A Framework for Enterprise Knowledge Discovery from Databases

S. Wesley Changchien and Wen-Jie Lee
Department of Information Management
Chaoyang University of Technology
168 GiFeng E. Rd., WuFeng, Taichung County, Taiwan
E-mail: swc@cyut.edu.tw, Kevin_Lee@mail.adi.com.tw

Abstract

Knowledge discovery from large databases has become an emerging research topic and application area in recent years primarily because of the successful introduction of large business information systems to enterprises in the electronic business era. However, transferring subjects/problems from managerial perspective to data mining tasks from information technology perspective requires multidisciplinary domain knowledge. This paper proposes a practical framework for enterprise knowledge discovery in a systematical manner. The six-step framework employs the cause-and-effect diagram to model enterprise processes, tasks and attributes corresponding diagram to define data mining tasks, and multi-criteria method to assess the mined results in the form of association rules. This research also applied the proposed framework to a real case study of knowledge discovery from service records. The mining results have been proven useful in product design and quality improvement and the framework has demonstrated its applicability of guiding an enterprise to discover knowledge from historical data to tackle existing problems.

Keywords*: knowledge discovery, data mining, association rules, cause-and-effect diagram, MCDM*

1. Background and Motivation

With the maturity of technology in business information systems, enterprises successively introduced large systems such as ERP, EC, SCM, CRM, etc. [1]. As information technology advances, it is a major issue an enterprise has to face to enhance competitiveness by acquiring useful analyzed information or knowledge to develop a more effective competition pattern or improvement strategy. But while enterprises obsess large bunches of data, they need a practical framework to analyze the data and discover knowledge using data mining methods [2].

In the process of enterprise decision-making, it is common to rely on decision-makers' experiences. But it may lead decision-makers to misjudgments not only because of the shortage of information in objective environment but also because of personal subjective preferences. Solving problems with traditional expert systems requires the establishment of enormous data and traditional discussions to locate problems and suitable solutions. The process of collecting data and searching feasible solutions, however, is very complicated. And the knowledge base is often incomplete and soon to become outdated.

This research presented a framework to help enterprises analyze interesting subjects, either operational or managerial, locate subjects to be explored, and discover knowledge from existing databases through systematical steps with data mining and other problem analysis technologies.

2. Literature Review

Processes that consist of tasks are building blocks of an enterprise from the perspective of operations management, both for manufacturing and service industries. The competitiveness of an enterprise is therefore heavily dependent on the efficient operations and management of the processes. Over the decades, various technologies and researches have been devoted to improve it, such as total quality management (TQM), just in time (JIT), business process reengineering (BPR), and enterprise resource planning (ERP). Many of them are operations research and quantitative approaches, while some are managerial or philosophical approaches. And these approaches all attempt to make the processes fulfilled efficiently and correctly at lowest costs and highest quality. Along with the introduction of ERP systems, all the data about the historical, ongoing, or future processes are all stored in the databases. However, there are more than a thousand tables in the database for a large ERP system, and the significant amount of transactions in the database make it impossible to analyze efficiently solely with traditional statistical methods. Data mining is such a new emerging technique that has gained the significant attentions and applications of practitioners.

The purpose of data mining is to apply different data analysis methods to enormous data to discover significant rules or knowledge. By doing so, we can help enterprises get a better understanding of subjects interested and offer enterprises related knowledge when they face problems [3].

This research aims to utilize association rules analysis to discover related knowledge on subjects/ problems that an enterprise greatly concerns. First of all, we have to understand the definition of association. Take an example of maintenance data, if the replacement of material B is 90% co-existed with the replacement of material A in maintenance records, the relation between them is an association[4].

We further make a more specific definition on the association rules [5]: Suppose that $I = \{ i_1, i_2, \ldots, i_m \}$ is a

set of items. Let D be a set of database transactions where each transaction T contains a set of items such that $T \subseteq I$. Each transaction is associated with an identifier, called TID. Let X be a set of items. A transaction T is said to contain X if and only if $X \subseteq T$. As association rule, represented as $X \Rightarrow Y$, where $A \subset I$, $Y \subset I$, and $X \cap Y = f$. The rule $X \Rightarrow Y$ holds in the transaction set D with support *s*, where *s* is the percentage of transactions in D that contain $X \cup Y$. The rule $X \Rightarrow Y$ has confidence *c* in the transaction set D if *c* is the percentage of transactions in D, containing X, that also contain Y.

