The 2009 International Conference on Digital Content, Dec. 17-18, 2009,
Chung-Li, Taiwan, R.O.C.

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

Shigeichi Hirasawa

Cyber University, Japan, and
Waseda Research Institute for Science and Engineering, Japan
hira@waseda.jp

A part of this paper was presented at the 2008 International Conference in Management Science and Decision Making, Taipei, R.O.C., June 28, 2008.

Contents of Talk

1. Introduction
2. Methods for Analysis
   - 2.1 Models
   - 2.2 Algorithms
3. Performance Evaluation of Algorithms
   - 3.1 Classification
   - 3.2 Clustering
4. Student Questionnaire Analysis
   - 4.1 Design of Student Questionnaire
   - 4.2 Verification of Class Model by IQ ･･･Class Partition Problem
   - 4.3 Verification of Class Model by IQ and FQ
   - 4.4 Clustering of Students in Japan and R.O.C.
5. Concluding Remarks

1. Introduction

- Class management
- Faculty development

Student questionnaire, class model

- Object class:
  “Introduction to Computer Engineering”
- Students of management and information department at:
  - □ Waseda University (Japan)
  - □ Leader University (Taiwan, R.O.C.)
  - □ Tamkang University (Taiwan, R.O.C.)

Technology:

(1) Classification or clustering for documents with fixed formats (items) and free formats (texts),

(2) Extraction of important sentences or feature sentences and words from texts which helps us to briefly understand the contents of the texts,

(3) Interpretation of characteristics of the set of documents by traditional statistical techniques.

- Problems of partitioning students of the class into a few subclasses
- to improve the degree of satisfaction of the students and to increase the effectiveness of education.

！！NOTE！！

Students in the 2nd academic year do not awake what kind of job they will take in future.

Two types of graduated students:

(a) Techically professional engineer

(b) General and economical anaysist, sales engineer

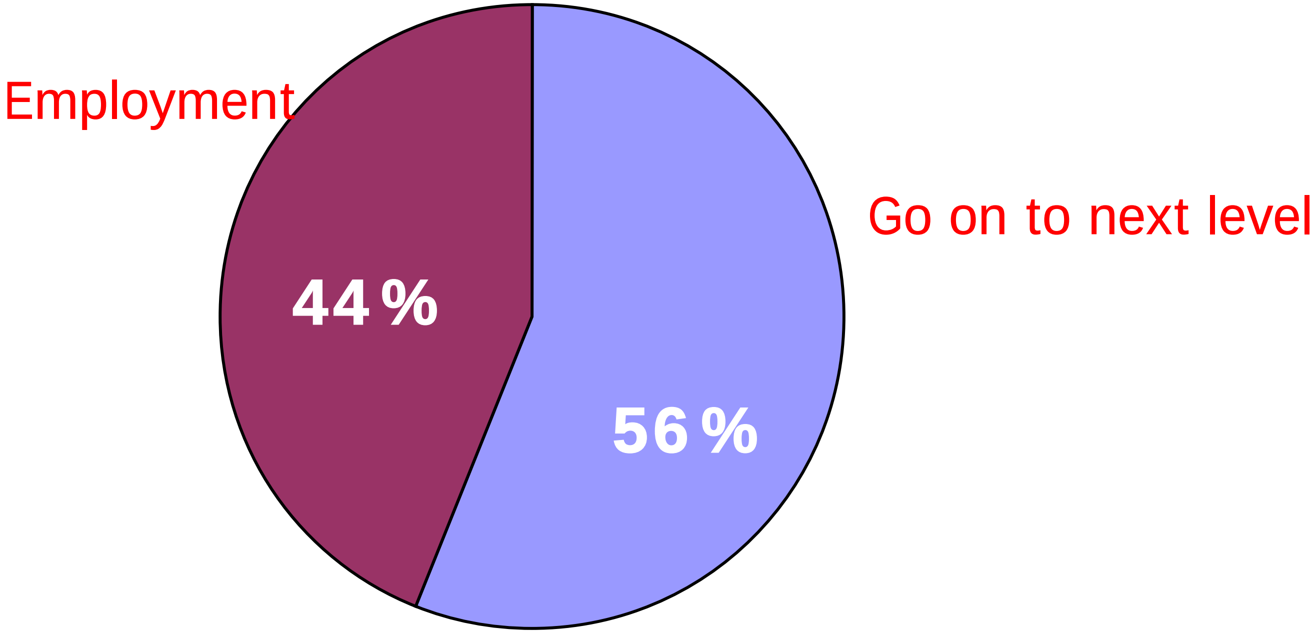
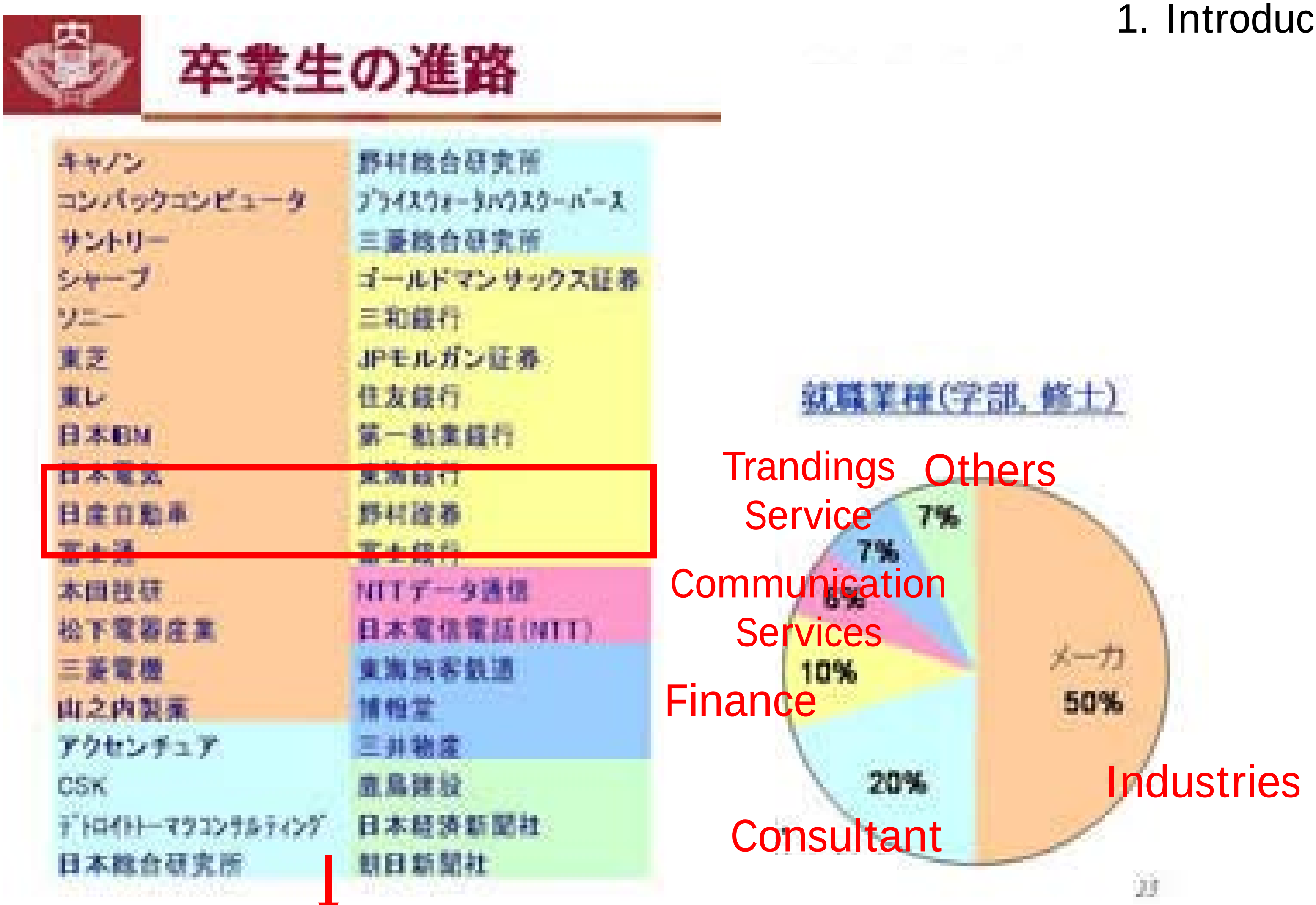


Fig. 1.1: Example of future path of undergraduate students (Waseda University)

1. Introduction



卒業生の進路

キャノン
コンパックコンピュータ
サントリー
シャープ
ソニー
東芝
東レ
日本IBM
日本電気
日産自動車
富士通
本田技研
松下電器産業
三菱電機
山之内製薬
アクセンチュア
CSK
デトロイトトーマツコンサルティング
日本総合研究所

野村総合研究所
プライスウォーターハウスクーパース
三菱総合研究所
ゴールドマンサックス証券
三和銀行
JPモルガン証券
住友銀行
第一勧業銀行
東海銀行
野村證券
富士銀行
NTTデータ通信
日本電信電話(NTT)
東海旅客鉄道
博報堂
三井物産
鹿島建設
日本経済新聞社
朝日新聞社

就職業種(学部、修士)

Major companies

Fig. 1.2: Example of jobs of undergraduate and graduate students (Waseda University)

Major companies:

[Industries]

- Canon Inc.
- Nihon Unisys, Ltd.
- Suntory Limited
- Sharp Inc.
- Sony Corp.
- Toshiba Corp.
- TORAY Ltd.
- IBM Japan Ltd.
- NEC
- Nissan Motor Co., Ltd.
- Fujitsu Ltd.
- Honda Motor Co., Ltd.
- Matsushita Electric Industrial Co., Ltd.
- Mitsubishi Electric Corp.
- Astellas Pharma Inc.

[Consultants]

- Accenture
- CSK Systems Corp.
- Deloitte Touche Tohmatsu. Japan Inc.
- The Japan Research Institute, Ltd.
- Nomura Research Institute, Ltd.
- Pricewaterhouse Coopers, International Ltd.
- Mitsubishi Research Institute, Inc.

[Finance]

- The Goldman Sachs Group, Inc.
- The Bank of Tokyo-Mitsubishi UFJ Ltd.
- Sumitomo Mitsui Banking Corp.
- Mizuho Bank, Inc.
- Nomura Secureties Co., Ltd.

[Communication Services]

- NTT Data Corp.
- Nippon Telphone and Telegraph East Corp.

[Tradings and Services]

- East Japan Railway Company
- Hakuhodo Inc.
- Mitsui and Co. Ldt.

