Data Mining for Hazard Elimination through Text Information in Accident Report

Shujiro Murayama\textsuperscript{a,*}, Koji Okuhara\textsuperscript{b}, Junko Shibata\textsuperscript{c}, Hiroaki Ishii\textsuperscript{b}

\textsuperscript{a}Department for Information Systems in Business, Hiroshima University of Economics, Japan
\textsuperscript{b}Department of Information and Physical Sciences, Graduate School of Information Science and Technology, Osaka University, Japan
\textsuperscript{c}Faculty of Economics, Kobe Gakuin University, Japan.

Received 6 February 2009; Received in revised form 9 September 2009; Accepted 4 November 2010

Abstract

Risk management is a management technique to process unexpected damage from various dangers at a minimum cost. It is important for the success of a project. In this paper, we try to extract the information concerning the risk by using the accident data in an actual enterprise. Our purpose is to extract the rule that avoids accidents from the accident data as a rule. Concretely, after making the scenario by using the KeyGraph, we extract the rule by using the rough set.

Keywords: Knowledge discovery, risk management, text data mining

1. Introduction

Business management is requested to invent the customer value and to bring the profit while holding various risks. Recently, problems of a lot of corporate scandals, the bankruptcies, and the unprofitable operations, etc. occur. It is necessary to management of risk for taking measures beforehand and preventing such a problem. It is defined with the risk, “Possibility that a loss, damage, and other events not preferable happen”. The meaning of the risk analysis decreases the relapse in investigating the cause of the risk, and kneading, and executing the action plan on the other hand, and it prevents, and the loss and damage are decreased at least. The obtained action plan is shared as knowledge. It is thought that it is risk management that the above-mentioned is systematized as an organization.

In several researches, the methodology of risk management is difficult for an insufficient tool, when actually using it for industry. Some researches pay attention to the risk which might produce a profit on the other hand though the loss when actualizing it occasionally occurs. There are a lot of researches from global standpoint of management, and the research of the case base is a little. In this paper we assume that definition of risk can be considered by the product of probability of occurrence and impact of event occurring (refer to Figure 1).

On the other hand, the rough set and self-organizing map (SOM) are used as a new method of the knowledge discovery by the high-speed automatic operation analysis of large-scale data by the computer. The rough set has been proposed by Pawlak (1991), and has developed as a method of the knowledge acquisition. The decision rule of the rough set gives the combination of minimum attribute values to identify a target achievement level in data as a conditional part of the rule. Unknown data is inductively inferred by the conditional part of the rule. In this paper, we use the rough set in which the ordinal between the attribute values is considered for the qualitative data. The SOM is one of the most popular artificial neural

* Corresponding author. Email: Sj-mura@hue.ac.jp
networks. The SOM is a type of unsupervised learning, designed to build a neural network where nearby neurons represent similar inputs. It is also called the topology-preserving map. SOM goes organizing them for representing the samples. Units are also allowed to change themselves by learning in order to fire more like samples in hopes of winning the next competition. This learning process makes the weights organize themselves into a map representing similarities.

The definition of the risk management is described as follows (Alexander, 2008; David, 2002; Kammen and Hassenzahl, 1999; Shen, 1997; Toshiaki, 1984; Uher and Toakley, 1999; Bedford and Cooke, 2001); In ISO/IEC Guide 73(2002), the risk management is defined “Activity that commands the organization for the risk and controls”. Moreover, the risk is defined “Combination of the probability of an event and its consequence”. In general, the quantification of the risk is defined as of “Risk = probability of occurrence * influence level”. In this paper, it applies to this definition. Moreover, as shown in Figure 1, the procedure of the risk management is described as:

(a) The event is chosen;
(b) The risk analysis is done;
(c) As a result of the analysis, the risk evaluation works;
(d) Measures and the crisis evasion are settled on if the evaluation is within a tolerance level.

So, we investigate the problem that occurs in the enterprise, clarify the cause of the risk, and propose the method of new risk management that considers the risk prevention by using these techniques. In this study we propose the framework of data mining based on regression analysis and rough set analysis and SOM in risk management. Especially, we focus on knowledge discovery based on real accident report of a company. In addition, we discover the rule to take precautions against possible accidents after significant information is extracted from it. Figure 2 shows the method of solving the proposed problem (Murayama et al., 2010).
2. Risk management proposal for Food Company

Food Company is a food-processing industry. This risk management aims at improving “quality control” and it is in line with the direction of the whole management of the company. The quality control items of the Food Company are as follows (called 6M):

(a) Machines / equipments,
(b) Man (Human resources),
(c) Materials / resources,
(d) Methods,
(e) Management,
(f) Money.