Association rules mining algorithms such as Apriori [5] generally begin with single items and then join frequent single items to obtain multiple item sets. The method of Apriori requires a lot of times of scanning on the database , thus a great amount of time and space are required. This research adopts CIT algorithm [6], which only requires a scan on database at the beginning of the mining process, and thereafter association rules can be generated based on the constructed index tree.

Though we could thoroughly unearth numbers of association rules by data mining methods, it's uneasy to determine which are constructive association rules by algorithms. Therefore, the key technology lies on how to produce association rules and determine constructive and meaningful rules to avoid worthless effort on meaningless association rules [7]. Determining whether an association rule is constructive may reply on more than a single criterion, and the criteria vary with domain areas too. To consider multiple criteria simultaneously, a multi-criteria decision making (MCDM) method [8] is needed, of which utility theory is an effective method when the decision makers' preferences are under consideration [9].

3. Proposed Framework for Enterprise Knowledge Discovery from Databases Using Data Mining

This research proposes a framework for enterprise knowledge discovery, which features on practical and systematical data mining procedure and extracting important and meaningful knowledge from the numerous mined association rules. The framework contains two stages. The first stage is to look into an enterprise's processes where the subject areas the decision makers feel interested or the problems had occurred. By mapping the tasks of the selected processes to the corresponding attributes of the database tables, data mining tasks can be defined and conducted. The second stage is the multi-criteria assessment on the results analyzed and mined at the first stage in order to assure the meaningful and important knowledge is discovered. The framework of this research is displayed in Figure1, which will be further introduced in the following sections.

3.1. Search for subjects/problems

The processes of an enterprise can be searched and examined to define the interested subjects or problems. Cause-and-effect diagram (or fishbone diagram) can be used to depict the overview of an enterprise's main processes [10][11]. Figure 2 shows a cause-and-effect diagram for the main management processes of a manufacture enterprise.

In constructing the diagram, main management processes have to be identified first, and the corresponding sub-processes for each main management process are listed next.

