JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

A DATA MINING APPLICATION: ANALYSIS OF PROBLEMS OCCURRING DURING A SOFTWARE PROJECT DEVELOPMENT PROCESS

S.Nandhini Devi, Assistant Professor, Department of Information Technology, Dhanalakshmi Srinivasan College of Engineering and Technology

Dr. E.Mohan, Professor,Department of Computer Science and Engineering, Dhanalakshmi Srinivasan College of Engineering and Technology

ABSTRACT;

A Data mining techniques provide people with new power to research and manipulate the existing large volume of data. A data mining process discovers interesting information from the hidden data that can either be used for future prediction and/or intelligently summarising the details of the data. There are many achievements of applying data mining techniques in various areas such as marketing, medical, and financial, although few of them can be currently seen in software engineering domain. In this paper, a proposed data mining application in software engineering domain is explained and experimented. The empirical results demonstrate the capability of data mining techniques in software engineering domain and the potential benefits in applying data mining in this area.

Keywords: Data Mining; Software Engineering, knowledge discovery

INTRODUCTION;

Nowadays software projects keep growing in both scale and complexity. Improvement of the software product quality is a challenge for every software project leader. A project leader has to manage a project with several issues involved such as negotiation with customers, project planning and scheduling, code implementation, test and release. It is difficult for a project leader to precisely estimate the project duration before it starts. He can not possibly analyse all possible causes accurately if a problem is beyond his own background knowledge or previous experience. Data Mining (DM) techniques can assist software engineers to conduct the estimation and cause analysis of a project. A project leader can make the accurate estimation of a new project by learning from the information gained by applying data mining methods to previous projects. A leader can eliminate potential problems when a similar pattern appears in the current project to the one that caused problems in previous projects. Various DM techniques have been developed based on the works of Statistic, Artificial Intelligence, Machine Learning and Database system for knowledge discovery. DM is the process of facilitating decision-making by identifying valid, novel, potentially useful, and ultimately understandable structures in data [7]. Usually this kind

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

of extracted knowledge is classification rules, characteristic rules, association rules, functional relationships, functional dependencies, causal rules, temporal knowledge or clusters according to the chosen DM technique and operation. DM techniques can analyse many kinds of data such as relational, object-oriented, text, temporal, spatial, combinatorial, web and multimedia. Data mining tools and applications have generated positive results, and are continuously stimulated by exploring new areas due to the benefits brought by this technology. There are many achievements of applying data mining techniques to various areas such as marketing, medical, and financial, although few of them can be currently seen in software engineering domain. There exist several difficulties in this domain, such as hard to find a data model to put through mining process and/or no suitable mining tools. This brings out the need to investigate the efficacy of data mining techniques applied in the software engineering domain. Data Mining and Software Engineering Domain 3 This paper explores the software engineering domain by applying DM techniques to a set of data collected from a software project process within a software development company. This paper starts by introducing a real software engineering problem, and data mining. After a data model isestablished for the underlying problem, data mining techniques and softwares such as CBA [1], C5.0 [2] and TextAnalyst [3] are introduced and experimented. Interesting findings from different data mining operations are discussed together with issues appearing during the mining process. Finally some suggestions are given about the future applications within software engineering domain.

DATA MINING PROCESS: DEFINING GOALS ;

Two areas of interest that may be suitable for mining tasks are identified. The first one appears at the early Estimation and Planning stage of a software project. To ensure software products quality, MASC engineers need to make estimations on many aspects of a project. This includes (1) the number of lines of code to be developed, (2) the kinds of document to be delivered to customer, and (3) the time required to accomplish each software engineering stage for the project. Currently, although there are several tools that can help a programmer to do the implementation jobs during the software design stage, but there is few or even no tool that can be used on both estimation and project problem reasoning stage. Project team members can only give estimations based on their own experience from previous projects. If the current project is not within their familiar topics, theaccuracy of the estimation may be worse. A PR (problem * MASC is a division of a global telecommunication company. Due to security reason, detail of the information source is removed. report) fixing work may become a time consuming task when a problem appears but the responsible person can not estimate the time to fix the problem. This directly affects the success of a software project. Finding a precise estimation figures on bug fixing or estimation work at the early stage of a project will bring great cost savings and accurate progress control to the whole development team as well as to the

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

whole organization. The second problem relates to the GNATS itself. Currently the system has limited ability of information retrieval. There is no actual database management system being implemented. If a PR is closed, it is just statically stored in GNATS and no further analysis is performed. This limits the potential benefits to software engineers who can obtain valuable information if the existing PR data is being analysed. This will be useful especially when a programmer is struggling with a bug while a resolution may already hide behind the knowledge that can be derived from the previous similar problems. The above two problems can be relieved by utilising data mining techniques. The format of a PR data is semi-structured, and every field of it contains a simple type of data, such as short text. These features make the PR data set an ideal candidate for mining task. Mining results of the PR data will bring benefits such as accurate project estimation and planing, improved control over the PR fixing and deduction of the project cycle time.

