International Journal of Innovative Computing, Information and Control
Volume **9**, Number **9**, September **2013**

ICIC International ©2013 ISSN 1349-4198
pp. 3573–3584

KNOWLEDGE DISCOVERY BASED ON FUZZY, ENTROPY AND DOMINANCE RELATION

Koji Okuhara[1], Chien-Hsing Wu[2], Hiroshi Tsuda[3]
Hiroe Tsubaki[4] and Noboru Sonehara[5]

[1]Department of Information and Physical Sciences
Graduate School of Information Science and Technology
Osaka University
1-5, Yamadaoka, Suita, Osaka 565-0871, Japan
okuhara@ist.osaka-u.ac.jp

[2]Department of Information Management
National University of Kaohsiung
No. 700, Kaohsiung University Rd., Nan-Tzu Dist., Kaohsiung 81148, Taiwan

[3]Department of Mathematical Sciences
Doshisha University
Kyotanabe, Kyoto 610-0321, Japan

[4]Department of Data Science
The Institute of Statistical Mathematics
10-3, Midori-cho, Tachikawa, Tokyo 190-8562, Japan

[5]Information and Society Research Division
National Institute of Informatics
2-1-2, Hitosubashi, Chiyoda-ku, Tokyo 101-8430, Japan

Received August 2012; revised December 2012

Abstract. *Most intelligent evaluation systems generally involve gaining information. Inference knowledge (e.g., decision rules) can be elicited through the gained information that is usually modeled as a set of data (e.g., Kansei ergonomic-based empirical survey). While the induction groups of technique are being used to automatically discover knowledge with respect to decision rule generation, the approach that can remove uncommon data, discover less reliable analyzed data, and detect inconsistent rules is complex, but important to the final outputs. In this paper, we present a model for decision rule discovery that utilizes (1) fuzzy quantification to normalize and group collected data, (2) entropy to detect unreliable data, and (3) dominance relation to return decision rules without contradiction. An example that deals with the decision rule discovery for website design quality is illustrated to demonstrate the applicability of the proposed model.*

Keywords: Decision rule, Fuzzy quantification, Entropy, Dominance relation, Website design quality

1. **Introduction.** With the increased use of information processing technology, rule-based systems are being broadly introduced to support decision making. While stress is being put on their applications, a critical factor that affects the performance and reliability is the quantity and quality of the rules that are stored in the rule-based format of IF condition, and THEN action. Consequently, rule discovery in databases becomes one of the major tasks in the rule-based system development life cycle [1, 2, 3, 4, 5]. Crucial issues including methods, processes, and techniques for knowledge acquisition are therefore being discussed in the applied artificial intelligence and machine learning community. Domain experts, domain facts and empirical survey data are the major knowledge sources that rule

discovery is concerned with. Essentially, rule discovery begins when knowledge engineers interview domain experts and then translate the gained information into the form of rule bases without changing the initial meaning. However, due to the low discovery efficiency, researchers developed the discovery tools to help elicit domain knowledge.

Particularly, the rule base of a rule-based system is to accurate the reflection of the current knowledge at the time the system is placed in service. However, the collected source usually comes with perturbation because of the human cognition that may cause contradiction or low reliability. For example, the Kansei ergonomics is one of the techniques used to reflect human feelings for a subject design (e.g., a product, a system, a plan), [6, 7, 8, 9]. It deals with the human perception acquisition that is measured by the scaled questionnaire of semantic differential method to reflect their impression, preference and individual differences [10]. The collected data from questionnaires are generally assumed to be involving psychological ambiguity at the time the survey is performed, and in consequence may result in unreliable or conflicting decision rules. Therefore, the knowledge source collected should be regularly reanalyzed while performing rule generation to keep the rule base current and reliable.

The data one has to deal with is not always given in a very precise and consistent form. It is very important that numerical data is taken into account while collecting domain information. The data collected from questionnaire usually contains contradiction because of fuzziness of human perception. Consequently, transformation of data value is necessary due to information consistency requirements to reduce the perturbation. The innovation and the significance of this paper are as follows: development of a knowledge discovery approach that employs fuzzy quantification to normalize and group survey data, secondly, entropy to detect unreliable data, and thirdly, dominance relation to return decision rules without contradiction to help generate decision rules in Kansei ergonomics-based survey data for any domain. Finally, utilization of semantic membership function for data conversion to reduce the irrelevance of information representation is implemented in the proposed model. The major strengths of the proposed model are the enhancement of data manipulation and the elimination of discovered decision rule contradiction.

In this research, therefore, we take three steps to present the discovery of decision rule in empirical survey data based on Kansei ergonomics: i.e., (1) calculate information utility maximization of collected data by applying the nonlinear mapping borrowed from the method II of fuzzy quantification [11], (2) remove unreliable samples by maximizing entropy [12], and (3) discover decision rule by dominance-based rough set analysis to avoid contradictions [13]. Details are described in the latter sections. We also present an illustration to demonstrate the applicability of our proposed model.

2. Data Analysis for Rule Discovery.

The rapid growth of the number of websites over the virtual cyberspace has brought together the Internet users who have familiar interests, want to circulate and share information or knowledge, the desire to learn things, and to do business. Website design quality is one of the most important factors that affect the users' willingness to participate [14]. A survey from [15] indicates that information quality, ease-of-use, aesthetics, and service quality are the common facets in evaluating websites. Studies have also shown that website quality significantly influences the users' perceptions of the abilities and credibility of e-businesses [16, 17, 18]. For example, system quality issues such as information quality, impression, and attractiveness all significantly influence the perceptions about an online business and highly increase the percentage of buyers [19, 20]. The website quality assessment criteria have been proposed in various studies. Our algorithm developed in this paper can be applied to such problems.