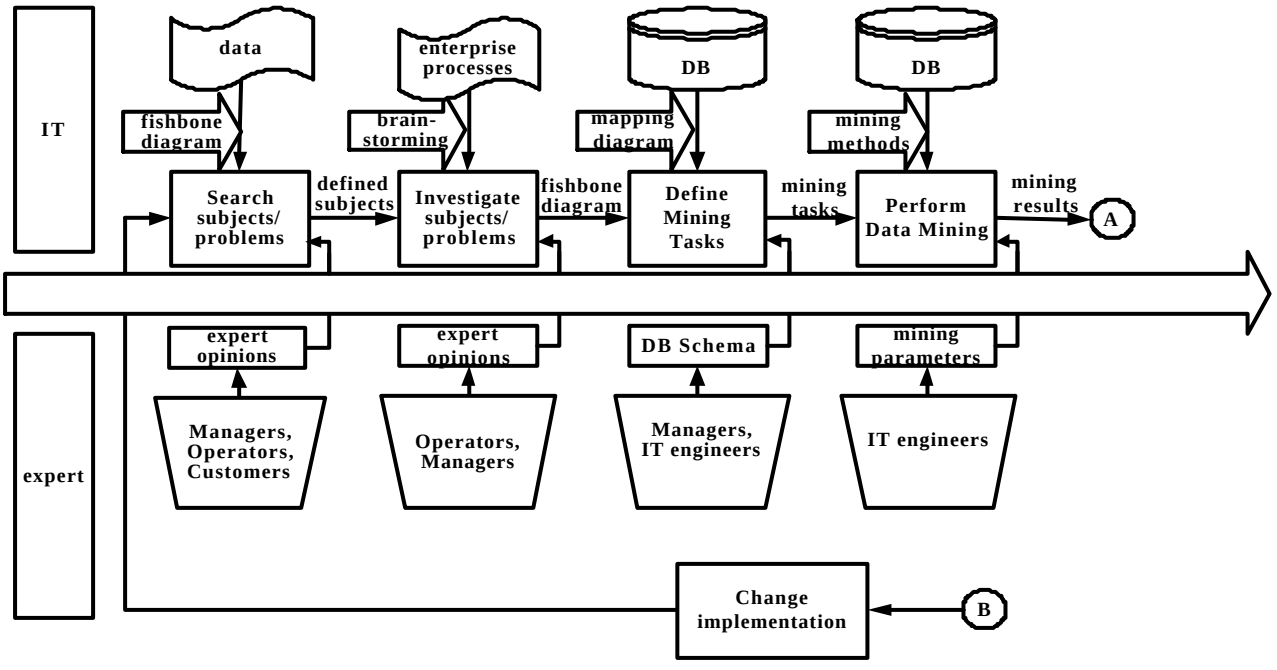

Figure1. A proposed framework for enterprise knowledge discovery from databases using data mining

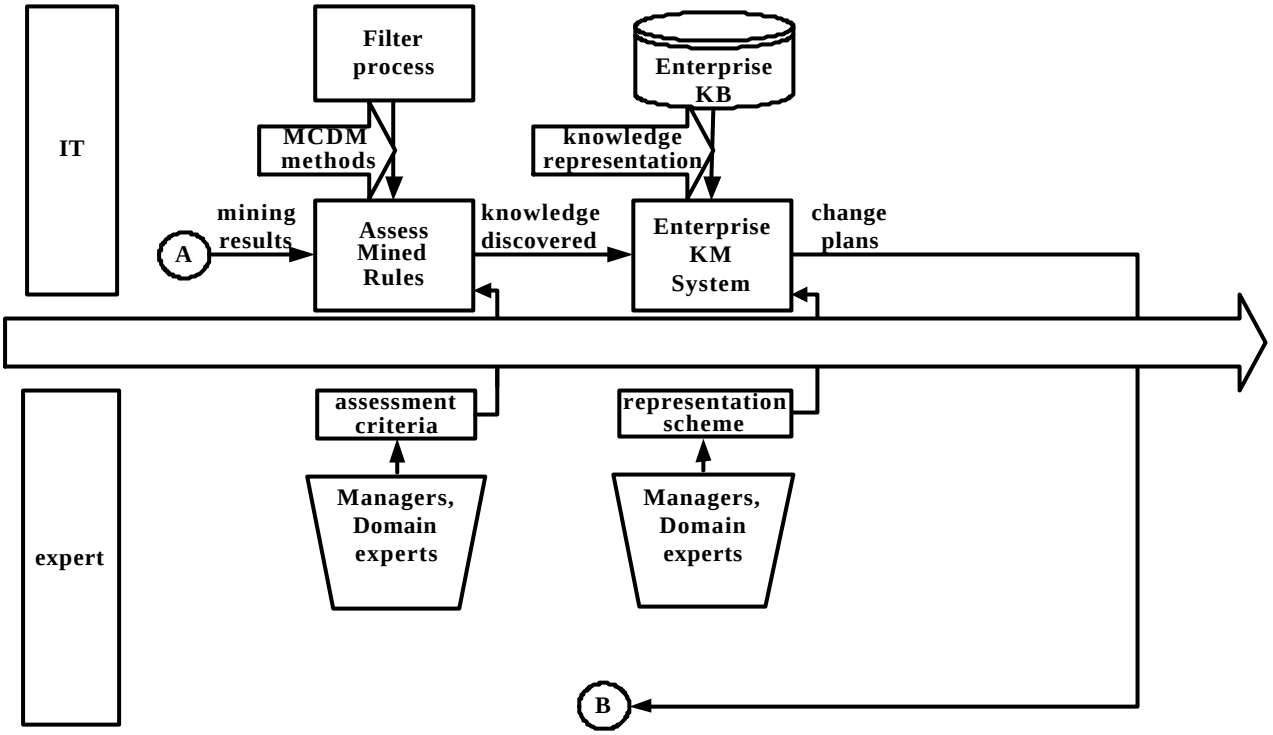

Figure1. A proposed framework for enterprise knowledge discovery from databases using data mining (Cont.)