[Others]

- Kashima Corp.
- Nikkei Corp.
- The Mainichi Newspapers

2. Methods for Analysis

2.1 Models

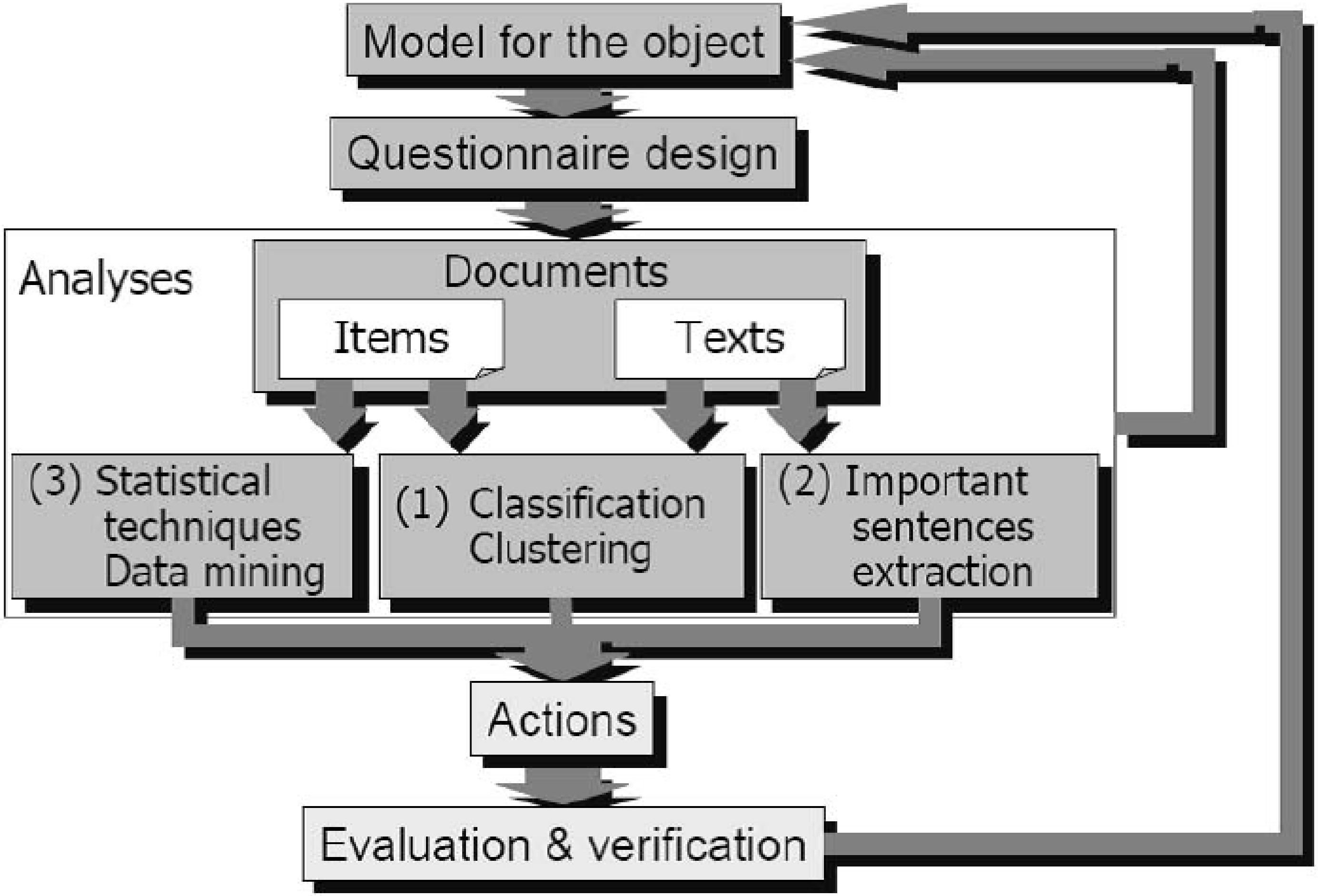

Model for the object
Questionnaire design
Analyses
Documents
Items
Texts
(3) Statistical techniques Data mining
(1) Classification Clustering
(2) Important sentences extraction
Actions
Evaluation & verification

Fig. 2.1: Questionnaire analysis model

Objects:Service level evaluation :

e.g.

hospital (patient) model

overseas student model

consumer model

job matching model

market model

ticket purchase model

etc.

Analyses phase:

(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.

We have proposed the algorithm based on the probabilistic latent semantic indexing (PLSI) model [2], [7].

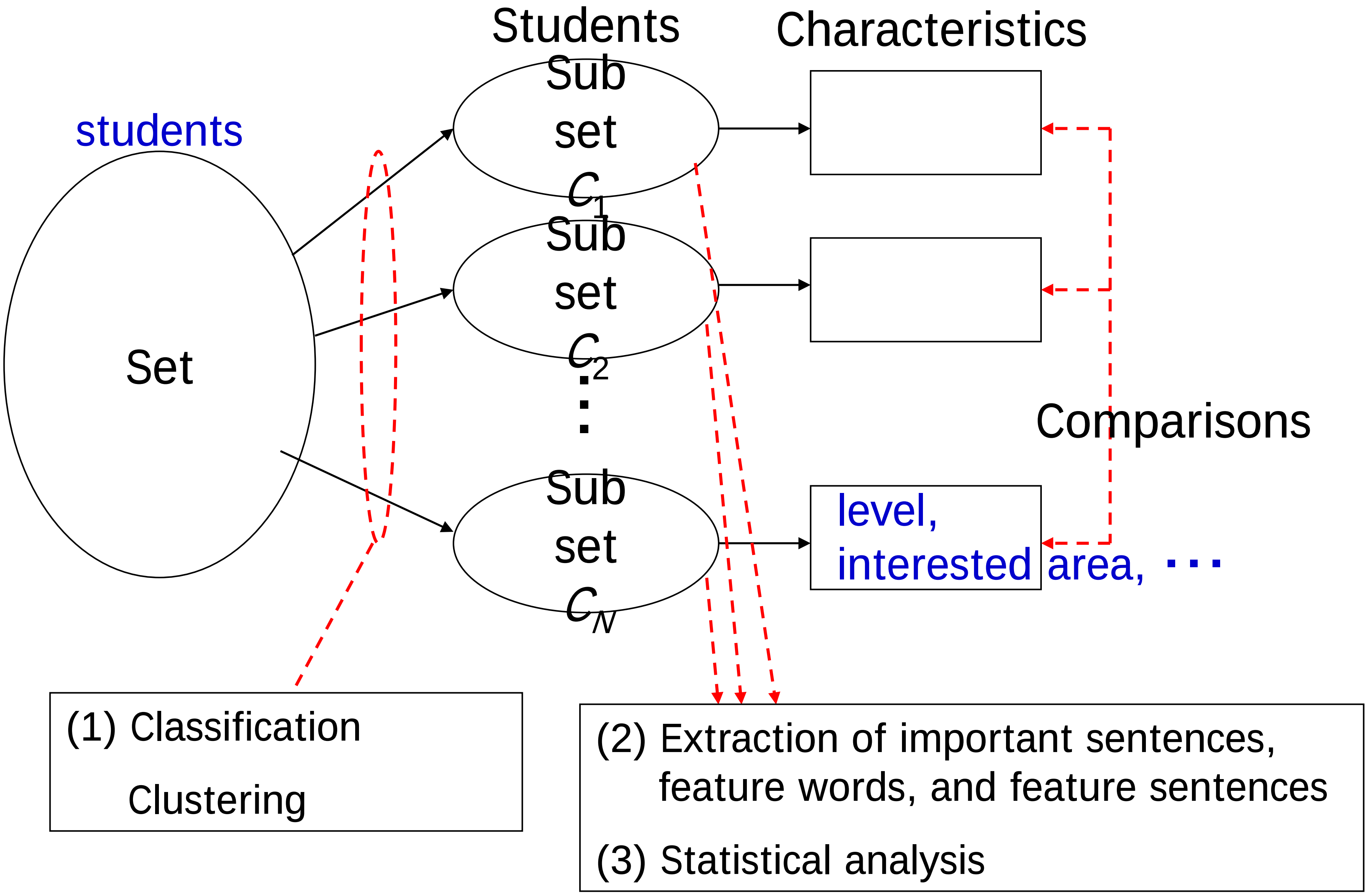
(2) For the texts only, important sentences, or feature sentences and words are extracted from the documents by the algorithms for extracting important information [11], [13], [16], [17].

These results are helpful to easily understand the opinions and directly give useful information of the classes (categories) or clusters.

(3) For the items only, statistical techniques such as multiple linear regression analysis, and discriminated analysis, are used to analyze the characteristics of each set of members.

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.



(1) Classification

Clustering

(2) Extraction of important sentences, feature words, and feature sentences

(3) Statistical analysis

Fig. 2.2: Outline of analysis

2.2 Algorithm

Information Retrieval Model

Text Mining:

- Information Retrieval including
- Clustering
- Classification

Table 2.1: Mathematical model of information retrieval

Base	Model
Set theory	(Classical) Boolean Model Fuzzy Extended Boolean Model
Algebraic	(Classical) Vector Space Model (VSM) [BYRN99] Generalized VSM Latent Semantic Indexing (LSI) Model [BYRN99] Neural Network Model
Probabilistic	(Classical) Probabilistic Model Extended Probabilistic Model Probabilistic LSI (PLSI) Model [Hofmann99] Inference Network Model Bayesian Network Model

Document

Table 2.2: Formats of questionnaire

Format		Example in paper archives		matrix
Fixed format	Items	- The name of authors - The name of journals - The year of publication - The name of publishers	- The name of countries - The year of publication - The citation link	$G \in \{0,1\}^{I \times D}$
Free format	Texts	The text of a paper - Introduction - Preliminaries .... - Conclusion		$H \in \{0,1,2,\cdots\}^{T \times D}$

$G = [\ g_{mj}\ ]$: An item-document matrix

$H = [\ h_{ij}\ ]$: A term-document matrix

d_j : The j-th document

t_i : The i-th term

i_m : The m-th item

g_{mj} : The selected result of the m-th item (i_m) in the j-th document (d_j)

h_{ij} : The frequency of the i-th term (t_i) in the j-th document (d_j)

The Probabilistic LSI (PLSI) Model

A) $$A = [a_{ij}] = \begin{bmatrix} \lambda G \\ (1-\lambda)H \end{bmatrix}, \; a_{ij} = tf(i,j) \qquad (2.1)$$

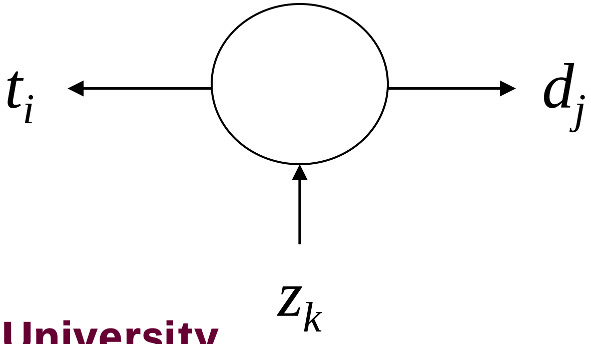
the number of term t_i in document d_j

B) Reduction of dimension by latent class (similar to SVD)

C) Latent class (state model based on factor analysis)

(i) an independence between pairs (t_i, d_j)

(ii) a conditional independence between t_i and d_j



z_k: state

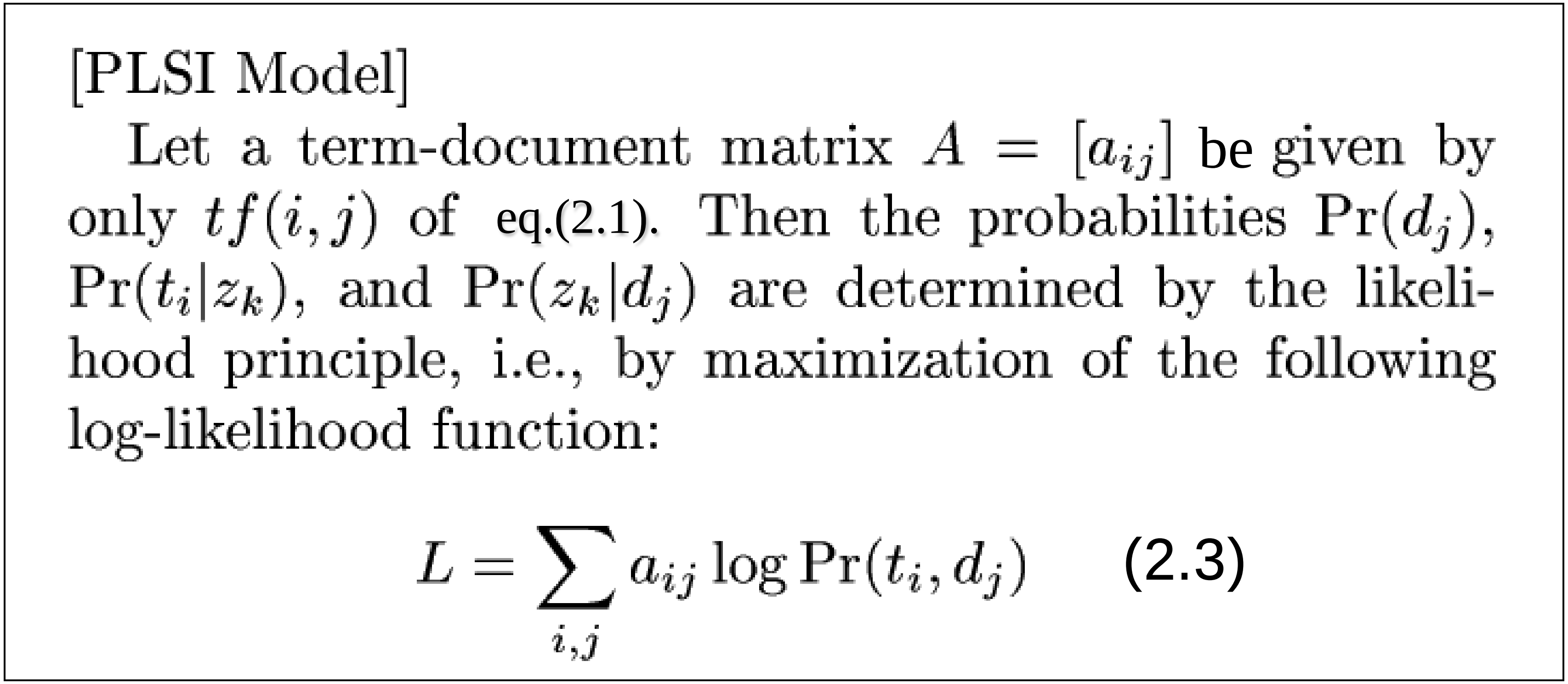
The Probabilistic LSI (PLSI) Model

Similarity function:

$$s(z_k, z_{k'}) = \sum_i \Big\{ h\big[\alpha \Pr(t_i|z_k) + (1-\alpha)\Pr(t_i|z_{k'})\big] - \alpha h\big[\Pr(t_i|z_k)\big] - (1-\alpha)h\big[\Pr(t_i|z_{k'})\big] \Big\} \quad (2.2)$$

