Electrical Engineering in Japan, Vol. 140, No. 2, 2002
Translated from Denki Gakkai Ronbunshi, Vol. 121-C, No. 8, August 2001, pp. 1289–1295

Applying Data Mining to a Field Quality Watchdog Task

SATOSHI HORI,[1] HIROKAZU TAKI,[2] TAKASHI WASHIO,[3] and HIROSHI MOTODA[3]
[1]Monotsukuri Institute of Technologists, Japan
[2]Faculty of Systems Engineering, Wakayama University, Japan
[3]I.S.I.R., Osaka University, Japan

SUMMARY

This article describes a watchdog program that discovers "meaningful" repair cases from a field service database. "Meaningful" cases are those judged worth probing further to prevent an epidemic of quality problems. Our system has employed the apriori algorithm, a data mining technique that efficiently performs the basket analysis. Our system proves that this data mining technique is not only useful in knowledge discovery but is also capable of performing the database watchdog task. The apriori algorithm automatically generates frequent itemsets from a large set of records. A frequent itemset is an arbitrary combination of values that appear more often than a threshold "minimum support." The algorithm often generates too many itemsets for quality engineers to review carefully in their daily work. Many itemsets do not provide sufficient information to investigate further. Hence, in order not to generate these valueless itemsets, the apriori algorithm is modified in two ways. One way is "basket analysis on objective and explanatory attributes" and the other is "itemset reduction." The advantage of our method is demonstrated with some experimental results. © 2002 Wiley Periodicals, Inc. Electr Eng Jpn, 140(2): 18–25, 2002; Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/eej.10034

Key words: field service; quality control; basket analysis; data mining; apriori algorithm; watchdog program.

1. Introduction

There is a keen demand for a watchdog program that is able to discover important field quality problems from a large service report database. We developed this kind of watchdog program using a data mining technique.* The system has been working at a display monitor factory and proved that the data mining technique can be successfully applied to the watchdog task.

This section describes the background and purpose of the watchdog program development. It also briefly mentions that the apriori algorithm, a data mining technique, is potentially able to solve the field quality watchdog problem.

1.1 Field quality control

Field quality is recognized as being very important in order to improve customer satisfaction and to compete successfully with global competition. The cost of field service, however, is becoming a major factor in the decreasing profits of manufacturers. Therefore, it is necessary to discover quality problems and quickly feed them back to design and manufacturing divisions in order to improve product quality and reduce repair expense. All service incidents, which are carried out at service stations, are documented in a service report. The service reports include the information shown in Table 1.

Nowadays the service reports are computerized. They are transmitted to and stored in a relational database at a manufacturing factory. Hence, quality engineers are now able to process thousands of service reports with their personal computers. However, it is a very time-consuming task to analyze the service reports because many combinations of product models, parts, symptoms, and so on must be aggregated. Therefore, there is a keen demand for a watchdog program which is able to analyze newly arrived service reports and automatically discover important service incidents which are or might become product quality epidemics.

*This work was done while the first author worked for Mitsubishi Electric Corporation.

Table 1. Service report attributes

Service Report	
Report ID No	Date of Service
End-user's name	End-user's address
Product model	Product Serial No.
Date of Manufacturing	
Symptoms	Cause of malfunction
Replaced parts List	.....

1.2 Watchdog program

Our factory has been using a conventional database system that collects weekly service reports and draws Pareto charts and defect trend graphs every month. This system eliminated a lot of paper work such as writing, faxing, and filing service reports. In addition, it helps the quality engineers to analyze and review quantitatively the service reports. However, some important service matters are still found and reported from service technicians. Why are they able to discover significant quality problems even though their experience is limited to only a portion of all repair incidents? The reasons are:

1. This conventional analysis program cannot analyze data in a manner flexible enough to discover new findings. The program is able to process and analyze data in a fixed pattern; for example, it quickly counts up how many service parts P_i were replaced to fix breakdowns of a specified product model. However, the program cannot report that parts P_i, P_j are often replaced together. What programs do is only counting the records in a database under a condition that a human engineer specifies to confirm his assumption, for example, that part P_i might be a major problem in field quality.

2. Service technicians have a lot of background knowledge, such as design changes and expected defect probabilities, and hence they can infer real causes behind these defects. That is, they can judge the significance of service reports.