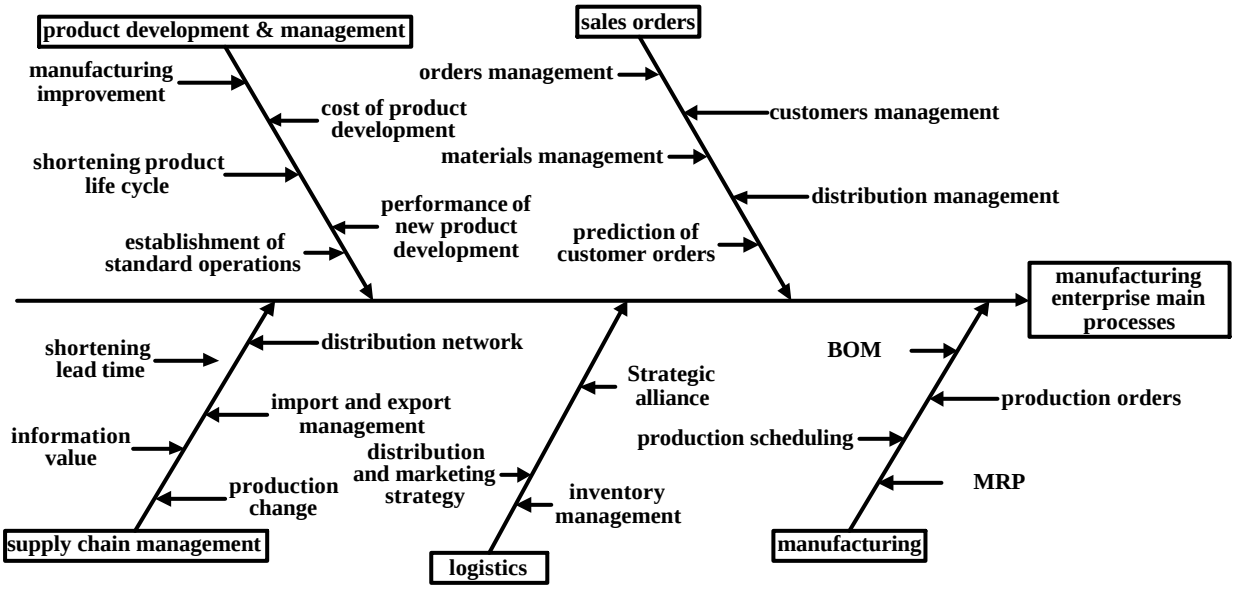

Figure 2. A cause-and-effect diagram for enterprise main functions

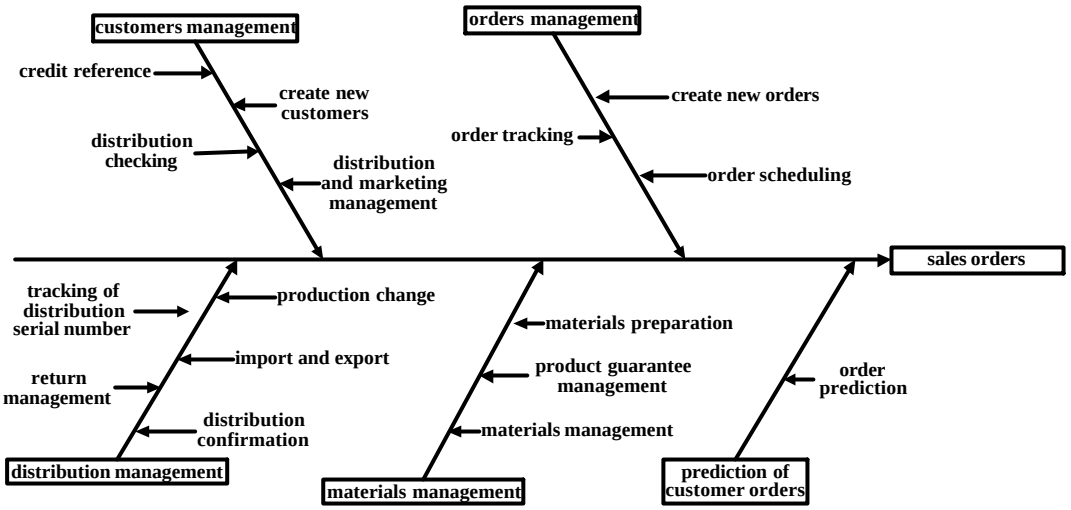

Figure 3. A cause-and-effect diagram for customer orders

The sub-processes can be further decomposed into sub-processes at a more detailed level. The cause-and-effect diagram of sales orders in Figure 2 is shown in Figure 3. Similarly other main processes can be decomposed and depicted as separate cause-and-effect diagrams at a more detailed level. With the processes of an enterprise depicted using cause-and-effect diagrams in detail at different levels, the interested subject areas or existing problems can be easily identified [12].

3.2. Investigate subjects/problems

After we locate the interested subjects/problems, the next step is to separate the associated processes into two parts: one that can be analyzed by data mining or the other that cannot [13]. Those processes to be mined are further investigated. Operators, managers, and domain experts all need to join the brainstorming team and opinions can be exchanged to better understand the subjects/problems, and finally the directions of data mining tasks can begin to take shape.

3.3. Define mining tasks

The function of Data Mining is to discover all kinds of hidden patterns from the data set. The purpose of this step is to search for the data for mining that are related to the target enterprise subjects/problems. Therefore, as defined in prior sections, the selected enterprise processes are compared with the attributes in the database. Finally compose a *tasks and attributes correspondence diagram* for Data Mining [14]. Figure 4 is an example of tasks and attributes correspondence diagrams. The diagram corresponds the enterprise processes interested, from managerial perspective, to the attributes in the databases, from data mining perspective. According to the tasks and attributes correspondence diagram, data mining tasks can be defined to discovery knowledge related to the selected enterprise processes from the databases. Before data mining is performed, data has to be preprocessed. Data preprocessing generally includes the following steps:

1. Conceive the data source and commence data collection
2. Acquire related knowledge and technology
3. Integrate and check data
4. Remove false or inconsistent data

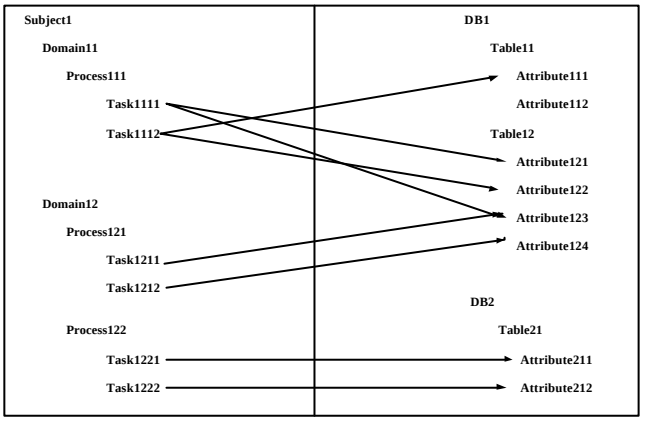

Figure 4. A tasks and attributes correspondence diagram

3.4. Data mining

By the progress of information technology and the entering of electronic business era, enterprises have already focused on making the best use of databases to acquire useful information or knowledge rather than only collecting and storing data. It becomes increasingly important nowadays for enterprises to effectively acquire data mining tools for database data mining when competing with competitors. Figure 5 displays the data mining project for enterprise knowledge discovery.

The feature of data mining is to search for meaningful patterns from giant databases. It acquires meaningful patterns from databases and transforms them into information or knowledge for supporting enterprise decision making. Different kinds of technologies, such as Genetic Algorithms, Neutral Networks, Fuzzy Logic, Case-Based Reasoning, etc., could be used for data

mining. This research adopts an association rules mining algorithm, CIT Algorithm [6] to look for possible results. The advantage of it is that CIT algorithm allows for mining with causality. Take the problem of quality control for example: we could have the materials replaced for cause attribute and the malfunctioning conditions for effects to search for specific associations between replacement materials and malfunctioning types. Mined association rules may provide valuable reference information or knowledge for product research and development.