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

PLSI Model



[PLSI Model]

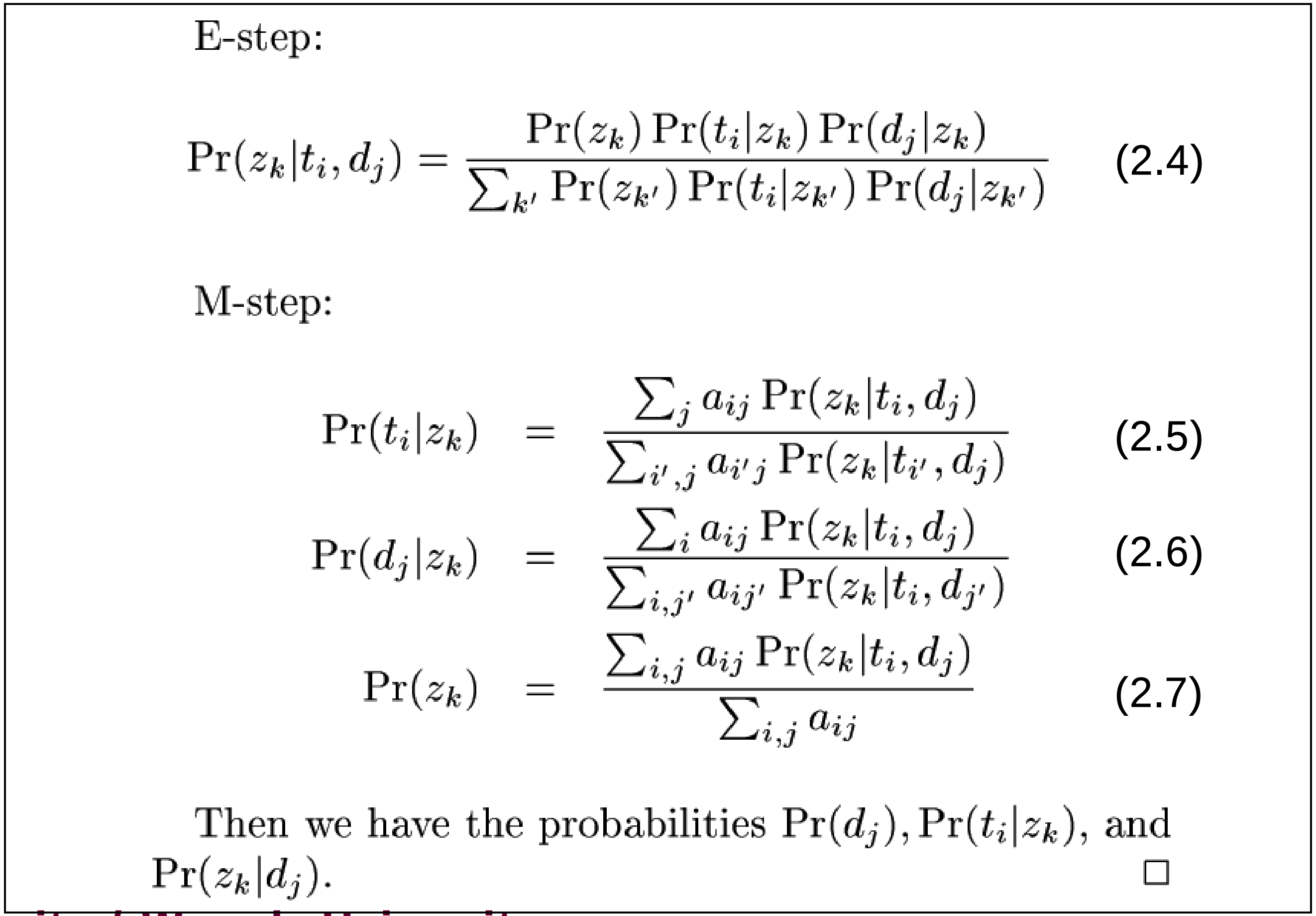
Let a term-document matrix $A = [a_{ij}]$ be given by only $tf(i,j)$ of eq.(2.1). Then the probabilities $\Pr(d_j)$, $\Pr(t_i|z_k)$, and $\Pr(z_k|d_j)$ are determined by the likelihood principle, i.e., by maximization of the following log-likelihood function:

$$L = \sum_{i,j} a_{ij} \log \Pr(t_i, d_j) \qquad (2.3)$$

EM Algorithm

[EM algorithm]

According to eq.(2.1), the maximum value of eq.(2.3) is computed by alternating E-step and M-step until it converges.



E-step:

$$\Pr(z_k|t_i,d_j) = \frac{\Pr(z_k)\Pr(t_i|z_k)\Pr(d_j|z_k)}{\sum_{k'}\Pr(z_{k'})\Pr(t_i|z_{k'})\Pr(d_j|z_{k'})} \qquad (2.4)$$

M-step:

$$\Pr(t_i|z_k) = \frac{\sum_j a_{ij}\Pr(z_k|t_i,d_j)}{\sum_{i',j} a_{i'j}\Pr(z_k|t_{i'},d_j)} \qquad (2.5)$$

$$\Pr(d_j|z_k) = \frac{\sum_i a_{ij}\Pr(z_k|t_i,d_j)}{\sum_{i,j'} a_{ij'}\Pr(z_k|t_i,d_{j'})} \qquad (2.6)$$

$$\Pr(z_k) = \frac{\sum_{i,j} a_{ij}\Pr(z_k|t_i,d_j)}{\sum_{i,j} a_{ij}} \qquad (2.7)$$

Then we have the probabilities $\Pr(d_j), \Pr(t_i|z_k)$, and $\Pr(z_k|d_j)$. □

A. Classification Algorithm [5]

The EM algorithm usually converges to the local optimum solution from starting with an initial value.

K: The number of categories (C_1, C_2, ... , C_K)

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

$$\vec{d}_k^* = \frac{1}{n_k} \sum_{\vec{d}_j \in C_k} \vec{d}_j \qquad (2.8)$$

where n_k is the number of selected documents to compute the representative document vector from C_k and $\vec{d}_j = (a_{1j}, a_{2j}, \cdots, a_{Dj})^T$, 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 TEM algorithm, where $|\mathscr{Z}| = K$

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

$$\max_k \Pr(z_k|\vec{d}_j) = \Pr(z_{\hat{k}}|\vec{d}_j) \Rightarrow d_j \in z_{\hat{k}} \qquad (2.9)$$

If we can obtain the K representative documents prior to classification, they can be used for $\vec{d}_k^*$ in eq. (2.8).

□

B. Clustering Algorithm [10]

S : The number of clusters ($c_1, c_2, \ldots , c_S$)

(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 maximizes the log-likelihood function $|\mathscr{Z}| = K$ corresponding to the matrix A by the TEM algorithm, where

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

$$\max_k \Pr(z_k | \vec{d}_j) = \Pr(z_{\hat{k}} | \vec{d}_j) \Rightarrow d_j \in z_{\hat{k}} \qquad (2.10)$$

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. (2.2). 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.

3. Performance Evaluation

Document sets

Table 3.1: Document sets

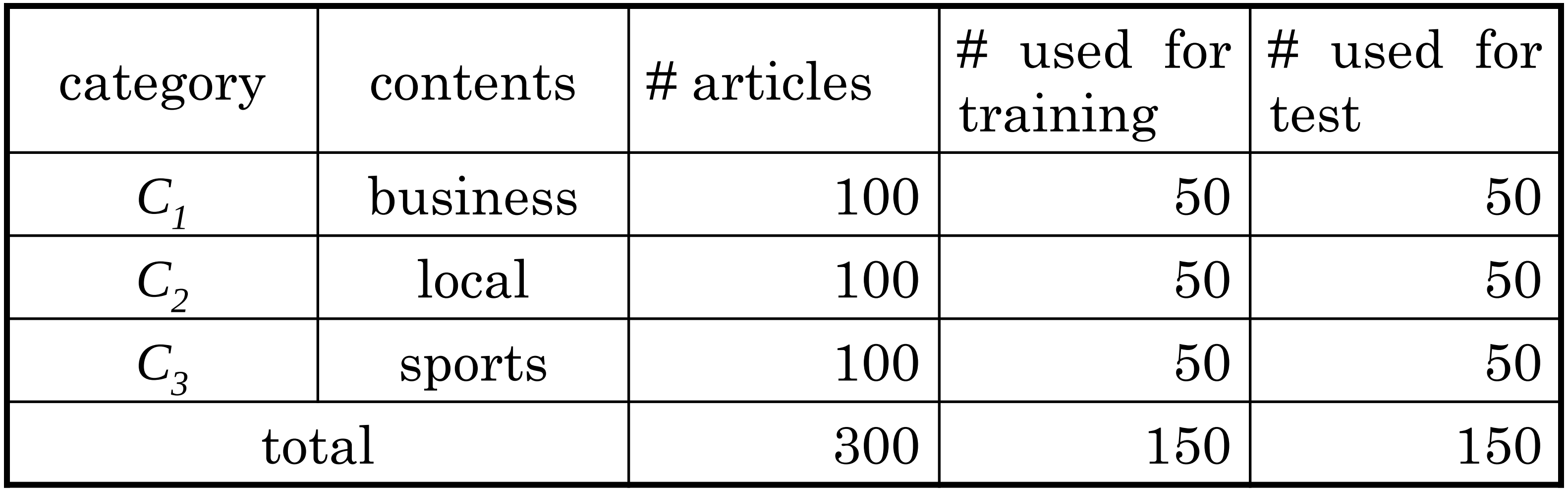
	contents	format	amount	categorize
(a)	articles of Mainichi news paper in ’94 [Sakai99]	Free (texts only)	101,058 (see Table 3.2)	Yes (9+1 ategories)
(b)	Questionnaire (see Table 3.6 in detail)	fixed and free (see Table 3.9)	135+35	Yes (2 categories)
(c)			135	no

3.1 Classification

Conditions of (a)

- Experimental data: Mainichi Newspaper in ‘94 (in Japanese) 300 article, 3 categories (free format only)

