Student Questionnaire Analyses for Class Management by Text Mining both in Japanese and in Chinese

Shigeichi Hirasawa, *Fellow*, *IEEE*, Fu-Yih Shih, and Wei-Tzen Yang

***Abstract*— By combining statistical analyses and information retrieval techniques, an efficient way for knowledge discovery from questionnaires is discussed. Since usual questionnaires include questions answered by a fixed format and those by a free format, it is important to introduce the methods by both data mining and text mining. The answers by the fixed format are called "items", and those by the free format, simply "texts". In this paper, using an algorithm for processing answers with both the items and the texts and that for extracting important sentences from texts combined with statistical techniques, a method for analyzing the questionnaires is established. The method is applied to a case of improvements for the quality of education by which the student questionnaire is executed to a class, and we obtain useful knowledge which leads to faculty development.**

I. Introduction

Universities are expected to develop useful and effective programs for class management and to improve the quality of education at all times. To perform these activities, student questionnaires are often used. By establishing a class model, we have evaluated student characteristics, the degree of satisfaction and final scores, and their relationships for a set of students or subsets of them [9]. One of the present authors has collected data from the student questionnaire in Japanese for the past five years, where the class considered is "Introduction to Computer Engineering", in the second academic year, Department of Industrial and Management Systems Engineering, School of Science and Engineering, Waseda University. The other authors have also applied it to students for corresponding classes in Taiwan, R.O.C. by translating it into Chinese.

We have developed techniques for (1) classification or clustering for documents with fixed formats and free formats [5], [12], and (2) extraction of important sentences or feature sentences and words from texts [11], [13], [16] which helps us to briefly understand the contents of the texts. Using the traditional statistical techniques, (3) interpretation of characteristics of the set of documents.

In this paper, we establish a class model used for the class of "Introduction to Computer Engineering". The model has as inputs information of students such as implicit characteristics, a prior knowledge, interested areas, intention, scores, etc. and it has as outputs such as final score and degree of satisfaction returned by the students. We apply the above techniques into student questionnaire analyses for both in Japan and in R.O.C.

Problems of partitioning students of the class into a few subclasses by their characteristics are evaluated. The purposes of these problems are to improve the degree of satisfaction of the students and to increase the effectiveness of education.

In this paper, we show a questionnaire analyses model in section II. The student questionnaire is mainly discussed in section III. In section IV, classification and clustering techniques are briefly described (See [10] in detail). Section V discusses results of analyses. Conclusions are given in Section VI.

II. Questionnaire Analyses Model

The method for analyzing the questionnaire is shown in Fig. 1 as a questionnaire analyses model.

First, a model for the object for which a questionnaire will be applied is presented. For example, we shall show a class model as the object in this paper.

Second, a questionnaire is designed based on this model, which includes both the items and the texts as the answers. We refer to them collectively as documents. The number of the documents equals that of examinees, i.e., students in this paper.

Next, analyses are executed as follows:

(1) The set of documents is classified or clustered by the algorithms [5], [10], [12]. Note that both the items and the texts are simultaneously processed, not separately.

(2) For the texts only, important sentences, or feature sentences and words are extracted from the documents by the algorithms for extracting important information [11], [13], [16], [17]. These results are helpful to easily understand the opinions and directly give useful information of the classes (categories) or clusters.

(3) For the items [1] only, statistical techniques such as multiple linear regression analysis, discriminant analysis, are used to analyze the characteristics of each set of members. If the amount of the data is extremely large, a data mining technique is also used to analyze them.

In (1), we have proposed the algorithm based on the probabilistic latent semantic indexing (PLSI) model [2], [7]

The work leading to this paper was partially proceeded during visiting of S.H. at Leader University from February 25 through March 17, 2006 and was supported by Ministry of Education, Taiwan, R.O.C.

S. Hirasawa is with Department of Industrial mad Management Systems Engineering, School of Science and Enginnering, Waseda University, 3-4-1, Ohkubo, Shinjuku, Tokyo 169-8555 Japan. E-mail: hira@waseda.jp

F-Y. Shih is with Department of Information Communication, Leader University, Taiwan, R.O.C. E-mail: fuyih@mail.leader.edu.tw

W-T. Yang is with Institute of Management, Tamkang University, Taiwan, R.O.C. E-mail: 018467@mail.tku.edu.tw

[1] Information investigated attribute of the categorical data, e.g., the scores of examinations for students, is added to a sort of the items.

