Application

Data mining for customer service support

S.C. Hui*, G. Jha

Nanyang Technological University, School of Applied Science, Nanyang Avenue, Singapore 639798, Singapore

Received 9 August 1999; accepted 24 November 1999

Abstract

In traditional customer service support of a manufacturing environment, a customer service database usually stores two types of service information: (1) unstructured customer service reports record machine problems and its remedial actions and (2) structured data on sales, employees, and customers for day-to-day management operations. This paper investigates how to apply data mining techniques to extract knowledge from the database to support two kinds of customer service activities: decision support and machine fault diagnosis. A data mining process, based on the data mining tool DBMiner, was investigated to provide structured management data for decision support. In addition, a data mining technique that integrates neural network, case-based reasoning, and rule-based reasoning is proposed; it would search the unstructured customer service records for machine fault diagnosis. The proposed technique has been implemented to support intelligent fault diagnosis over the World Wide Web. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Data mining; Knowledge discovery in databases; Customer service support; Decision support; Machine fault diagnosis

1. Introduction

Customer service support is becoming an integral part of most multinational manufacturing companies that manufacture and market expensive machines and electronic equipment. Many companies have a customer service department that provides installation, inspection, and maintenance support for their worldwide customers. Although most of these have some engineers to handle day-to-day maintenance and small-scale troubleshooting, expert advice are often required from the manufacturing companies for more complex maintenance and repair jobs. Prompt response to a request is needed to maintain customer satisfaction. Therefore, a hot-line service centre (or help desk) is usually set up to answer frequently encountered problems from the customers.

Fig. 1 shows the workflow in a traditional hot-line service centre. The service centre is responsible for receiving reports on faulty machines or enquiries from customers via telephone calls. When a problem is reported, a service engineer will suggest a series of checkpoints for customers using the hot-line advisory system. Such suggestions are based on past experience. This has been extracted from a Customer Service Database, which contains previous service records that are identical or similar to the current problem. The customer can then try to solve the problem and subsequently confirm, with the service centre, if the problem is resolved. If the problem still persists, the centre will dispatch a service engineer to the customer’s premise for an on-site repair. During such trips, the service engineer will take past records of the customer’s machine, related manuals, and spare parts

---

* Corresponding author: Fax: +65-792-6559.
E-mail address: asschui@ntuvax.ntu.ac.sg (S.C. Hui).
that may be required to carry out the repair. Such a process is inconvenient.

At the end of each service cycle, a customer service report is used to record the new problem and the proposed remedies or suggestions taken to rectify it. This database is used for billing purposes, as well as for maintaining a corporate knowledge base. The service centre stores the customer service report in the database.

Apart from maintaining a knowledge base on common faults and its remedies, the customer service database also stores data on sales, employees, customers and service reports. These data are not only used for day-to-day management operations, but help the company in decision making on job assignment and promotion of service engineers, and marketing, manufacturing, and maintenance of different machine models.

The customer service database serves as a repository of invaluable information and knowledge that can be utilized to assist the customer service department in supporting its activities. The objective of this paper is to discuss how to apply data mining techniques to extract knowledge from the customer service database to support two types of activities: decision support and machine fault diagnosis.

The work was carried out as a collaborative work between a multinational company and the School of Applied Science, Nanyang Technological University, Singapore. The company manufactures and supplies insertion and surface mount machines for use mainly in the electronics industry.

2. Data mining

Data mining, also known as knowledge discovery in databases (KDD) [7,9], is a rapidly emerging field. This technology is motivated by the need of new techniques to help analyze, understand or even visualize the huge amounts of stored data gathered from business and scientific applications. It is the process of discovering interesting knowledge, such as patterns, associations, changes, anomalies and significant structures from large amounts of data stored in databases, data warehouses, or other information repositories. It can be used to help companies to make better decision to stay competitive in the marketplace. The major data mining functions that are developed in commercial and research communities include summarization, association, classification, prediction and clustering. These functions can be implemented using a variety of technologies, such as database-oriented techniques, machine learning and statistical techniques [10].