Almost all the production processes have been automated, but some processes, especially inspection and mixing ones, can be done only by humans. This research takes up mixing process. Accident recurrence prevention reports are issued about 50 times (in the report rule due to the quality accident, there is an obligation of the report of 50 minutes or more the line stop time when the amount of money of damage is 50,000 yen or more) throughout the year per factory. But under the slogan of “preventing an accident from happening again”, the company has been engaged in risk management, focusing on quality control section. We put together a risk management proposal for a company, as outlined in Figure 3 and suggested and explained it and gained approval.

Figure 2. Method of solving problems.
Accident recurrence prevention project was organized and four employees except executives were selected to proceed with the operations in the future. The key points of the proposal are as follows:

(a) Text mining is applied to the accident causes parts stated in the accident recurrence prevention reports and problems are extracted. After group discussions by experts, a solution is developed;

(b) A developed solution and the solution proposed by the section that had an accident are compared and reviewed;

(c) Reviewing the two solutions brings about the difference and from it, new knowledge is gained;

(d) A new solution to the problem is developed by reviewing;

(e) After the risk is analyzed, the risk is evaluated as a procedure of the risk management (refer to Figure 1). It judges the permissible range of the evaluation result, and if it is a permissible range, measures, maintenance, and the crisis evasion and making to localizing are performed;

(f) The new knowledge will be converted into explicit knowledge in the future, and knowledge will be shared.

The explicit knowledge newly found is shared by Food Company’s group, and the method becomes a horizontal affiliation as a group enterprise, and will spread. As a result, the quality accident number will decrease.
3. Proposed process for finding out problems

In order to execute the proposal, a research plan was drawn, and tools were selected. The selected tools are as follows:
(a) “ChaSen” is applied for a morpheme analysis (Chasen HomePage);
(b) “KeyGraph” is applied for extracting key words (Ohsawa et al., 1999);
(c) A regression analysis and a rough set analysis are applied for extracting a rule;
(d) SOM is applied for analyzing the spread of the process on pecuniary loss.

Based on the above-mentioned premises, a new suggestion is presented. A table is made from problem/solution keywords in order to make an accident clearly understood. This table is named “decision table” meaning one can make a judgment or a decision. A decision table is filled with data extracted from accident recurrence prevention reports. A rough set analysis and a multivariate analysis are used as a means of picking out rules and causes from the plotted data. These extracted rules and causes are translated into measures against accident recurrence.

If such measures are implemented at a factory, that would prevent an accident from happening. If the decision table is added to the daily report which is written when an accident happens, it would serve as a checklist, helping clarifying the surroundings of the accident and it would also be utilized for the accident analysis. As time goes by, the decision table is repeatedly reviewed and ameliorated. From it, new knowledge is found and it is formalized and shared. After I explained these points, approval was given. The policy modification of a company was used as the occasion for reconfirming they would continue to be engaged in an accident recurrence prevention project.

3.1 Supporting risk management by regression analysis

In the data treated by this regression analysis, the explanatory variable is qualitative variable. It is necessary to encode dummy variable to analyze it. It becomes available as independent variable because it encodes it. It is quantification theory I application. Moreover, to find the best item (explanatory variable) group in this analysis, explanatory variable selection criterion (RU) was applied.

Using multiple regression analysis, we analyze the factors that affect the land. The equation of the multiple regression analysis on the land is given as

\[ \ln y = \alpha_1 \ln x_1 + \alpha_2 \ln x_2 + \cdots + c + \varepsilon \]  

where \( y \) is explained variable, \( x \) is explanatory variable, \( a \) is regression coefficient, \( c \) is constant, \( \varepsilon \) is residual. The selection criteria and explanatory variables are

\[ Ru = 1 - \frac{(1 - R^2)^* (n + k + 1)}{n - k - 1} \]  

where \( R \) is correlation coefficient, \( n \) is, number of data, \( k \) is degrees of freedom of the regression analysis of variance table.