The objective of the watchdog program is to automatically discover "meaningful" itemsets from a large number of service reports. The "meaningful" itemsets are worth probing further to prevent quality problem epidemics. This intelligent program supports the quality engineers in taking actions of quality improvement before the experienced service technicians make a report.

Data mining [3] has been attracting attention because it can discover interesting knowledge and/or rules from a large database. Basket analysis, a data mining technique, seemed effective for solving our problem because it generates frequent itemsets and association rules without any background knowledge. Hence, we have decided to employ the apriori algorithm [1], which is a popular data mining technique and which can efficiently perform the basket analysis. The apriori algorithm can be a powerful tool that can automatically discover significant quality problems in the form of frequent itemsets. The frequent itemset is a tuple of values (items) that appears so frequently in a database that it could be interpreted as an important service incident. However, the apriori algorithm often generates too many frequent itemsets. Most of them are not valuable for the quality engineers. In order to avoid generating valueless itemsets, we have introduced two modifications to the apriori algorithm: "basket analysis on objective and explanatory attributes" and "itemset reduction." Our method can generate a small set of the frequent itemsets including the itemsets that are proven important quality problems by further investigation.

The next section briefly explains the Apriori algorithm and its limitation. "Basket analysis on objective and explanatory attributes" is described in Section 3. The overview of our watchdog program is also given in Section 3 with some experimental results.

2. Data Mining and Watchdog Task

Discovering "meaningful" itemsets, that is, important service incidents, should be performed without interaction with a human analyst so that we can save human power for probing further. Therefore, we need an algorithm that can pick up a small number of meaningful itemsets from the large database without carefully tuned criteria and background knowledge. This task has been researched as "knowledge discovery from databases," and, in particular, basket analysis seems effective in order to solve our problem. This section briefly explains the basket analysis and apriori algorithm. Then the advantages and limitations of applying the apriori algorithm to the field quality watchdog task are discussed.

2.1 Apriori algorithm

The basket analysis derives all itemsets and association rules that have greater support and confidence levels than given thresholds. Assume there are a set of shopping records, and each of the records is a set of purchased items.

Then the basket analysis proceeds as follows:

1. Generate frequent itemsets that have support values greater than a threshold.

The itemset is a tuple of purchased items, such as {bread, milk, coffee, . . .}. The frequent itemsets are those whose occurrence in the database exceeds a threshold that is called "minimum support."

2. Generate association rules from the frequent itemsets.

An association rule has the form $A \Rightarrow B$ where A, B are sets of items and $A \cap B = \phi$, $A \cup B = C$.

Example: Assume we have a frequent itemset C = {bread, milk, butter}. If the likelihood that "milk" and "butter" appear in the records containing "bread" is greater than a threshold "minimum confidence," we derive an association rule: {bread} $\Rightarrow$ {milk, butter}.

The apriori algorithm [1] efficiently performs the basket analysis. Figure 1 briefly illustrates the apriori algorithm.

Step-1. Collect frequent 1-itemsets L_1 = {{$item_i$}, {$item_j$}, . . . , {$item_j$}}. Each frequent 1-itemset {$item_i$} occurs more than a threshold **minsup** (minimum support) in the database D.

Step-2. In this loop, frequent k-itemsets such that each itemset has k items and occurs more than **minsup** times in the database.

Step-3. The **apriori-gen** function takes as an argument L_{k-1}, the superset of all frequent k-1-itemsets. It returns a superset of all k-itemsets.

Step-4. This loop counts the frequency of each k itemset $c \in C_k$.

Step-5. The subset function returns a subset of k-itemsets $C_t \subset C_k$ where $c \in C_t$ is included in a transaction (a database record) t.

```
Step - 1  L1 = {Frequent 1 - item - sets};
Step - 2 for {k = 2; Lk-1 ≠ f; k + +} do begin
Step - 3    Ck = apriori – gen(Lk-1);
Step - 4       forall transactions t ∈ D do begin
Step - 5          Ct = subset(Ck, t);
Step - 6          forall candidates c ∈ Ct do
Step - 7                c.count + +;
Step - 8       end
Step - 9     Lk = {c ∈ Ck | c.count ≥ minsup}
Step - 10 end
Step - 11 Answer = ⋃k Lk;
```

Fig. 1. Apriori algorithm.