Table 3.2: Selected categories of newspaper

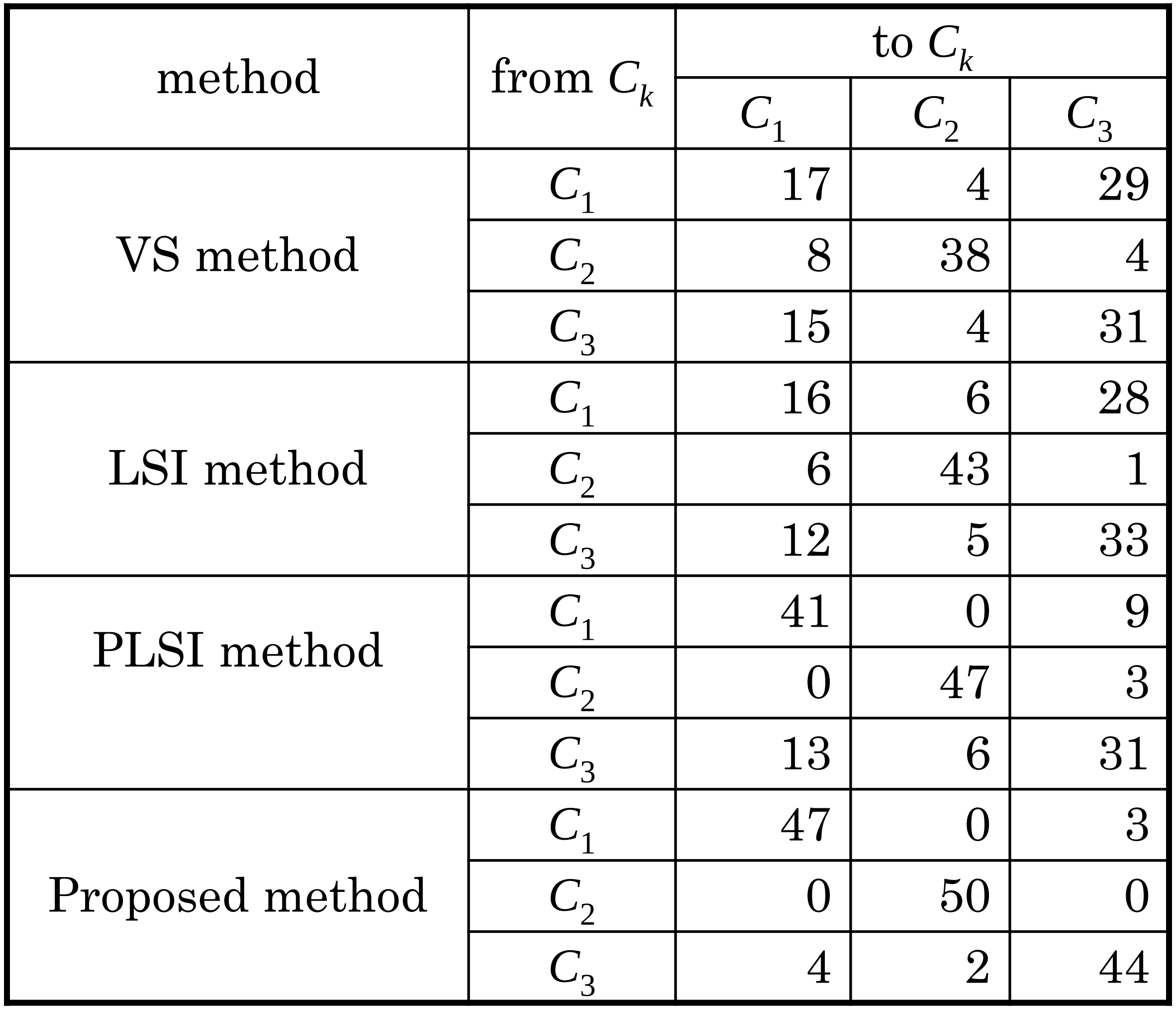


category	contents	# articles	# used for training	# used for test
C_1	business	100	50	50
C_2	local	100	50	50
C_3	sports	100	50	50
total		300	150	150

- LSI : K = 81
  PLSI: K = 10

Results of (a)

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%

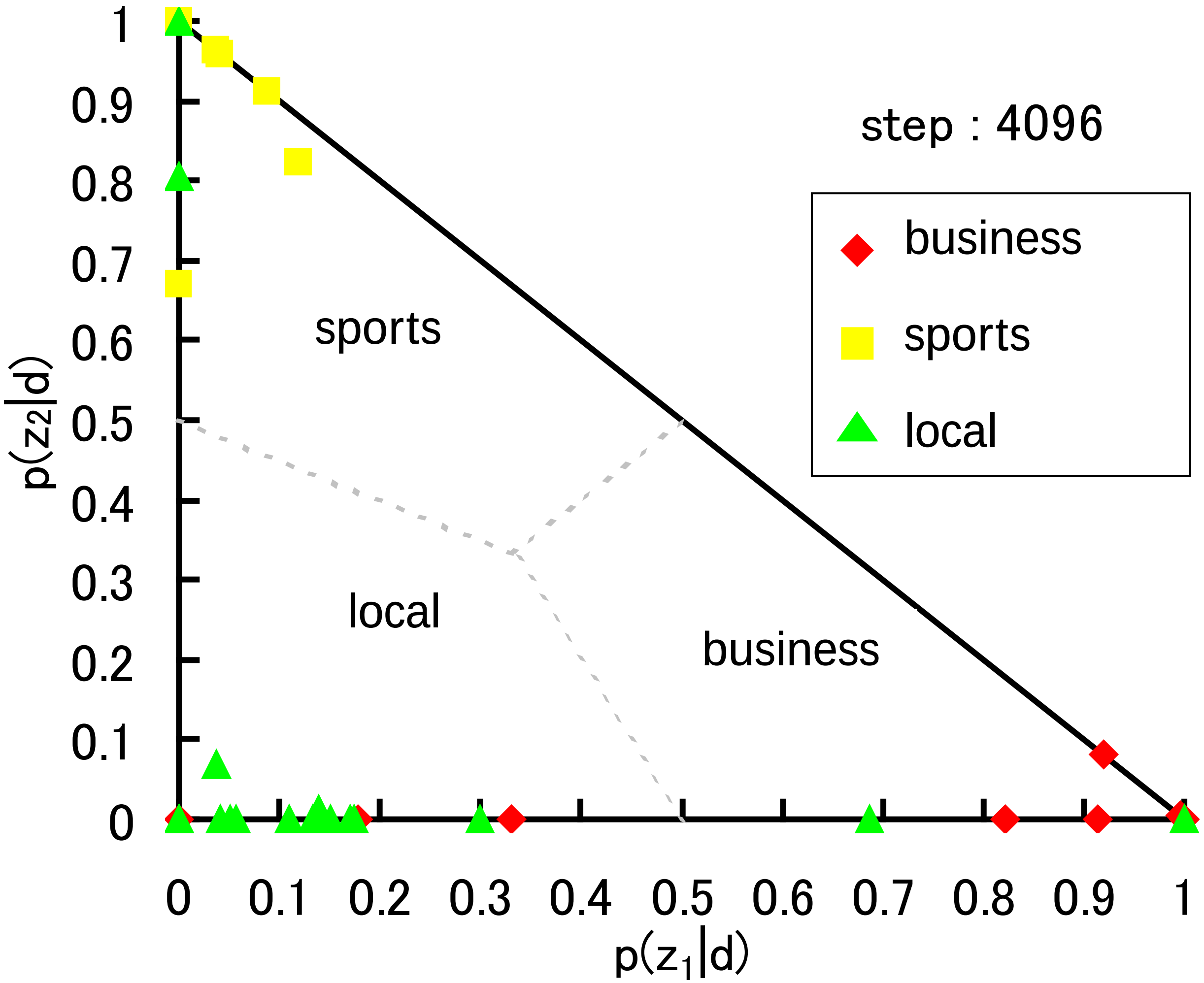


Fig. 3.1: Clustering process by EM algorithm

3.2 Clustering

Student Questionnaire

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?

[2] Each question has 4-21 minor questions.

[3] Each text is written within 250-300 Chinese and Japanese characters.

[4] There is a possibility to improve the performance of the proposed method by elimination of these items.

Object classes

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

Condition of (b)

I) First, the documents of the students in Class CS and those in Class IS are merged.

II) Then, the merged documents are divided into two class (S=2) by the proposed method.

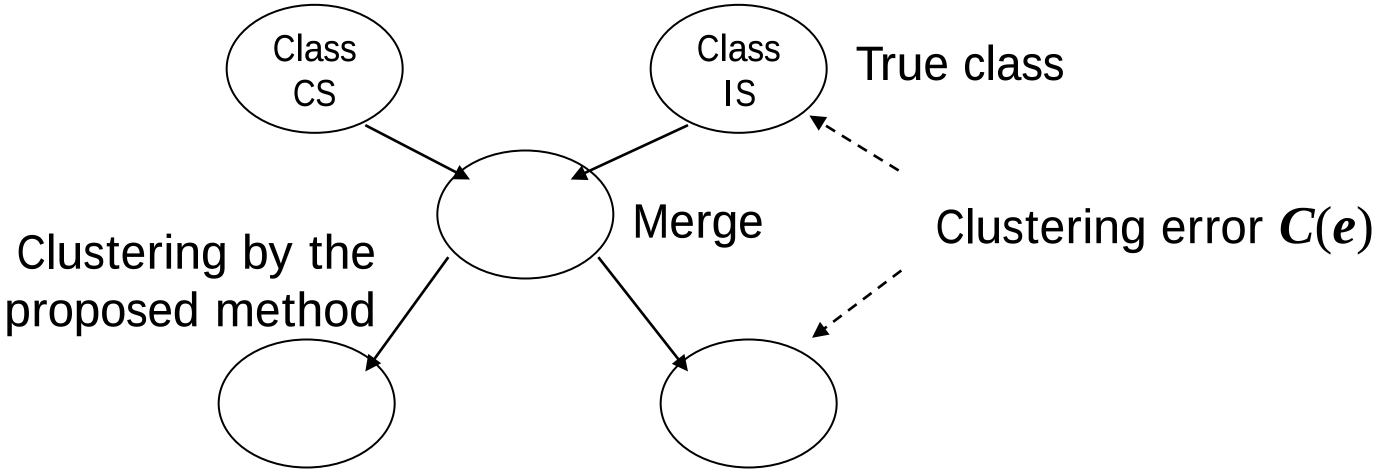


Fig.3.2 Class partition problem by clustering method

Results of (b)