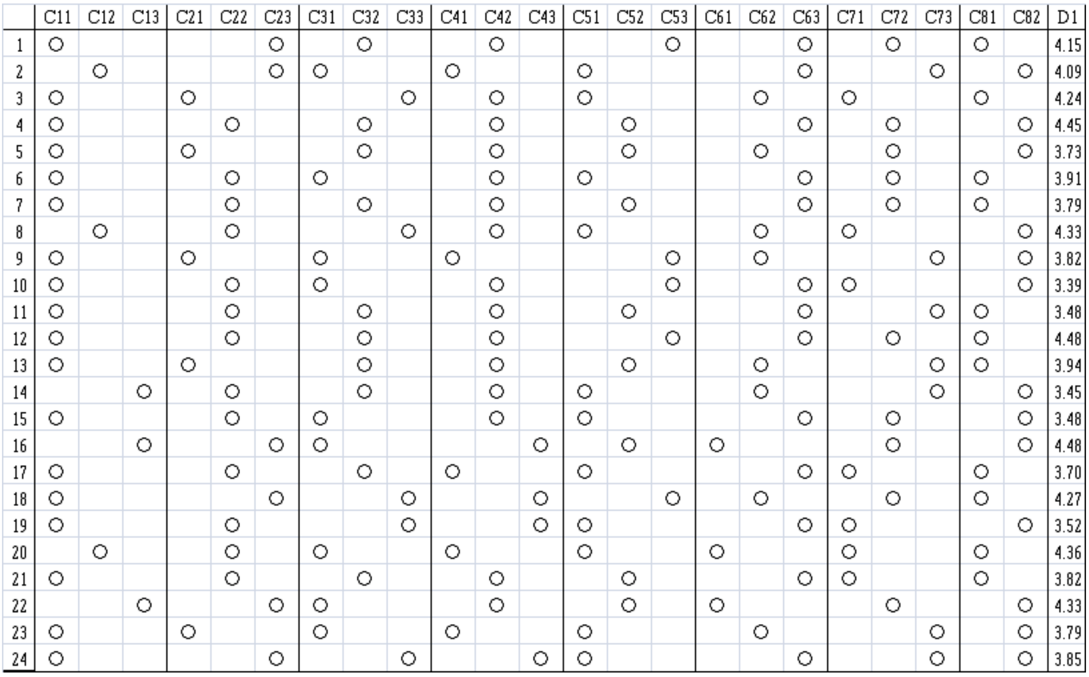
They can be further decomposed into eight dimensions, including information quality, completeness, interaction, visualization, response time, navigation, confidence and privacy. In the Kansei ergonomics based data collection, a subject designed to meet the customer's satisfaction requires an analysis of information which is gathered through questionnaire for human perception. The analyzed information can be further processed into decision rules for further use (e.g., rule-based systems). The questionnaire is listed in Table 1 that contains 8 items. In our example, the questionnaire design used a 5-digit rating scale (from 1 to 5) with bi-polar descriptors for each question was constructed for users to answer.

Items from I_1 to I_8 and D_1 were measured by a 5-point Likert scale where 1 represents "very disagree" and 5 "very agree". Moreover, to further improve the designed questionnaires, domain specialists in the area of website design were consulted to improve the validity, readability and reliability of the questionnaires. The respondents were 40 experts to review the questionnaire. In the survey, we first show the website design of online store to the test subjects, and then let them answer the questionnaire. After three weeks, there were 24 valid questionnaires returned, indicating a 60% valid response rate.

Table 1. Item and description of questionnaires

Item	Description
I_1	I think the structure of the website is easy to follow and functions diverse.
I_2	I think the website provides me sufficient information with multi-media format.
I_3	I think I can freely interact with other members via the community built in the website.
I_4	I can use my own way to read information in the website.
I_5	I can easily contact the web manager via website.
I_6	I can easily link to other related website.
I_7	I can easily access information or buy goods in the website and quickly get response.
I_8	I think the website has the privacy policy.
D_1	Overall, I will use the website.

Table 2. Results in category classification and the average of evaluation



	C11	C12	C13	C21	C22	C23	C31	C32	C33	C41	C42	C43	C51	C52	C53	C61	C62	C63	C71	C72	C73	C81	C82	D1
1	○					○		○			○				○			○		○		○		4.15
2		○				○	○			○			○					○			○		○	4.09
3	○			○					○		○		○				○		○			○		4.24
4	○				○			○			○			○				○		○			○	4.45
5	○			○				○			○			○			○			○			○	3.73
6	○				○		○				○		○					○		○		○		3.91
7	○				○			○			○			○				○		○		○		3.79
8		○			○				○		○		○				○		○				○	4.33
9	○			○			○			○					○		○				○		○	3.82
10	○				○		○				○				○			○	○				○	3.39
11	○				○			○			○			○				○			○	○		3.48
12	○				○			○			○				○			○		○		○		4.48
13	○			○				○			○			○			○				○	○		3.94
14			○		○			○			○		○				○				○		○	3.45
15	○				○		○				○		○					○		○			○	3.48
16			○			○	○					○		○		○				○			○	4.48
17	○				○			○		○			○					○	○			○		3.70
18	○					○			○			○			○		○			○		○		4.27
19	○				○				○			○	○					○	○				○	3.52
20		○			○		○			○			○			○			○			○		4.36
21	○				○			○			○			○				○	○			○		3.82
22			○			○	○				○			○		○				○			○	4.33
23	○			○			○			○			○				○				○		○	3.79
24	○					○			○			○	○					○			○		○	3.85

Table 2 summarizes the survey result from the received questionnaires. The category classification for items can be obtained by derivation based on minimum and maximum from the average value of each sample. The average value for each item from I_1 to I_7 can be divided into three levels such as C_{i1}: lower class, C_{i2}: medium class and C_{i3}: upper class. Also the average for item I_8 can be divided into C_{81} which represents smaller than 2.5 and C_{82} larger than or equal to 2.5. It should be noticed that based on the suggestion from domain specialists there are only two levels (lower and upper) for item 8 (privacy policy). The lower represents (not satisfactory) and upper (satisfactory) while decision rule is generated in the final stage. The last column in Table 2 shows the averaged results of D_1 by questionnaire answer. From Table 2, we constructed the decision table to discover decision rules.

3. **Consideration of Ambiguity and Detection of Outliers.**

3.1. **Grouping samples by method II of fuzzy quantification.**

The multivariate analysis of variance or a canonical correlation analysis is well known as an effective preprocessing technique for the separation of the collected data. In this research, we apply the method II of fuzzy quantification after transforming the defined data into the fuzzy data, because it is assumed that the collected data involves perturbation. Furthermore, the membership function used in fuzzy allows defining for a domain variable the strength of participation of an entity. Among the main concerns is, consequently, the usage of transformation mechanism that uses a nonlinear type fuzzy membership function to standardize the collected data. Table 3 shows the fuzzy data which can be analyzed by the method II of fuzzy quantification.

Table 3. Data for method II of fuzzy quantification

Sample	Category				External criteria			
n	1	2	$\cdots$	K	1	2	$\cdots$	R
1	$\mu_1(x_1^1)$	$\mu_2(x_1^2)$	$\cdots$	$\mu_K(x_1^K)$	$\mu^1(y_1)$	$\mu^2(y_1)$	$\cdots$	$\mu^R(y_1)$
2	$\mu_1(x_2^1)$	$\mu_2(x_2^2)$	$\cdots$	$\mu_K(x_2^K)$	$\mu^1(y_2)$	$\mu^2(y_2)$	$\cdots$	$\mu^R(y_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	$\mu_1(x_N^1)$	$\mu_2(x_N^2)$	$\cdots$	$\mu_K(x_N^K)$	$\mu^1(y_N)$	$\mu^2(y_N)$	$\cdots$	$\mu^R(y_N)$

In case there are some variables which cannot be assumed to follow only local linear formulae to the defined variables, then the research should apply the the nonlinear type membership function rather than the linear one. However, it is important to adjust both center and width of membership function by applying such a nonlinear one to fuzzy data. In order to execute such analysis in more detail for the defined variables later on, we divide all variables into I categories, small region, medium region, large region and so on.

