Student Questionnaire Analyses for Class Management based on Document Clustering and Classification Algorithms

Shigeichi Hirasawa

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 algorithms for processing answers with both the items and the texts and those for extracting important sentences from texts combined with statistical techniques, a method for analyzing the questionnaires is completed. Especially, we discuss on class partition problem for the students of the class: "Computer Engineering" to drive an effective class management, where the students are portioned into two classes: Class for Specialist (S) and Class for Generalist (G). From 2002 through 2008, student questionnaires with the same contents have been performed, and they were analyzed pursuing after occupations of graduated students at their companies. Although the accuracy of analyzed data is not enough satisfied, we can still obtain useful knowledge for class management which leads to faculty development.

Keywords: student questionnaire, classification algorithm, faculty development, probabilistic latent semantic indexing (PLSI) model, important sentences

1. Introduction

In recent two decades, declining birth rate brought competitive intensity among Japanese universities not only in research activities but in educational activities. For many universities in Japan, the evaluation of the quality assurance of education programs by Japan Accreditation Board for Engineering Education (JABEE) has become important.

---

*This paper was presented at the 2009 International Conference of Digital Contents as key note speech on December 18, 2009 at Yuan Ze University, Taiwan, R. O. C.

サイバー大学 IT 総合学部・学部長，サイバー大学 IT 総合学部・教授

原稿受付日：2010年10月 4 日
原稿受理日：2011年 2 月10日

To improve the class management, our university introduced the student questionnaire system using Web-site. It is, however, not enough for improving in detail with taking into account the special situation of the class. Especially, there have been existing difficult problems for class management in the class: "Computer Engineering" at the 2nd academic year, Department of Industrial and Management Systems Engineering, Waseda University, since the students in this class have different qualities of a prior knowledge, interested areas, motivation, experiences, and levels in computer skills [4]. Moreover, their jobs after graduation have many kinds of fields in business. Therefore, we have designed the student questionnaire only applicable to this class [4], [9], [10]. The student questionnaire has two types of replies: item type and text type [5]. The former is fixed format such as answered by selected numbers, symbols, and yes or no. The latter is free format answered by text.

In this paper, we focus upon the class partition problem. It arises to this class, because we intend to perform effective education by supplying different contents of topics to different sorts of students. We discuss on partition into two classes which have two contents of topics and are represented as Class G (generalist): wide and shallow technical topics, and Class S (specialist): deep technical and professional topics, by only using the initial questionnaire.

Based on the probabilistic latent semantic indexing (PLSI) model [8], we have proposed the classification[1] and clustering algorithms [5], [13]. These methods exhibit good performances for a small set of documents [6]. Besides the proposed algorithm, we have also developed the extraction algorithm of important sentences, feature words, and feature sentences used for the text part [12], [14], [16]. The traditional statistical techniques can also be applied to the item part.

We have applied the student questionnaire for these seven years (2002–2008), where the contents of the student questionnaire are the same except for the first year (2002). The students at the 2nd year in 2003 graduated in March 2006 as Bachelors, and those in part, in March 2008 as Masters. Then we know the type of business after their graduation. Estimating the kind of their jobs from the type of business, we would try to guess their true choice from Class G or Class S.

By analyzing the student questionnaire, coincident rates between results by the automatic partition, those by student's own choice, and those by student's estimated choice are derived. Although the results obtained by the automatic partition method can not explain enough their future jobs, it would be still useful to assist or guide the students. As a result, we can obtain useful information which leads to faculty developments.

In Section 2, we represent the methods for analysis, i. e. models and algorithms.

The performance evaluation of algorithms is briefly described in Section 3. The student questionnaire analysis is discussed in Section 4. Concluding remarks are given in Section 5.

2. Methods for Analysis

2.1 Models

The method of analyzing the questionnaire is shown in Fig. 2.1 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], [11], [13]. 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 [12], [14], [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[2] only, statistical techniques such as multiple linear regression