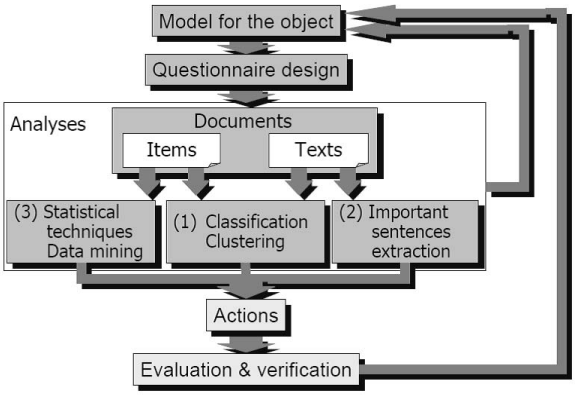

Fig. 1. Questionnaire analysis model

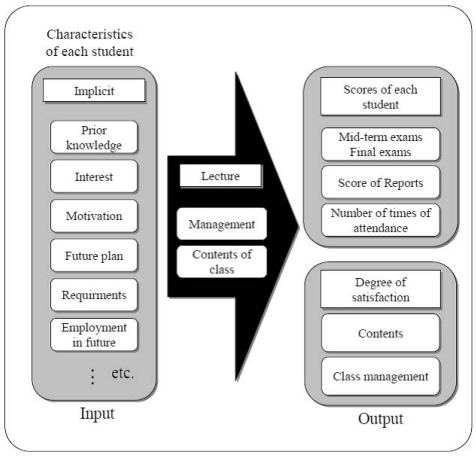

Fig. 2. Class model

which is known to be one of the most powerful model in information retrieval systems. The proposed algorithm based on PLSI model exhibits good performance in classification or clustering especially for a small size of the document set [5], [6], [10]. In (2), we have also presented the algorithm to select important sentences by extracting representative sentences based on Japanese language processing.

The results obtained by combining (1) and (3) give the profile of each class (category) or cluster by the characteristics of the members. Combining (2) and (3) is also used for understanding the characteristics of the members of each class or cluster and these results give us useful information to manage the mass or improve the conventional systems.

Finally, actions are made based on the analyzed results. The actions are evaluated from the standpoint of their effectiveness, and a new model for the object is generated by the feedback loop if necessary.

III. Student Questionnaire

A class model for this object is shown in the Fig. 2. A technique to find out requirements of the students from the questionnaire is discussed by applying the questionnaire analyses model. First, relationships between the degree of satisfaction, scores and the characteristics of the students are presented as a class model. Next, the questionnaire is designed to verify the hypothesis given by this class model. Finally, according to the results of this questionnaire analyses together with the score of each student, we evaluate the degree of satisfaction, that of achievement in learning, and characteristics of students. This knowledge is useful to manage the class. In many Japanese universities, the quality assurance of the education program by Japan Accreditation Board for Engineering Education (JABEE) has recently become important for improving the classes management.

A. Class model

We have proposed a class model for the class "Introduction to Computer Engineering" as shown in Fig. 2 [4].

The implicit characteristics of each student are essentially measured by questionnaire. While explicit characteristics are objectively given by numerical data of the class. These characteristics generate explanatory variables. Then each student yields his or her final score and the degree of satisfaction as the result of the class. The degree of satisfaction is also measured by questionnaire. The final score and the degree of satisfaction play as criterion variables in this model. Besides these variables, we can expect to get some information regarding to class management such as partition of the class. Usually there exist many differences between each student in level, interested area, experiences, and motivation before beginning the class. Hence a proper partition of the class depending on features such as the future plan of each student is desirable. Later, the partitions shown in Table III can be considered depending on the contents of topics of the lecture, where G stands for a generalist course, S, for a specialist course by estimating his or her future job. According to this model, we can effectually design the questionnaire. In Fig. 2, note that explicit characteristics can act as both input variables and output variables for the analyses[2].

B. Design of Questionnaire

A questionnaire was applied to the class: "Introduction to Computer Engineering". It consists of the initial questionnaire (IQ) and the final questionnaire (FQ). Scores of technical report (TR) submitted every week, and those of the midterm exam (ME) and final exam (FE) are explicit characteristics of each student. We analyze them by using statistics, data mining, and information retrieval techniques which include classification and clustering. The contents of a questionnaire and their examples are shown in Table I and in Table II respectively. The time schedule for the class is depicted in Fig. 3.

[2]We choose variables carefully so that they do not generate loops for an analysis.

TABLE I

DATA OF CLASS

Exercise	Contents
Initial Questionnaire (IQ)	
Item type	7 questions (4-20 sub-questions each)
Text type	5 questions (250-300 characters in Japanese and 100 in Chinese each)
Midterm Exam (ME)	5 subjects
Technical Reports (TR)	11 times (each 1-2 subjects)
Final Exam (FE)	5 questions
Final Questionnaire (FQ)	6 questions (6-21 sub-questions each)
Item type	5 questions (250-300 characters in Japanese and 100 in Chinese each)
Text type	

TABLE II

CONTENTS OF QUESTIONNAIRE

Exercise		Examples (sub questions)
IQ	Item-type	✓ For how many years have you used computers? ✓ Do you have a plan to study abroad? ✓ Can you assemble a PC? ✓ Do you have a qualification related to information technology? ✓ Write 10 technical terms in information technology which you know.
	Text-type	✓ Write about your knowledge and experience on computer. ✓ What kind of work will you have after graduation? ✓ What do you imagine from the name of this class subject name?