Step-7. Increment the frequency of $c \in C_t$.

Step-9. Create frequent k-itemset L_k whose members, k-itemsets, appear more than a threshold **minsup**.

2.2 Field quality analysis and apriori algorithm

Field quality has been conventionally analyzed using a relational database and Pareto charts. The Pareto analysis requires us to specify attributes and/or values to be analyzed. This is a tedious drill-down analysis. The apriori algorithm has the following advantages over this conventional quality analysis. The apriori algorithm is able to:

1. Pick up automatically any combinations of attributes and attribute values. Without specifying product models or attributes to analyze, a computer reports frequent itemsets, that is, significant sets of items.

2. Aggregate the frequency of each item of a list-value. It is hard for a relational database query to count a list-value and values stored in separate attributes.

The apriori algorithm has large advantages over the conventional quality analysis, especially SQL (Structured Query Language) in relational databases. However, it generates so many frequent itemsets that a human analyst cannot review them easily. If the minimum support value is set at 5, because most of the important service matters occurred around 10 times a month, the apriori algorithm generates 759 frequent itemsets from 1479 service reports. This size of generated itemsets is too large for browsing and taking actions. Many of the generated frequent itemsets are valueless for quality engineers because:

1. Many of the frequent itemsets do not include a product model and/or Part-IDs. Quality engineers cannot infer any defect causes for these itemsets nor take any actions to solve the problems.

2. Itemsets, which can be included in another itemset, are generated. For example, if an itemset is generated, its subsets are also generated as frequent itemsets. However, the quality engineers often need only the larger itemset because it contains more information than others.

In order to derive only valuable itemsets/rules, several methods have been proposed. Introduced item constraints [5] derive valuable itemsets and association rules by providing a taxonomy of interesting item values. Matsuura [4] proposed a new principle that provides maximal guesses from minimal facts and deletes less-valued association rules.

Field service report analysis is similar to many cases of relational database analysis in that human analysts know what attributes are vital and the causality relations among the attributes. Therefore, our "basket analysis on objective

and explanatory attributes" utilizes this background knowledge to derive only meaningful frequent itemsets rather than providing the taxonomy of item values for analysis. In our problem, frequent itemsets rather than association rules are interesting, because human engineers can easily infer causality among itemsets. Hence, our "itemset reduction" simply merges itemsets that share the same values of the objective attributes.

3. Field Quality Watchdog Program

There were two major development issues in building the watchdog program. These are:

1. To reduce the size of the frequent itemsets generated by the apriori algorithm. Hence, human analysts can easily review and investigate them.

"Basket analysis on objective and explanatory attributes" successfully discovers a small number of the frequent itemsets from a service report database.

2. To provide supplemental information together with the frequent itemsets. Hence, it makes it easier for human analysts to judge the importance of each frequent itemset.

Our watchdog program can draw a trend chart, and Weibull graph for each frequent itemset.

3.1 Objective and explanatory attributes

One of the reasons why the apriori algorithm generates many frequent itemsets is that the algorithm ignores the meaning of attributes. However, when we analyze data stored in a relational database, we know the meanings, relations, and importance of their attributes, that is, columns of a table. For example, if we need to know the quality trend of a product, product models and replaced parts are indispensable but end-user's name and service technicians' names are not. In our method, the attributes of the records are classified into two categories: objective and explanatory attributes. The definitions of the two categories are:

Objective attributes are key attributes. Frequent itemsets ought to contain at least one value for each objective attribute. In the case of field quality control, the product model and replaced part columns of the service report table are objective attributes. Quality engineers cannot take any action from an itemset that lacks these values. Note that each objective attribute can be a column or a set of columns of the relational database tables. This is because in the real world, values of a single attribute, such as service parts, are often stored in several columns.

Explanatory attributes appear in a frequent itemset only if their values have close relations with the objective attributes in the itemset. Explanatory attributes, such as defect-cause and date of manufacturing, convey important information when they are associated with the objective attributes.

Figure 2 depicts the relation among purpose, necessary information, and service report's attributes. Model name and parts ID in service reports are required to identify an electric part. Date of manufacturing can be vital to judge if the lot quality of the part is in doubt. Therefore, in this example, model name and parts ID are classified as the objective attributes, and date of manufacturing is treated as the explanatory attribute. Table 2 explains what attributes are necessary for what purpose.

