the structure of software system is completely understood by developers.

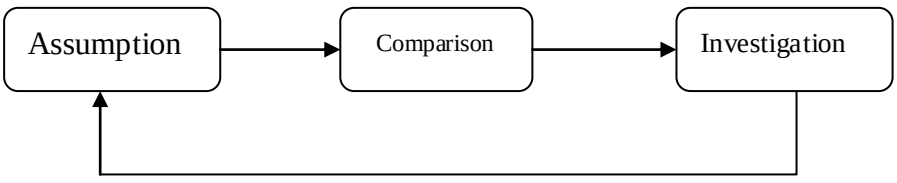

Figure 3. The process of understanding the structure of software system

V. CONCLUSION

The data mining technology is becoming more and more mature with an increase of data in the software engineering database, researchers have paid more and more attention to studying algorithm suitable for database mining. In the software database, there are information about the project's procedures, tasks' divisions and resources. Source code of software development as well as historical modification versions is also stored in the database. There are also documents of demands, designs and tests. Besides, there are also reports on the bugs. Study of data mining for software engineering is a hot topic in today's society. Many scholars are studying and finding potential valuable algorithm or technologies from the database by employing mature theories of data mining. By this way, they can provide powerful support for software developers and maintainers with mining rules or models, and meanwhile the cost for development and maintenance can also be lowered.

In this paper, it studies software database combined with data mining technologies with discovery of potential models for developers' better understanding so as to strengthen the stability of the system of short development span. Meanwhile, it will help maintainers to reduce the maintenance cost with timely discovery and solution of bugs and so on.

REFERENCE

[1]. Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques. (Machinery Industry Press) 2007.03.

[2]. L.Breiman,JH Friedman, RA Olshen and CJ Stone. Classification and Regression Trees. Wadsworth International Group,1984.

[3]. Hussein Almuallim, Thomas G and Dietterich. Learning With Many Irrelevant Features.Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI-91), July 1991.

[4].R.Agrawal, J. Gehrke, D. Gunopulos and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In Proc.1998 ACM SIGMOD Intl. Conf. Management of Data,pages94-105,Seattle,Washington,June 1998.

[5]. A.Arning,R.Agrawal and P.Raghavan.A liner method for deviation detection in large databases.In Proc.1996 Intl.Conf.Data Mining and Knowledge Discovery, Portland, Oregon, August 1996.

[6].V. Barnett and T. Lewis.Outliers in Statistical Data. John Wiley&Sons,1994.

[7].Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, et al. Identifying Density-Based Local Outliers. In: Proc.ACM SIGMOD 2000 Int. Conf. On Management of Data, Dalles, TX,2000.

[8].Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, et al. Identifying Density-Based Local Outliers. In: Proc.ACM SIGMOD 2000 Int. Conf. On Management of Data, Dalles, TX,2000.

[9].Edwin M. Knorr and Raymond T. Ng. Algorithms for Mining Distance-Based Outliers in Large Datasets. In: Proceedings of the 24th Interrmnational Conference on Very Large Data Bases, pages 392-403,1998.

[10].Roger S. Pressman, Software Engineering: A Practitioner's Approach (Fourth Edition),McGraw-Hill, 1997.

[11].Ian Sommerville, Software Engineering, Addison-Wesley, 1992.

[12].Dong Shao, Bin Luo. 2006, Design of the Curricula in Software Engineering. CEIS-SIOE 2006 Proceedings.