A Survey: Data Mining Techniques in Software Engineering

Inderjeet Singh
Computer Science department, Chandigarh University, Mohali, India

ABSTRACT

There are various software engineering activities and a lot of data mining techniques available. Main objective of this paper is to know data mining techniques that are used in software engineering. There are various steps in software development life cycle. All those activities has been discussed which could be related to data mining for their better performance. Various methods include the different techniques of data mining such as clustering, classification, association rule, etc. It has been found that the activities of software engineering had flaws. Flaws included complexity and vast amount of database. Complexity resulted in high cost in handing large database and it also resulted in more bugs. These bugs were harder to find. These flaws could be removed using data mining tools and techniques. Data mining tools, techniques and algorithms have been discussed which could increase the efficiency of software engineering activities. Application of this is that we can make software engineering process in a fast, cost effective and accurate way using data mining techniques.

Key words: Software Engineering; data mining; clustering, classification

INTRODUCTION

Software engineering refers to an approach for the development, proper working and later maintenance of software. Developers develop software. Software are mainly categorized into seven forms, i.e. System software, application software, engineering and scientific software embedded software, product-line software, web-applications, artificial intelligence software. As the time passes, software can deteriorate due to changes. Software is built custom. Soon after development of software, failure rate of software decreases a lot because of proper software engineering processes but with the passage of time it start increasing very slowly because of newer software or because of new viruses or because of newer needs which the current software couldn’t fulfill. With the passage of time there could be some side effects which increases the chance or can cause sudden failure of the software. At the starting of software’s life there is highest failure rate, which needs to be decreased. The failure curve shows that initially there is a very large chance of failure of software. It could be decreased in earlier phases using data mining techniques. It is simply because software engineering contains a large amount of database because of different software engineering activities. As the data is more, there are more chances of defects due to high complexity. Data mining can decrease it to a very small level. Data mining mainly refers to collection of important data from a large amount of database using various techniques such as clustering, classifications.
etc. Different data mining techniques can be employed using different data mining tools such as Weka, Rapid Miner, etc. Software engineering is itself a layered technology which includes tools to employ different methods, which helps in working of different processes so as to increase the quality of focus.

DATA MINING IN SOFTWARE ENGINEERING

Documentation process goes on throughout the software engineering. It results in creation of a very large amount of database. It is important to pre-process the data. Processing refers to working on data and pre-process refers to mining the important data and discarding the useless data. It results in beneficial data to be worked on. It can be done using data mining.

First of all there is Feasibility study. It includes initial study to know if software is feasible to make or not. To measure cost we can use clustering data mining technique to compare numerical data. Each cluster contains a different type of data items than the other clusters. [1] In software engineering clustering algorithms used are construction algorithms, optimization algorithms, hierarchical algorithms and graph-theoretical algorithms. For scheduling and requirement analysis, metaheuristic data mining technique can be used. Metaheuristic algorithms can be used to find near optimal solutions to a problem. In requirement, scheduling cost and effort estimation techniques like tabu search technique, simulated search technique and hill climbing technique can be used.

Second there is design phase. For it association rule of data mining can be used. It can be helpful in catalogue design decision making process.

Implementation phase includes coding. Text mining technique of data mining can be used to find duplicate code and bug duplicate reports. Dynamine is a data mining tools which can be used to find possible errors in a coding. It uses apriori algorithm.

Testing phase is used to find bugs, errors and other left changes to be made. Association rule can be used to work on large variable datasets to find bugs or severe defects. Classification technique using trees can track bugs.

Maintenance is the last phase of software development life cycle. It is also the longest phase. Classification technique of data mining can be used for maintenance of risky modules using
graphical diagrams. ROSE is a data mining tools which helps the developers to perform maintenance tasks by suggesting the changes and by predicting likely changes that can be made [2]; prevention of errors because of changes which are incomplete; program analysis for coupling and cohesion. It has been observed that software engineering include various activities; some software development life cycle activities have been shown in figure 1.

![Software Development Life Cycle](image)

Figure 1: Some basic activities in software engineering

A simple diagram showing how data mining used in software engineering has been shown in figure2. [3]

![Data Mining Techniques](image)

Figure 2: Mining activities in software engineering
Apart from these basic activities of software engineering there are also other activities which use data mining tools and techniques. It includes software configuration management, repositories, etc.

REPOSITORIES

Repositories contain a lot of database of different types. Software engineering contains different types of repositories, it includes following important repositories:

1. Source control repositories: - These are used to record development history of project.
2. Bug repositories: - These are used to track bug report history.
3. Archived communication: - These repositories track discussion on various aspects of SDLC. It includes emails, IRL, etc.
4. Deployment logs: - It records deployment execution process.
5. Code repositories: - It is used to archive source code for different kinds of projects.

To mine needed and accurate data in small amount of time from these extremely large databases, data mining performs a very crucial task. [4]

CONCLUSION

Software engineering is a vast field in itself. It contains a large number of activities. Diverse activities store a vast and diverse kind of data. Because of large amount of data there is large amount of complexity. It results in time consuming activities, which results in higher cost on services. More complexity results in higher amount of bugs. This could also result in loss of perfection because of some hidden bugs or repeated code. Repeated code will result in large storage space which further results in more cost. All these problems could be solved using data mining techniques. The method for the solution of these problems has been discussed. Various tools, techniques, algorithms and activities which involve data mining in software engineering have been discussed. This proves that data mining is a vital part of software engineering process if we want to perform work with higher perfection, cost effective and time saving way.

REFERENCES