This analysis method has a feature of being able to analyze even a certain level of small amount of data. The results tend to be comprehensive but this method is very helpful as an analyzer can obtain convincing results.
3.2 Supporting risk management by rough set analysis

To support the risk management, their understanding of the current situation and the methods of mission and work were confirmed by the following process: (a) Requirements for improvement, (b) How to see KeyGraph and creation of a scenario, and (c) Extraction of keywords - determining conditional attributes for a decision table (hereafter referred to decision items) and so on.

Concerning (a) and (b), the work has been being continued. As for (c), along with the project members, looking at the KeyGraph, we examined whether or not the first edition of a decision table was consistent. What the members discussed was put into the KeyGraph and they implemented the Chance Discovery Double Helix Process. After that, they began to determine decision items and categories. As a result, the third edition of the decision table was created as shown in Table 1.

Table 1. Decision table (the third edition).

<table>
<thead>
<tr>
<th>Decision table at the time of reporting an accident</th>
<th>reporter:</th>
</tr>
</thead>
<tbody>
<tr>
<td>process</td>
<td></td>
</tr>
<tr>
<td>item</td>
<td>Category 1</td>
</tr>
<tr>
<td>the number of workers</td>
<td>1: one person</td>
</tr>
<tr>
<td>years of service</td>
<td>1: within a half year</td>
</tr>
<tr>
<td>production items</td>
<td>1: one item</td>
</tr>
<tr>
<td>production frequency</td>
<td>1: once a week</td>
</tr>
<tr>
<td>working shift</td>
<td>1: not exist</td>
</tr>
<tr>
<td>procedure manual</td>
<td>1: not exist</td>
</tr>
<tr>
<td>check list</td>
<td>1: not exist</td>
</tr>
<tr>
<td>checker</td>
<td>1: not exist</td>
</tr>
<tr>
<td>means of confirming the work</td>
<td>1: machine analysis</td>
</tr>
<tr>
<td>proficiency degree</td>
<td>1: low</td>
</tr>
<tr>
<td>accident outbreak time</td>
<td>1: start time</td>
</tr>
<tr>
<td>accident outbreak day of the week</td>
<td>1: Sunday</td>
</tr>
<tr>
<td>evaluation item</td>
<td>accident outbreak</td>
</tr>
<tr>
<td>the date of the previous accident</td>
<td></td>
</tr>
<tr>
<td>the amount of loss</td>
<td></td>
</tr>
<tr>
<td>comment</td>
<td></td>
</tr>
</tbody>
</table>

As a way of extracting rules, we proposed a rough set analysis taking into consideration of ordinal number and applied it (Okuhara et al., 2004). The existing rough set analysis had a possibility of picking out conflicting rules. What we proposed is an algorithm-based rough set analysis that can exclude such rules.

We explain the rule extraction by the rough set which took into consideration the ordinal number nature between attribute values to quantitative data. Assuming that data consists of $S$ samples, we consider the decision table which consists of $N$ condition attributes and $M$ determination attributes. It is called criteria what can assume ordinal number nature among condition attributes. Here $O_q$ is set as the out ranking relation based on the $q$th criteria. Namely, if based on the $q$th criteria, $x O_q y$, describes that $x$ is a good thing of same extent as $y$
at least. Furthermore, in the \( m \)th decision attribute, we consider the division 
\[ \mathcal{R}_m = \{ C^1_m, C^2_m, \ldots, C^R_m \} \] 
of \( U \) which has the properties that the arbitrary elements of \( C^s_m \) are more desirable than all the elements of \( C^t_m \) for \( s > t \). Here, the relation 
\( (x \in C^s_m, y \in O^t_m, s > t \Rightarrow x \mathcal{O}_m y \text{ and } \sim y \mathcal{O}_m x) \) is satisfied.

The \( m \)th decision attribute is classified into \( R \) classes
\[ C^s_m \cap C^t_m = \emptyset, \quad (s \neq t), \quad C^R_m > \cdots > C^t_m > C^1_m \] 
(3)

Then, if \( x \in U \) is given, we can define the bottom accumulation set \( C^{sr}_m \) which is a set of the element of \( U \) at least belonging to the class \( C^{sr}_m \), and the bottom accumulation set \( C^{dr}_m \) which is a set of the element of \( U \) at most belonging to the class \( C^{dr}_m \) as follows:
\[ C^{sr}_m = \bigcup_{s \geq r} C^s_m, \quad C^{dr}_m = \bigcup_{s \leq r} C^s_m \] 
(4)