Exercise		Examples (sub questions)
FQ	Item-type	✓ Could you understand the contents of this lecture? ✓ Was the midterm test difficult? ✓ Was it easy to read the handwritings on the white-board? ✓ Do you think the contents of this lecture to be useful to yourself? ✓ Do you want to finish this course even if it is optional? ✓ Which are you interested in applied technology or the fundamentals of computers? ✓ Which do you choose class (S) or class (G)?
	Text-type	✓ Do you want to be a member of laboratories related to the information technology? ✓ In the future, will you get a job in industries related to the information technology? ✓ Did your image on computers change after taking this lecture?

This questionnaire is made in WEB form, and it is on the following Web Site.
http : //hirasa.mgmt.waseda.ac.jp/users/comp-eng/

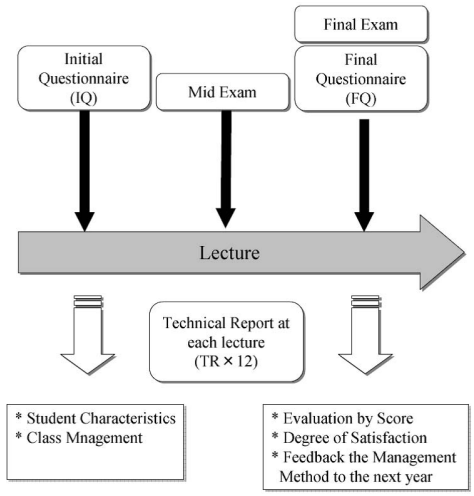

Fig. 3. Time schedule for class

IV. ALGORITHMS USED FOR ANALYSES

The documents with fixed formats are represented by an item-document matrix $G = [g_{mj}]$, where $g_{mj} \in \{0,1\}$ is the selected result of the item m (i_m) in the document j (d_j). The documents with free formats are also represented by a term-document matrix $H = [h_{ij}]$, where $h_{ij} \in \{0,1,2,\cdots\}$ is the frequency of the term i (t_i) in the document j (d_j). The dimensions of matrices G and H are $I \times D$, and $T \times D$, respectively, where the number of the total documents is D, that of the total items, I, and that of the total terms, T. Both matrices are compressed into those with smaller dimensions by the probabilistic decomposition in PLSI model [2], [7] similar to the single valued decomposition (SVD) in LSI (latent semantic indexing) model [1]. The (latent) states are denoted by $z_k \in \mathscr{Z}$ $(k = 1,2,\cdots,K)$. Introducing a weight λ $(0 \le \lambda \le 1)$, the log-likelihood function corresponding to the resultant matrix A:

$$A = \begin{bmatrix} \lambda G \\ (1-\lambda)H \end{bmatrix} = [a_{ij}] \quad (i = 1,2,\cdots,I+T, \quad j = 1,2,\cdots,D) \tag{1}$$

is maximized by the EM algorithm [7]. Then we obtain the probabilities $\Pr(z_k)(k = 1,2,\cdots,K)$, and the conditional probabilities $\Pr(t_i|z_k,i_m)$, and $\Pr(d_j|z_k)$. Using these probabilities, $\Pr(i_m,d_j)$ and $\Pr(t_i,d_j)$ are derived, and we decide the state for d_j depending on $\Pr(z_k|d_j)$.

The similarity function between z_k and $z_{k'}$, $s(z_k,z_{k'})$ is defined by [10]:

$$s(z_k,z_{k'}) = \sum_i \Big\{ h\big[\alpha \Pr(t_i|z_k) + (1-\alpha)\Pr(t_i|z_{k'})\big] - \alpha h\big[\Pr(t_i|z_k)\big] - (1-\alpha)h\big[\Pr(t_i|z_{k'})\big] \Big\} \tag{2}$$

where $0 \le \alpha \le 1$ and $h[x] = -x\log x$.

Assume that pairs (i_m,d_j) and (t_i,d_j) are generated independently, and also assume that i_m and t_i are generated independently of d_j conditioned on z_k. We construct the matrix A so that the above assumptions hold. Based on the good performance for a relatively small document set [3] discussed in the previous paper [5], [6], and the further improvement of it [10], we have used the following algorithms.

A. Classification algorithm [5]

The algorithm is strongly dependent on the fact and property that the EM algorithm usually converges to the local optimum solution from starting with an initial value. Hence

[3]Note that algorithms used in this paper are required to exhibit good performance to a set of a small number of documents, since the number of the students in a class will be usually at most 200.

we use a representative document as the initial value for the EM algorithm.