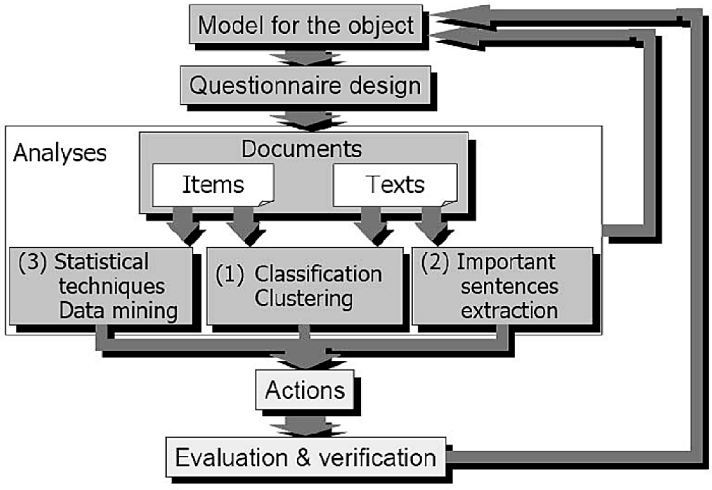

Fig. 2.1 Questionnaire analysis model

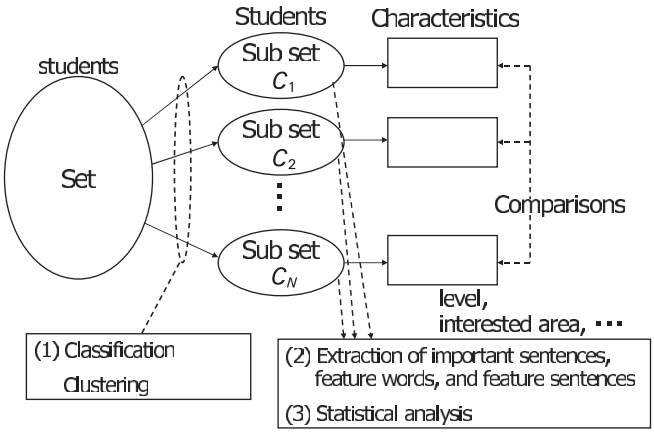

Fig. 2. 2 Outline of analysis

analysis, and 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], [8] which is known to be one of the most powerful models 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], [11]. 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. As the result, out line of the analysis can be represented by Fig. 2. 2.

2. 2 Algorithms

2. 2. 1 Document Set

Documents called in this paper imply the answers of student questionnaire.

The documents with fixed formats are represented by an item-document matrix $G=[g_{mj}]$, where $g_{mj} \in \{0, 1\}$ is the selected answer[3] 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 based on PLSI model [2], [8] similar to the single valued decomposition (SVD) based on LSI (latent semantic indexing) model [1]. The (latent) states are denoted by $z_k \in Z$ $(k=1, 2, \cdots, K)$. Introducing a weight $\lambda(0 \leq \lambda \leq 1)$, the log-likelihood function corresponding to the resultant matrix A:

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

is maximized by the EM algorithm [8]. Then we obtain the probabilities $\Pr(z_k)$ $(k = 1, 2, \cdots, K)$, and the conditional probabilities $\Pr(t_i \mid z_k, i_m)$, and $\Pr(d_j \mid z_k)$. 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 $\Pr(z_k \mid d_j)$.

The similarity function between z_k and $z_{k'}$, $s(z_k, z_{k'})$ is defined by [11]:

$$\begin{aligned} & s(z_k, z_{k'}) \\ & \quad = \sum_i \{h[\alpha \Pr(t_i \mid z_k) + (1-\alpha) \Pr(t_i \mid z_{k'})] \\ & \qquad - \alpha h[\Pr(t_i \mid z_k)] - (1-\alpha) h[\Pr(t_i \mid z_{k'})]\}, \end{aligned} \qquad (2)$$

where $0 \leq \alpha \leq 1$ and $h[x] = -x \log x$ $(0 \leq x \leq 1)$.

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 by 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 number of documents[4] discussed in the previous paper [5], [6], [13] and the further improvement of it [11], we have used the following algorithm.

2.2.2 Classification Algorithm [5]

The algorithm is constructed strongly depending on the fact and property that the EM algorithm usually converges to the local optimum solution from an arbitrary initial state. Hence we use a representative (pseudo) document as the initial value for the EM algorithm.

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

(1) Choose a subset of documents D^* ($\subset D$) which are already categorized. Compute representative document vectors $\boldsymbol{d}_1^*, \boldsymbol{d}_2^*, \cdots, \boldsymbol{d}_k^*$:

$$\begin{aligned}\boldsymbol{d}_k^* &= \frac{1}{n_k} \sum_{\boldsymbol{dj} \in C_k} \boldsymbol{d}_j \quad (k = 1, 2, \cdots, K) \\ &= (a_{1k}^*, a_{2k}^*, \cdots, a_{(I+T)k}^*)^{\mathrm{T}}, \end{aligned} \tag{3}$$

where n_k is the number of selected documents to compute the representative document vector from C_k and $\boldsymbol{d}_j = (a_{1j}, a_{2j}, \cdots, a_{(I+T)j})^{\mathrm{T}}$, where T denotes the transpose of a vector. Set the initial values as:

$$\Pr(z_k) = \frac{1}{K}, \tag{4}$$

$$\Pr(d_j \mid z_k) = \frac{1}{D}, \tag{5}$$

$$\Pr(t_i \mid z_k) = \frac{a_{ik}^* + \alpha}{\sum_{i'} (a_{i'k}^* + \alpha)}, \quad (\alpha > 0). \tag{6}$$

(2) Compute the probabilities $\Pr(z_k)$, $\Pr(d_j \mid z_k)$, and $\Pr(t_i \mid z_k)$ which maximize the respective log-likelihood functions corresponding to the matrix A by the Tempered EM (TEM) algorithm, where $|Z| = K$.