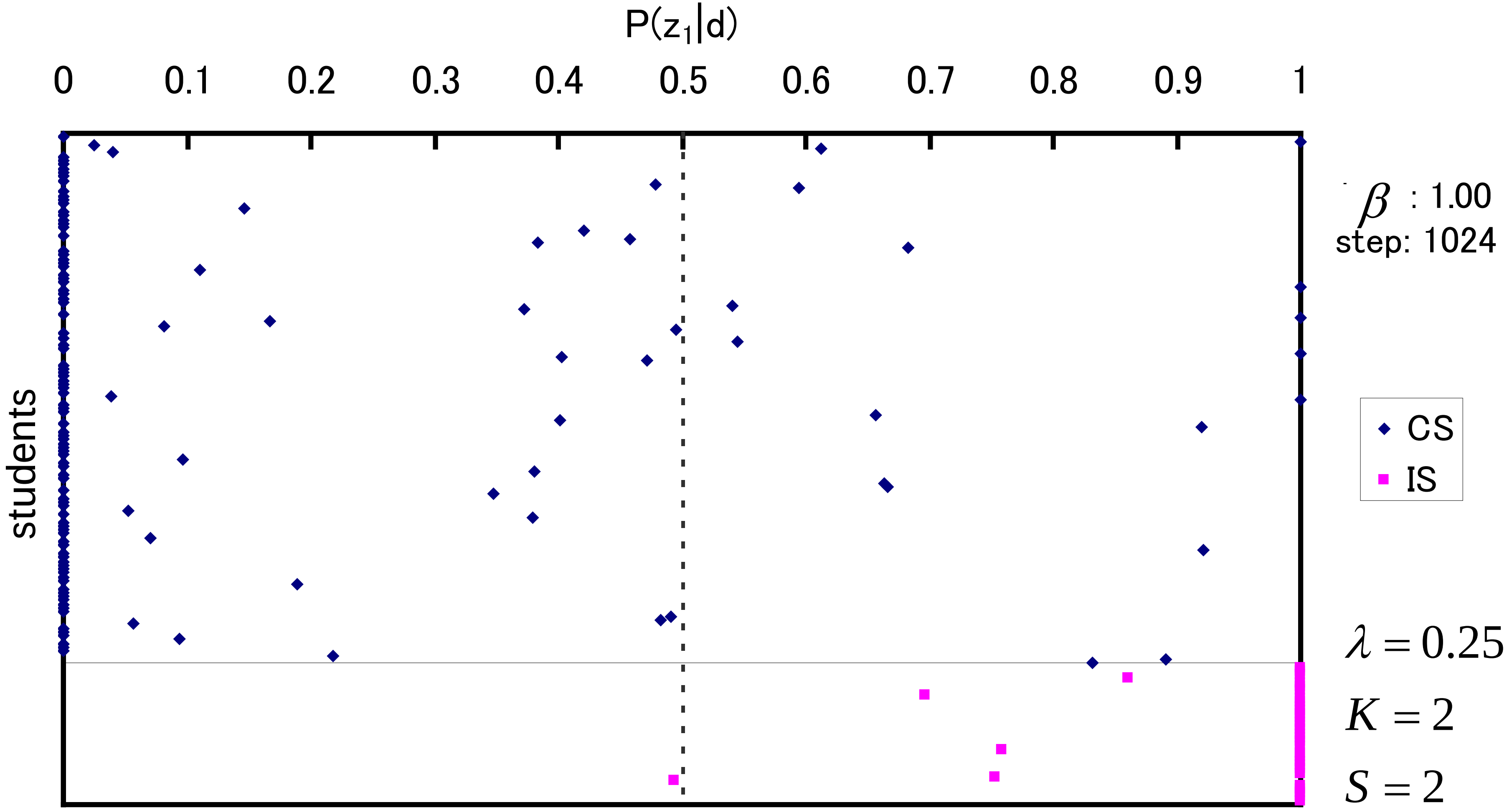


Fig.3.3: Clustering process by EM algorithm, K=2

similarity

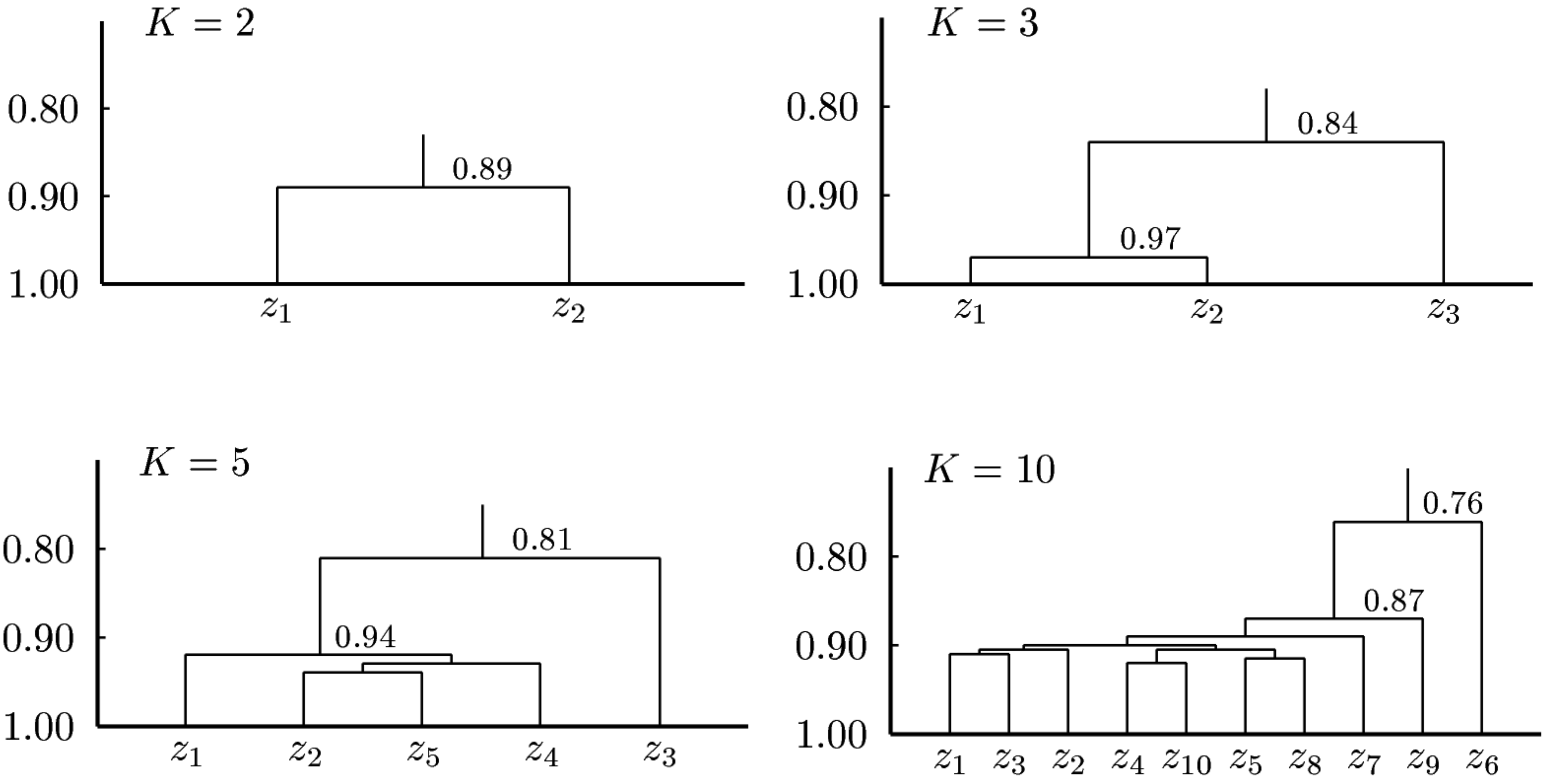
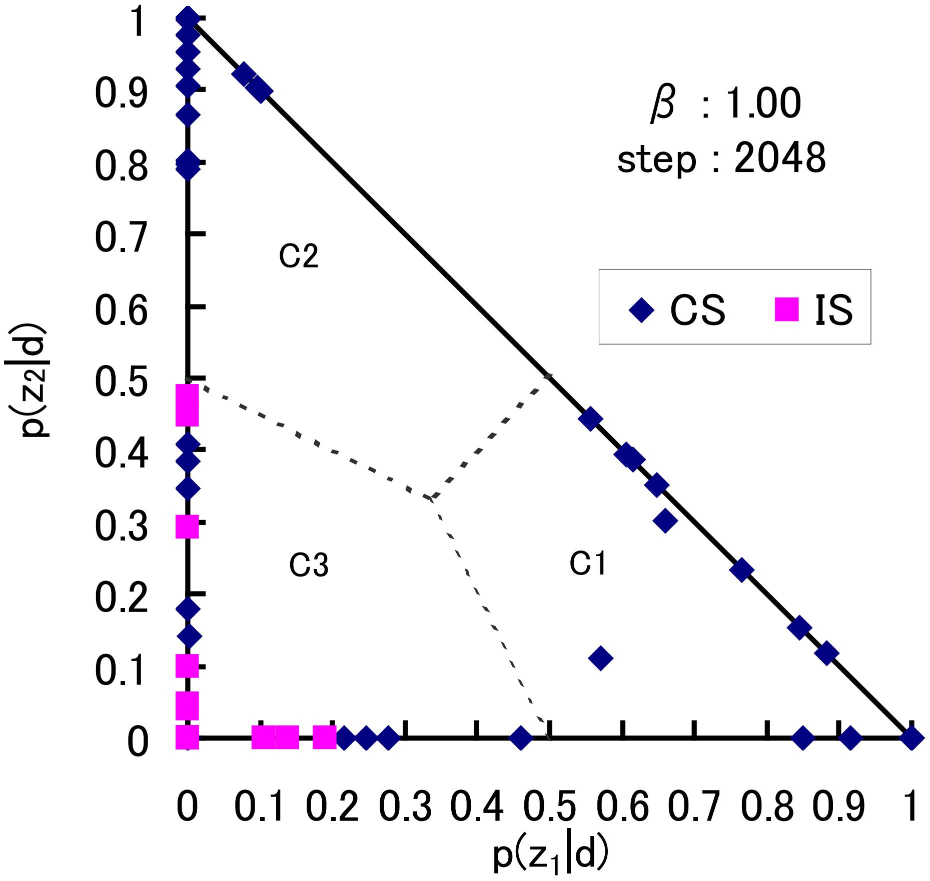


Fig. 3.4: Dendrogram of clusters



$\lambda = 0.25$

$K = 3$

$S = 2$

Fig.3.5 Clustering process for EM algorithm, K=3

K-means method

$S=K=2$ $C(e)=0.411$

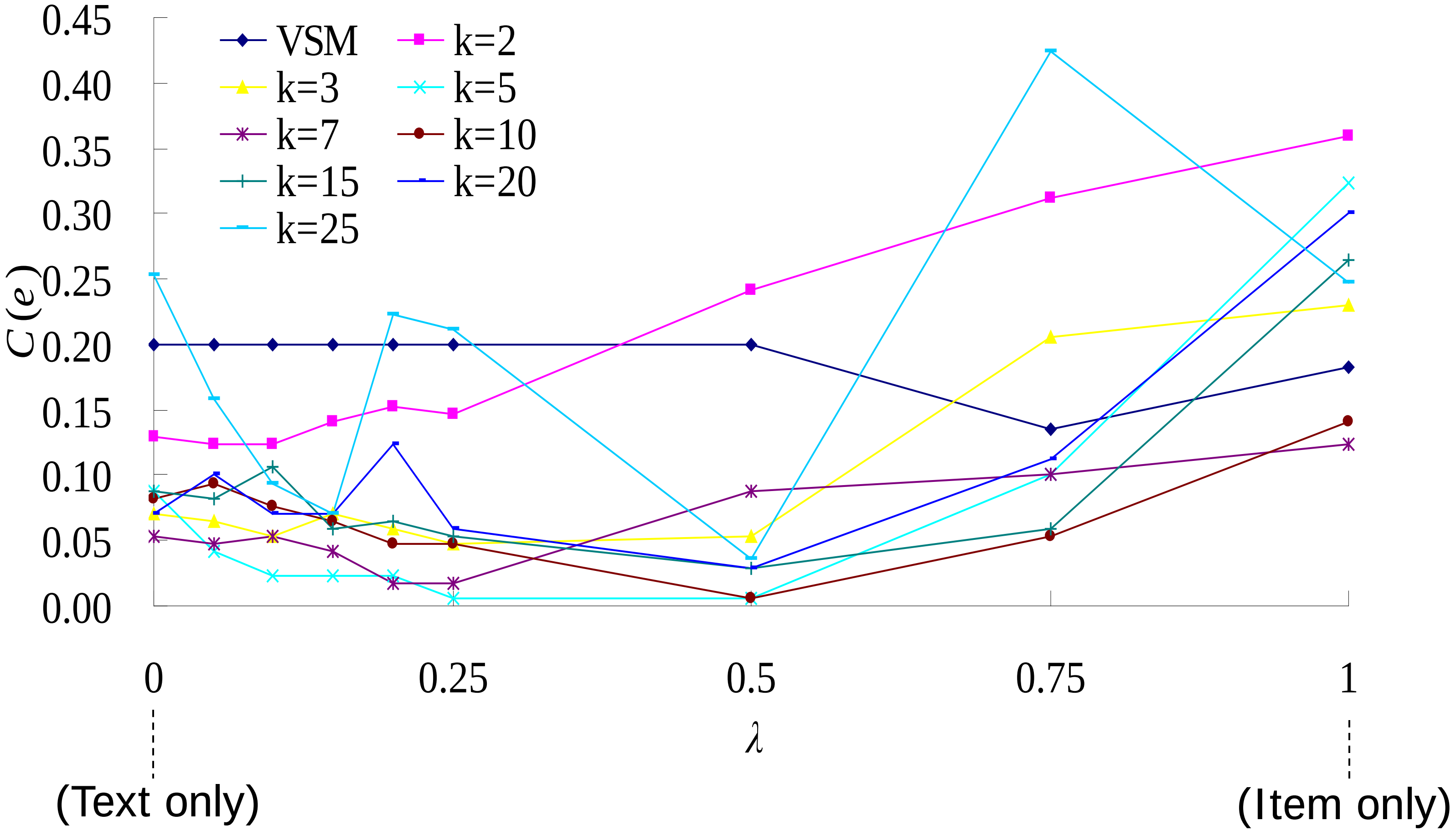


Fig. 3.6: Clustering error rate ***C*(*e*)** vs. *λ*

***C*(*e*)** : the ratio of the number of students in the difference set between divided two classes and the original classes to the number of the total students.

