A Study of Service Center Records Using Data Mining

Han Tong Loh, Wee Leong Koh, Rakesh MENON, Christopher K Leong

***Abstract*-- In many manufacturing companies, large databases containing data on failures and repairs of products are maintained. This paper reports on a case where information from such a database is extracted. Such information extracted could be useful for identifying opportunities for reliability improvements in the products. The database was obtained from a manufacturer of inkjet printers and contains both fixed-format and free-form text fields. At present, techniques developed in data mining and information retrieval are mainly designed to handle either fixed-format or free-form text fields, but not a combination of both. The approach taken in this paper is to first transform the free-form text fields into a number of fixed-format fields through analysis of the frequency of key words. Association analysis was then carried out on the resulting fixed-format fields. Results obtained from the analysis produced a number of associations that could contribute to improving product reliability.**

I. INTRODUCTION

It has been estimated that the amount of information in the world doubles every 20 months [1]. The size and number of databases probably increases even faster. The field of engineering is not exempt from this proliferation of data. The increasing computerization of all engineering endeavors has helped fuel this tremendous growth. From production data and experimental results to CAD/CAM designs and blueprints, from customer feedback and warranty information to published articles and research papers, the amount of data available to the modern engineer has never been so extensive, so complete, and so utterly overwhelming. The sheer volume of data being held in databases is now so large as to defy manual analysis.

Within this mass of data, it is statistically certain that there is at least one gleaming nugget of knowledge that will boost yields, improve quality, or revolutionize designs. The problem lies in finding it. Data mining offers a way of extracting such information.

One application is to use data mining to extract useful information from a warranty database. The warranty database in question was obtained from a multi-national corporation (MNC). The database is a collection of warranty repair information from service centers located worldwide. It contains records of repair actions, customer complaints and individual product details of inkjet printers. The database grows at a rate of several thousand records per month. It is hoped that a data mining analysis of the database would produce actionable patterns that can lead to improvements in product reliability.

The database is a hybrid of fixed-format fields and free-form text fields. Fixed format fields are fields that have strict formatting criteria for the type, range and precision for its contents. The “NRIC_Number”, “Sex” and “Date_of_Birth” fields in a personal particulars database are examples of fixed-format fields. Free-text fields are unstructured with no formatting requirements. The “Abstracts” field in an article database is a good example of a free-text field.

Most techniques used in data mining are only capable of dealing with fixed-format data, whilst information retrieval techniques are designed for free-form text. As yet, there is a lack of work on hybrid database that contain both types of data. One of the few works in the field is the project by Tan et al [2]. They state, “advanced knowledge discovery technologies have been developed in both research areas, but systems that can categorize or cluster records containing both kinds of data are still lacking.” In their work, they used feature extraction and clustering techniques from information retrieval and with classification algorithms from machine learning to categorize a free-text field in a hybrid database. They then incorporated the results of this categorization into their classification scheme for the entire record. The results show that by incorporating free-text information, the classification model can potentially be improved.

It is hoped that a method can be developed to assist in the analysis of the warranty database, both to improve the quality of information obtained from the database as well as to improve the efficiency of the analysis.

The approach taken is first to use keyword identification to

[1] Han Tong Loh, is with the Innovation in Manufacturing Systems and Technology(IMST), Singapore-MIT Alliance (SMA), N2-B2c-15, Nanyang Technological University, Nanyang Avenue, Singapore 639798

[2] Wee Leong Koh, student with the Innovation in Manufacturing Systems and Technology(IMST), Singapore-MIT Alliance (SMA), N2-B2c-15, Nanyang Technological University, Nanyang Avenue, Singapore 639798

[3] Rakesh Menon, is with Centre for Robust Design, Department of Mechanical Engineering, National University of Singapore,10 Kent Ridge Crescent, Singapore 119260

[4] Christopher K.Leong, is with the Innovation in Manufacturing Systems and Technology(IMST), Singapore-MIT Alliance (SMA), N2-B2c-15, Nanyang Technological University, Nanyang Avenue, Singapore 639798

transform the free-form text fields into fixed-format fields. The new set of fixed-format fields so generated can then be analyzed using association analysis. It is hoped that such analysis would generate high quality and actionable rules that would be able to improve the reliability of the printers, either through the manufacturing or design processes.

The initial results from the analysis were satisfactory. By analyzing the rules generated, it was possible to gain a better insight into the reliability of the products.

II. Overview of DATA MINING Procedure

In this section, an overview of the methods used to understand, prepare and analyze the database is presented. The process can be divided into 4 phases: data extraction, data pre-processing, association analysis and evaluation of rules generated. Each phase will be described in detail below.

A. Data Extraction

The data used for analysis is obtained from the service center warranty database. This database records all request for repairs carried out at the service centers. The main purpose of the database is to maintain transaction records for repair actions so to carry out billing and accounting purposes as well as to control the inventory of spare parts. The fields relevant to this reliability analysis are extracted from this database. This is a hybrid database, containing both fixed format and free-text fields:

Fixed-format Fields – These include the month and year of repair, the product model number, product serial number, repair office number and part number of replaced part.

Free-text Fields – There are two free text fields. They are the "*Customer_Comment_Tx*" field, which records the complaints that the customer has about the product, and the "*Repair_Details_Tx*" field, which records the actions taken by the technician attending to the customer.

Data was extracted from the company's database and stored in Microsoft Excel format for further analysis.