(3) Decide the class $C_{\hat{k}}$ for d_j as

$$\max_k \Pr(z_k \mid d_j) = \Pr(z_{\hat{k}} \mid d_j) \Rightarrow d_j \in C_{\hat{k}}. \tag{7}$$

□

2.2.3 Clustering Algorithm [11]

Suppose a set of documents to be clustered into S tentative 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 $\Pr(z_k)$, $\Pr(d_j \mid z_k)$, and $\Pr(t_i \mid z_k)$ which maximize the respective log-likelihood functions corresponding to the matrix A by the TEM algorithm, where $|Z| = K$.

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

$$\max_k \Pr(z_k \mid \boldsymbol{d}_j) = \Pr(z_{\hat{k}} \mid \boldsymbol{d}_j) \Rightarrow d_j \in z_{\hat{k}}. \tag{8}$$

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). □

2.2.4 Extraction Algorithm of Important Sentences [14]

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.

2.2.5 Extraction Algorithm of Feature Sentences and Feature Words [12]

Let $\Pr(t_i \mid z_k) - \Pr(t_i)$ be the score of t_i, and let 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 class.

3. Performance Evaluation of Algorithms

In this section, we show the performance of the proposed algorithm in general cases. The document sets which have used are summarized in Table 3.1.

Table 3.1 Document sets

<table>
<tr><th></th><th>contents</th><th>format</th><th>amount</th><th>categorize</th></tr>
<tr><td>(a)</td><td>articles of Mainichi news paper in '94</td><td>Free (texts only)</td><td>101,058 (see Table 3.2)</td><td>Yes (9+1 ategories)</td></tr>
<tr><td>(b)</td><td rowspan="2">Questionnaire (see Table 4.1 in detail)</td><td rowspan="2">fixed and free (see Table 4.2)</td><td>135+35</td><td>Yes (2 categories)</td></tr>
<tr><td>(c)</td><td>135</td><td>No</td></tr>
</table>

3.1 Classification

We first apply the proposed algorithm into the set of articles of newspaper. The experimental data is shown in Table 3.2.

Table 3.2 Selected categories of newspaper

category	contents	# articles	# used for training	# used for test
C_1	business	100	50	50
C_2	local	100	50	50
C_3	sports	100	50	50
total		300	150	150

The results obtained in this case are depicted in Table 3. 3 and 3. 4. We see that the error rate of the proposed algorithm is smaller than that of the conventional algorithms. The clustering process used for the classification are illustrated in Fig. 3. 1 (a) and 3. 1(b) for the initial step (=0) and the final step (=4095), respectively.

Table 3. 3 Classified number form C_k to $C_{\hat{k}}$ for each method

method	from C_k	to C_k		
		C_1	C_2	C_3
VS method	C_1	17	4	29
	C_2	8	38	4
	C_3	15	4	31
LSI method	C_1	16	6	28
	C_2	6	43	1
	C_3	12	5	33
PLSI method	C_1	41	0	9
	C_2	0	47	3
	C_3	13	6	31
Proposed method	C_1	47	0	3
	C_2	0	50	0
	C_3	4	2	44

Table 3. 4 Classification error rate

Method	Classification error
VSM	42.7%
LSI	38.7%
PLSI	20.7%
Proposed method	6.0%

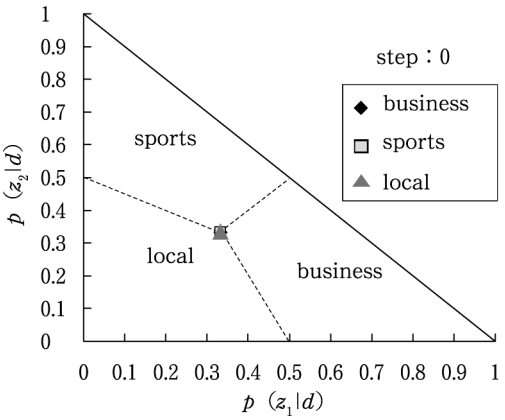

Fig. 3. 1 (a) Clustering process by EM algorithm (step: 0)

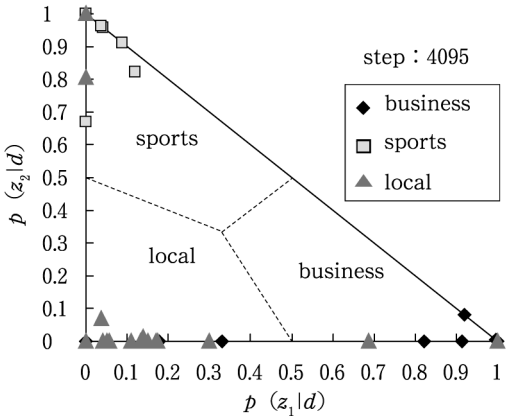

Fig. 3. 1 (b) Clustering process by EM algorithm (step: 4095)

3.2 Clustering

We have used the student questionnaire data as shown in Table 3.6 where contents of the questionnaire are shown in Table 3.5. First, the documents of the students in Class CS and those in Class IS are merged as shown in Fig. 3.2 Then, the merged documents are divided into two classes ($S=2$) by the proposed algorithm.