Suppose a set of documents $\mathscr{D}$ for which the number of categories is K, where the K categories are denoted by $C_1, C_2, \cdots, C_K$.

(1) Choose a subset of documents $\mathscr{D}^*$ $(\subset \mathscr{D})$ which are already categorized and compute representative document vectors $\vec{d}_1^*, \vec{d}_2^*, \cdots, \vec{d}_K^*$:

$$\vec{d}_k^* = \frac{1}{n_k} \sum_{\vec{d}_j \in C_k} \vec{d}_j \tag{3}$$

where n_k is the number of selected documents to compute the representative document vector from C_k and $\vec{d}_j = (a_{1j}, a_{2j}, \cdots, a_{Dj})^{\mathrm{T}}$, where T denotes the transpose of a vector.

(2) Compute the probabilities $\Pr(z_k)$, $\Pr(d_j|z_k)$ and $\Pr(t_i|z_k)$ which maximizes the log-likelihood function corresponding to the matrix A by the Tempered EM (TEM) algorithm, where $|\mathscr{Z}| = K$.

(3) Decide the state $z_{\hat{k}} (= C_{\hat{k}})$ for $\vec{d}_j$ as

$$\max_k \Pr(z_k|\vec{d}_j) = \Pr(z_{\hat{k}}|\vec{d}_j) \Rightarrow d_j \in z_{\hat{k}} \tag{4}$$

□

If we can obtain the K representative documents prior to classification, they can be used for $\vec{d}_k^*$ in eq. (3).

B. Clustering algorithm [10]

Suppose a set of documents to be clustered into S clusters, where the S clusters are denoted by $c_1, c_2, \cdots, c_S$.

(1) Choose a proper $K(\geq S)$ and compute the probabilities $\Pr(z_k)$, $\Pr(d_j|z_k)$, and $\Pr(t_i|z_k)$ which maximizes the log-likelihood function corresponding to the matrix A by the TEM algorithm, where $|\mathscr{Z}| = K$.

(2) Decide the state $z_{\hat{k}}$ $(= c_{\hat{k}})$ for $\vec{d}_j$ as

$$\max_k \Pr(z_k|\vec{d}_j) = \Pr(z_{\hat{k}}|\vec{d}_j) \Rightarrow d_j \in z_{\hat{k}} \tag{5}$$

If $S = K$, then $d_j \in c_{\hat{k}}$, and stop.

(3) If $S < K$, then compute a similarity measure $s(z_k, z_{k'})$ by eq. (2). Use the group average distance method with the similarity function $s(z_k, z_{k'})$ for agglomerative clustering the states z_k's until the number of clusters becomes S, then we have S clusters. Go to step (2). □

C. Extraction algorithm of important sentences [13]

A document is composed of a set of sentences. Measure the similarities between a sentence and the other sentences, and compute the score of the sentence by the sum of the similarities. Then choose a sentence which has the largest score as the important sentence in the document.

D. Extraction algorithm of feature sentences and feature words [11]

Let $\Pr(t_i|z_k) - \Pr(t_i)$ be the score of t_i, and the sum of the scores of t_i's which appear in a sentence be the score of the sentence. Then choose the words which have the larger scores as the feature words. Similarly, choose a sentence which has the larger scores as the feature sentence in the category or the cluster.

TABLE III
CONTENTS OF TOPICS

Class	Contents
Class G	- History of computers, fundamental concepts in computer - Basics of architecture - Basics of hardware - Basics of software - Applications of information technology etc.
Class S	- Architecture(stack machine, binary system, processor architecture) - Hardware(logic design, logical circuit, automaton) - Software(operating system, UNIX, language processor) etc.

V. QUESTIONNAIRE ANALYSES

A. Verification of class model by IQ

Before beginning of the class, we discuss a problem on the class management and the lecture plan. By using only IQ, the partition of students of the class is considered for students in Japan and in R.O.C.

The purpose of the partition of students is to improve the effect of education by adequately partitioning the students of the class based on their interested areas, levels, or intentions. Since the partition is made at the beginning of the class, we must make it by IQ only.

We discuss on partition by the contents of topics as shown in Table III.

Class G (generalist): wide and shallow technical topics
Class S (specialist): technical and professional topics.

A partition of Class G and Class S is examined by a classification algorithm using both the item-type and the text-type questionnaire of IQ only. Since the algorithm requires representative vectors (pseudo documents), they are obtained from the same questionnaire by students at graduate school (or the senior students whose jobs in future were decided). Then classification should be automatically generated. If we have the past documents of graduated students who got their jobs, we can use them as the categorized data in supervised learning.