Assuming \( V \subseteq W \) when \( W \) denotes set of all criteria. About arbitrary \( v \in V \), it is said that \( X \) dominates \( Y \) in \( V \) when \( x \mathcal{O}_m^v y \) is satisfied. It is represented by \( xD^v_m y \) and is defined as follows:
\[ xD^v_m y \leftrightarrow g(x, n) \succ g(y, n), \quad (\forall v \in V) \] 
(5)
where \( g(x, n) \) denotes the attribute value with respect to the attribute \( n \) of a sample \( x \). When \( g(x, n) \) and \( g(y, n) \) are represented by interval \( S[s_1, s_2] \) and \( T[t_1, t_2] \) respectively, \( xD^v_m y \) means \( S[s_1, s_2] > T[t_1, t_2] \leftrightarrow s_1 > t_1, s_2 > t_2 \) 
(6)

If \( x \in U \) is given in the \( m \)th decision attribute, then, the set \( D^v_m(x) \) of the element of \( U \) which dominates \( x \) in \( V \), and the set \( D^{-v}_m(x) \) of the element of \( U \) which is dominated by \( x \) in \( V \) can be defined by
\[ D^v_m(x) = \{ y \in U \mid yD^v_m x \}, \quad D^{-v}_m(x) = \{ y \in U \mid xD^v_m y \} \] 
(7)

The lower approximation set \( V^l(C^{2r}_m) \) and the upper approximation set \( V^u(C^{2r}_m) \) of the accumulation set \( C^{2r}_m \) by the domination set \( D^v_m(x) \) can be defined as follows:
\[ V^l(C^{2r}_m) = \{ x \in U \mid D^v_m(x) \subseteq C^{2r}_m \}, \quad V^u(C^{2r}_m) = \bigcup_{x \in C^{2r}_m} D^v_m(x) \] 
(8)

From this lower approximation set, the rule which the data \( x^* \) dominating \( x \) in \( V^l(C^{2r}_m) \) surely belongs to upper class than \( r \) is derived. That is, for the certain \( x^* \in C^{2r}_m \), the if-then rule can be obtained as follows:
\[ \text{If } g(x^*, n_1) \succ g(x, n_1) \text{ and } g(x^*, n_2) \succ g(x, n_2) \ldots \text{and } g(x^*, n_y) \succ g(x, n_y), \text{ then } x^* \in C^{2r}_m \] 
(9)

By the same way, for the certain \( x^* \in C^{2r}_m \), we can also obtain the if-then rule.
3.3 Supporting risk management by self organizing map

Self-organizing map (SOM) is one of the neural networks studied by the rival neighborhood study without the teacher proposed by Kohonen (1982). The map from a higher dimension space to the low level space is formed systematically, and it can be visualized on two dimensional planes from higher dimension data. Such image is called a feature map.

SOM is neural network composed by two layers such as the input layer and the output layer. The output layer is called a rival layer. It is assumed that input vector is \( x_j = (x_{j1}, x_{j2}, x_{j3}, \ldots, x_{jn}) \), and output is \( m_i = (m_{i1}, m_{i2}, m_{i3}, \ldots, m_{in}) \) at the node in output layer \( i \) \( (i = 1, 2, \ldots, I) \) like Figure 4.

Vector \( m_i \) is updated by unsupervised learning.

![Figure 4. Structure of self-assembler map.](image)

Algorithm of SOM (Okita et al., 2007)

Step 1: All the reference vectors are initialized.

Step 2: The input vector is sequentially selected from the head of the input data set. When the number of steps is larger than the number of input data, it selects it from making to the head.

Step 3: The selected input vector is compared with all the reference vectors, and the node with the reference vector in which Euclidean distance is minimized is made a node of the most much adjustment. The node of the most much adjustment is defined by the following expressions.

\[
C = \arg\min_i \{\|x_i - m_i\|\} \quad i = 1, 2, \ldots, I
\]  

Step 4: The node of the most much adjustment and the reference vector of the neighborhood node are renewed according to the following expressions.