DATA FIELD SELECTION;

There are several fields such as Confidential, Submitter-ID, Environment, Fix, Release Note, Audit Trail, the associated project name and PR number that are ignored during mining. These fields provide identification information about a PR, but they contain no data mining value. Instead, some of these values are used as support roles during pre-processing and post-processing stages to assist in the selection of data and a better understanding of the rules being found.

Data Mining and Software Engineering Domain 7 The aim of this data mining exercise is to find out useful knowledge from existing projects, all the existing projects should already be finished and all the corresponding PRs should also be closed. Otherwise, a PR can still be changed and is not stable for mining. Hence all PRs with a closed value in their State field are chosen.

Whenever a PR is raised, a project leader will have to find answers for the following questions before taking any action: How long it will take to fix? How many people were involved? How sever the problem is (customer impact)? What is the impact of the problem on project schedule (Cost & Team priority)? and What type of the problem it is (a Software bug or a design flaw)? According to these, attributes such as Severity, Priority, Class, Arrival-Date, Closed-Date, Responsible, Synopsis (table 1) are found important to describe the characteristics of a PR. The first three attributes describes how a PR is handled within a project, the next three attributes indicate how long a PR is fixed and who were responsible, and the last one lists the content in a PR. The attribute ‘class’ is chosen as the target attribute in order to find out any valuable knowledge among the type of a problem and the rest of the PR attributes. For example, when project leader comes to know of the relationship between the fix effort (time) and the PR class,

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

he can analyse the fix effort versus the human resources available and put it in the schedule and resource plan.

DATA CLEANING;

Data is further investigated to identify problems, such as missing values, inconsistent values, and mistaken values using graphical tools such as histogram for frequency distribution of the values or calculating maxima, minima and mean values. Histogram plots the contribution made by each value for the (categorical) attribute, and therefore helps to identify distribution skews and invalid values. The occurrence of these problems comes with several factors such as human mistakes and evolutions of the GNATS system. An example is the use of different terminologies over the time such as SW-bug or sw-bug as an input value for Class field (Example a, d in Figure 2). A Time-Zone field and other new input values have been added later in the system on management request based on feedback of users after several years of system running.

In unrecoverable group, the error cannot be recovered precisely and the PRs with error(s) have to be discarded. An example (Example a in Figure 2) shows that it has its closed time even earlier than the time being raised. Some PRs even lost the part of time related values, such as Example c in Figure 2. An example of inconsistent values shown in Figure 2 is the Time-Zone field that was only added in 1998, there was no input for the Time-Zone field in a PR recorded before 1998.

DATA TRANSFORMATION;

Data transformation is considered, in the way of converting attributes Arrival-Date and Closed-Date to a time-period - identifying the time spent to fix a PR - taking account the additional information Time-Zone and Responsible. This transformation resulted in the Time-to-fix attribute with continuous values. The Responsible attribute has the information about personnel engaged in rectifying the problem. We assume that the derived attribute Time-to-fix is

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

total time spent to fix a problem if there is only one person involved.

```
    Severity |Priority| Time-to-fix| Class |Synopsis

a. serious, high, 61, sw-bug, STI STR register not being reset at POR
b. serious, high, 56, support, sequence_reg varable in the RDR_CHL
   task is not defined
c. serious, low, ?, doc-bug, In URDRT2 of design doc, the word 'last'
   should be 'first'
```

During the data cleaning stage, all the recoverable errors are corrected after identification. The PR data is then transferred into several data files to meet different data mining operation formats shown in Figure 3. For unrecoverable errors, the corresponding PR records are either discarded or put in further process if the error can be ignored or irrelevant with a specific data mining operation. Example c in Figure 3 has a '?' symbol as its Time-to-fix value, it is unrecoverable error but is still available for text mining since its Synopsis field provides correct information.

```
PR_ID|Category|Severity|Priority|Class|Arrival-Date|Close-Date|Synopsis

a. 17358|bambam|serious|high|sw-bug|20:50 May 25 CST 1999|11:35 Mar
   24 CST 1999| STI STR register not being reset at POR

b. 17436|bambam|serious|high|support|18:10 Mar 30 CST 1999|12:00 May
   24 CST 1999| sequence_reg varable in the RDR_CHL task is not
   defined
…
c. 580 |bingarra|serious|low|doc-bug|10:10 May 31 May 1996|  | In
   URDRT2 of design doc, the word 'last' should be 'first'
…
d. 6205 |gali|serious|medium|SW-bug|14:30 Nov 5 1997|13:14 Dec 1
```

After data pre-processing, four attributes (time-to-fix, class, severity and priority) are chosen for mining process. Out of total 40000 PRs initially selected as the data set, we are left with 11,000 PRs after choosing only the PRs with state field as closed and after applying pre-processing steps. These 11,000 PRs (depends on different mining tasks, the numbers varies a little) cover more than 120 projects within MASC from 1996 to 2000. For example 11364 PR records have been applied with text-mining tools on the valid values in their Synopsis fields, as 364 records have no time values so could not be applied with classification tools.