The result of this classification compared with student's own choice is shown in Table IV. The characteristics extracted by discriminant analysis are shown in Table V.

B. Verification of class model by IQ and FQ

Let us try to interpret (1) the scores, (2) the degree of satisfaction, and (3) the favorite partition to students by the item-type questionnaire of IQ and FQ.

(1) Scores of students

We expect to explain the scores of the midterm exam (ME) and of the final exam (FE)(as intermediate criterion variables) by the item-type questionnaire (as explanatory variables) of IQ and FQ. Important sentences extracted from the text-type

TABLE IV
PARTITION OF CLASS G AND CLASS S

(i) Students in Japan

Automatic classification	A student's own choice		
	G	S	Total
G	22	24	46
S	17	35	52
Total	39	59	98

(ii) Students in R.O.C.

Automatic classification	A student's own choice		
	G	S	Total
G	13	3	16
S	9	7	16
Total	22	10	32

questionnaire of IQ and FQ based on the scores are shown in Table VI.

(2) Degree of satisfaction

Similar to the above experiment, the item-type questionnaire (as explanatory variables) of IQ and FQ can interpret the degree of satisfaction (as criterion variables) in terms of the contents of topics and in terms of class management by the multiple linear regression analysis as shown in Table VII. The degree of satisfaction is calculated as the weighted sum of the results of the item-type questionnaire.

(3) Partition by Class G and Class S

The reasons why the students choose Class G or Class S (as criterion variables) are shown in Table VIII.

C. Clustering of students

The clustering algorithm is applied to merged documents of both students in Japan and those in R.O.C. The results in the case $K = 2, 3$ are shown in Table IX. Extracted feature sentences in the case $K = 2$, $\lambda = 1.0$, and extracted feature words in the case $K = 3$, $\lambda = 0.5$, are shown in Table X, and XI, respectively.

D. Discussions

(1) It is shown that the degree of agreement between the student's own choice and automatic classification are approximately 60% by IQ only (Table IV). Although our method is probably not accurate enough to use automatic classification, but it would be still useful to assist and to guide students. We know that most of all students do not decide their future jobs yet in their second academic year. It is worth noting from our experience that the student's own choice is not always true. For example, it would be interesting whether a graduated student who is a member of staff at industry chose Class S or not. Past data by graduated students in their second year can be effectively used to this analysis. Automatic classification gives interesting tendency such that the students in Class S like to learn actively and wish to go to study abroad. There is almost no difference between students in Japan and in R.O.C.(Table V).

TABLE V
CHARACTERISTICS OF CLASS G AND CLASS S

(i) Students in Japan

	Characteristics x_i	Distinction coefficient a_i
Student's choice	You would like to attend this class and understand what it offers.	G S
	How long have you used email?	
	You are sciences-oriented, not literature-oriented.	
	Your grades last year were relatively good.	
	You would like to acquire some qualifications in the future.	
	As long as you receive a credit, you don't mind what your grades are.	
	You have looked at the syllabus.	
	How long have you used your own PC?	

Mis-discriminant ratio 30.5%

Automatic classification	You would like to study abroad.	
	This class should be mandatory for this school (department).	
	Have you ever expanded the memory of your PC?	
	How long have you used email?	
	How long have you used a computer?	
	You think you will learn to utilize a PC through this class.	
	You would like to attend this class and understand what it offers.	
	You have looked at the syllabus.	
	How many days per week did you come to the university last year?	
	You are sciences-oriented, not literature-oriented.	
	This class is necessary for the years to come.	

Mis-discriminant ratio 25.9%

(ii) Students in R.O.C

	Characteristics x_i	Distinction coefficient a_i
Student's choice	How long have you used the internet?	G S
	You would like to study abroad.	

Mis-discriminant ratio 30.2%

Automatic classification	You would like to study abroad.	
	You think you will learn to utilize a PC through this class.	
	You would like to acquire some qualifications in the future.	
	You would like to attend this class and understand what it offers.	
	How long have you used a computer?	
	You have a clear purpose of taking this class.	

Mis-discriminant ratio 10.7%

Discriminant analysis:

Discriminant function $z = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_p x_p$ $\begin{cases} z > 0 & d \in \text{class S} \\ z < 0 & d \in \text{class G} \end{cases}$

(2) Students in higher level both in Japan and in R.O.C. are interested in computer (Table VI). This would be quite natural.

(3) It is a little difficult to interpret the degree of satisfaction by the way of the class management, but easy, by the contents of the lecture by IQ and FQ (Table VI). This suggests that the degree of satisfaction depends on the contents of the lecture rather than the class management. The degree of satisfaction is influenced by interest of the field and motivation of learning. These are the important points for faculty development. The above discussion is useful to students in Japan, since the class is a required subject. A little difference between students in Japan and in R.O.C. exists such as motivation to qualification proceeded by the government (Japan) and to work abroad (R.O.C.).