\[
m_i(t + 1) = \begin{cases} 
m_i(t) + h_{ci}(t) \left[ x_i(t) - m_i(t) \right] & i \in N_c \\
m_i(t) & i \notin N_c \end{cases} 
\]

where \( h_{ci}(t) \) is a neighborhood function at time \( t \) \( (t = 1, 2, \ldots, T) \)

\[
h_{ci}(t) = \alpha(t) \exp \left( -\frac{||r_c - r_i||^2}{2\sigma^2(t)} \right) 
\]

where \( a(t) \) is a learning coefficient, \( \sigma^2(t) \) is a variance. \( r_c \) and \( r_i \) are vectors on 2 dimensional spaces.
4. Analysis results of daily reports

In fact, a company put data from the past accident recurrence prevention reports into the newly-made decision table at the time of reporting an accident (please see Table 1) as a sample. Making use of these data, they conducted an analysis in order to find out problems and then finally grope for a new solution to them. They chose 54 cases of Quality accident data of headquarters factory which have few missing value, did an analysis and obtained an interesting answer from the analysis results. In usual, amount of product is 10,000 lots per one year and 100 accidents occur in 4 factories totally, that is, 1%. Therefore 54 cases is not small amount it is equal to one year. We explain the case and its assessment.

4.1 Results of usual method by regression analysis

We calculated a regression coefficient by doing a regression analysis and by using it, found the incidence degree of all decision items. Table 2 shows the processing result of the regression analysis by SPSS. The combination when RU is calculated according to explanatory variable selection criterion (RU), and the value becomes the maximum becomes the best combination in this table (refer to Table 3). The item of 14 is set to an initial decision table (Table 1). However, the calculation result of RU was limited to 12 pieces. It means the interpretation of the explanatory variable selection criterion is enough by 12 items. Table 2 shows a regression analysis result in the chosen item according to explanatory variable selection criterion (RU).

Table 4 shows an incidence degree for each item. The calculation of the impact is a result of pulling minimum value from the maxima of the item category coefficient.

Figure 5 shows the Incidence degree each item in Table 4 in the graph. From the bar chart based on the results of the incidence degree analysis (see Figure 5), the items most likely causing an accident in descending order are (a) years of service, (b) accident outbreak day of the week, (c) accident outbreak time, (d) proficiency level, and (e) production frequency.

Above-mentioned results are quite a commonsense view and anybody would raise such reasonable factors.
Table 2. Results of regression analysis by SPSS.

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R2 power</th>
<th>Error margin of standard deviation estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.787</td>
<td>0.620</td>
<td>0.084</td>
</tr>
</tbody>
</table>

### Coefficient a

<table>
<thead>
<tr>
<th>Model</th>
<th>Standard deviation error margin</th>
<th>Standardization coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Selection item</th>
<th>1. worker</th>
<th>2. the number of workers</th>
<th>3. years of service</th>
<th>4. production items</th>
<th>5. production frequency</th>
<th>6. working shift</th>
<th>7. procedure manual</th>
<th>8. checker list</th>
<th>9. proficiency level</th>
<th>10. means of confirming the work</th>
<th>11. accident outbreak time</th>
<th>12. accident outbreak day of the week</th>
<th>RU</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Items</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.295209</td>
<td>0.919071</td>
</tr>
<tr>
<td>12 influence level high-ranking items</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.295209</td>
</tr>
</tbody>
</table>

Table 3. Selected items based on RU.
Table 4. Incidence degree for each item.

<table>
<thead>
<tr>
<th>Model</th>
<th>Item name</th>
<th>Incidence degree</th>
<th>MAX</th>
<th>MIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>the number of workers</td>
<td>0.152</td>
<td>0.000</td>
<td>-0.152</td>
</tr>
<tr>
<td>3</td>
<td>years of service</td>
<td>1.090</td>
<td>0.884</td>
<td>-2.06</td>
</tr>
<tr>
<td>4</td>
<td>production items</td>
<td>0.259</td>
<td>0.000</td>
<td>-0.259</td>
</tr>
<tr>
<td>5</td>
<td>production frequency</td>
<td>0.443</td>
<td>0.000</td>
<td>-0.443</td>
</tr>
<tr>
<td>6</td>
<td>working shift</td>
<td>0.121</td>
<td>0.000</td>
<td>-0.121</td>
</tr>
<tr>
<td>7</td>
<td>procedure manual</td>
<td>0.119</td>
<td>0.119</td>
<td>0.000</td>
</tr>
<tr>
<td>8</td>
<td>check list</td>
<td>0.292</td>
<td>0.292</td>
<td>0.000</td>
</tr>
<tr>
<td>10</td>
<td>worker judgment</td>
<td>0.236</td>
<td>0.000</td>
<td>-0.236</td>
</tr>
<tr>
<td>11</td>
<td>means of confirming the work</td>
<td>0.352</td>
<td>0.051</td>
<td>-0.301</td>
</tr>
<tr>
<td>12</td>
<td>proficiency level</td>
<td>0.645</td>
<td>0.591</td>
<td>-0.054</td>
</tr>
<tr>
<td>13</td>
<td>accident outbreak time</td>
<td>0.765</td>
<td>0.448</td>
<td>-0.317</td>
</tr>
<tr>
<td>14</td>
<td>accident outbreak day of the week</td>
<td>0.872</td>
<td>0.869</td>
<td>-0.003</td>
</tr>
</tbody>
</table>