3.5. Assess mined rules

This step assesses patterns from the mined results according to the predefined criteria and unimportant or minor association rules will be ignored. To effectively determine the assessment criteria, this research proposes a concept which lays emphasis both on objective and subjective factors in the process of filtering. There are two basic elements in the process of assessment: feasible alternatives and decision-making criteria. Criteria either objective or subjective, taking quality issue for example, such as the cost of processing, the effect on goodwill, the improvement on productivity, the violation on policies, and the decrease in the number of service can be applied based on the criteria importance and decision maker's judgment.

Since multiple criteria are being considered in assessing the mined association rules, weighted sum model (WSM) is used to formulate the assessment equation:

$$Rule\_imp(j) = \sum_{i=1}^{m} w_i * x_i, \ for\ i = 1,2,...m.$$

where
Rule_imp(j) is the importance value for rule *j*,
x_i is performance value for criterion *i*,
w_i is the weight for criterion *i*, and
m is the total number of criteria considered.

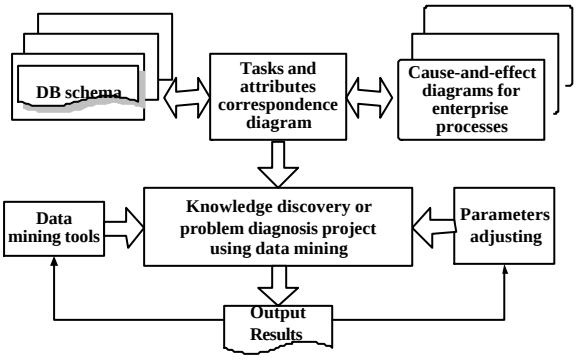

Figure 5. Data mining project

3.6. Enterprise knowledge base

Personal know-how through knowledge acquisition process and knowledge discovery using information technology both contribute to the creation of enterprise knowledge base. When decision makers are dealing with semi-structured or unstructured problems, enterprise knowledge base can be referenced to improve the quality of enterprise decision making. To make the knowledge acquired accessed and shared, an enterprise knowledge management system is needed. Besides, knowledge representation scheme is necessary for systematically storing the acquired knowledge. The associated managers and operators to improve the current performance or to solve the problem can reference the discovered knowledge related to the selected subjects/problems. If the mining results cannot satisfy the associated personnel, the whole knowledge discovery process needs to be initiated all over again. If the knowledge discovered provides insight in solving the problem or improving performance, the resulting change plan then can be implemented, and it ends the knowledge discovery process.

4. A Case Study on the Quality Improvement in a Manufacturing Company

The proposed knowledge discovery framework and the associated methodology were applied to quality improvement on product design in a manufacturing company. In product development, it will be, however, a Herculean task for the R&D center to reassemble parts and test to find all the defective items provided by the manufacturing bases located world wide. Usually the purpose of these tests is only to locate the incompatibility among some electronic items to reduce the occurrence of uncertain operational breakdowns after the goods left the factory.

Owing to the numerous combinations in product items, how to prevent this kind of problems from happening again and again has always been a tough task for industries in analyzing the maintenance data. The burden of analyzing the incompatibility among electronic items is always untaken by experienced experts. This kind of approach is however time-consuming yet ineffective. In fact, the industry already has established a database and a complete collection of all detailed maintenance data from service centers world wide. We should be able to make good use of the database, accumulated for years, for analysis and data mining.

The proposed methodology was applied to the R&D problem. A cause-and-effect diagram was generated as shown in Figure 6.

Next we constructed the tasks and attributes correspondence diagrams and defined the association rules mining tasks. CIT algorithm was applied to more than twenty thousand real maintenance records of products being returned for service due to malfunctioning with six attributes from two database tables. We employed different values of *minimum confidence* and *minimum support* and obtained different number of association rules as shown in Table 1.