(4) Comparing to IQ only (Table V), it is more clear to

TABLE VI

IMPORTANT SENTENCES EXTRACTED FROM TEXT-TYPE QUESTIONNAIRE FOR SCORES OF STUDENTS

(i) Students in Japan

Score	Example of sentence
High Over 70	I am interested in information security, network and internet technologies. We are to learn how the computer is used, not how it works. Now I'd like to know much more about the computer. Attendance checking makes much incentive to me.
Low Under 69	I rarely used a computer or a PC until now, except for the internet, so I have no special knowledge. Attendance checking should be done properly and should be reflected on the grades. I browsed through the textbook - as difficult as I had anticipated. I never really cared much about any of the computer-related areas.

(ii) Students in R.O.C

Score	Example of sentence
High Over 80	I'd like to take on a computer-related job. I'd like to learn about the computer and then to research on it. To me, the computer is nothing but a processor and an application. I'd like a class that actually uses a computer hands-on.
Low Under 79	I understand almost nothing about the computer. I know very little about the computer. The computer always makes me suffer. I'd like the class to actually use a computer in order to teach the theory behind it.

interpret better partition to students by IQ and FQ (Table VIII). This suggests that proper partition to the next year should take causal relations obtained in this year into account. The students who are classified to Class S like sciences rather than literature, and wish to go to the graduate school.

(5) In the case of $\lambda = 0.0$ (texts only), students are completely separated into students in Japan and those in R.O.C. by the clustering algorithm (Table IX). This would be dependent on the difference in:

- used languages themselves and
- national characteristics which can be seen in the extracted feature sentences (Table X).

Text processing is strongly influenced by the translation methods of Chinese into Japanese, since the questionnaire analyses system was developed for the Japanese language. There are automatic translation method [15] and human translation method. In this paper, human translation is used quoted by automatic translation.

In the case of $\lambda = 1.0$ (items only), the difference of used languages does not affect to clustering. Clusters are constructed by only characteristics of students. Extracted feature sentences exhibit the characteristics of students in Japan and in R.O.C. (Table X).

In the case of $K = 3$, $\lambda = 0.5$, extracted feature words represent that the cluster z_3 contains more professional students (Table XI).

(6) It is also possible to realize the system for Chinese language, where we use automatic indexing by N-gram or morpheme in Chinese. Table XII shows important sentences extracted from text-type questionnaire (IQ only) for high or low scores of students in R.O.C. The (i) in this table corresponds to (ii) of Table VI, where

TABLE VII

INTERPRETATION OF DEGREE OF SATISFACTION BY ITEM-TYPE QUESTIONNAIRE

(i) Students in Japan

Satisfaction in terms of contents of the lecture

Explanatory variable x_{jl}	Partial regression coefficient b_l
This class should use a PC in every possible way.	
This class should be mandatory for this school (department).	
Did you understand the lecture every time within the class hour?	
Are you willing to attend the class?	
How long have you used a computer?	
The computer will be an important tool for corporate management.	
You think you will learn to utilize a PC through this class.	
You want to work hard in every class and get good grades.	
You are sciences-oriented, not literature-oriented.	
You have looked at the syllabus.	
You would like to acquire some qualifications in the future.	
Do you think there should be a registration for this class?	
How long have you used your own PC?	

Contribution ratio = 0.766

Satisfaction in terms of class management

Explanatory variable x_{jl}	Partial regression coefficient b_l
Did you find the entire course difficult?	
How was the progress within the class?	
How was the volume of the reports?	
Were the lectures useful every time?	
You would like a mid-term exam.	
Was class registration handled appropriately?	
You want to work hard in every class and get good grades.	
This class should be mandatory for this school (department).	
You plan to attend this class every week.	
As long as you receive a credit, you don't mind what your grades are.	

Contribution ratio = 0.782

(ii) Students in R.O.C

Satisfaction in terms of contents of the lecture

Explanatory variable x_{jl}	Partial regression coefficient b_l
Were the lectures useful every time?	
Do you feel fulfilled, now that you have finished the course?	
Did you find the lectures useful?	
Was the final exam difficult?	
I'd like to attend this lecture and understand what it offers.	
How long have you used email?	
How long have you used a computer?	
Are you interested in the applications of the computer, or its basic principles?	
You would like to work actively abroad after you graduate.	

Contribution ratio = 0.893

Satisfaction in terms of class management

Explanatory variable x_{jl}	Partial regression coefficient b_l
Was the final exam difficult?	
Did you find the entire course difficult?	
Was class registration handled appropriately?	
Did you try to solve the problems for your report on your own every time?	
Do you think this class is necessary for you?	
Do you feel fulfilled, now that you have finished the course?	
If you like a class, you work especially hard for it.	
You would like to study abroad.	
As long as you receive a credit, you don't mind what your grades are.	