Figure 5. Incidence degree analysis by each category.

As for the items that have high incidence degree cited in figure, we calculated the incidence degree by each category. The results are shown in the graph below:
(a) The incidence degree by years of service.

If you take a look at the incidence degree by years of service in the Figure 6, it is highly likely that a worker having within a half year of service after joining the company causes an accident. This is also a very sensible conclusion.
According to the incidence degree by accident outbreak day of the week in the Figure 7, it is highly likely that more accidents do happen on Tuesday and Friday. The tension on Monday that is the work start might cause the accident cause on Tuesday. It seems that the result of concentrated level's decreasing because the tiredness of one week is accumulated on Friday in the weekend. The line of the last under line displays the extracted rule.

(c) The incidence degree by accident outbreak time.

When seeing the incidence degree by accident outbreak time as shown in Figure 8, it can be said that more accidents happened in the start time, the lunchtime break operation, and the afternoon. It is clear that a lot of accidents occur in the operation in the afternoon. Lunch break operation is thought that hungry dulls concentrated level to work. Tiredness piles up, and there is a possibility of causing the accident in the afternoon. It is possible to increase the rest frequency in the afternoon as measures.
4.2 Rough set result of analysis

Table 5 shows the proposed rough set analysis results. This table is a part of the output slip of the result of analysis in rough set analysis program. The numeral after the character of (Rule) of the first line shows the probability of occurrence. It is 96.29% in this analysis. The content of actual data is displayed from the second line. The line of the last under line displays the extracted rule. The display of an analysis of this content is shown on the first line of Figure 6.

Table 5. Proposed rough set analysis results.

<table>
<thead>
<tr>
<th>Rule(0.962963)_10332</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) IF (3) = 0 0 0 0 1 (4) = 1 0 0 (5) = 0 0 0 1 0 (7) = 0 1 (12) = 0 0 0 0 0 0 1 (14) = 0 1 0 0 0 0 THEN (1) = 1</td>
<td></td>
</tr>
<tr>
<td>2) IF (3) = 0 0 0 0 1 (4) = 1 0 0 (5) = 1 0 0 0 0 (7) = 1 0 (12) = 0 0 0 0 0 1 1 (14) = 0 0 0 1 0 0 THEN (1) = 1</td>
<td></td>
</tr>
<tr>
<td>3) IF (3) = 0 0 0 1 (4) = 1 0 0 (5) = 1 0 0 0 0 (7) = 1 0 (12) = 0 0 0 0 0 1 1 (14) = 0 1 0 0 0 0 THEN (1) = 1</td>
<td></td>
</tr>
<tr>
<td>4) IF (3) = 0 0 0 1 0 (4) = 0 1 0 (5) = 0 0 0 0 1 (7) = 0 1 (12) = 0 0 0 0 0 1 0 1 (14) = 0 0 0 0 0 1 0 THEN (1) = 1</td>
<td></td>
</tr>
<tr>
<td>5) IF (3) = 0 0 0 0 1 (4) = 0 1 0 (5) = 0 0 0 0 1 0 (7) = 0 1 (12) = 0 0 0 0 0 1 0 1 (14) = 0 1 0 0 0 0 THEN (1) = 1</td>
<td></td>
</tr>
<tr>
<td>6) IF (3) = 0 0 0 0 1 (4) = 0 1 0 (5) = 0 0 0 0 1 0 (7) = 0 1 (12) = 0 0 0 0 0 1 1 0 (14) = 0 0 0 0 0 0 1 0 THEN (1) = 1</td>
<td></td>
</tr>
</tbody>
</table>

In this analysis, two rules have been extracted (refer to Tables 6 and 7). Only one rule for each is explained in full. From the proposed rough set analysis, very clear-cut rules were presented. Rules when damage doesn't occur are as follows (refer to Table 6):
(a) Workers have five or more years of service;
(b) The production items are two;
(c) The production frequency is once a half month;
(b) The procedure manual exists;
(e) The accident outbreak times are unknown. It is possible that the data may be missing;
(f) The accident outbreak day of the week is Tuesday.
Table 6. Obtained result of rule when damage doesn’t occur.