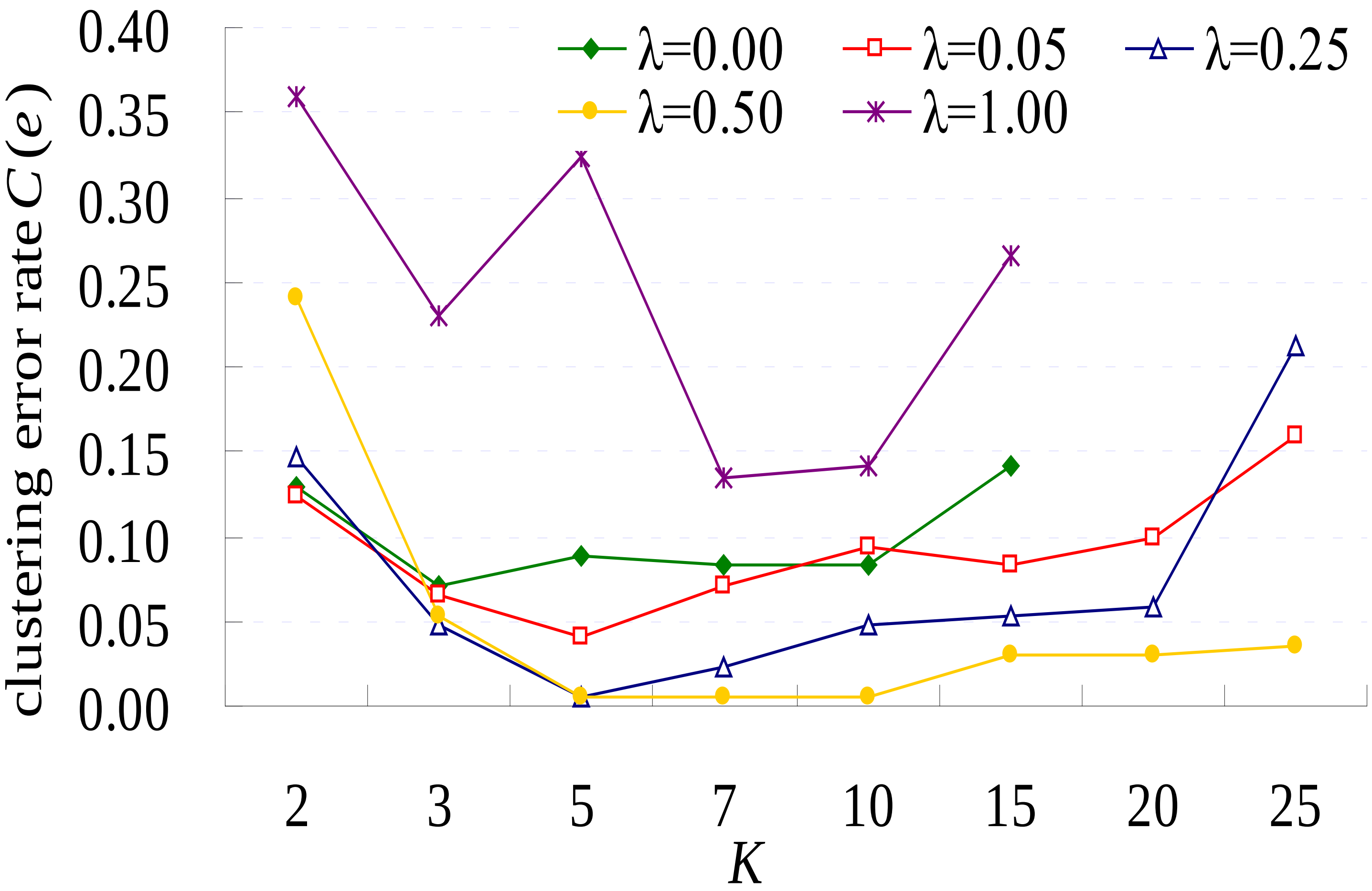


Fig. 3.7: Clustering error rate ***C*(*e*)** vs. λ

Results of (b)

Statistical analysis by discriminant analysis

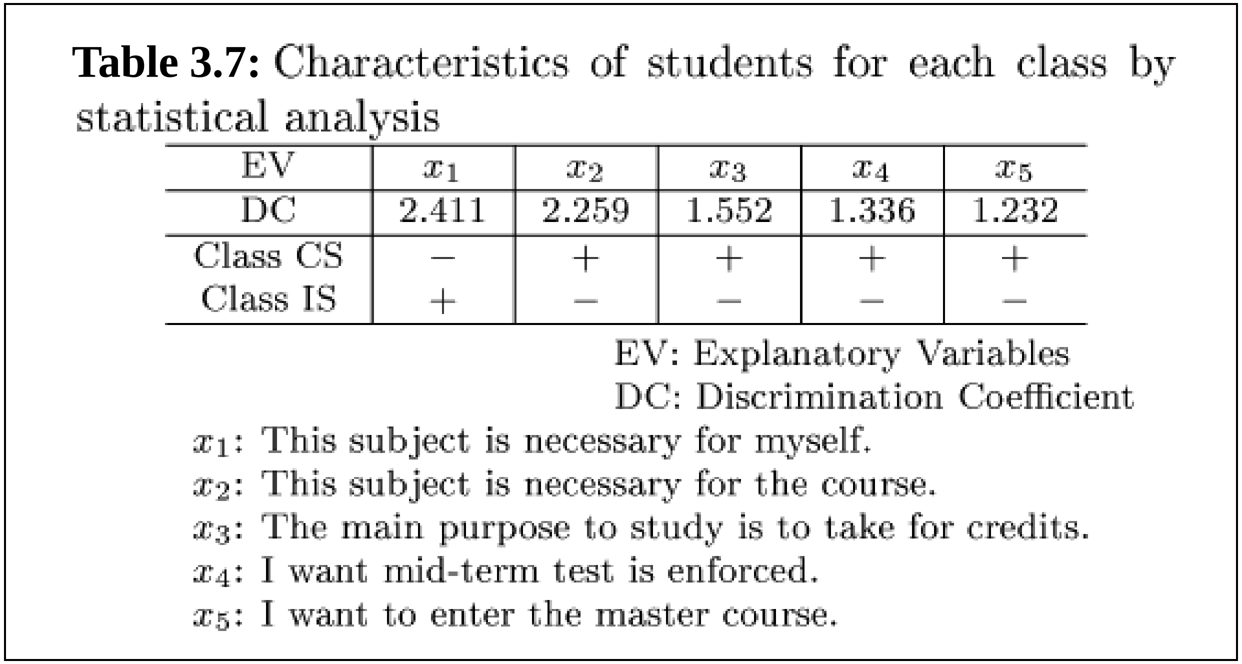


Table 3.7: Characteristics of students for each class by statistical analysis

EV	x_1	x_2	x_3	x_4	x_5
DC	2.411	2.259	1.552	1.336	1.232
Class CS	−	+	+	+	+
Class IS	+	−	−	−	−

EV: Explanatory Variables
DC: Discrimination Coefficient

x_1: This subject is necessary for myself.
x_2: This subject is necessary for the course.
x_3: The main purpose to study is to take for credits.
x_4: I want mid-term test is enforced.
x_5: I want to enter the master course.

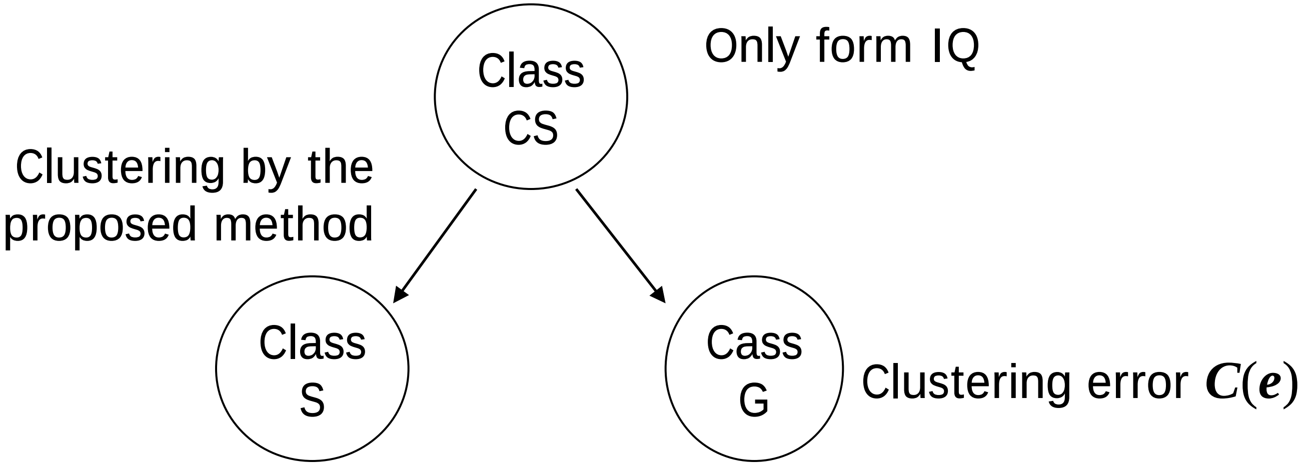
$$z = a_0 + a_1 x_{1j} + a_2 x_{2j} + \cdots + a_5 x_{5j}$$

$$z \geq 0: \quad d_j \in \text{ Class CS}$$

$$z < 0: \quad d_j \in \text{ Class IS}$$

Another Experiment

Clustering for class partition problem



S: Specialist

G: Generalist

Fig. 3.8: Another Class partition problem by clustering method

(1) Member of students in each class

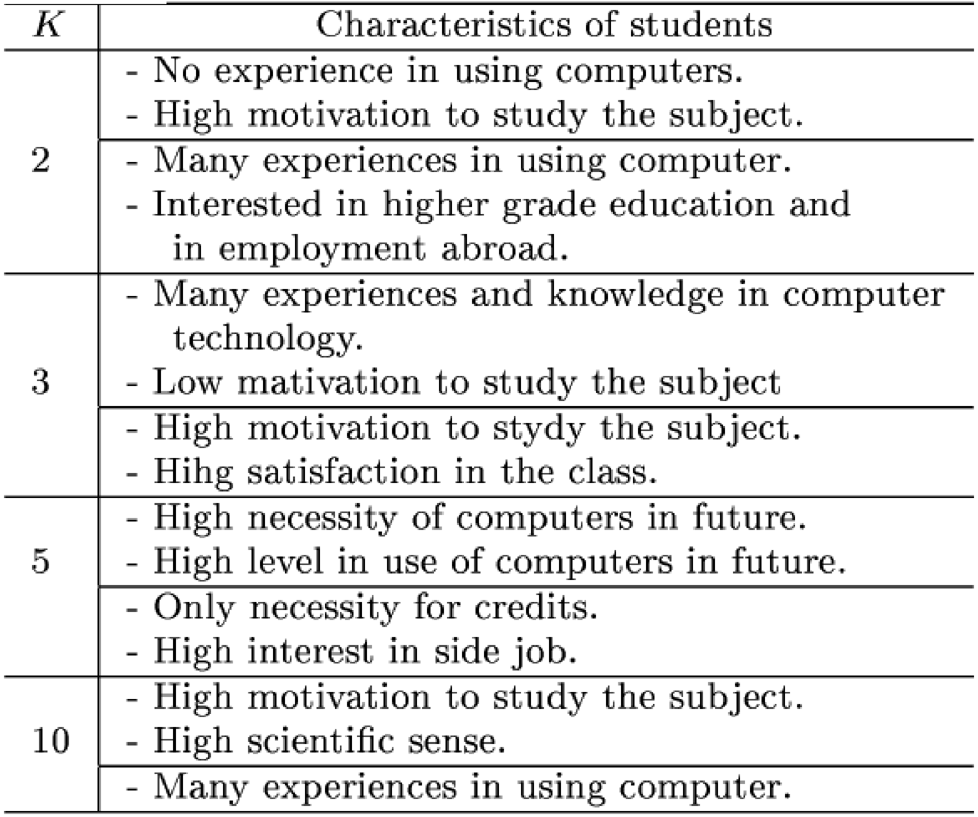
Table 3.8: Difference between SOC and AC

	class	Characteristics of students
student's selection	S	- Having a good knowledge of technical terms - Hoping the evaluation by exam
	G	- Having much interest in use of a computer
Clustering	S	- Having much interest in theory - Having higher motivation for a graduate school
	G	- Having much interest in use of a computer - Having a good knowledge of system using the computer

SOC: Student’s own choice
AC: Automatic clustering

(2) Member of students in each class

Table 3.9: Characteristics of students for each class



K	Characteristics of students
2	- No experience in using computers. - High motivation to study the subject.
	- Many experiences in using computer. - Interested in higher grade education and in employment abroad.
3	- Many experiences and knowledge in computer technology. - Low mativation to study the subject
	- High motivation to stydy the subject. - Hihg satisfaction in the class.
5	- High necessity of computers in future. - High level in use of computers in future.
	- Only necessity for credits. - High interest in side job.
10	- High motivation to study the subject. - High scientific sense.
	- Many experiences in using computer.