To do this, we use one bell-shaped membership function on each variable for the category k $(k = 1, 2, \cdots, K)$. Each membership function is given by

$$\mu_{k_i}(x) = \exp\left\{-\frac{(x - b_{k_i})^2}{2s_{k_i}^2}\right\}, \quad i = 1, 2, \cdots, I \tag{1}$$

where $i \in \{$ 1FSmallC2FMedium$_1$C$\cdots$, $I - 1$: Medium$_{I-2}$, IFLarge $\}$. For the membership function of Small or Large, it is realized by considering the conditions $\mu_{k_1}(x) = 1$, $(x < b_1)$ and $\mu_{k_I}(x) = 1$, $(x > b_I)$ where b_{k_i} and s_{k_i} denote the center and the width of category k_i, respectively. Figure 1 shows the bell-shaped membership functions, $\mu_{k_1}(x)$, $\mu_{k_2}(x)$, $\cdots$, $\mu_{k_{I-1}}(x)$, $\mu_{k_I}(x)$ where $\{$ 1FSmallC2FMedium$_1$C$\cdots$, $I - 1$: Medium$_{I-2}$, IFLarge $\}$, for the kth category.

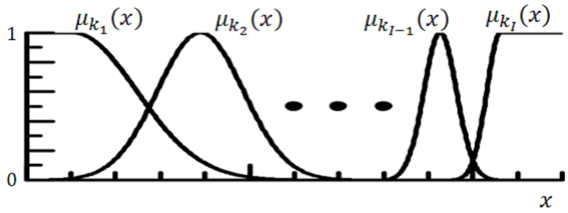


Figure 1. Membership functions for category

For the external criteria, we assume a nonlinear type membership function of the shape after transforming the evaluation data which is given by the integral of the normal distribution function as follows:

$$\mu(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}s} \exp\left\{-\frac{(x-b)^2}{2s^2}\right\} dx \tag{2}$$

where b and s denote a mean and a standard deviation for each factor respectively.

These three classes are derived by dividing the average value of D_1 into three levels. Circular indication describes the samples that belong to upper class U ($4.12 < y_n, n \in \{1, 3, 4, 8, 12, 16, 18, 20, 22\}$), the rectangular indication represents medium class M ($3.76 < y_n \leq 4.12$, $n \in \{2, 6, 7, 9, 13, 21, 23, 24\}$), and triangular indication represents lower class L ($y_n \geq 3.76, n \in \{5, 10, 11, 14, 15, 17, 19\}$).

Now it can be assumed to be $K \times I$ category weights, and we write them as w_{k_i}, the evaluation scale for environment valuation is given by

$$y_n = \sum_{k=1}^{K} \sum_{i=1}^{N} w_{k_i} \mu_{k_i}(x_n^{k_i}). \tag{3}$$

To be illustrative, the center b_{k_i} and width s_{k_i} of the membership function and the category weight w_{k_i} must be determined. As criteria to determine these parameters, we consider maximizing the fuzzy variance ratio η^2.

The fuzzy variance ratio η^2 is defined by the ratio of the sum of squares between treatment of factor B to the total sum of squares T, which is given by

$$\eta^2 = \frac{B}{T} \left(\equiv \frac{\eta_1^2}{\eta_2^2}\right) \tag{4}$$

where

$$T = \sum_{r=1}^{R} \sum_{n=1}^{N} \left(y_n - \sum_{l=1}^{N} S_l y_l\right)^2 \mu^r(y_n), \tag{5}$$

$$B = \sum_{r=1}^{R} \sum_{n=1}^{N} \left(\sum_{l=1}^{N} S_l^r y_l - \sum_{l=1}^{N} S_l y_l\right)^2 \mu^r(y_n). \tag{6}$$

If the nth sample y_n of the external criteria r ($r = 1, 2, \cdots, R$) and its membership function $\mu^r(y_n)$ of fuzzy external criteria are given, then S_n^r and S_n, in the fuzzy external criteria r are defined by

$$S_n^r = \frac{\mu^r(y_n)}{\sum_{n=1}^{N} \mu^r(y_n)}, \tag{7}$$

$$S_n = \frac{\sum_{r=1}^{R} \mu^r(y_n)}{\sum_{n=1}^{N} \sum_{r=1}^{R} \mu^r(y_n)}. \tag{8}$$

The method II of fuzzy quantification is applied to determining the category weight vector $\mathbf{w}$. In order to determine the center b_{k_i} and width s_{k_i} of the membership function, we use the steepest ascent method

$$\frac{d\theta^p_{k_i}}{dt} = \epsilon \frac{\partial \eta^2}{\partial \theta^p_{k_i}} = \epsilon \frac{(\partial B/\partial \theta^p_{k_i})T - B(\partial T/\partial \theta^p_{k_i})}{T^2} = \frac{\epsilon}{T}\left(\frac{\partial B}{\partial \theta^p_{k_i}} - \eta^2 \frac{\partial T}{\partial \theta^p_{k_i}}\right) \tag{9}$$

where ϵ is a positive constant and the parameter $\theta^p_{k_i}$, $p \in \{b, s\}$ is an element of the set $\{b_{k_i}, s_{k_i}\}$, $k = 1, 2, \cdots, K$, $i = 1, 2, \cdots, I$. For example, $\theta^s_{k_i}$ denotes s_{k_i}.

Then, Equation (9) can be rewritten by

$$\frac{\partial \eta_c^2}{\partial \theta^p_{k_i}} = 2w_{k_i} \sum_{r=1}^{R} \left\{\sum_{n=1}^{N} \mu^r(y_n)\right\} \left\{\sum_{k'=1}^{K} \sum_{j=1}^{I} w_{k'_j} \sum_{n'=1}^{N} \xi_1^{cn'r} \mu_{k_i}(x_{n'}^{k'_j})\right\} \left\{\sum_{n'=1}^{N} \xi_1^{cn'r} \xi_2^{pn'i}\right\} \tag{10}$$

where the parameter c is an element of the set $\{1, 2\}$. For example, η_1^2 denotes B, and

$$\xi_1^{cn'r} = \begin{cases} S_{n'}^r - S_{n'} & (c = 1) \\ \dfrac{1}{N} - S_{n'} & (c = 2) \end{cases} \tag{11}$$

$$\xi_2^{pn'i} = \begin{cases} \dfrac{(x_{n'}^{k_i} - b_{k_i})}{s_{k_i}} \mu_{k_i}(x_{n'}^{k_i}) & (p \in b) \\ \dfrac{(x_{n'}^{k_i} - b_{k_i})^2}{4s_{k_i}^3} \mu_{k_i}(x_{n'}^{k_i}) & (p \in s) \end{cases} \tag{12}$$

We must consider the following condition for the membership function of Small $\xi_2^{pn'1} = 0$, $(x_{n'}^{k_1} < b_{k_1})$ or Large $\xi_2^{pn'I} = 0$, $(x_{n'}^{k_I} > b_{k_I})$.