Table 3.5 Contents of initial questionnaire

Format	Number Of questions	Examples
Fixed (item)	7 major questions[2]	- For how many years have you used computers? - Do you have a plan to study abroad? - Can you assemble a PC? - Do you have any license in information technology? - Write 10 terms in information technology which you know[4].
Free (text)	5 questions[3]	- Write about your knowledge and experience on computers. - What kind of job will you have after graduation? - What do you imagine from the name of the subject?

Table 3.6 Object classes

Name of subject	Course	Number of students
Introduction to Computer Science (Class CS)	Science Course	135
Introduction to Information Society (Class IS)	Literary Course	35

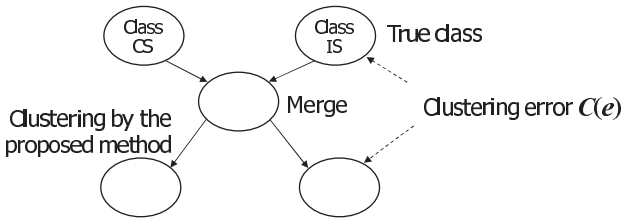

Fig. 3.2 Class partition problem by clustering method

The results obtained are illustrated in Figs. 3.3, 3.4, and 3.5. Fig. 3.3 shows the division into two classes is successfully performed, since the one class always appears with high similarity independent of K as seen in the dendrogram. We also see that the clustering error rate is low enough by choosing the parameters λ and K form Table 3.3 and 3.4.