Recently, a number of data mining applications and prototypes have been developed for a variety of domains [4] including marketing, banking, finance, manufacturing and health care. In addition, data mining has also been applied to other types of data such as time-series, spatial, telecommunications, web, and multimedia data. In general, the data mining process, and the data mining technique and function to be applied depend very much on the application domain and the nature of the data available.

3. Customer service support

Service records (or reports) are currently defined and stored in the customer service database. Each service record consists of customer account information and service details, which contain two types of information: fault-condition and checkpoint information. The former contains the service engineer’s description of the machine fault, while the later indicates the suggested actions or services to be carried out to repair the machine, based on the actual fault-condition given by the customer. Checkpoint information contains checkpoint group name, and checkpoint description, with priority and an optional help file. The checkpoint group name is used to specify a list of group checkpoints. Each checkpoint is associated with
a priority that determines the sequence in which it can be exercised and a help file that gives visual details on how to carry out the checkpoint. An example of a fault-condition and checkpoint information for a service record is given in Fig. 2.

In addition, the customer service database also stores data related to sales, customers, and employees: six major tables are defined in the customer service database for this. Two, namely, MACHINE_FAULT and CHECKPOINT, are used to store the knowledge base on common machine fault-conditions and their checkpoints. These are unstructured textual data. The remaining four tables are used to store information on customers (CUSTOMER), employees (EMPLOYEE), sales (MACHINE) and maintenance (SERVICE_REPORT). These four store only the structured data. There are over 70,000 service records. Since each of the fault-conditions has several checkpoints, there are over 50,000 checkpoints. Information on over 4000 employees, 500 customers, 300 different machine models and 10,000 sales transactions are also stored.

### 3.1. Mining structured data

A list of most popular data mining tools available commercially or in public domain is given in the KDNuggets website [17]. These tools can mine the structured data of sales, maintenance, and particulars of employees and customers in the customer service database. It is interesting to see that a number of tools support multiple approaches; i.e., more than one data mining techniques. For example, Darwin from Thinking Machine Corp. supports neural networks, regression tree (CART), k-means algorithm, and case-based reasoning for classification, prediction, and clustering functions. There are also some tools that only aim at a specific data mining function. This provides flexibility; the users can select different data mining tools for their problem domains to achieve the best results.

The choice of data mining tool must be based on the application domain and its supported features. Certain applications may require only one data mining function; others may require more than one. In this research, DBMiner [12] was chosen. This system was developed by the DBMiner Research Group from the Intelligent Database Systems Research Laboratory at Simon Fraser University in Canada. The system, which integrates data warehousing, on-line analytical processing (OLAP) [6] and data mining techniques, supports the discovery of various kinds of knowledge at multiple conceptual levels from large relational databases. The DBMiner system supports most of the major functions. It was implemented using many advanced data mining techniques. In addition, it provides multidimensional data visualization support and interacts with standard data sources through open database connectivity (ODBC) interface.

### 3.2. Mining unstructured data

Although DBMiner is an excellent data mining tool for large databases with structured data, it is unsuitable for extracting knowledge from the textual data of the customer service database. As the information or knowledge on common faults and their suggested remedies are stored in textual format as fault-condi-
tions and checkpoints, new techniques are needed to extract knowledge from this database for machine fault diagnosis. This is known as text mining [2,19,22].

Traditionally, case-based reasoning (CBR) has been successfully applied to fault diagnosis for customer service support [20,23,25]. CBR systems rely on building a large repository of diagnostic cases (or past service reports) in order to circumvent the difficult task of extracting and encoding expert domain knowledge [1]. It is one of the most appropriate techniques for machine fault diagnosis, as it learns by experience gained in solving problems and hence emulates human-like intelligence. However, the performance of CBR systems critically depend on the adequacy as well as the organization of cases and the algorithms used for retrieval from a large case database. Most CBR systems [28] use the nearest neighbour algorithm for retrieval from the flat-indexed case database; this is inefficient, especially for large case database. Other CBR systems use hierarchical indexing such as CART [5], decision trees [26], and C4.5 [27]. Although this performs efficient retrieval, building a hierarchical index needs the knowledge of an expert during the case-authoring phase.