The obtained parameters weight, center and width for each item and category by Equation (9) are listed in Table 4 after learning the algorithm stopped, and in fact this algorithm was used to produce Figure 2.

The obtained mapping of D$_1$ from the collected data is illustrated in Figure 2 by applying method II of fuzzy quantification. There are 24 indications (9 circular indications, 8 rectangular indications and 7 triangular indications) representing the collected data obtained by Equation (1) after stopping learning the algorithm.

Table 4. Weight, center and width for each item and category

Item	Category $\{w_{k_1}, b_{k_1}, s_{k_1}\}$		
I$_1$	C$_{11}$ { 0.69, 1.03, 0.71 }	C$_{12}$ { 0.62, 0.44, 0.08 }	C$_{13}$ { 0.61, 0.53, 0.17 }
I$_2$	C$_{21}$ { 0.58, 0.59, 0.35 }	C$_{22}$ { 0.55, 1.06, 0.66 }	C$_{23}$ { 0.60, 0.48, 0.07 }
I$_3$	C$_{31}$ { 0.60, 0.49, 0.10 }	C$_{32}$ { 0.57, 0.73, 0.52 }	C$_{33}$ { 0.64, 0.59, 0.15 }
I$_4$	C$_{41}$ { 0.57, 0.57, 0.32 }	C$_{42}$ { 0.54, 0.61, 0.32 }	C$_{43}$ { 0.60, 0.50, 0.09 }
I$_5$	C$_{51}$ { 0.61, 0.78, 0.56 }	C$_{52}$ { 0.62, 0.50, 0.17 }	C$_{53}$ { 0.60, 0.48, 0.08 }
I$_6$	C$_{61}$ { 0.60, 0.51, 0.09 }	C$_{62}$ { 0.59, 0.49, 0.09 }	C$_{63}$ { 0.50, 0.97, 0.74 }
I$_7$	C$_{71}$ { 0.59, 0.47, 0.10 }	C$_{72}$ { 0.60, 0.50, 0.06 }	C$_{73}$ { 0.64, 1.12, 0.77 }
I$_8$	C$_{81}$ { 0.59, 0.48, 0.05 }	C$_{82}$ { 0.65, 0.96, 0.61 }	

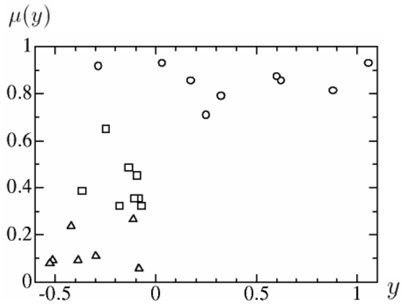


FIGURE 2. Mapping by fuzzy quantification

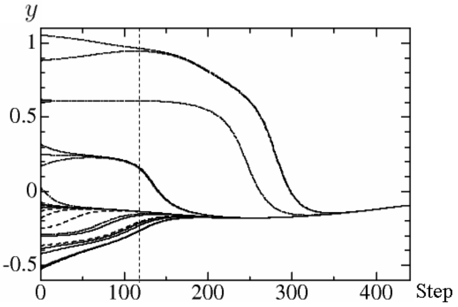


FIGURE 3. Merged paths by melting

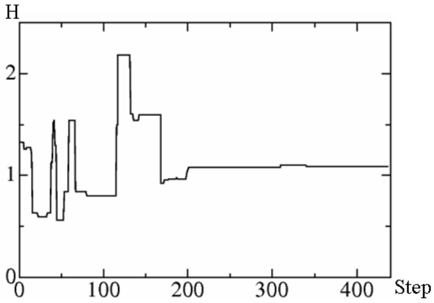


FIGURE 4. Entropy by melting

3.2. **Detection of doubtful data based on entropy.** There may be data which can not be classified clearly into a class. The detection of such doubtful data for rule selection is described below. We consider the maximization of entropy,

$$S_\beta(\mathbf{x}) = -\sum_{i=1}^{N} p(y_i|\mathbf{x}) \log p(y_i|\mathbf{x}) \tag{13}$$

under the normalization condition of $p(y_i|\mathbf{x})$ and

$$\epsilon^2(\mathbf{x}) = \sum_{i=1}^{N} (y_i - \phi_i(\mathbf{x}))^2 p(y_i|\mathbf{x}) \tag{14}$$

The optimal solution is given by the conditional probability

$$p(y_i|\mathbf{x}) = \exp^{\beta F_\beta(\mathbf{x})} \exp^{-\beta(y_i-\phi_i(\mathbf{x}))^2} \tag{15}$$

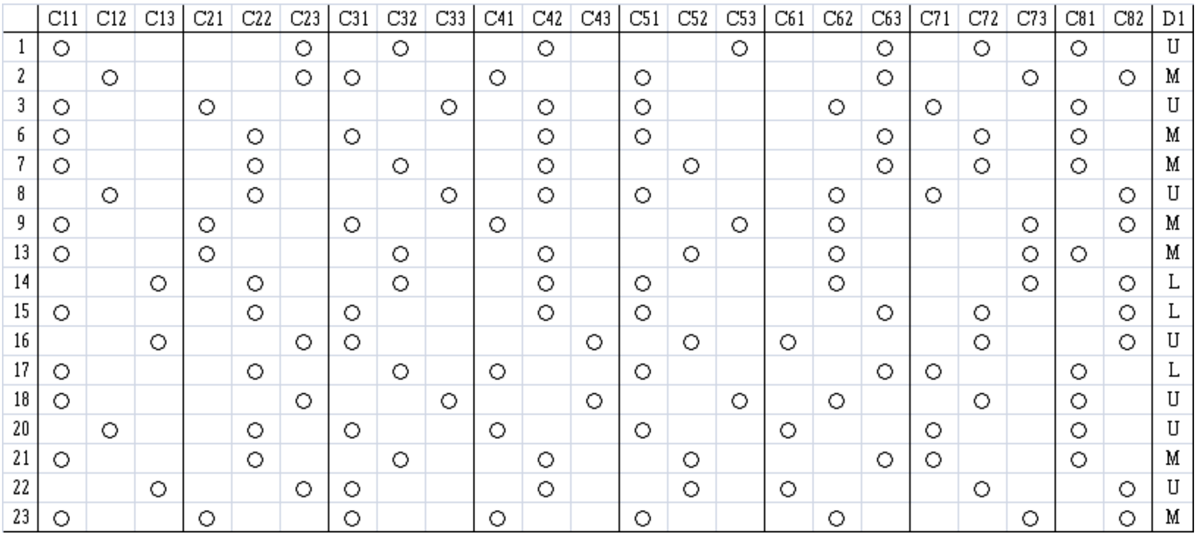
where $F_\beta(\mathbf{x})$ is the free energy.