<table>
<thead>
<tr>
<th>Rule when damage doesn’t occur</th>
<th>3. years of service</th>
<th>4. production items</th>
<th>5. production frequency</th>
<th>7. procedure manual</th>
<th>9. checker</th>
<th>13. accident outbreak time</th>
<th>14. accident outbreak day of the week</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF (3) = 0 0 0 0 1 (4) = 0 1 0 (5) = 0 0 1 0 0 (7) = 0 1 (13) = 0 0 0 0 0 1 (14) = 0 0 1 0 0 0 THEN (1) = 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IF (3) = 0 0 0 0 1 (4) = 0 1 0 (5) = 0 0 1 0 0 (9) = 0 1 (13) = 0 0 0 0 0 1 (14) = 0 0 1 0 0 0 THEN (1) = 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Rules when damage occurs are as follows (refer to Table 7):
(a) Workers have within a half year of service;
(b) The production items are two;
(c) The production frequency is one at half month;
(d) The Procedure manual exists;
(e) The accident outbreak time is afternoon;
(f) The accident outbreak day of the week is Tuesday.

Table 7. Obtained result of rule when damage occurs.

<table>
<thead>
<tr>
<th>Rule when damage occurs</th>
<th>3. years of service</th>
<th>4. production items</th>
<th>5. production frequency</th>
<th>7. procedure manual</th>
<th>9. checker</th>
<th>13. accident outbreak time</th>
<th>14. accident outbreak day of the week</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF (3) = 1 0 0 0 0 (4) = 0 1 0 (5) = 0 0 1 0 0 (7) = 0 1 (13) = 0 0 0 1 0 0 0 (14) = 0 0 1 0 0 0 0 THEN (1) = 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IF (3) = 1 0 0 0 0 (4) = 0 1 0 (5) = 0 0 1 0 0 (9) = 0 1 (13) = 0 0 0 1 0 0 0 (14) = 0 0 1 0 0 0 0 THEN (1) = 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8 shows the selected items by the rough set analysis. Moreover, it is a comparison table of the item selected by the item (refer to Table 3) chosen by explanatory variable selection criterion (RU) and the rough set analysis. Item group in the rough set is a least common multiple when the rule is extracted. Item group requested by the regression analysis result as a result of comparing these two item groups by explanatory variable selection criterion (RU) is growing.

Because the item to which item group chosen by explanatory variable selection criterion (RU) is extracted by the rough set has been included it is possible to say as a result, it can be judged that there is a conformance in this verification.
Table 8. Selected items by rough set analysis.

<table>
<thead>
<tr>
<th>Selection item</th>
<th>1. worker</th>
<th>2. number of workers</th>
<th>3. years of service</th>
<th>4. production items</th>
<th>5. production frequency</th>
<th>6. working shift</th>
<th>7. procedure manual</th>
<th>8. check list</th>
<th>9. worker judgment</th>
<th>10. means of confirming the work</th>
<th>11. proficiency level</th>
<th>12. accident outbreak time</th>
<th>13. accident outbreak day of the week</th>
<th>RU</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item group selected by RU</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.29521</td>
<td>0.813717</td>
<td></td>
</tr>
<tr>
<td>Item in rough set</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.20670</td>
<td>0.45797</td>
</tr>
</tbody>
</table>

4.3 Feature map obtained from SOM

In addition, we make the feature map of the working process in the amount of money loss of the accident based on the above-mentioned result. And, we examine the distribution of the working process by the amount of money loss. Here, the amount of money loss is divided into five stages. Figure 9 shows the feature map of the working process obtained by the simulation. The number of parentheses shows the stage of damage in figure. The high number shows that the amount of money loss is large.

Figure 9. The feature map of the working process obtained by the simulation.