Our algorithm consists of two stages, apriori on objective attributes and apriori on explanatory attributes. The detail of the algorithm is as follows.

Apriori on objective attributes

STEP-1. Starting with N = 1.

STEP-2. Derives the frequent itemsets from Objective-N attribute.

STEP-3. Next, expand the frequent itemsets that are generated in the previous step, so that they include at least one value of Objective-N+1. Branches that cannot grow to this layer are pruned.

STEP-4. Go back to STEP-2 until all objective attributes are processed. The frequent itemsets generated are denoted L_o.

Apriori on explanatory attributes

STEP-5. Execute the apriori on explanatory attributes for the records of each $l_o^k \in L_o$.

STEP-5.1. Select the records that include $l_o^k = \{o_1, \ldots, o_m\}$.

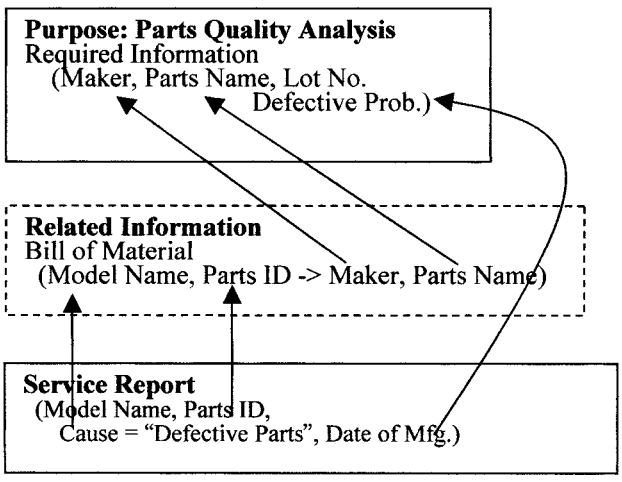

Fig. 2. Quality analysis versus required information.

Table 2. Objective and explanatory attributes

	Quality Analysis on			
Attribute	Parts	Assembly	Design	Type of Attr.
Product Name	*Required*	*Required*	*Required*	Objective
Replaced Parts	*Required*	*Required*	*Required*	Objective
Cause of defect	*Important*	*Important*	*Important*	Explanatory
Symptom			*Important*	Explanatory
Date of Mfg.	*Important*	*Important*		Explanatory
End User's Name			*Important*	Explanatory

STEP-5.2. Generate the frequent itemsets L_e^k.
STEP-5.3. Combine l_o^k and L_e^k. That is,

$$l^k = l_o^k \cup \bigcup_{i=1}^{n} l_{e,i}^k$$

3.2 System overview

Figure 3 shows an overview of the watchdog program, which analyzes the service reports weekly for 4 weeks and generates only meaningful itemsets. The frequent itemset generation is executed as follows:

Step-1. Select the service reports of the latest 4 weeks from the service report database.
Step-2. Execute the apriori on "objective" attributes and generate the frequent itemsets of the objective attributes, L_o.
Step-3. Execute the apriori on "explanatory" attributes of the records that contains each member of L_o and generate the frequent itemsets of the explanatory attributes, L_e. L_o and L_e are combined. Let L denote the result.
Step-4. Reduce the size of L by the itemset reduction rule explained in Section 3.2.
Step-5. Display the frequent itemsets L.

The derived frequent itemsets are displayed on the Frequent Itemsets screen shown in Fig. 4. Each itemset consists of objective attribute values and a list of explanatory attribute values. By clicking the buttons of each itemset, human analysts can also see service reports, the trend chart, Weibull graph, and service and manufacturing date table to probe further. These charts and table provide a year-long trend for a certain defect and statistical evidence as to whether it is epidemic or not. With these functions they can easily judge if further investigation or action is necessary.