By discriminant analysis, two classes are evaluated for each partition which are interpreted in table 5. The most convenient case for characteristics of students should be chosen.

4. Student Questionnaire Analysis

4.1. Design of Student Questionnaire

To find out requirements of the students from the questionnaire by the questionnaire analyses model:

- We show relationships between the degree of satisfaction, scores and the characteristics of the students by a class model.
- We design the questionnaire to verify the hypothesis (the class model).
- 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.

Student Questionnaire

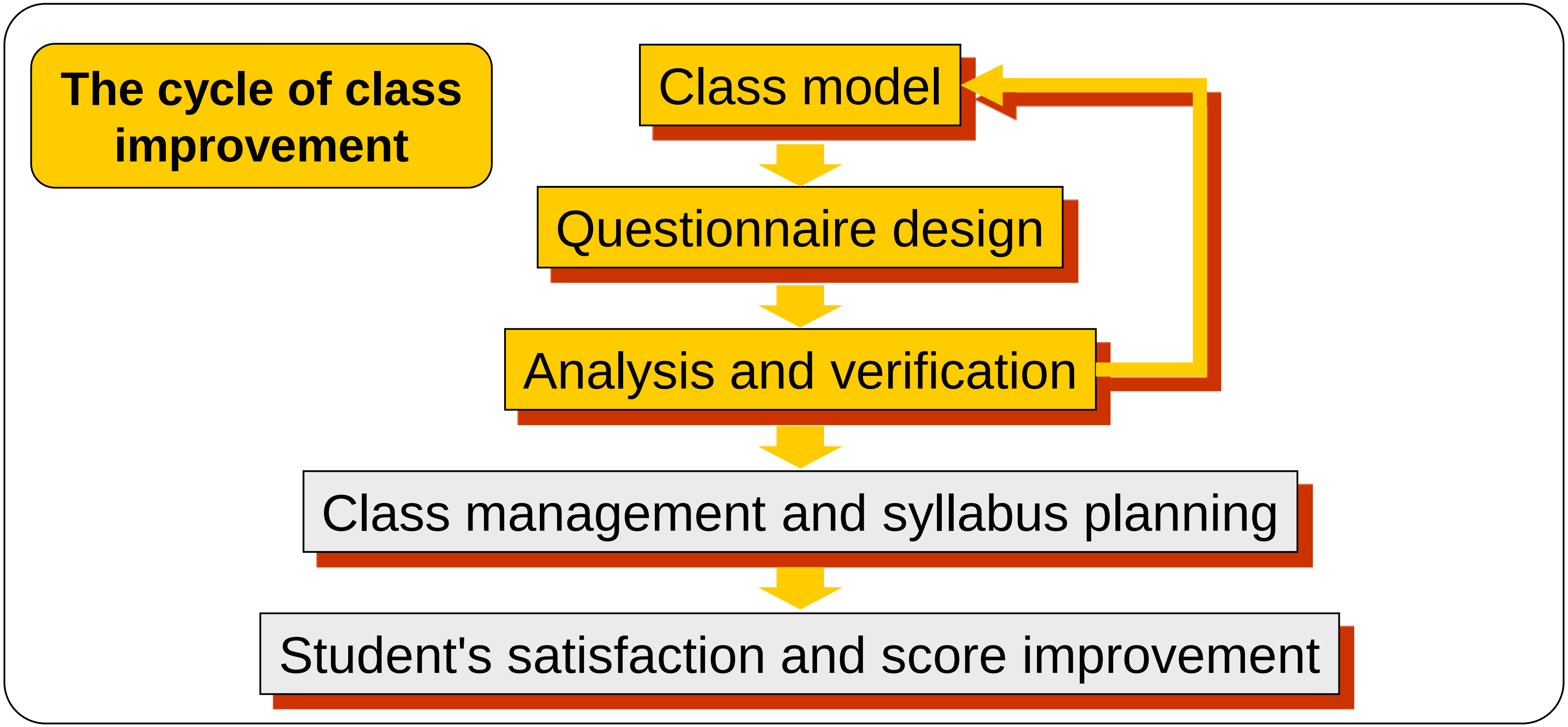

The cycle of class improvement
Class model
Questionnaire design
Analysis and verification
Class management and syllabus planning
Student's satisfaction and score improvement

Fig. 4.1: Faculty Development by Student Questionnaire [10]

Questionnaire

- Fixed format (multiple choice questions: Items)
- Free format (Texts)

A. Class Model

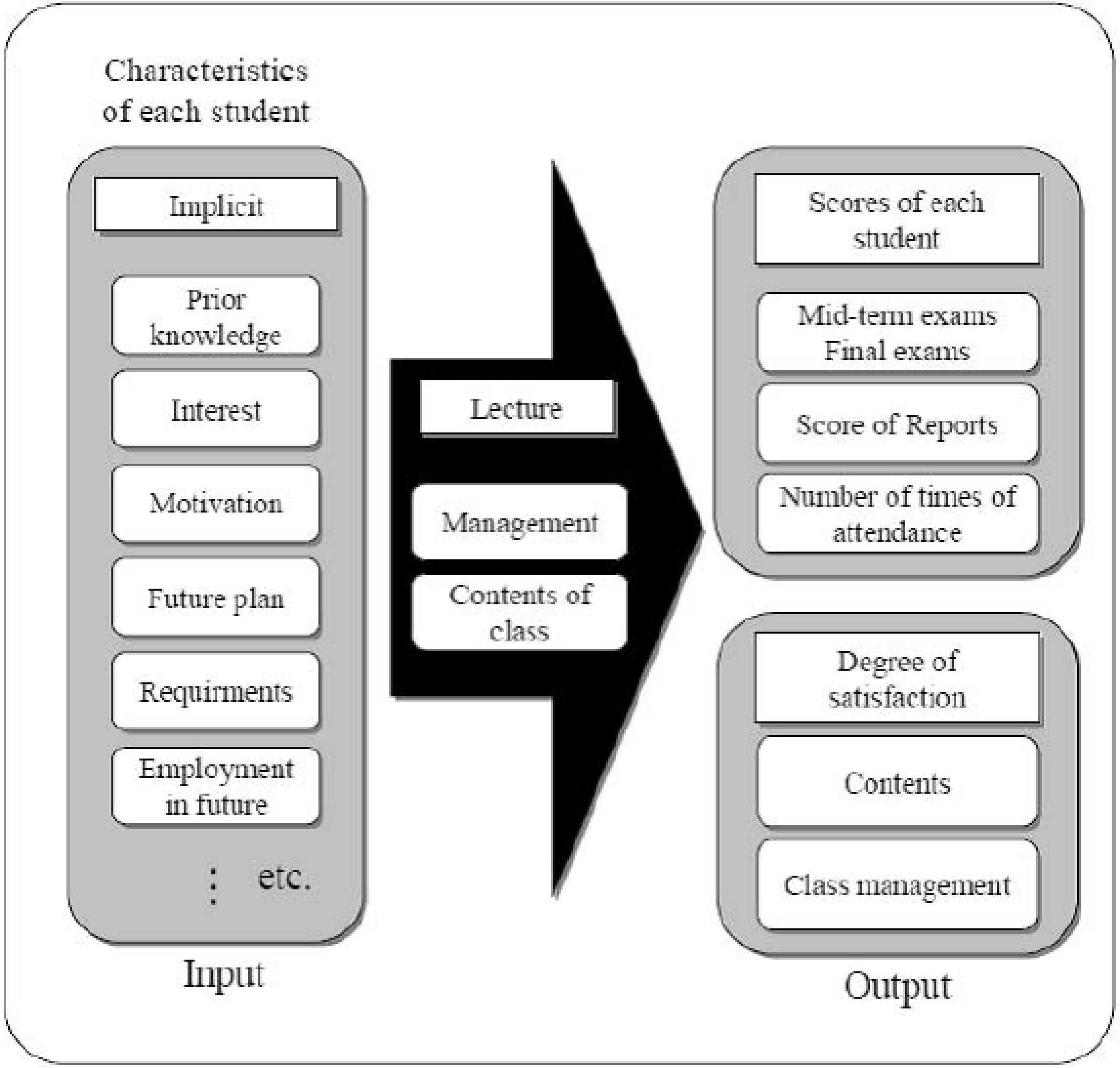


Fig. 4.2: Class model for the class “Introduction to Computer Engineering

B. Design of Questionnaire

4. Student Questionnaire Analysis

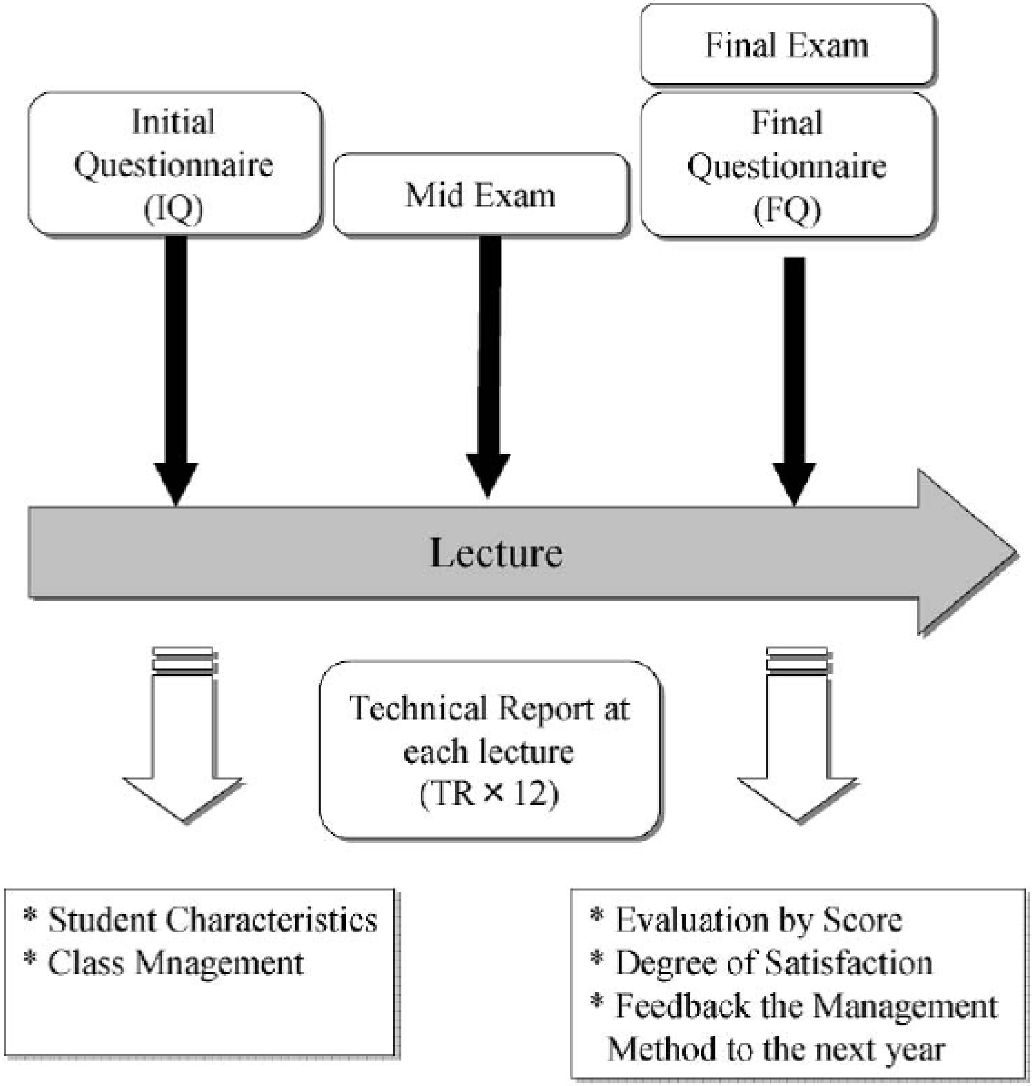

Final Exam
Initial Questionnaire (IQ)
Mid Exam
Final Questionnaire (FQ)
Lecture
Technical Report at each lecture (TR × 12)
* Student Characteristics
* Class Mnagement
* Evaluation by Score
* Degree of Satisfaction
* Feedback the Management Method to the next year

Fig. 4.3: Time schedule for class

B. Design of Questionnaire

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 each）
Midterm Test（MT）	5 subjects
Technical Reports （TR）	11 times（each 1-2 subjects）
Final Test （FT）	5 questions
Final Questionnaire （FQ）	
Item type	6 questions （6-21 sub-questions each）
Text type	5 questions （250-300 characters in Japanese each）

B. Design of Questionnaire

Table 4.2 (a) : Contents of a questionnaire (IQ)

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?