The neural network approach [8] provides an efficient learning capability when provided detailed examples. Neural networks may be either supervised or unsupervised, depending on the method of training. It performs retrieval based on nearest neighbour matching, since it stores the weight vectors as the code-book or exemplar vector for the input patterns. The matching is based on a competitive process that determines the output unit that is the best match for the input vector, similar to the nearest neighbour rule. However, the search space in a neural network is greatly reduced because of the generalisation of knowledge through training. In contrast, the CBR systems need to store all the cases in the case database in order to perform accurate retrieval. The CBR systems that store only relevant cases for an efficient retrieval lack the accuracy as well as the learning feature. Thus, neural networks are very suitable for case indexing and retrieval.

Other data mining techniques include rule-based reasoning, fuzzy logic, genetic algorithms, decision trees, inductive learning systems, and statistical pattern classification systems [3]. In addition, hybrid approaches, such as hybrid case-based reasoning and neural network [21,24], have also been developed.

Here, a data mining technique that integrates case-based reasoning, neural network and rule-based reasoning is defined. These two are incorporated into the framework of the CBR cycle. Instead of using the nearest neighbour technique of traditional CBR systems, a neural network is used for indexing and retrieval of most appropriate service records based on user’s fault description. Rule-based reasoning is used to guide the reuse of checkpoint solutions.

4. Data mining for decision support

Information, such as the best selling machines, the customers of a particular machine, a comparison of sales among different machines, and the performance of different service engineers are highly desirable for the management team.

4.1. Data mining process

Fig. 3 depicts the data mining process for extracting hidden knowledge from large databases. The process focuses on finding interesting patterns that can be interpreted as useful knowledge. It consists of seven steps.

4.1.1. Establishing the mining goals

This involves the understanding of the customer service support process, its database, and the administrative procedures of the company. A number of mining goals were identified:

- Marketing: identify the machine models with poor sales and determine possible reasons; then improve the design and reliability of those machine models in order to increase sales. Target the customers with mail campaigns of the machine models in which they are likely to be interested.
- Customer support: provide the best possible service to customers based on the machine model, the nature of the problem, and geographical location.
- Resource management: assign duties to service engineers based on their expertise and past experience. Promote service engineers based on their performance.
4.1.2. Selection of data

This step identifies a subset of variables or data samples, on which mining can be performed. There are many tables in the database. However, not all are suitable for mining, since they are not sufficiently large. After an initial study, the structured data tables EMPLOYEE and CUSTOMER were found unsuitable for mining, while MACHINE and SERVICE_REPORT were considered suitable for mining.

4.1.3. Data pre-processing

This step removes the noisy, erroneous, and incomplete data. The presence of too many different categories of categorical data makes visualization of the displayed information very difficult. Hence, those categories with only a few records are eliminated. Moreover, all the records with missing values are deleted to avoid problems in visualization. Since the proportion of such records is quite small, their deletion will have little effect on the results.

4.1.4. Data transformation

The data stored in the various tables are in a specified format (defined during the construction of the database). Sometimes, it is useful to transform the data into a new format in order to mine additional information. For example, a new column ‘svc_repair_time’ (service repair time) is created by calculating the difference, measured in number of days, between ‘svc_start_dt’ and ‘svc_end_dt’ in the SERVICE_REPORT table. This new attribute is useful in analyzing the performance of the service engineers.

4.1.5. Data warehousing

Data warehousing is the process of visioning, planning, building, using, managing, maintaining and enhancing databases. The data suitable for mining are collected from the various tables of the customer service database and stored in DBMiner’s data warehouse. OLAP data marts are then generated from the data warehouse, which contains customized data at a higher level of summarization. Data cubes can be constructed from data marts to provide multi-dimensional views of the data. On-line analytical mining [13] can then be performed using the multi-dimensional data cube structure for knowledge discovery.