Then maximizing entropy is equivalent to minimizing free energy, because of $S_\beta(\mathbf{x}) = -F_\beta(\mathbf{x}) + \beta\epsilon^2(\mathbf{x})$. Thus, we derive the clustering method (called melting) by applying the steepest descent method to the free energy $F_\beta(\mathbf{x})$

$$\phi_j^{t+1}(\mathbf{x}) = \phi_j^t(\mathbf{x}) - \rho \frac{\partial F_\beta(\mathbf{x})}{\partial \phi_j(\mathbf{x})} = \phi_j^t(\mathbf{x}) - \rho \sum_{i=1}^{N} (y_i - \phi_j(\mathbf{x})) p(y_i|\mathbf{x}). \tag{16}$$

The behavior of categorized samples by melting for the collected data is shown in Figure 3. In fact, the vertical axis of Figure 3 corresponds to the horizontal axis of Figure 2. Melting started from $\phi_j(\mathbf{x}) = y_i$ by assuming $i = j$, and the parameter β is changed by $100 \exp^{-0.01 \times Step}$. On one hand, when the parameter β becomes larger, the number of path P comes closer to N. On the other hand, when β becomes smaller, the number of path P comes closer to 1.

Table 5. Samples rough set analysis applies to



	C11	C12	C13	C21	C22	C23	C31	C32	C33	C41	C42	C43	C51	C52	C53	C61	C62	C63	C71	C72	C73	C81	C82	D1
1	○					○		○			○				○			○		○		○		U
2		○				○	○			○			○					○			○		○	M
3	○			○					○		○		○				○		○			○		U
6	○				○		○				○		○					○		○		○		M
7	○				○			○			○			○				○		○		○		M
8		○			○				○		○		○				○		○				○	U
9	○			○			○			○					○		○				○		○	M
13	○			○				○			○			○			○				○	○		M
14			○		○			○			○		○				○				○		○	L
15	○				○		○				○		○					○		○			○	L
16			○			○	○					○		○		○				○			○	U
17	○				○			○		○			○					○	○			○		L
18	○					○			○			○			○		○			○		○		U
20		○			○		○			○			○			○			○			○		U
21	○				○			○			○			○				○	○			○		M
22			○			○	○				○			○		○				○			○	U
23	○			○			○			○			○				○				○		○	M

Moreover, in order to distinguish the data regarded as misclassified samples, we provide the following procedure by using entropy

$$H = -\sum_{p=1}^{P}\sum_{r=1}^{R} p(x_{pr}) \log p(x_{pr}). \tag{17}$$

Assume that N_{pr} presents the number of path belonging to class r in merged path p. So merged path p has N_p $(= \sum_{r=1}^{R} N_{pr})$ paths, the total sample number N is equivalent to $\sum_{p=1}^{P}\sum_{r=1}^{R} N_{pr}$, and $p(x_{pr}) = N_{pr}/N_p$. In each merged path, we decide doubtful samples by applying majority decision logic. If they become the same number, then we do not employ those samples. It should be noted that the function H is different from $S_\beta(\mathbf{x})$, because the function H is the path combination of entropy.

The profile of entropy H for the melting of the collected data in Figure 3 is shown in Figure 4. The entropy H takes a maximum value 2.19 at the 117th step, the dot line in Figure 4 represents the state of 117th step. We found that there are 9 paths. Each merged path involves $\{1[U], 3[U]\}$, $\{2[M], 5[L], 6[M],$ $7[M], 9[M], 10[L],$ $12[U], 13[M], 21[M], 23[M]\}$, $\{4[U], 19[L]\}$, $\{8[U]\}$, $\{11[L], 24[M]\}$, $\{14[L]\}$, $\{15[L], 17[L]\}$, $\{16[U]\}$ and $\{18[U], 20[U],$ $22[U]\}$. From these results, it is realized that samples $\{5, 10, 12, 4, 19, 11, 24\}$ are the doubtful data because they delayed on making a classification.

We therefore employ the method II of fuzzy quantification and the clustering to obtain the samples without doubtful data. The results are shown in Table 5. It is found that samples $U' = \{4, 5, 10, 11, 12, 19, 24\}$ are removed. Each value of the decision attribute is grouped into three classes because we would like the discovered rules be reliable and consistent. In order to derive order relations from intervals, we define $S[s_1^{k_i}, s_2^{k_i}]$ by $s_1^{k_i} = \min_{n'} \left\{w_{k_i}\mu_{k_i}(x_n'^{k_i})\right\}$ and $s_2^{k_i} = \max_{n'} \left\{w_{k_i}\mu_{k_i}(x_n'^{k_i})\right\}$ where n' denotes difference set $\{U/U'\}$. These values are calculated from Table 4.

3.3. **Rule extraction without contradiction based on dominance relation.** The rule extraction by the rough set analysis which takes into consideration the dominance relation from the attribute values to the quantitative data given by Table 5 is described below. Assuming that the data consists of S $(= 23)$ samples, we consider the decision table which consists of N $(= 8)$ condition attributes and M $(= 1)$ decision attributes. The combination of condition attributes used for constructing the rule is regarded as criteria.

For example, there are 8 condition attributes that are defined to be items to extract questionnaire answers' perception for a subject. In such a case, there are 256 $(= \sum_i^8 {}_8C_i)$ combinations of condition attributes. The mth decision attribute is classified into R classes as $C^s \cap C^t = \phi$, $(s \neq t)$ and $C^R \succ \cdots \succ C^r \succ C^1$.

For the given sample $x \in U$, we can define the bottom accumulation set $C^{\geq r}$ which is a set of the element of U at least belonging to the class C^r by $C^{\geq r} = \bigcup_{s \geq r} C^s$, and the bottom accumulation set $C^{\leq r}$ which is a set of the element of U at most belonging to the class C^r by $C^{\leq r} = \bigcup_{s \leq r} C^s$.