Here, we pay attention about the accident with a large amount of money loss such as the process Pectin measure room (R_P (5)), Tank 3(Tank3 (5)), Storage before disinfect process (Sto (4)), and Cleaning (Clean (4)). These processes are shown by the bold type. It is thought that these are similar accidents because these accidents exist in the short distance. Especially, we note the accident of M(1) (Material process) and F4(2) (Filling process) that exists near them. Both loss levels are low. However, it is understood that the level of the amount of a loss might have risen if making a mistake by one step. Moreover, it is thought that the possibility of the accident of M(1) that the amount of a loss becomes a large accident was higher than that of F4(2) from the rule obtained ahead.
5. Evaluation of analysis results

5.1 Comparison of analysis technique

Figure 10 shows the evaluation of the analysis technique. Three kinds of analytical procedures were applied in the verification in Food Company. It turned out that the result was able to supplement of each. In the overall judgment, it is thought that a more adequate result of analysis is obtained by combining two analytical procedures.

<table>
<thead>
<tr>
<th>Methods of analysis</th>
<th>Regression analysis</th>
<th>Proposed rough set</th>
<th>Self organizing map (SOM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domination point</td>
<td>The result can be put out even by little data. And because the operation is easy, the proficiency is high. Moreover, “Rule” can begin to be seen from the relativity between data.</td>
<td>It is easy for the point to become clear, and to set up the countermeasure because high “Rule” can be extracted stochastically.</td>
<td>It is comprehensible in the user because it is expressible by visual.</td>
</tr>
<tr>
<td>Note</td>
<td>There is a possibility that software is separately needed when the number of items and the number of categories increases. When the number of items and the number of categories increase, a lot of numbers of data are needed based on it. Number of data &gt; Number of item * Number of categories</td>
<td>It is necessary to evaluate the category (Normalize by weighting etc.).</td>
<td></td>
</tr>
<tr>
<td>Overall judgement</td>
<td>The result of being possible to consent commonsensibly is obtained though a total flower. It is effective in making of the decision table proper.</td>
<td>It becomes a persuasive analytical procedure by the combinations such as the regression analysis, SOM, and rough sets and SOM.</td>
<td></td>
</tr>
</tbody>
</table>

Figure 10. Comparison and evaluation of analysis result.

5.2 Evaluation and results for food company

Wisdom obtained from the result of analysis in Food Company is as follows: (a) The factor to cause the accident became clear; (b) Factorial experiment of the quality accident that suits the site by devising the decision table can be done; (c) Because the cause was clarified, the decision of the countermeasure became possible. The following operation was led being able understanding, and the confirmation of these. To make the original decision table, the item and the category item were reviewed in each manufacturing premise; Moreover, (d) revision of incompleteness of accident relapse prevention report.

It is evaluated that the effect as a new approach on each factory is large. On the other hand, a new approach as the headquarters was born. The importance of the manufacturing worker education became remarkable. The case database has already been made and sharing information progresses. Exactly, the innovation of the fabrication sector started. A current period required two years. The briefing to manufacturing place section reaches ten several-time in total. The manufacturing premise has confidence to the technology that absolutely becomes it. It has the custom of not esteeming others' opinions. Especially, I do not want to hear the opinion of the accident from the outside. The accident factor was concretely specified by this approach. The content obtained the understanding of the person in charge of manufacturing. It is a result that the experience value of the person in charge of manufacturing and the validity of the result of analysis were able to be understood. The features of this proposal are given as follows:
(a) There is handiness of the registration of the accident data. In a word, the data requirement becomes clear because it uses the decision table, and the report leakage etc. can be prevented. Moreover, the speedup of the report can be achieved;
(b) The analytical work can be done enough even by a little education;
(c) The scenario of the solution can be drawn from the result of analysis.

Actually, the Performance is not alleged though there is a desire to the contingency planning. The accident often works the profit and loss of the enterprise right and left.

Ameliorable even a little is a condition that the future can remain live.

6. Conclusions

In this paper, we focused on knowledge discovery based on a real accident report of a company. We extracted the significant information from the accident report using the proposed method. From computer simulations, it was understood that as workers with up to a half year of service have more accidents, the plans for the on-site training programs are now being worked out. Moreover, we showed the data that the level of the amount of a loss might have risen if making a mistake by one step.

References

Chasen HomePage (http://chasen.naist.jp/hiki/ChaSen/)