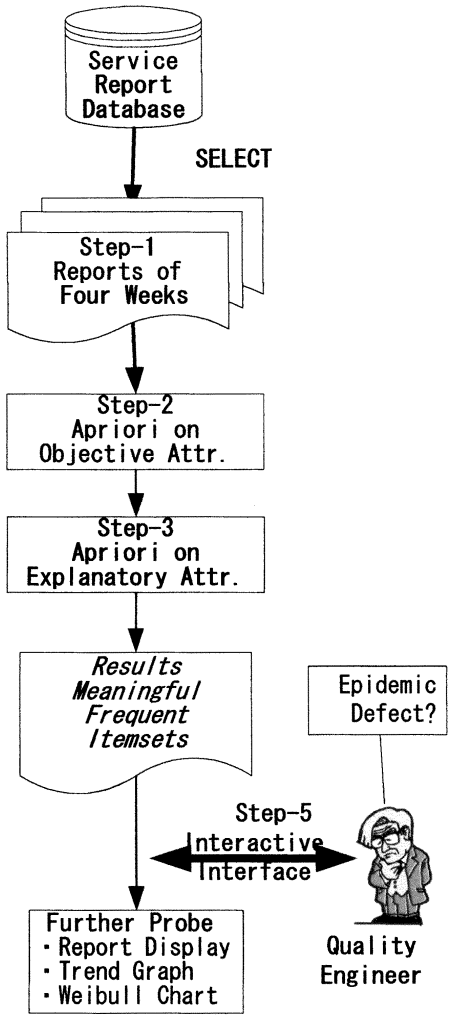

Fig. 3. Overview of watchdog program.

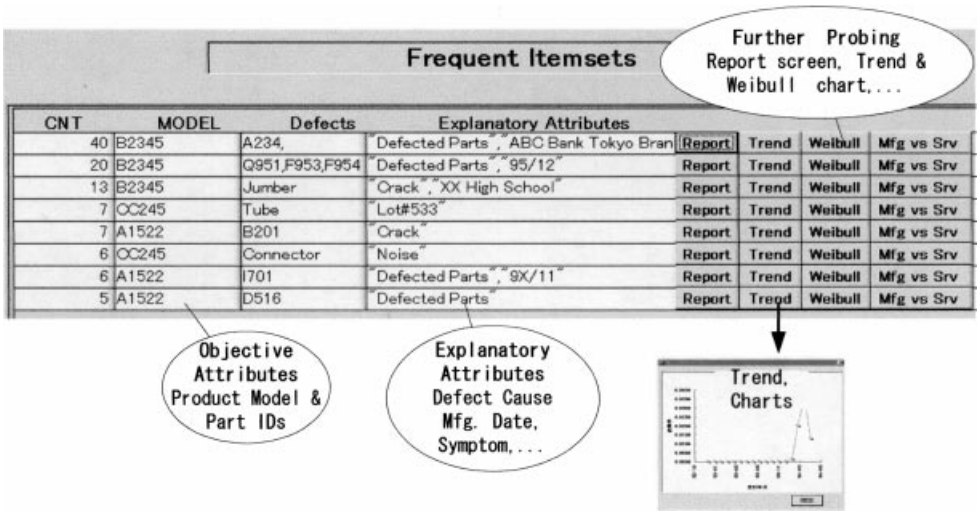

Fig. 4. Frequent itemsets screen.

3.3 Experimental results

We conducted an experiment on 1479 service reports whose repairs were performed within a month. Each service report consists of six attributes: product model, service date, symptom, defect cause, replaced Part-IDs, date of manufacturing.

The experiment was conducted in three steps:

(a) Generate frequent itemsets from the given service reports using the apriori algorithm with the minimum support record size 5.

(b) Generate frequent itemsets with the objective attributes of {Product model, Replaced Part-IDs} and the explanatory attributes of {Date of manufacturing, Defect cause, Symptom, User's name}.

Table 3 shows the experimental results. The apriori algorithm generated 759 frequent itemsets from 1479 service reports. On the other hand, only 136 frequent itemsets are derived using the objective and explanatory attributes. All of these itemsets contain both the product model and Part-IDs. After deleting overlapping items, we obtained 28 frequent itemsets.

The three epidemic incidents, which were considered epidemic and for which actions were taken to fix them, are ranked 7th, 13th, and 16th in these 27 frequent itemsets. On the other hand, they are ranked 251st, 414th, and 492nd, respectively, in the frequent itemsets generated by the apriori algorithm (Table 4). This experiment shows that the apriori generates too many frequent itemsets. Our method can dramatically reduce the size of the output without missing important service incidents. Therefore, human experts are able to easily scan the result and take action to improve product quality.