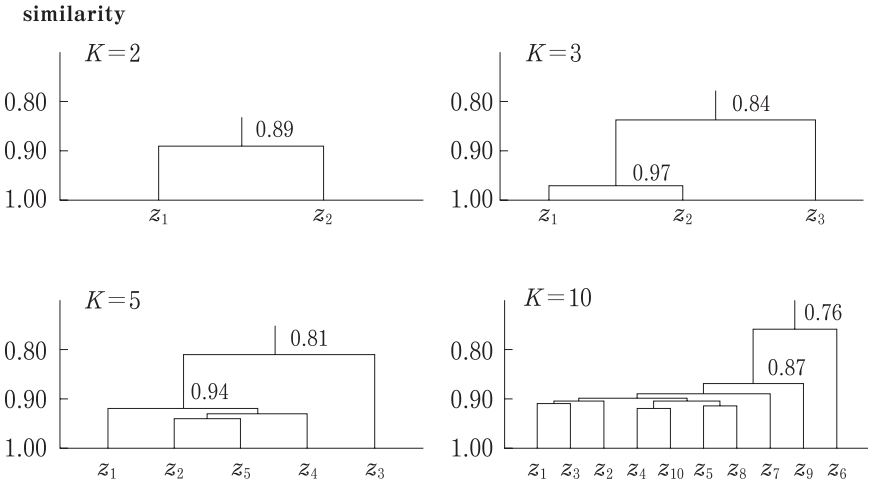

Fig. 3. 3 Dendrogram of clusters

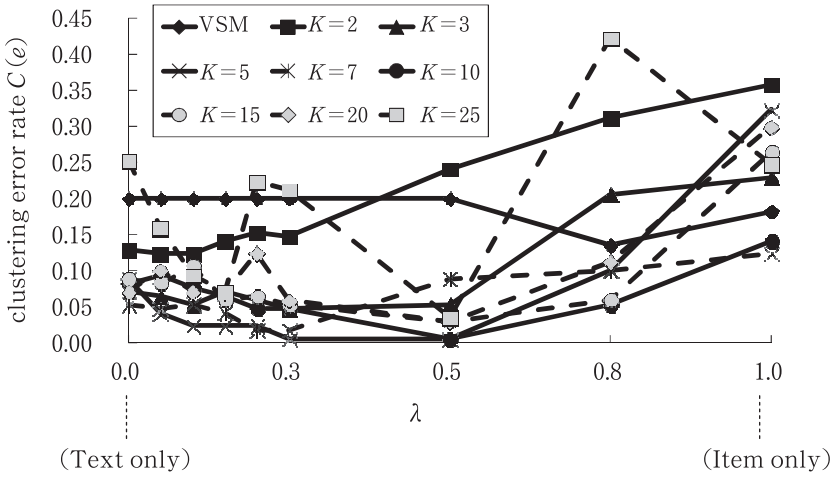

Fig. 3. 4 Clustering error rate $C(e)$ vs. λ

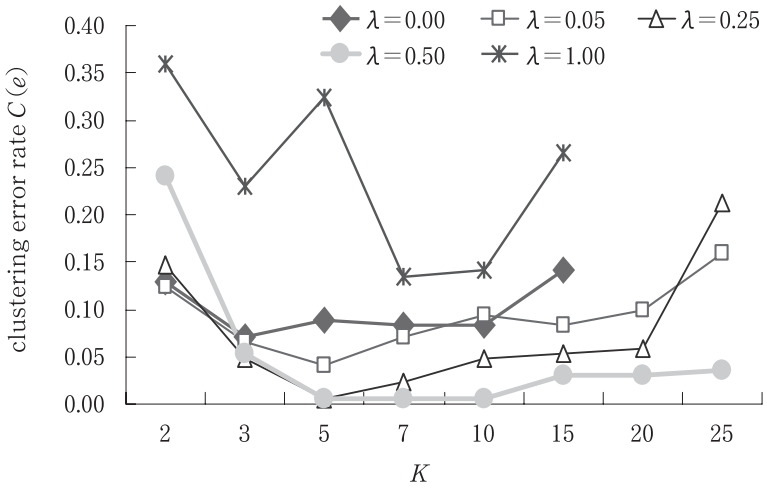

Fig. 3. 5 Clustering error rate $C(e)$ vs. K

4. Student Questionnaire Analysis

A class model for this object is shown in the Fig. 4. 1. A technique to find out requirements of the students from the questionnaire is discussed by applying the questionnaire analyses model shown in Fig. 2. 2. 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.

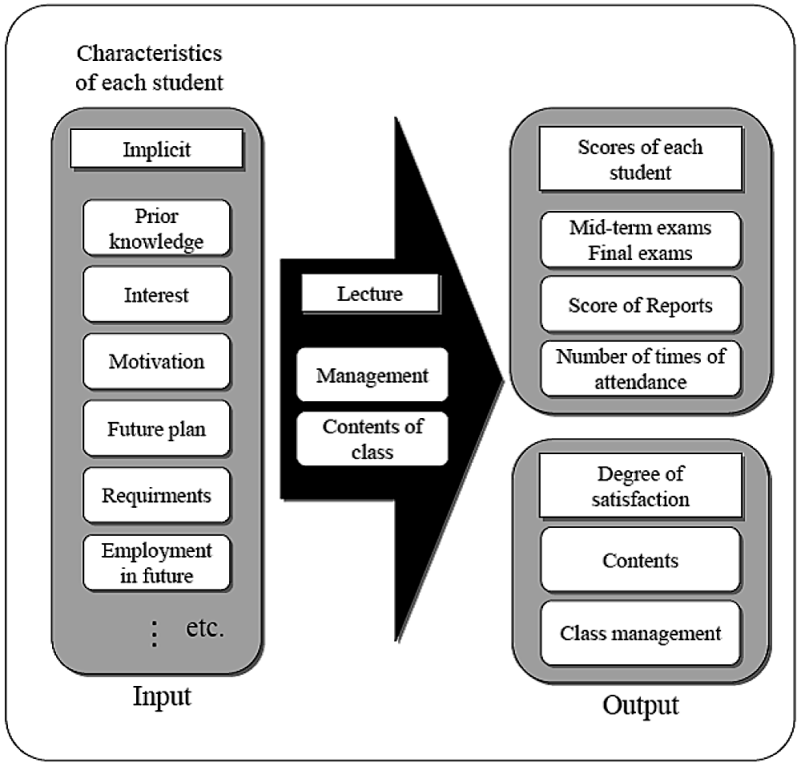

Fig. 4. 1 Class model

4. 1 Design of Student Questionnaire

A questionnaire is applied to the class: “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. The data of the class, and the

contents of a questionnaire and their examples are shown in Table 4. 1 and in Table 4. 2, respectively. The time schedule for the class is depicted in Fig. 4. 2.

We analyze them by using statistics, data mining, and information retrieval techniques which include partition of a set of documents.

Table 4. 1 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 4. 2 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 any qualification related to information technology? ✓ Write 10 technical terms in information technoloqy 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?

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)?
	Texp-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 imaqe on computers change after taking this lecture?

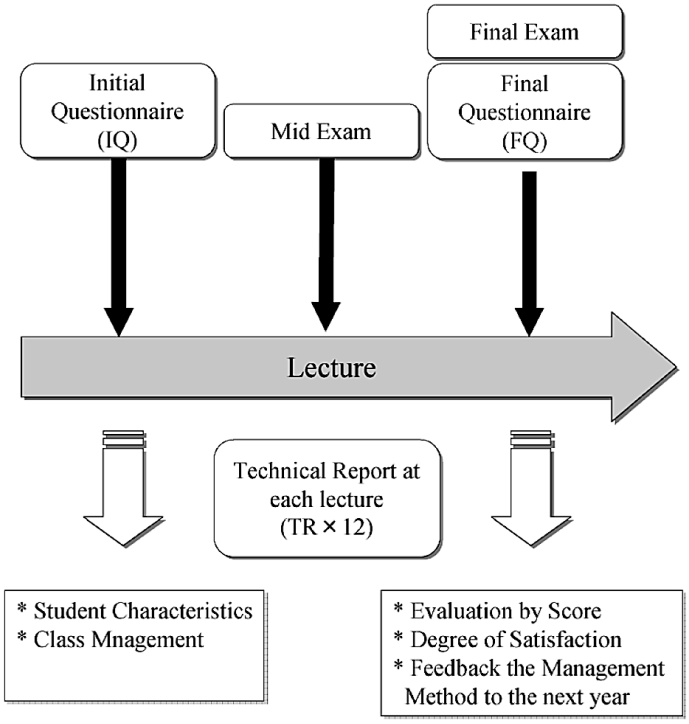

Fig. 4. 2 Time schedule for class

4. 2 Verification of Class Model by IQ

(1) Class Partition Problem

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.[5]

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 4. 3.