The set of all criteria is denoted by W, and the subset of W is denoted by $V \subseteq W$. For the arbitrary subset $v \in V$, the relationship in V such that sample x dominates sample y is represented by $xD_m^{'V}y \leftrightarrow g(x,n) \succeq g(y,n)$, $(\forall v \in V) \leftrightarrow \min_{s_1 \in g(x,n)} s_1 \geq \min_{t_1 \in g(y,n)} t_1$ and $\max_{s_2 \in g(x,n)} s_2 \geq \max_{t_2 \in g(y,n)} t_2$, where $g(x,n)$ denotes the value of attribute n at sample x. The order relation of two intervals $S[s_1, s_2]$ and $T[t_1, t_2]$ is defined by $S[s_1, s_2] \succeq T[t_1, t_2] \ \leftrightarrow\ s_1 \geq t_1$ and $s_2 \geq t_2$.

For the given sample $x \in U$, the set $D_V^+(x)$ of the element of U which dominates x in V, and the set $D_V^-(x)$ of the element U which is dominated by x in V can be defined by $D_V^+(x) = \{y \in U | yD_Vx\}$ and $D_V^-(x) = \{y \in U | xD_Vy\}$. The lower approximation set $V_*(C^{\geq r})$ and the upper approximation set $V^*(C^{\geq r})$ of the accumulation set $C^{\geq r}$ by the domination set $D_V^+(x)$ can be defined by $V_*(C^{\geq r}) = \{x \in U | D_V^+(x) \subseteq C^{\geq r}\}$ and $V^*(C^{\geq r}) = \bigcup_{x \in C^{\geq r}} D_V^+(x)$.

From this lower approximation set, the if-then rule can be obtained as follows:

$$\begin{aligned} &\text{IF } g(x^*, v_1) \succeq g(x, v_1) \text{ and } g(x^*, v_2) \succeq g(x, v_2) \\ &\qquad \cdots \text{ and } g(x^*, v_N) \succeq g(x, v_N), \text{ THEN } x^* \in C^{\geq r}. \end{aligned} \tag{18}$$

In a similar way, the lower approximation set $V_*(C^{\leq r})$ and the upper approximation set $V^*(C^{\leq r})$ of the accumulation set $C^{\leq r}$ by the domination set $D_V^-(x)$ can be defined by $V_*(C^{\leq r}) = \{x \in U | D_V^-(x) \subseteq C^{\leq r}\}$ and $V^*(C^{\leq r}) = \bigcup_{x \in C^{\leq r}} D_V^-(x)$. From this lower approximation set, the rule by which the data x^* dominating x in $V_*(C^{\leq r})$ is derived surely belongs to lower class than r. That is, for the certain $x^* \in C^{\leq r}$, the if-then rule can be obtained as follows:

$$\begin{aligned} &\text{IF } g(x^*, v_1) \preceq g(x, v_1) \text{ and } g(x^*, v_2) \preceq g(x, v_2) \\ &\qquad \cdots \text{ and } g(x^*, v_N) \preceq g(x, v_N), \text{ THEN } x^* \in C^{\leq r}. \end{aligned} \tag{19}$$

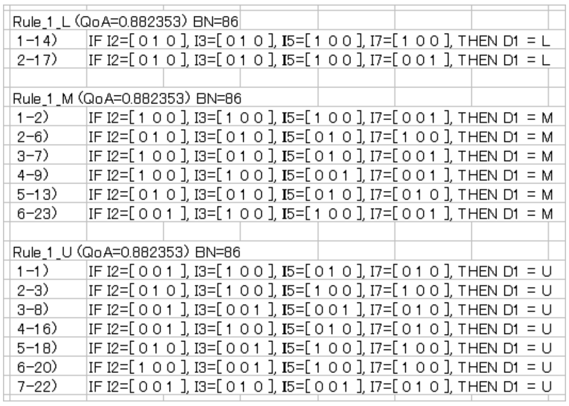
The boundary between $C^{\geq r}$ and $C^{\leq r}$ is obtained by $B_V(C^{\geq r}) = V^*(C^{\geq r}) - V_*(C^{\geq r})$ and $B_V(C^{\leq r}) = V^*(C^{\leq r}) - V_*(C^{\leq r})$, respectively. Thus the quality of approximation (QoA), such that the ratio of targets can be classified exactly against division τ by the partial criteria V, is given by

$$\beta_V(\tau) = \frac{\left|U - \left(\bigcup_{r=1}^n B_V(C^{\geq r}) \cup \bigcup_{r=1}^n B_V(C^{\leq r})\right)\right|}{|U|}. \tag{20}$$

The minimum set $V \subseteq W$ which satisfies $\beta_V(\tau) = \beta_W(\tau)$ is called the reduction. Two or more reductions can exist at the same time, and the common set of them is called core. We can simplify the decision table without decreasing the QoA by using the attributes belonging to the reduction.

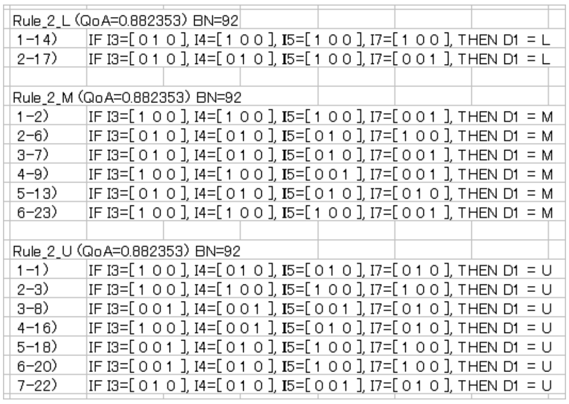
Furthermore, in the conditional set of rules which are obtained by the lower approximation set $V_*(C_m^{\geq r})$ and $V_*(C_m^{\leq r})$, we can derive the following if-then rule which explains that the sample $x^* \in U$ certainly belongs to class r because the set takes the same category in