Fig. 4 shows a piece of the multi-dimensional (3D) view of a data cube using the three dimensions, ‘mc_fault_gp’, ‘months’ and ‘svc_member_id’ for the three axes. The size of each individual cube represents its number of records, whereas the color of the cube indicates the total value of the attribute ‘svc_repair_time’ for the records contained in the cube. Pivoting, drilling, slicing and dicing operations can be performed on the data cube for further exploration. As higher service repair time indicates longer machine down time, this may result in customer dissatisfaction. The company should look into those cases with high service repair time in order to enhance service efficiency.

4.1.6. Data mining

DBMiner is used to perform the data mining functions, including summarization, association, classification, prediction and clustering. Two examples of data mining functions are now described:

The first example illustrates the use of a summarization function. Fig. 5 presents a summary of the
Fig. 4. OLAP mining.

Fig. 5. Data summarization.
machine models serviced by service engineers using a bar chart. This summarization is very useful for understanding the expertise of service engineers. This information can be used to assign appropriate engineers to servicing particular machine models. From the figure, it can be seen that the machine model ‘AVK_2013S’ has been serviced only by service engineer ‘KL006.’ However, this also shows that service engineer ‘KL006’ has not worked on any other machine model.

Another example is given in Fig. 6 to illustrate the use of association rules mining. The association rules determine how the various attributes are related. Here, there is a strong association among the attributes in a textual format. The rules have a minimum support of 8% and a minimum confidence of 98%. High confidence rules represent distinctive patterns within a database: the first two association rules state that the customer ‘TAIT’ has reported all the faults during the year ‘1996’ and it was serviced by the service engineer ‘KL006’ only. This shows that the service engineer ‘KL006’ would be the most suitable person to be assigned to serve ‘TAIT’ in future. The next two rules show that the machine model ‘AVK_2013S’ is serviced by the service engineer ‘KL006’ only and the faults were reported in the year ‘1996’. In addition, Rule 8 indicates that the service engineer ‘10530’ had resolved all the fault problems within a day.

4.1.7. Evaluating the mining results

Different data mining functions have been exercised, providing data. The information obtained is next analyzed. The results are:

- Marketing: OLAP analysis and summarization have been applied to identify machine models with poor sales and high frequency faults. Clustering is used to identify customers suitable for cross sales.
- Customer support: association rules, classification, and clustering are used to identify those who have recently reported many faults. Better quality of service can then be provided based on customer reports.
geographical location and machine model purchased.

- Resource management: summarization can be used to identify the expertise of different service engineers. Association rules provide useful information on efficient engineers who can repair machine faults within a day. Prediction analysis can be used to compare different service engineers who have repaired machine faults under the same conditions. With this, the company can assign job duties to service engineers based on their expertise and efficiency.

5. Data mining for machine fault diagnosis

The unstructured textual data of fault-condition and checkpoint information of the customer service database provides useful machine service information. A data mining technique based on the integration of neural network, case-based reasoning, and rule-based reasoning has been applied to the customer service database to support intelligent machine fault diagnosis.

5.1. Data mining process

Fig. 7 shows the framework of the data mining process. It consists of two major processes: the off-line knowledge extraction process and the on-line fault diagnosis process. The first extracts knowledge from the customer service database to form a knowledge base that contains the neural network models and a rule-base. The neural network models and the rule-base work within the CBR cycle to support the second, which uses the four stages of CBR cycle (retrieve, reuse, revise, and retain) to diagnose customer reported problems. It accepts user's problem descrip-

![Fig. 7. Data mining process for machine fault diagnosis.](image-url)
tion as input, maps the description into the closest fault-conditions of the faults previously stored from the knowledge base, and retrieves the corresponding checkpoint solutions for the user. The user’s feedback on the fault diagnosis process is used to revise the problem and its solution. The new result is ultimately retained as knowledge for enhancing performance of future problems.

5.2. Knowledge extraction process

Fig. 8 shows the knowledge extraction process for retrieving information from the unstructured textual data of the fault-conditions and checkpoints in the customer service database. There are two major generation steps: neural network model and rule base.