DATA MINING PROCESS: DATA MODELLING;

The GNATS system provides some simple methods to retrieve basic information from the PR data set, such as the PR numbers related to a particular person, etc. Besides this, some general data base techniques, such as SQL can also give some more useful information, i.e., the

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

average time spent for fixing a PR in a project. But all these methods cannot perform in situations like performing queries over a large number of records with high dimensional structures,summarising a large data set to facilitate data based on the existing rules, and visualising simplified extracted local structures. On the contrary, data mining techniques perform well on above situations and are able to reveal the deeper characters of the PR data. Some questions can only be answered after applying data mining techniques on the data set, such as:

• What type of project documents that needs a lot of time to have the development team fix its associated bugs in a project?

• If a PR being raised, how long should it be fixed according to the contents/values in its Synopsis, Severity, Priority and Class? The selection of data mining operations and techniques is one of the most important things that directly affect the progress and the accomplishment of any DM applications. We have chosen:

• Prediction modelling on the time consuming patterns of the PR data and helping a project team to make estimation.

• Link analysis to discover association among the contents/values of the variables being selected.

• Text Mining to analyse Synopsis field.

TEXT MINING IN PR DATA;

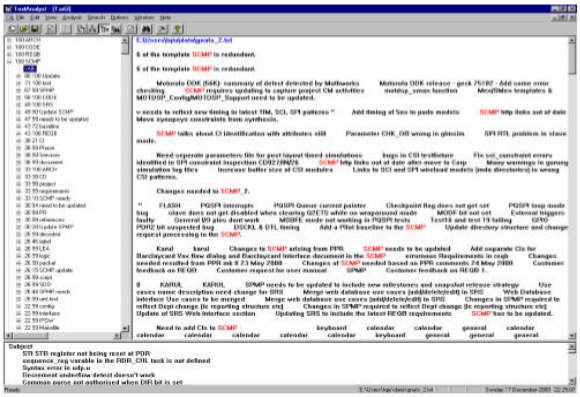
In order to find valuable knowledge from thousands lines of text, we categorise the pure text into several document types based on certain background knowledge. For example, the following are PR values copied directly from the original PR data set. serious high change-request 3.95 "SRS - missing requirements "

Data Mining and Software Engineering Domain 17 serious high change-request 5.05 “SRS - ability to turn off sending of SOH serious high change-request 5.92 “RT_014_SRS_3.0.0” serious high change-request 6.7 “SRS - misc changes needed” serious high change-request 15.71 “Changes needed to SCMP_2.0.0” The pure text value of synopsis field appears on themost right side between two double quotes in PRs. It can be observed that ‘SRS ‘ document (Software Requirement Specification) is the main target in 4 PRs, but SCMP (Software Configuration Management Plan) only appears once. To reveal the deep character of the PR data, it is not enough to only investigate what happens if the class, time or priority of a PR changes. It is also necessary to study what happens if the PR talking about different sources. Based on the above example, the analysis can reveal that SRS document costs more time than a SCMP document when fixing PRs. The question is: can this information be a valuable rule and apply in general? To answer this kind of question, we need a method and a tool that is able to

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

automatically summarise the pure text data and extract some valuable rules. TextAnalyst tool has been used to do the text-mining job in this research. It builds up a semantic network for the investigation over the PR data. Each element of the semantic network is characterised by a weight value and a set of relationship of this element to other elements in the network. Every relationship between elements is also assigned a weight value. The semantic network can then provide a concise and accurate summary of the analysed text. The good thing is that, TextAnalyst does not need the user to specify any predefined rules for

building the semantic network. It can automatically create the semantic network based on the structure, vocabulary and volume of the analysed text. As seen from the simple graphical user interface window in TextAnalyst (Figure 4), a semantic network tree of the PR data is shown in the upper-left window. The network contains a set of the most important words or word combinations, called concepts, extracted from the PR data set. The relations among those concepts together with the semantic weights of concepts and relations are also shown. The values of the weights range from 0 to 100, which correspond to the probability that the associated concept is characteristic for the whole PR data. Let us take the second concept “test” below “SCMP” as an example. “71”is the weight of the relation between the concept “test” and its parent concept “SCMP”. “100” is the semantic weight of “test” itself. “+” sign means “test” node can be further opened to view related concepts in the network, and the “-“ sign beside “SCMP” means the “SCMP” node is already opened. The upper-right window shows all the text related to a concept if the user clicks the corresponding concept on the upper-left window with the text displayed in red colour. Currently all pure text value related to concept “SCMP” are shown in the window. The third window on the bottom contains the whole value of synopsis in pure text form.