Class G (generalist): wide and shallow technical topics

Class S (specialist): technical and professional topics.

(2) Estimation of job

It is easy to decide by students themselves whether they choose Class G or Class S. The reason why we apply student questionnaire (by initial questionnaire (IQ)) is to extract the student characteristics which are not awaked by them and to try to adequately partition the classes depending their latent properties at the beginning of the class.

Table 4. 3 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.

We use the term “job” as the kind of occupation such as:

(S) circuit design, mechanical design, electric design, production management, quality control, software development, system engineering, R & D, and so on, and
(G) sales, accounting, personal management, services, and so on.

The former (S) is a type of engineering or technology, while the latter (G) is not the type of them. Hence (S) would require professional skills in computer, and (G), does not so much.

On the other hand, we use the “business” as the kind of company such as

(a) trading, finance, banking, service, securities market, consultation, general construction, and so on, and
(b) electric manufacturing, automobile manufacturing, precision instrument manufacturing, system integration, software development, and so on.

After graduation, most of the students join companies. Although we know only the name of companies in which they joined, the job of each student is estimated by the author according to his experience. As a result, there is a difficulty to estimate the job from the name of company, which means estimation of (G) or (S) from (a) or (b), where (a) and (b) are classified by the name of companies such as Canon Inc., IBM Japan Ltd., NEC, Toyota Motor Corp., Acceture, Nomura Research Institute Ltd., East Japan Railway Co., Kashima Corp., and so on. This difficulty arises an ambiguity of decision in the label (G) and (S). In other words, the name of company does not directly fix on the job. For example, we know one of the graduated students joins

Sony Corp., but we don't know his or her job. His or her job may be production management, or may become sales. As another example, he or she joins Tokyo Mitsubishi UFJ Bank, but his or her job may be a teller, or may be a staff at information service division. The author's experience avoids an error in estimation by his knowledge on the jobs of graduated students as possible as he can.

Anyway, some errors may occur in estimated jobs as shown in Fig. 4. 3.

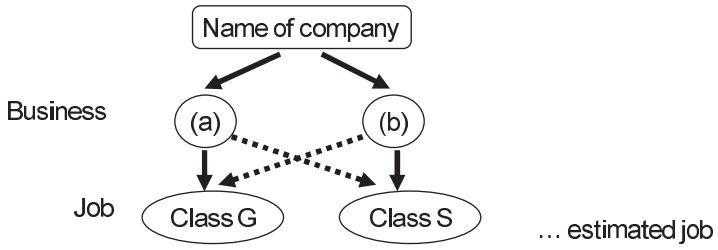

Fig. 4. 3 Transition of students

(3) Results of partition

By using IQ, the partition algorithm is performed, where the representative document for Class G and Class S are generated by those of graduated students in March 2007. The results obtained by it are indicated as "Automatic Partition (AP)". By using FQ, each student replies his or her own choice of Class G or Class S. The results obtained by it are indicated as "Student's Own Choice (SOC)". By the estimation of the job of each graduated student, the results obtained by it are indicated as "Student's Estimated Choice (SEC)".

Table 4. 4 shows the list of the number of partitioned students between AP and SEC. Table 4. 5 also shows that between SOC and SEC.

Table 4. 4 Numbers of partitioned students between AP and SEC

		SEC		Total
		G	S	
AP	G	20	19	39
	S	17	30	47
	Total	37	49	86

AP: Automatic Partition
SEC: Students Estimated Choice

Table 4. 5 Numbers of partitioned students between SOC and SEC

		SEC		Total
		G	S	
SOC	G	30	24	54
	S	7	28	35
	Total	37	52	89

SOC: Student's Own Choice

(4) Results of extracted important sentence

The extraction of the important sentences of the documents is examined for the students appeared in the elements except for diagonal elements in Table 4. 4. The results are shown in Table 4. 6.

Table 4. 6 Extracted important sentences

(a) AP vs. SEC

(AP, SEC) = (Class G, Class S)

[IQ]	– I think that what is necessary is just to be able to master a computer. – What I am reminded of from the term "computer" is a personal computer. – I would like to be able to master a computer.
[FQ]	– It was meaningful that the knowledge of the computer was able to be acquired. – In the future, I think that I will associate with a computer for a long time. – I thought that it was not so difficult to understand the structure of a computer.

(AP, SEC) = (Class S, Class G)

[IQ]	– I would like to decompose by myself or to set up a personal computer. – I am very interested in the content of the class.
[FQ]	– I did not think that this class was not much important for myself. – I was not able to acquire the impression that this field was interesting. – Although it is not interested in a computer, I think that knowledge is required.