Table 6. Extracted rule 1



Rule_1_L (QoA=0.882353) BN=86	
1-14)	IF I2=[0 1 0], I3=[0 1 0], I5=[1 0 0], I7=[1 0 0], THEN D1 = L
2-17)	IF I2=[0 1 0], I3=[0 1 0], I5=[1 0 0], I7=[0 0 1], THEN D1 = L
Rule_1_M (QoA=0.882353) BN=86	
1-2)	IF I2=[1 0 0], I3=[1 0 0], I5=[1 0 0], I7=[0 0 1], THEN D1 = M
2-6)	IF I2=[0 1 0], I3=[0 1 0], I5=[0 1 0], I7=[1 0 0], THEN D1 = M
3-7)	IF I2=[1 0 0], I3=[0 1 0], I5=[0 1 0], I7=[0 0 1], THEN D1 = M
4-9)	IF I2=[1 0 0], I3=[1 0 0], I5=[0 0 1], I7=[0 0 1], THEN D1 = M
5-13)	IF I2=[0 1 0], I3=[0 1 0], I5=[0 1 0], I7=[0 1 0], THEN D1 = M
6-23)	IF I2=[0 0 1], I3=[1 0 0], I5=[1 0 0], I7=[0 0 1], THEN D1 = M
Rule_1_U (QoA=0.882353) BN=86	
1-1)	IF I2=[0 0 1], I3=[1 0 0], I5=[0 1 0], I7=[0 1 0], THEN D1 = U
2-3)	IF I2=[0 1 0], I3=[1 0 0], I5=[1 0 0], I7=[1 0 0], THEN D1 = U
3-8)	IF I2=[0 0 1], I3=[0 0 1], I5=[0 0 1], I7=[0 1 0], THEN D1 = U
4-16)	IF I2=[0 0 1], I3=[1 0 0], I5=[0 1 0], I7=[0 1 0], THEN D1 = U
5-18)	IF I2=[0 1 0], I3=[0 0 1], I5=[1 0 0], I7=[1 0 0], THEN D1 = U
6-20)	IF I2=[1 0 0], I3=[0 0 1], I5=[1 0 0], I7=[1 0 0], THEN D1 = U
7-22)	IF I2=[0 0 1], I3=[0 1 0], I5=[0 0 1], I7=[0 1 0], THEN D1 = U

Table 7. Extracted rule 2



Rule_2_L (QoA=0.882353) BN=92	
1-14)	IF I3=[0 1 0], I4=[1 0 0], I5=[1 0 0], I7=[1 0 0], THEN D1 = L
2-17)	IF I3=[0 1 0], I4=[0 1 0], I5=[1 0 0], I7=[0 0 1], THEN D1 = L
Rule_2_M (QoA=0.882353) BN=92	
1-2)	IF I3=[1 0 0], I4=[1 0 0], I5=[1 0 0], I7=[0 0 1], THEN D1 = M
2-6)	IF I3=[0 1 0], I4=[0 1 0], I5=[0 1 0], I7=[1 0 0], THEN D1 = M
3-7)	IF I3=[0 1 0], I4=[0 1 0], I5=[0 1 0], I7=[0 0 1], THEN D1 = M
4-9)	IF I3=[1 0 0], I4=[1 0 0], I5=[0 0 1], I7=[0 0 1], THEN D1 = M
5-13)	IF I3=[0 1 0], I4=[0 1 0], I5=[0 1 0], I7=[0 1 0], THEN D1 = M
6-23)	IF I3=[1 0 0], I4=[1 0 0], I5=[1 0 0], I7=[0 0 1], THEN D1 = M
Rule_2_U (QoA=0.882353) BN=92	
1-1)	IF I3=[1 0 0], I4=[0 1 0], I5=[0 1 0], I7=[0 1 0], THEN D1 = U
2-3)	IF I3=[1 0 0], I4=[1 0 0], I5=[1 0 0], I7=[1 0 0], THEN D1 = U
3-8)	IF I3=[0 0 1], I4=[0 0 1], I5=[0 0 1], I7=[0 1 0], THEN D1 = U
4-16)	IF I3=[1 0 0], I4=[0 0 1], I5=[0 1 0], I7=[0 1 0], THEN D1 = U
5-18)	IF I3=[0 0 1], I4=[0 1 0], I5=[1 0 0], I7=[1 0 0], THEN D1 = U
6-20)	IF I3=[0 0 1], I4=[0 1 0], I5=[1 0 0], I7=[1 0 0], THEN D1 = U
7-22)	IF I3=[0 1 0], I4=[0 1 0], I5=[0 0 1], I7=[0 1 0], THEN D1 = U

all attribute.

$$\begin{aligned} &\text{If } g(x^*, v_1) = g(x, v_1) \text{ and } g(x^*, v_2) = g(x, v_2) \\ &\qquad \cdots \text{ and } g(x^*, v_N) = g(x, v_N), \text{ THEN } \ x^* \in C_m^r. \end{aligned} \tag{21}$$

The data satisfying this extracted rule certainly belong to the class r. We can extract rules which have more consistency by considering the dominance relation.

Finally, the discovered decision rules are obtained from Table 5, as shown in Tables 6 (rule 1) and 7 (rule 2) in a reduction manner. The QoA is about 0.88. Both rules consist of 4 condition attributes such as $\{\mathrm{I}_2, \mathrm{I}_3, \mathrm{I}_5, \mathrm{I}_7\}$ for rule 1 and $\{\mathrm{I}_3, \mathrm{I}_4, \mathrm{I}_5, \mathrm{I}_7\}$ for rule 2, respectively. The notation of BN represents the decimal display of each reduction. For example, the binary display of condition attributes for rule 1 is [I_8, I_7, I_6, I_5, I_4, I_3, I_2, I_1] = [001010110], thus the BN is 86.

Furthermore, we find that two samples $\{15, 21\}$ are not involved in both Tables. Thus the QoA becomes $15/17 \approx 0.88$. In Table 6, it can be seen that these two samples are different at I_6. So, we can distinguish samples 15 and 21 based on category I_6. However, by rough set analysis which considers the dominance relation, the proposed model could produce a more reliable analysis by avoiding contradictory situations. In addition, it is found that the condition attributes $\{\mathrm{I}_3, \mathrm{I}_5, \mathrm{I}_7\}$ are regarded as important items. Therefore, we suggest that website designers should consider items such as interaction, the response time, and the confidence so as to meet the users' perception and satisfy their requirements.

4. **Algorithm of Knowledge Discovery.** The approach that can remove uncommon data, discover less reliable analyzed data, and detect inconsistent rules is complex, but

important to the final outputs. Furthermore, in order to view the steps of our proposed model of decision rule discovery in a concise manner, we present the algorithm as follows:

S1 Calculate a mean, b, and a standard deviation, s, of defined data, and transform the collected data, y, of each defined variable to fuzzy data, $\mu(y)$, by using a normal distribution function (refer to Equation (2)).

S2 Give the initial value of parameters relating to their center, b_{k_i}, width, s_{k_i}, and category weight, w_{k_i}, for explanatory data (refer to Equations (1) and (3)).

S3 Transforming the collected data of explanatory data into the fuzzy data by Equation (1).

S4 Derive the category weight maximizing the fuzzy variation ratio by Equation (4).

S5 By using the updated rule based on Equation (9), change the parameters relating to the center and the width of the membership function.

S6 Calculate the fuzzy variation ratio again, if it is smaller than the fuzzy variation ratio which is obtained from **S4**, then the learning algorithm is stopped, otherwise return to **S3**.

In addition, the algorithm of detection of doubtful data after separation criteria by the method II of fuzzy quantification is described as follows:

D1 Divide samples into some classes, for example, upper level, middle level and lower level based on the decision attribute of the Kansei evaluation experiment.