Text mining is applied on a total 11226 cases. An interesting relation is researched with “SCMP”, Software Configuration Management Plan, a support document in every MASC project. Since SCMP is not a main design document for a project with just tens pages, it has never been considered as a trouble making item. But amazingly, it is clear from the figure that SCMP has 71 percentage probability of Data Mining and Software Engineering Domain appearing in test related PR records, and 58 percentage

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

probability of appearing in Code related PR records. This result is even higher than the result associated with SRS ("Software Requirements Specification", a main development document directly related to software). Although we can not say SCMP causes more problems than SRS, but the higher appearance of SCMP inside test related PR definitely shows a warning. It is worth for software engineers to be more careful when dealing with SCMP document, and hence reducing the total cost of fixing SCMP related PRs. Another analysis shows that a test related PR has even a higher weight (36, 100) in document related bugs than a SRS related PR (35, 99) does. Which suggests a better project management should not only focus on the quality of product related documents, but also pay attention on the quality of testing related documents. In general, TextAnalyst is a useful tool to perform analyses on data in pure text format. Its automatic mechanism of semantic network creation is very attractive, and saves time for data preparation.

EXISTING PROBLEMS IN PERFORMING MINING;

The error rates of testing data sets in both CBA and C5 are higher than expected. Although several approaches are attempted to reduce the error such as uniform distribution of values, cross validation, boosting, different size of training set, etc. Unfortunately, the average error rate is only fallen down by 5% from 47% to 42%. The best result is 9% decrease from 46% to 37%. These results indicate that some amount of noise is still existent in data after dealing with the noise during the pre-processing phase. For example, the relationship between PRs and human resources within a particular project plays a great impact. The time needed to fix a bug is different depending upon the actual human resources available. We have used only the attribute 'Responsible' to indicate the human resource available. Truly, the relationship with the human resources available for past projects whose data was analysed is needed to use time patterns to help project leaders to predict time consummations more accurately. The use of additional data source 'Change Request data set' that records all customer request process data may rectify this problem.

FUTURE DIRECTIONS AND CONCLUSION;

During the life cycle of a software project development, there are many problems such as negotiation with customers, project estimation, project planning and scheduling have to be dealt efficiently. Resolutions to these problems are time consuming and costly. This paper considers the use of data mining techniques to analysethese problems, and further find valuable rules to reduce the effort of fixing those. The data mining result may bring some relief of the difficulties in planing, estimation and bug fixing activities for software project teams. The time patterns rules may help a project leader to understand knowledge in numeric values about the PR fixing process, the leader can then estimate or predict time consummations more accurately than before. Another finding suggests that bug fixing efforts have more probabilities to be spent on test related PR. This could cost the project team a lot of time in fixing non-product-related problems. By reducing such type of PR, the total time in bug fixing is then reduced and project efficiency

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

will be improved. Several data mining operations are executed on a data set collected from the software engineering process under a real software business environment. Some useful rules and background knowledge are extracted on the time patterns of the PR fixing and the relationship between the content and the type of a PR in the form of association rules, classification rules or semantic trees. Results of the application indicate that data mining techniques are capable in software engineering domain even though the scale of the data mining task is limited. As in many other domains, the benefits and capabilities brought by data mining in software engineering domain areworth of further investigations.

REFERENCES ;

1. CBA, http://www.comp.nus.edu.sg/~dm2/ 2. C5.0, http://www.rulequest.com/see5-info.html 3. TextMiner, http://www.megaputer.com/company/index.html 4. Tlearn, http://crl.ucsd.edu/innate/tlearn.html 5. P.Cabena, P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi, “Discovering Data Mining: From Concept to Implementation”, ISBN 0-13-743980-6, 1997 6. H. Edelstein, “Mining for gold. (Selecting data mining tools)”, Information Week, April 21, 1997 n627 p53 (6). 7. U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “'From Data Mining to Knowledge Discovery: An Overview”, In Advances in Knowledge Discovery and Data Mining. Eds. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAI Press, 1996 8. J. Han and M. Kamber, “Data mining: concepts and techniques”, ISBN 1-55860-489-8, 2001 9. C. Westphal, and T. Blaxton, “Data Mining Solutions: Methods and Tools for Solving Real-World Problems”, ISBN 0471-253847, John Wiley & Sons, Inc, 1998. 10. J Furnkranz and M. Kamber, “First-Order Knowledge Discovery in Database”, In Applied Artificial Intelligence, 0883-9514/98, 1998