(b) SOC vs. SEC

(SOC, SEC) = (Class G, Class S)

[IQ]	– I would like to be able to master a computer. – Since I was imagining that I used a personal computer in this lesson, it differed from prior imagination.
[FQ]	– My view about a computer changed by having studied the principle of the computer. – From now on, I will associate with a computer for a long time. – The content of the class was difficult. – It was serious to have understood the content of the class. – I am interested in how to use a computer.

(SOC, SEC) = (Class S, Class G)

[IQ]	– I would like to understand the principle of a computer. – It is required to understand a principle, in order to master a computer.
[FQ]	– I would like to study a computer more and to obtain a deeper understanding. – In order to master a computer, it is helpful to know the structure.

(5) Discussions

(1) It is shown that the coincident rate between AP and SEC is approximately 58.1 % by IQ only (Table 4. 4), and that between SOC and SEC, 65.1% by FQ (Table 4. 5). The method for partitioning the class is probably not accurate enough, although the rate of the latter is slightly improved.

(2) It can be explain that the above improvement is brought by learning the subjects, since FQ is performed at the end of the class.

(3) Table 4. 5 suggests us that the student at the 2nd academic year do not decide their future jobs. Hence they would not awake whether professional skill is

required or not in future.

(4) From the viewpoint of the hypothesis testing, under the hypothesis H_0: Two variables are independent, H_0 for Table 4. 4 cannot be rejected, while H_0 for Table 4. 5 can be rejected [7].

(5) Although the coincident rates are not large, partition is still useful to guide the students by the suggestions: There are cases such as

(i) Even though the student becomes a generalist, he or she who is interested in computers, would chose Class S (Table 4. 6 (a)).

(ii) There are many cases such that if the student wanted to learn only the method for using computers, he or she who graduated as a Master, will join an industry as a specialist (Table 4. 6 (a)).

(iii) If the student who wanted to be a specialist, and could not be interested in computers, he or she will become a generalist (Table 4. 6 (a)).

(iv) In contrast to (iii), there is a case such that the student who was interested in such as the structure of computers, will go to professional in engineering (Table 4. 6 (a)).

(v) If the student who chose Class G, change his or her idea by learning the principle of computers, he or she becomes a specialist (Table 4. 6 (b)).

(vi) Even if the student felt that the lecture was difficult, he or she will become a specialist (Table 4. 6 (b)).

(vii) Since recent students usually chose easy way, there is a case that he or she who wants to become a specialist, joins the Class G.

(6) Most of all students state that they will satisfy fruitful and interested contents of the lecture, and their choice of the Class S or Class G depends on the topics. Therefore, the contents of topics are very important.

As additional experiments, if we use FQ, we can partition the students into Class G or Class S with high coincident rates by weighting the following items.

1. [IQ] Prior knowledge (technical term).
2. [FQ] The range of the theme is suitable?
3. [FQ] I would like to study about a logic circuit.
4. [FQ] I would like to study about cache memory.

4. 3 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 questionnaire of IQ and FQ based on the scores are shown in Table 4. 7.

Table 4. 7 Important sentences extracted from text-type questionnaire for scores of students

(i) Students in Japan

Score	Example of Sentences
High over 70	I'm interested in Information security, network and Internet technology. We are to learn how the computer works, not how to work with it. Now I'd like to know much more about the computer. How the class registration is done makes much sense to me.
Low under 69	I rarely used a computer or a PC until univercity, except for the Internet, so I have no special knowledge. Class registration 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 Sentences
High over 70	I'd like to take on a computer-related job. I'd like to learn about the computer and then do a 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 69	I understand 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.

(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 4. 8. 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 4. 9.

Table 4. 8 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 4. 9 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 $z = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_p x_p$ $\begin{cases} z > 0 & d \in \text{class S} \\ z < 0 & d \in \text{class G} \end{cases}$

(4) Discussions

From Table 4. 7:

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

(2) Students in lower level do not have prior knowledge in computer.

From Table 4. 8:

(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.

(4) This suggests that the degree of satisfaction depends on the contents of the lecture rather than the class management.

(5) The degree of satisfaction is influenced by interest of the field and motivation of learning. These are the important points for faculty development.

(6) The above discussion is useful to students in Japan, since the class is a required subject.

(7) 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.).

From Table 4. 9:

(8) Comparing to IQ only, it is more clear to interpret better partition to students by IQ and FQ. This suggests that proper partition to the next year should take causal relations obtained in this year into account.

(9) The students who are classified to Class S like sciences rather than literature, and wish to go to the graduate school.

4. 4 Clustering of Students in Japan and R.O.C.

(1) Difference between students in Japan and in R.O.C.

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 4. 10. 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 4. 11 and 4. 12, respectively.

Table 4. 10 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 4. 11 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.

(2) Discussions

From Table 4. 10:

(1) 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 4. 12 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

(2) This would be dependent on the difference in:
- used languages themselves and
- national characteristics which can be seen in the extracted feature sentences.