Contribution ratio = 0.810

Multiple linear regression analysis:

Criterion variable (score) $y_j = b_0 + b_1 x_{j1} + \cdots + b_p x_{jp} + N(0, \sigma^2)$

TABLE VIII

INTERPRETATION OF PARTITION FOR CLASS G OR CLASS S

(i) Students in Japan

Characteristics x_l	Distinction coefficient a_l G	S
You are sciences-oriented, not literature-oriented.		
Did you find the lectures interesting?		
You work hard for a class even if you are not interested in it.		
You would like to acquire some qualifications in the future.		
Did you find the entire course difficult?		
You have a clear purpose of taking this class.		
Do you think this class is necessary for you?		
How long have you used the internet?		
You would like to study abroad.		
You would like to go on to graduate school.		

Mis-discriminant ratio 21.5%

(ii) Students in R.O.C.

Characteristics x_l	Distinction coefficient a_l
You would like to acquire some qualifications in the future.	
How long have you used a computer?	
You think you will learn to utilize a PC through this class.	
You would like to study abroad.	
Did you find the entire course difficult?	
Do you think this class is necessary for you?	
This class should use a PC in every possible way.	
Were the lectures useful every time?	
You would have taken this class even if it was optional.	
Because you took this class, now you would like to study more in this field.	
How long have you used the internet?	
Was class registration handled appropriately?	
Do you think that you don't need to know how the computer works as long as you know how to use it?	

Mis-discriminant ratio 10.7%

Discriminant analysis:

Discriminant function $z = a_0 + a_1x_1 + a_2x_2 + \cdots + a_px_p$ $\begin{cases} z > 0 & d \in \text{class S} \\ z < 0 & d \in \text{class G} \end{cases}$

TABLE IX

RESULTS OF CLUSTERING

$K = 2$

λ	0.0		0.5		1.0	
z_k	z_1	z_2	z_1	z_2	z_1	z_2
Japan	0	144	0	144	118	26
R.O.C.	90	3	102	5	24	83

$K = 3$

λ	0.0			0.5			1.0		
z_k	z_1	z_2	z_3	z_1	z_2	z_3	z_1	z_2	z_3
Japan	0	83	61	0	86	58	15	68	61
R.O.C.	85	4	4	90	4	13	79	19	9

TABLE X

EXTRACTED FEATURE SENTENCES ($K = 2$, $\lambda = 1.0$)

	Feature sentences
z_1 (Japan)	I am willing to learn about UNIX. I will learn about network technology. I learn about information retrieval. I will learn about information and communication technology.
z_2 (R.O.C.)	I plan to attend this class every week. I am willing to learn about making web pages. I will learn about EXCEL and WORD. I will learn about network technology. I will work hard for classes that I am interested in. I would like to understand the lecture.

TABLE XI

EXTRACTED FEATURE WORDS ($K = 3$, $\lambda = 0.5$)

	Feature words
z_1 (R.O.C.)	computer, field, professor, introduction, program, design, course, work
z_2 (Japan A)	PC, interest, class, management, area, study, computer, myself, system, employment, internet, engineering, information filtering
z_3 (Japan B)	report, information, network technology, information and communication technology (IT), information security, software and hardware

TABLE XII

IMPORTANT SENTENCES EXTRACTED FROM TEXT-TYPE QUESTIONNAIRE (IQ ONLY) FOR SCORES OF STUDENTS IN R.O.C.

(i) By translating Chinese into Japanese

Score	Example of sentence
High Over 80	I'd like to learn much about computers, especially OS. I wish I not only use computers, but improve them. I wish I have my own computer. I hope that computers are practical tools. I'd like to learn computers, because I did not know about them.
Low Under 79	I notice that there are many terms related to computers. I'd like to assemble a computer and to learn knowledge about it. I wish I can learn computers by Q&A. I wish I can catch up my classmate.

(ii) By directly Chinese text processing

Score	Example of sentence
High Over 80	When I faced to computers, I feel that I will enter in the IT age. This class teaches us the history of computer development and introduces basic computer systems. I wish I have my own computer.
Low Under 79	If I choose one interested area on computers, I'd like to learn hardware. Computers, especially networks are very useful for me. If everything is running well, I wish I will be able to enter to the IT society.

the translation of Chinese into Japanese is used and processing of text to extract important sentences is done by Japanese language. While the (ii) in this table uses Chinese text processing, where we have used morpheme of Chinese and used the same algorithm as for Japanese discussed in Sec. IV, C.

We will generally estimate the performance of the methods for processing Chinese [14].

VI. CONCLUSIONS AND FUTURE WORKS

It can be concluded that we obtain useful information to improve the class management by student questionnaire with both the item-type and the text-type. The result shows verification of the class model for "Introduction to Computer Engineering".