The major reason why the apriori generated many itemsets is that the minimum support was set at five. The epidemic incidents are hidden among meaningless itemsets. This suggests that we cannot discover valuable and

Table 3. Experimental results

	Freq.
Service Reports	1479
(a) Frequent Itemsets (Apriori)	759
(b) Frequent Itemsets (Our Method)	28

Table 4. Discovered epidemic incidents

Epidemic Incidents	Freq.	Order (a)	Order (b)
Case-A	11	251st	7th
Case-B	7	414th	13th
Case-C	6	492nd	16th

meaningful itemsets with frequency alone. However, our experience shows that exploiting background knowledge regarding attributes, our method can derive meaningful itemsets with the small minimum support threshold.

4. Conclusions

It is shown that the basket analysis—the apriori algorithm—is very effective in discovering important service incidents from a service report database and it has large advantages over the conventional quality analysis. However, the apriori program generates so many frequent itemsets that a human analyst cannot review them easily. This article showed that "basket analysis on objective and explanatory attributes" and "itemset reduction" are able to derive only valuable frequent itemsets. The experimental results demonstrated the effectiveness of the new method.

The watchdog program is deployed at a factory and used in daily routines. The quality engineers' response to this system is quite positive. The main reason for this success is that the system targeted extraction and reduction of data to a size and format suitable for human inspection [6]. The idea of our algorithm, exploiting the background knowledge regarding attributes, can be applied to the analysis of information stored in a relational database because the table columns—attributes—convey great meaning for the records.

Acknowledgments

Special thanks to Niall Murtagh for reviewing this article. Thanks to Toru Yukimatsu and Yoshitaka Kawashima for programming the watchdog.

REFERENCES

1. Agrawal R, Srikant R. Fast algorithms for mining association rules. Proc 20th VLDB Conference, 1994, p 487–499.
2. Cai Y, Cercone N, Han J. Attribute-oriented induction in relational databases. In Piatesky-Shapiro (editor). Knowledge discovery from databases. MIT Press; 1991. p 214–228.
3. Fayyad UM, et al. (editors). Advances in knowledge discovery and data mining. AAAI Press; 1996.
4. Matsuura H, Washio T, Motoda H. A principle and its implementation to extract association rules for estimation and prediction in data mining. Proc SICE System/Information Symposium '97, p 103–108.
5. Srikant R, Vu Q, Agrawal R. Mining association rules with item constraints. Proc 3rd Int Conf on Knowledge Discovery and Data Mining, 1997, p 67–73.
6. Huber PJ. From large to huge: A statistician's reactions to KDD & DM. Proc 3rd Int Conf Knowledge Discovery and Data Mining, 1997, p 304–308.

AUTHORS (from left to right)

Satoshi Hori received his B.E., M.E., and D.Eng. degrees from Tokyo Institute of Technology, and an M.S.E.E. degree from Purdue University. He worked at the manufacturing engineering center of Mitsubishi Electric Corporation, Japan. He is currently an associate professor at Monotsukuri Institute of Technologists. His research interests include artificial intelligence, statistics in the fields of diagnosis, maintenance, quality control, and field service. He is a senior member of IEEE, JSAI, IEICEJ, and IEEJ.

Hirokazu Taki received his B.E., M.E., and D.Eng. degrees from Osaka University. He worked at Mitsubishi Electric Corporation and researched knowledge acquisition at ICOT from 1986 to 1989. He is currently a professor of systems engineering at Wakayama University. His research interests include artificial intelligence, intelligent CAD, and robotics. He is a member of IEEE, JSAI, and IEEJ.

AUTHORS (continued) (from left to right)

Takashi Washio received his D.Eng. degree from Tohoku University. He was a visiting researcher at MIT in 1988. He worked at Mitsubishi Research Institute from 1990 to 1996. He is currently an associate professor at I.S.I.R, Osaka University. His research interests include artificial intelligence, diagnosis, and data mining. He is a member of AAAI, SICE, JSAI, and the Japanese Society of Fuzzy Theory.

Hiroshi Motoda is currently a professor at I.S.I.R, Osaka University. Before joining Osaka University, he worked at the central laboratory of Hitachi Ltd. His research interests include artificial intelligence, especially machine learning, knowledge discovery. He is a member of IEEE, AAAI, and the Japanese Society of Artificial Intelligence.