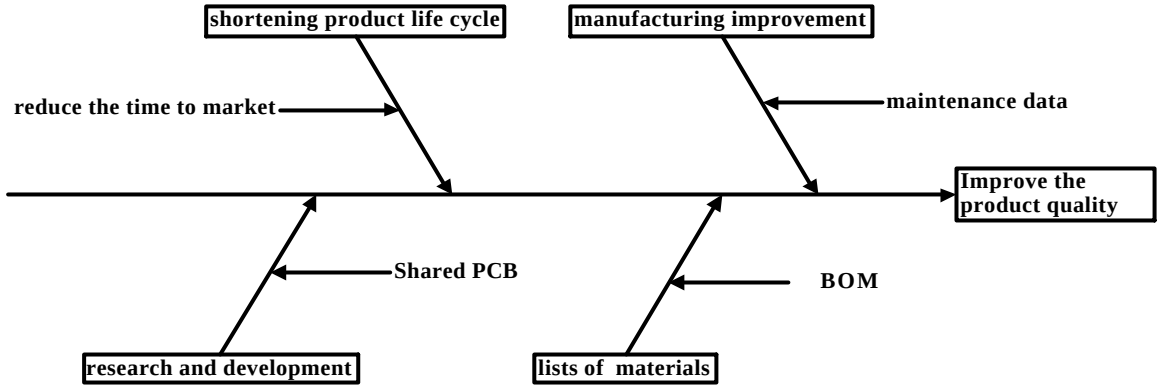

Figure 6. A cause-and-effect diagram for improvement of product quality

Table 1. Numbers of association rules extracted using CIT algorithm for different values of support and confidence

Support / Confidence	5	10	20	50
99%	376531	40566	2288	58
80%	505550	43767	2878	58
50%	515960	48907	2878	93

The association rules generated from maintenance records as in Table 1 contain a great deal of meaningless association rules. After eliminating these meaningless ones, we can better understand the quality problems due to defective parts or incompatibility in the assemblies. Defining the rules of eliminating meaningless association rules requires domain knowledge about maintenance and design processes and three rules are shown below:

Rule 1 For an association rule "items A and B together cause the failure of item C," if the confidence of the rule is not 100%, it will be eliminated. This is to make sure defective items A and B simultaneously do relate to the failure of item C.

Rule 2 Only defective electronic parts are considered in discovering knowledge about incompatibility problems. If the defective parts contain non-electronic items, the rules are eliminated.

Rule 3 By checking with the BOM structures, if different electronic items from different products are found in an association rule, it will be eliminated. This is because the incompatibility problem only considers items on the same product.

In the data mining process, for the circumstance of the *minimum confidence* set to be 99.9% and the *minimum support* set to be 10, 40566 association rules were acquired. 2788 association rules were left after elimination based on rules 1 and 2. Finally, 82 crucial association rules were obtained when the third elimination rule was performed. Some of the final association rules are listed in Table 2.

The 82 eventually gained association rules were assessed in terms of five criteria and assigned with separate weights. Five criteria cost, goodwill, productivity, policy, and service are assigned with weights of 30%, 30%, 20%, 10%, and 10%, respectively. Cost criterion refers to direct and indirect costs spent related to the service; reputation refers to the level of enterprise's reputation was affected due to malfunctioning types; productivity refers the level of enterprise's productivity can be improved by redesign; policy means importance from the perspective of enterprise's policies from decision makers (e.g. the phased out models will not be processed); and service indicates the times of services required can be reduced after redesign.

Table 2. Examples of association found by data mining on incompatibility problem

Part ID	Part ID	Part ID	=>	Related Part ID
BA015J1	BP022A4	BY400B0	=>	BA015J1
BC176G0	BL003W0	BN012A5	=>	BC176G0
…				

In terms of these 5 criteria, we used weighted sum model (WSM) to calculate the rule importance for each association rule. The importance can be regarded as the priority for each association rule. Some of the results are shown in Table 3.

After testing and verifying the proposed framework on enterprise knowledge discovery on quality problem, it is found that the mined association rules provide frequent associations of multiple electronic items occurring on the same products being serviced. These rules may provide interesting and valuable information in product research and development, they can also be provided for inspection of a new model design to reduce the possibility of the same failure to happen again.