Student questionnaire analyses systems always require effective algorithms for a set of small number of documents, since the class is usually consisted by 30–150 students. To solve this problem, it is necessary to develop new information retrieval techniques, hence we are considering to apply Bayesian decision theory into information retrieval systems [3].

We have developed the questionnaire system by Japanese language. We would like to expand our system so that we can handle other languages such as Chinese.

Questionnaires must be carried out to collect data for several years, and their time series analysis and the review of the model also remain as further studies.

ACKNOWLEDGEMENT

The authors would like to thank Mr. T. Ishida, Mr. M. Nagao, Mr. T. Sakaguchi, Dr. T. Sakai, Dr. M. Goto, Dr. Y-C. Tsai, and Dr. M. Suzuki for their helpful and fruitful discussions. This work was partially supported by the Grant of the Telecommunications Advancement Foundation (TAF).

REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, ACM Press, New York, 1999.

[2] D. Cohn, and T. Hofmann, "The missing link - A probabilistic model of document content and hypertext connectivity," *Advances in Neural Information Processing Systems* (NIPS $\times$13), MIT Press 2001.

[3] M. Goto, T. Ishida, and S. Hirasawa, "Representation method for a set of documents from the viewpoint of Bayesian statistics," *Proc. IEEE 2003 Int. Conf. on SMC*, pp.4637-4642, Washington DC, Oct. 2003.

[4] M. Gotoh, T. Sakai, J. Itoh, T. Ishida, and S. Hirasawa, "Knowledge discovery from questionnaires with selecting and descripting answers," (in Japanese) *Proc. of PC Conference*, pp.43-46, Kagoshima, Aug. 2003.

[5] S. Hirasawa, and W. W. Chu, "Knowledge acquisition from documents with both fixed and free formats," *Proc. IEEE 2003 Int. Conf. on SMC*, pp.4694-4699, Washington DC, Oct. 2003.

[6] S. Hirasawa, and W. W. Chu, "Classification methods for documents with both free and fixed formats," *Proc. 2004 Int. Conf. Management Sciences and Dexision Making*, pp.427-444, Taipei, R.O.C., May 2004.

[7] T. Hofmann, "Probabilistic latent semantic indexing," *Proc. of SIGIR'99*, ACM Press, pp.50-57, 1999.

[8] T. Ishida, J. Itoh, M. Gotoh, T. Sakai, and S. Hirasawa, "A model of class and its verification," (in Japanese) *Proc. of 2003 Fall Conference on Information Management, JASMIN*, pp.226-229, Hakodate, Nov. 2003.

[9] T. Ishida, M. Gotoh, and S. Hirasawa, "Analysys of student questionnaire in the lecture of computer science," (in Japanese) *Computer Education, CIEC*, vol.18, pp.152-159, July 2005.

[10] M. Nagao, H. Yagi, and S. Hirasawa, "Document clustering methods based on probabilistic latent semantic indexing with feature of words," (in Jappanese) *The 29th Symposium on Information Theory and its Applications (SITA2006)*, Hakodate, Hokkaido, Japan, Nov. 28 - Dec. 1, 2006.

[11] J. Itoh, T. Ishida, M. Gotoh, and S. Hirasawa, "A method for extracting important sentences using co-occurrence similarities between words ," (in Japanese) *Forum on Information Technology 2002*, pp.83-84, Tokyo, Sept. 2002.

[12] J. Itoh, T. Ishida, M. Gotoh, T. Sakai, and S. Hirasawa, "Knowledge discovery in documents based on PLSI," (in Japanese) *Forum on Information Technology 2003*, pp.83-84, Ebetsu, Sept. 2003.

[13] J. Itoh, T. Sakai, and S. Hirasawa, "A method for extracting parts of important sentences from Japanese documents using dependency trees," (in Japanese) *IPSJ, Tech. Rep. Natural language processing*, 158-4, pp.19-24, Nov. 2003.

[14] H. Hamada, T. Ishida, and S. Hirasawa, "Classification and clustering methods for Chinese text based on PLSI model" *(in preparation)*.

[15] J-Beijing Chinese-Japanese Machine Translation System, http://www.kodensha.jp/soft/jb/

[16] T. Sakai, J. Itoh, M. Gotoh, T. Ishida, and S. Hirasawa, "Efficient analysis of student questionnaires using information retrieval techniques," (in Japanese) *Proc. of 2003 Spring Conference on Information Management, JASMIN*, pp.182-185, Tokyo, June 2003.

[17] T. Sakai, T. Ishida, M. Gotoh, and S. Hirasawa, "A student questionnaires analysis system based on natural language expressions," (in Japanese) *Forum on Information Technology 2004*, N-021, pp.325-328, 2004.