(3) Text processing is strongly influenced by the translation methods of Chinese into Japanese, since the questionnaire analyses system was developed for the Japanese language.

(4) There are automatic translation method [15] and human translation method.

(5) In this paper, human translation is used aided by automatic translation.

(6) In the case of $\lambda = 1.0$ (items only), the difference of used languages does not affect to clustering.

From Table 4. 11:

(7) Clusters are constructed by only characteristics of students. Extracted feature sentences exhibit the characteristics of students in Japan and in R.O.C.

From Table 4. 12:

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

5 Concluding Remarks

Collecting documents obtained by student questionnaire for these six years, we analyze the graduated student questionnaire by trace back to their 2nd academic year. It is necessary to collect data at least four years for taking account the

estimated their jobs. The results obtained in Section 4 are not accurate enough to use automatic partition of the class, but it is still useful to assist and to consult the students. We know that almost all students do not decide their future jobs yet in their 2nd academic year. It proves, however, that students are sound and have some robustness in their future plans, in a sense that they are going to learn not only their future jobs but their unsophisticated thirst for knowledge.

(1) The reason for the choice of the course is strongly dependent on their contents of interested topics. This corresponds to the previous result, i.e., the degree of satisfaction depends on the contents of the lecture [7].
(2) The degree of satisfaction for 90% of the students is in high (including in very high). This suggests us that we have to update the topics so that we let the students be always interested in.
(3) Two-thirds of the students support the introduction of the course system. This leads us to introduce the class partition into Class G and Class S.

Based on the results of the above trial, we will introduce the class partition in next year, although careful and detailed investigations are required as further works.

By development of the questionnaire analysis system in Chinese [7], we can apply directly to students in R.O.C. This is also remained as a future study.

Acknowledgment

The author would like to thank late Professor Fu-Yih Shih at Leader University, R.O.C. and Professor Wei-Tzen Yang at Tamkang University, R.O.C. for their helpful activities for applying the student questionnaire in R.O.C.

He also thank Professor M. Goto and Dr. T. Ishida at Waseda University for their valuable works to this study.

A part of this research was supported by the Grant of the Telecommunication Advancement Foundation (TAF).

Notes

1. We use the term in this paper “partition” rather than “classification”, since the students must choose a subset of the class, and cannot choose two or more subsets of the class. Strictly speaking, partition implies mathematically as follows: Let a set A has its subsets A_i, where A iff $A_i \cap A_j = \phi$ or $A_i = A_j$ for $i \neq j$.
2. Information investigated attribute of the categorical data, e.g., the scores of examinations for students, is added to a sort of the items.
3. The attribute values of n-level ($n > 2$) are orthogonalized so that each value has 0 or 1.
4. 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.

5. The student questionnaires translated into Chinese are also applied to students at Leader University and Tamkang University in R.O.C. [7].

Bibliography

[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*) *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 Decision Making,* pp. 427–444, Taipei, R.O.C., May 2004.

[7] S. Hirasawa, F–Y. Shih, and W–T. Yang, "Student questionnaire analyses for class management by text mining both in Japanese and in Chinese," *Proc. 2007 IEEE International Conference on System, Man and Cybernetics,* pp. 398–403, Montreal, Canada, Oct. 2007.

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

[9] 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, the Japan Society for Management Information* (*JASMIN*), pp. 226–229, Hakodate, Nov. 2003.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[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, the Japan Society for Management Information* (*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.

文書クラスタリングと文書分類アルゴリズムに基づく授業運営のための学生アンケート分析

平　澤　茂　一

本論文は，授業に関する学生のアンケート結果から，授業運営に役立つ情報が得られることを示した一連の研究成果をまとめたものである。確率型潜在的意味インデキシングモデルに基づく文書クラスタリング・文書分類アルゴリズム，重要文抽出アルゴリズムを提案し，統計的手法を併せ用いて選択型と記述型の混在するアンケートを分析している。

授業アンケートは，早稲田大学理工学部経営システム工学科の「コンピュータ工学」を履修した学生を対象に実施してきた。2002 年から 2008 年にわたり同一内容で実施しており，特にクラス分割問題を扱っている。コンピュータに対する事前知識・興味・成績などを収集し，将来コンピュータのスペシャリスト（S）かジェネラリスト（G）かにクラスを分割するための分析を行った。さらに，卒業生の就職先の業務を推定し，分割した場合の妥当性・有効性についても検証を行った。残念ながら分割結果は，精度において十分満足できるものではなかったが，授業運営に有効な示唆を得た。

なお，この論文は，著者が 2009 年 12 月に台湾で開催された The 2009 International Conference of Digital Contents において口頭発表した key note speech をフルペーパーとしたものである。予稿集には要旨のみ掲載されている。スライドは下記 URL 参照。
http://www.it.mgmt.waseda.ac.jp/hirawork/09 list.htm

キーワード：学生アンケート，分類アルゴリズム，授業改善，PLSI モデル，重要文