Table 3. Rules importances calculated for mined association rules

Part ID	Part ID	Part ID	=>	Related Part ID	Rule importance
BA015J1	BY400B0		=>	BP022A4	2
BA015J1	BY400B0	BY809B0	=>	BP022A4	1
BP022A4	BY400B0	BY809B0	=>	BA015J1	3
BU002S5	BN012A5		=>	BR008B5	5
BR008B5	BU002S5		=>	BN012A5	7
BR008B5	CR107C0		=>	BN012A5	6
RR229F	TP156Z0	RR229F1	=>	BN012A5	4
…					

5. Conclusions

This research has proposed a practical framework for enterprise knowledge discovery using data mining. The six steps of the framework are described in detail along with associated techniques or methods. By cause-and-effect diagrams, the processes of an enterprise can be modeled and analyzed. With tasks and attributes correspondence diagram, data mining tasks can be defined. Collect the right data from the enterprise database, then proceed data mining to discover the association rules relevant to the subjects/problems selected. In the case study, we extracted relevant association rules to incompatibility issue out of more than 20,000 maintenance records. After filtering process, fewer significant and more meaningful association rules are left. With multi-criteria decision making method, the rule importance can be calculated for each association rule. The association rules then can be referenced based on the rule importance. The association rules discovered have been found very useful in improving the product design and to reduce the failure in products due to incompatibility. This proposed framework that integrates multiple techniques provides a very practical approach and guidance for practitioners in knowledge discovery and problem solving. The case study also demonstrated the applicability of the framework in the industry.

References

[1] Saxena, K.B.C. and Sahay, B.S., "Managing IT for world-class manufacturing: the Indian scenario," International Journal of Information Management, Vol.20, No.1, pp. 29-59, 2000.

[2] Deslandres, V. and Pierreval, H., "Knowledge acquisition issues in the design of decision support systems in quality control," European Journal of Operational Research, Vol.103, No.1, pp. 296-311, 1997.

[3] Lin, F. Y. and McClean, S., "A data mining approach to the prediction of corporate failure," Knowledge-Based System, Vol.14, Nos.3-4, pp.189-195, 2001.

[4] Dabbas, R. M. and Chen, H.-N., "Mining semiconductor manufacturing data for productivity improvement – an integrated relational database approach," Computers in Industry, Vol. 45, No.1, pp.29-44, 2001.

[5] Han, J. and Kamber, M., Data mining: concepts and techniques, Morgan Kaufmann Publishers, Academic Press, 2001.

[6] Changchien, S.W. and Lu, T.C., "Mining Association Rules from Databases Using Class Inheritance Tree (CIT)," submitted for publication in Data and Knowledge Engineering.

[7] Guillaume, S. and Charnomordic, B., "Knowledge discovery for control purposes in food industry databases," Fuzzy Sets and Systems, Vol.122, No.3, pp.487-497, 2001.

[8] Triantaphyllou, E., Multi-Criteria Decision Making Methods: A Comparative Study, Kluwer Academic Publishers, 2000.

[9] Hatush, Z. and Skitmore, M., "Contractor Selection Using Multi-criteria Utility Theory: An Additive Model," Building and Environment, Vol.33, Nos.2-3, pp.105-115, 1998.

[10] Tseng ,H. C., Ip, W.H. and Ng, K.C., "A model for an integrated manufacturing system implementation in China: a case study," Journal of Engineering and Technology Management, Vol.16, No.1, pp.83–101, 1999.

[11] Wu, B. and Ellis, R., "Manufacturing strategy analysis and manufacturing information system design: Process and application," International Journal Production Economics, Vol.65, No.1, pp.55-72, 2000.

[12] Coronado, M., Adrian, E., Sarhadi, M. & Millar, C. "Defining a framework for information systems requirements for agile manufacturing," International Journal Production Economics, Vol.75, Nos.1-2, pp.57-68, 2002.

[13] Gillenwater, E., Conlon, S. and Hwang, C., "Distributing Manufacturing Support System: the Integration of Distributed Group Support System with Manufacturing Support System," Omega, Vol.23, No.6, pp.653-665, 1995.

[14] Tsechansky, M. S., Pliskin, N., Rabinowitz, G. and Porath, A., "Mining relational patterns from multiple relational tables," Decision Support Systems, Vol.27, Nos.1-2, pp.177–195 1999.