The neural network model generation phase extracts the knowledge from the fault-conditions to train the neural network to build neural network models for classification and clustering. The fault-conditions in the customer service database are first pre-processed to extract keywords. The pre-processing uses a word-list, stop-list, and algorithms from Wordnet [11]. The extracted keywords are used to form weight vectors to initialize the neural networks. Then, the neural networks are trained to generate the neural network models.

Two types of neural networks were investigated: the supervised learning vector quantization (LVQ3) neural network and the unsupervised Kohonen self-organizing map (KSOM) neural network [18]. LVQ3 and KSOM are used as classification and clustering techniques for intelligent fault diagnosis, respectively. Classification techniques are used to put an instance of a new fault description into one of the known classes of faults and then use the suggested solution of the known fault for the current problem. Clustering technique can be used to extract information from the customer service database to form groups of similar faults and then define a new problem instance in one of the clusters. The classification into a specific fault-condition can be determined based on the closest match of the fault-condition with the input pattern within the cluster. The clustering technique has better efficiency, but lower accuracy compared to the classification approach.

The rule base generation process involves the extraction of knowledge from the checkpoint solutions of the fault-conditions to generate a rule-base to guide the reuse of checkpoint solution in the most effective way. The rule-base consists of control and checkpoint rules. Control rules are coded manually to specify the diagnostic procedure for the firing of checkpoint rules so that the checkpoints can be exercised one by one procedure according to its priority. As each service record in the database contains a fault-condition and its prioritized checkpoints to be exercised for diagnosing the fault-condition, checkpoint rules can be generated automatically to provide specific diagnostic
instructions towards solving a particular machine fault. Therefore, a rule-base is generated from the checkpoint information of the CHECKPOINT table. Using these two types of rules, the rule-based inference engine under the CLIPS environment [16] can provide a step-by-step guide to the user in diagnosing a fault-condition.

5.3. Fault diagnosis process

The fault diagnosis process consists of four phases:

5.3.1. Pre-processing of user input

The user input to the fault diagnosis process is a ‘free-text fault description’. However, the user can also input keywords to provide fault information. In addition, the user can also provide the names of machine components and their states as input. If the user input contains keywords that are not in the keyword list, synonyms of keywords will be retrieved for user confirmation as input keywords. This information is then combined to form an input vector during the retrieval process.

5.3.2. Neural network retrieval

The neural network retrieval process recalls similar fault-conditions experienced in the past and ranks them, based on the score that signifies the closeness of the retrieved fault-condition to the user input fault description. The neural network performs retrieval by computing the winner through a competitive learning process. The winner is the fault returned that corresponds to the weight vector with minimum distance from the input vector. For the supervised LVQ3 neural network, the winner node corresponds to a fault-condition with known checkpoint solution. In the case of unsupervised KSOM neural network, the winner node represents a cluster. The retrieval of a specific fault-condition is based on the nearest Euclidean distance of all the fault-conditions in the retrieved cluster.

Fig. 9 shows the fault-conditions retrieved by the LVQ3 neural network when the user enters the fault description. It can be observed that the fault-condition displayed at the top of the screen is the one closest to the fault description provided by the user. They are ranked according to their score. The lower frame of the display shows a checkpoint solution given to the
user for the fault-condition retrieved by the LVQ3 neural network.

5.3.3. Reuse of service records

The checkpoints are presented in the order according to the checkpoint rules fired. They operate in a competitive manner to display the checkpoints in the order of their priority in solving the fault-condition. Help can also be obtained by clicking the ‘Help’ option. This will load the help file for the corresponding checkpoint to help the user to carry out the remedial task.

5.3.4. Revise and retain with user feedback

The neural network indexing database, the checkpoint rules, and the service records in the customer service database are next updated, based on user feedback about the effectiveness of the fault diagnosis process. The input problem description and its past checkpoint solutions are revised through user’s feedback and updated into the relevant databases. As shown in Fig. 10, the user states whether the problem is resolved or not. If the problem is resolved, then the neural network indexing database and the checkpoint rule-base are updated.