B. Data Preprocessing

This phase involves transforming the free-form text fields into fixed-format fields as well as carrying out transformations on the fixed-format fields.

For the fixed-format fields, four tasks were carried out: aggregating the repair office and the parts changed fields, decoding the serial number field and calculating the time to failure. For the free-text fields, the approach taken is to associate a number of categories or groups to each free-form text record. The groups are identified by the occurrence of keywords in the free-form text record. Therefore two steps are required. First, the keywords and groups are identified from the database. Second, each field is associated with the keywords present in the field.

Before the identification of keywords and groups can be carried out, the free-text fields must first be cleaned. This is done to reduce the number of terms that would need to be analyzed in subsequent steps. It involves the removal of the following types of terms: purely numerical terms, punctuations, except for apostrophes, and combination alphabetic and numeric terms.

Another precursor to the identification step is to determine the frequency of occurrence in the free-text fields of the database of 1-word, 2-word and 3-word phrases. These phrases form the basis for the identification of the keywords and their frequency of occurrence in the database will help to determine their importance as keywords.

The next step is to determine new keywords and groups. A huge number of phrases were identified in the previous step. This needs to be screened with phrases that have less than 0.01% support being eliminated. Phrases were then selected based on the relevancy of the phrases to the failures reported or the repair actions.

The next step is to associate each record to the keywords and groups. As many keywords as are present in each field should be identified so as to fully describe the information within the record.

C. Association Analysis

Association analysis, or association rule mining, finds interesting association or correlation relationships among a large set of data items [3]. Through association analysis on the transformed database, it is hoped that useful rules with high quality and actionable information can be obtained. These can then be applied to improve the manufacturing processes or to modify the product design.

The association analysis was carried out using Enterprise Miner, the data-mining module of SAS Institute Inc's SAS System. The fields of interest were extracted from the database and stored as flat files for processing using Enterprise Miner. For the purpose of the association analysis, each record with the extracted fields was considered to be one transaction. Thus the information contained in one transaction is heterogeneous, as opposed to the classic association problem where only one type of information is used for the analysis.

The following analysis were carried out:

- Association with all fields in the Complete Warranty Database
- Association between Parts for one repair
- Association between Country and Part Numbers for each repair
- Association between Customer Comments and Repair Details
- Association between Customer Comments, Repair Details and Country
- Association between Customer Comments, Repair

Details and Part

- Association between Part and Date of Manufacture
- Association between Part and Time to Failure

D. *Evaluation of Rules Generated*

To evaluate the rules generated is to determine which rules are of interest. This is determined by the levels of support and confidence that the rule has. Using the subset table function in the association node, it is possible to filter the rules by confidence and support to obtain only those rules with high support and confidence. These filtered rules are referred back to the MNC's engineers for further evaluation. With their knowledge of the function of the printers and the warranty repair procedures, the engineers were able to sieve through the numerous rules and evaluate whether the rules are useful.

III. Results

At the beginning of the project, it was hoped that a data mining analysis of the warranty database would generate some rules that might prove useful in improving product reliability. After evaluating the rules with the engineers from the MNC, some interesting ones were found. However, the majority of the rules were trivial, as they contained knowledge that was already obvious to the engineers.

Most of the interesting rules were from associations carried out between small numbers of fields. The better-interpreted results tend to be from analyses utilizing a small number (two to three) of fields, for example, between parts for one repair, or between parts and country. When the analysis is carried out with more fields, the results are difficult to analyse. This is due in part to the huge number of associations being generated as well as it being difficult for an analyst to intuitively interpret the correlations between items in several different fields.

For similar reasons to those in the situation described above, the number of items in a rule was limited to four. This reduced the number of rules generated because rule generation increases exponentially with the number of relations. This is more important when using many fields in the analysis because, generally, having more fields in the analysis means that there are more items in each transaction.

The engineers at the MNC already carry out a simple form of association analysis using Pivot Tables in Microsoft Excel. But using Pivot Tables requires the analyst's input at every step as this method can only associate one item with items in another field - for example, one country, Singapore, with all parts changed in Singapore. Association Analysis in data mining, on the other hand, can generate all association rules present in the database, subject to confidence and support constraints. Even for those cases which involve only a few fields, and therefore lend themselves to manual analyses, using Association Analysis makes it possible for all items in the fields to be inter-associated. In this way, interesting rules can be located faster and more thoroughly. All these accentuate the power of Association Analysis using a data mining software.

IV. Conclusion

The study of service centre records from a manufacturing firm using data mining resulted in the extraction of knowledge, in the form of association rules, on the firm's products. It is expected that the results obtained could contribute to product quality and reliability improvements. The method of transforming free-form text fields into fixed-format fields is one way of addressing hybrid databases that contain both fixed-format and free-form text fields. This method is applicable to many other such instances of extracting information from such hybrid databases.

References

[1] W J Frawley, G Piatetsky-Shapiro, C J Matheus; Knowledge Discovery in Databases: An Overview; Knowledge Discovery in Databases 1-27; AAAI Press, 1991

[2] Pang-Ning Tan, Hannah Blau, Steve Harp, Robert Goldman; Textual Data Mining of Service Center Call Records; Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 417-423

[3] Jiawei Han, Micheline Kamber; Mining Association Rules in Large Databases; Data Mining: Concepts and Techinques 225-277; Academic Press, 2001