D2 Change parameter β and move the point by minimizing free energy.

D3 Calculate entropy of the merged path and find the maximum entropy.

D4 Extract the paths involved in each merged path.

D5 Detect samples which are held off making a classification by applying majority decision.

The induction-based approach used to generate decision rules from observations and/or a set of survey data has been successfully used in automated rule discovery. It is believed that the key success factor that affects the results of the generated decision rules is the techniques used in the data preprocessing, data manipulation, and discovery process. That is the motivation of practical use of this study.

Compared with the typical tools viewed in AQUINAS [21], ALTO [22], ICONKAT [23], KASE [24] and ID3RGS [25], our method can solve a major drawback revealed in these tools by using approaches such as machine learning, neural network, and explanation-based strategies. We can illustrate the domain dependent, complex acquisition process, low reliability, and lack of dealing with the rule contradiction as one of major drawbacks. However, it is understood that our method can present a model for decision rule discovery with consideration of ambiguity by fuzzy, detection of outliers by entropy and removal of contradiction by dominance relation from the main results of Tables 6 and 7.

5. Conclusion.

In this paper, we propose a knowledge discovery approach that employs fuzzy quantification to group survey data, entropy to detect unreliable data, and dominance relation to return decision rules without contradiction to help generate decision rules. We find in this research that the set of data, the identification of the semantic membership function and the method used to determine the number of classes are the major factors that strongly affect the generated decision rules. It is also realized that no single combination of these issues can be used to generate the most precise and reliable decision rules for all domain problems. Novel contribution of this paper is a proposal of sequential data handling for such a particular problem which may need a specific combination of these factors and much involvement of domain experts.

REFERENCES

[1] C. Garg-Janardan and G. Salvendy, A conceptual framework for knowledge elicitation, in *Knowledge Acquisition Tools for Expert Systems*, J. Boose and B. Gains (eds.), Academic Press, New York, 1988.

[2] S. Marcus, *Automating Knowledge Acquisition for Expert Systems*, Kluwer Academic Pub., Boston, 1988.

[3] H. Motoda, R. Mizoguchi, J. Boose and B. Gaines, Knowledge acquisition for knowledge-based systems, *IEEE Expert*, vol.6, pp.53-63, 1991.

[4] M. A. Musen, An overview of knowledge acquisition, in *Second Generation Expert Systems*, J.-M. David, J.-P. Krivine and R. Simmons (eds.), New York, Springer-Verlag, 1993.

[5] C. H. Wu, S. C. Kao and K. Srihari, GDKAT: A goal driven knowledge acquisition tool for knowledge base development, *Expert Systems: The International of Knowledge Engineering and Neural Networks*, vol.17, no.2, pp.90-105, 2000.

[6] M. Nagamachi, Kansei engineering: A new ergonomic consumer-oriented technology for product development, *International Journal of Industrial Ergonomics*, vol.15, no.1, pp.3-11, 1995.

[7] Y. Matsubara and M. Nagamachi, Hybrid Kansei engineering system and design support, *International Journal of Industrial Ergonomics*, vol.19, no.2, pp.81-92, 1997.

[8] N. Mori, Rough set and Kansei engineering, *Journal of Japan Society for Fuzzy Theory and Systems*, vol.13, no.6, pp.52-59, 2001.

[9] M. Nagamachi, Kansei engineering as a powerful consumer-oriented technology for product development, *Applied Ergonomics*, vol.33, no.3, pp.289-294, 2002.

[10] C. E. Osgood, G. J. Suci and P. H. Tannenbaum, *The Measurement of Meaning*, University of Illinois Press, 1957.

[11] K. Okuhara, T. Tanaka and M. Sakawa, Cultivation environment evaluation and kind selection system for crops using the method II of fuzzy quantification, *Trans. on Institute of Electronics, Information and Communication Engineers*, vol.J85-A, no.8, pp.887-894, 2002.

[12] Y. Wong, Clustering data by melting, *Neural Computation*, vol.5, pp.89-104, 1993.

[13] K. Okuhara, Y. Matsubara, K. Sugihara and H. Ishii, Rule selection by rough set considering ordinality in attributes for Kansei evaluation, *Trans. on Institute of Electronics, Information and Communication Engineers*, vol.J87-A, no.7, pp.1045-1053, 2004.

[14] H. W. Webb and L. A. Webb, SiteQual: An integrated measure of web site quality, *The Journal of Enterprise Information Management*, vol.17, no.6, pp.430-440, 2004.

[15] D. G. Gregg and S. Walczak, Dressing your online auction business for success: An experiment comparing two EBay business, *MIS Quarterly*, vol.32, no.3, pp.653-670, 2008.

[16] C. Shchiglik and S. Barnes, Evaluating web quality in the airline industry, *Journal of Computer Information Systems*, vol.44, no.3, pp.17-25, 2004.

[17] J. V. Iwaarden, T. V. D. Wiele, L. Ball and R. Millen, Perceptions about the quality of web sites: A survey amongst students at northeastern university and erasmus university, *Information and Management*, vol.41, no.8, pp.947-959, 2004.

[18] S. Kim and L. Stoel, Dimensional hierarchy of retail web quality, *Information and Management*, vol.41, no.5, pp.619-633, 2004.

[19] J. Park, Y. Lee and R. Widdows, Empirical investigation on reputation and product information for trust formation in consumer market, *Journal of the Academy of Business and Economics*, vol.3, no.1, pp.231-239, 2004.

[20] V. Venkatesh and V. Ramesh, Web and wireless site usability: Understanding differences and modeling use, *MIS Quarterly*, vol.30, no.1, pp.181-205, 2006.

[21] J. Boose and J. W. Bradshaw, Expertise transfer and complex problem: Using AQUINAS as a knowledge acquisition workbench for knowledge-based systems, in *Knowledge Acquisition Tools for Expert Systems*, J. Boose and B. Gains (eds.), New York, Academic Press, 1988.

[22] N. Major and H. Reichgelt, ALTO: An automated laddering tool, in *Current Trends in Knowledge Acquisition*, B. Wielinga, J. Boose, B. Gaines, G. Schreiber and M. V. Someren (eds.), Washington, 1990.

[23] K. Ford, A. Canas and J. Jones, ICONKAT: An integrated constructivist knowledge acquisition tool, *Knowledge Acquisition*, vol.3, no.2, pp.215-236, 1991.

[24] D. Araki, S. Kojima and T. Kohno, KASE project toward effective diagnosis system development, *Knowledge Acquisition*, vol.4, pp.323-346, 1992.

[25] C. H. Wu and S. C. Kao, An induction-based approach to rule generation using membership function, *International Journal of Computer Integrated Manufacturing*, vol.15, no.1, pp.86-96, 2002.