If the problem persists after trying all the checkpoints for all the retrieved fault-conditions, the user reports a failure by filling in a service request form via the web. In this case, a service engineer will rapidly contact the user. If the problem is too complicated, a service engineer will be dispatched for on-site repair. The service report subsequently generated will be updated in the customer service database.

5.4. Performance evaluation

Performance evaluation has been carried out to compare the retrieval performance of the LVQ3 neural network for classification and KSOM neural network for clustering with the k nearest neighbour (kNN) technique used in the traditional CBR systems [15]. Two popular variations of kNN techniques were chosen for comparison. The first variation, known as kNN1, stores cases in a flat memory structure, extracts keywords from the textual descriptions and uses normalized Euclidean distance for matching. The second variation, known as kNN2, uses the fuzzy-trigram technique [14] for matching. The performance comparison of the neural network techniques with the two

![Fig. 10. Revise and retain with user feedback.](image-url)
variations of the kNN techniques was measured in terms of efficiency and precision. Two experiments were carried out. The first was based on the testing set of service records, while the second was based on user fault descriptions. By doing this, a better evaluation can be obtained. The results are given in Table 1.

For the first experiment, the testing set of 15,850 service records was used. Both LVQ3 and KSOM perform better than both variations of the kNN methods for retrieval in both the speed and accuracy because of their ability to generalize information through training. In kNN1 technique using Euclidean distance for matching, equal weights are assigned to the individual attributes (i.e., keywords). Therefore, the retrieval is less accurate. In kNN2 using fuzzy-trigram matching, positive scores are assigned for every sequence of three-letters. Although this technique may be useful in checking spelling errors and grammatical variations, the retrieval is quite inaccurate compared to the neural network techniques. Moreover, the major drawback in both of these kNN techniques is that new cases are indexed separately into the flat memory structure and thus the search space keeps on increasing, thereby decreasing the efficiency.

In the second experiment, a total of 50 fault descriptions were taken from non-expert users. The purpose of this experiment was to test the performance when the input was less precise in describing the fault-condition. Unlike the service records tested in the first experiment, user input was less accurate than that of service engineers. In this test, all the four retrieval techniques were found to have lower retrieval accuracy due to the impreciseness and grammatical variation of the user input. The retrieval accuracy of the LVQ3 neural network and the KSOM neural network were 88 and 86%, respectively. Fuzzy-trigram matching has a better performance than the Euclidean distance matching because of its ability to handle spelling mistakes and grammatical variations in the user input. However, its retrieval accuracy is lower than that of the two neural networks.

6. Conclusions

As a collaborative research project with a multinational company, this research investigated the application of data mining techniques to extract knowledge from the customer service database for two kinds of customer service activities: decision support and machine fault diagnosis. The information stored in the customer service database are classified as structured and unstructured textual data. The structured data are mined to enhance the decision making process for better management of resources and marketing of products. The unstructured data are mined to support intelligent diagnosis of machine faults over the World Wide Web.

In order to mine the structured data in the customer service database, a data mining process based on the data mining tool, DBMiner was proposed. To support machine fault diagnosis, a data mining technique based on the integration of neural network, case-based reasoning, and rule-based reasoning is incorporated. This data mining technique can operate within a system to provide efficient on-line machine fault diagnosis over the World Wide Web.

References

[1] A. Aamodt, E. Plaza, Case-based reasoning: foundational issues, methodological variations, and system approaches,
References


S.C. Hui is an Associate Professor in the School of Applied Science at Nanyang Technological University, Singapore. He received his B.Sc. Degree in Mathematics in 1983 and a Ph.D. Degree in Computer Science in 1987 from the University of Sussex, UK. He worked in IBM China Hong Kong Corporation as a system engineer from 1987 to 1990. His current research interests include data mining, internet technology, and multimedia systems.

G. Jha is currently a software engineer in eGain Communications Corp., Sunnyvale, USA. He received his Bachelor of Technology Degree in Computer Science and Engineering from Indian Institute of Kanpur, India, and his M.A.Sc in Computer Engineering from the School of Applied Science, Nanyang Technological University, Singapore. His research interests include data mining, machine learning and information management.