B. Design of Questionnaire

Table 4.2 (b) : Contents of a questionnaire (FQ)

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?

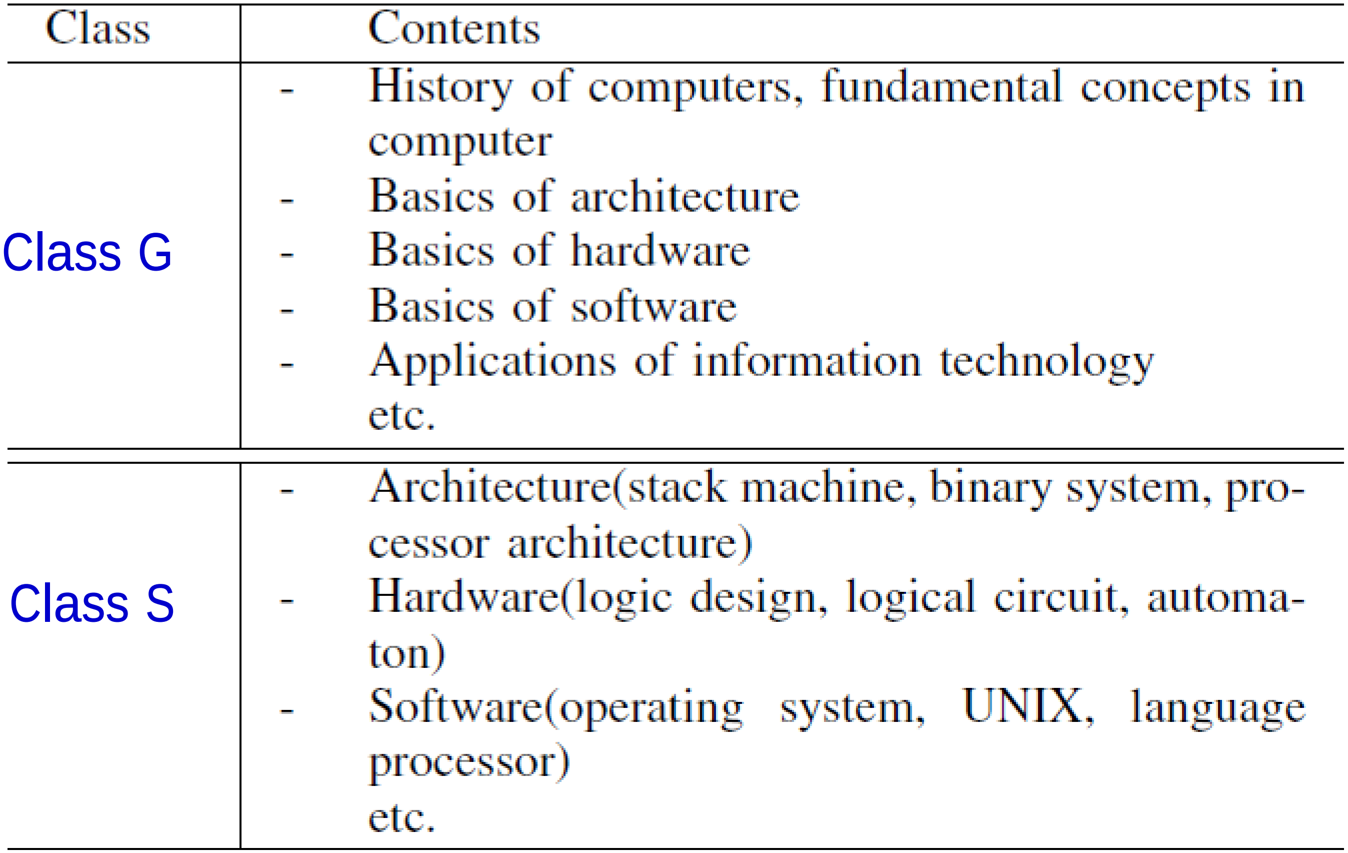
This questionnaire is made in WEB form, and it is on the following Web Site.
http : //www.hirasa.mgmt.waseda.ac.jp/users/comp-eng/

4.2 Verification of class model by IQ

Class G (generalist): wide and shallow technical topics

Class S (specialist): technical and professional topics

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.

4. Student Questionnaire Analysis

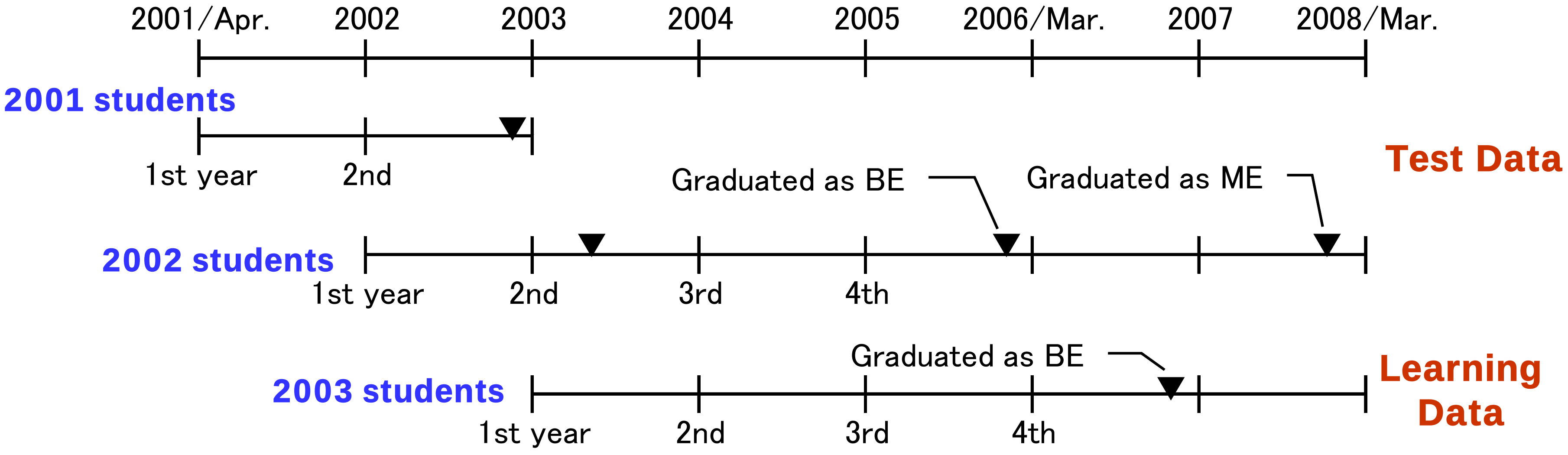
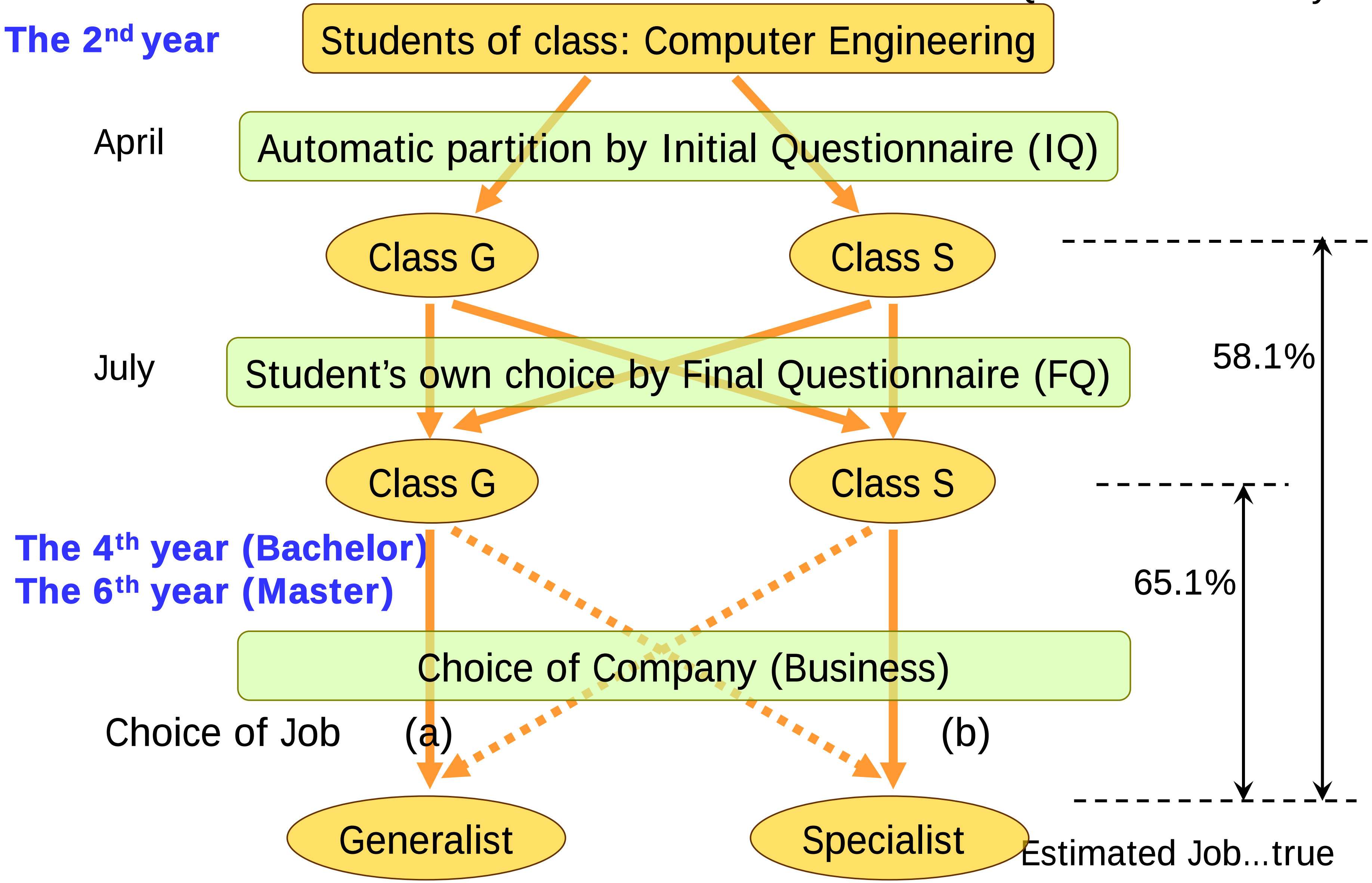


Fig. 4.4: Collected data



The 2nd year

Students of class: Computer Engineering

April

Automatic partition by Initial Questionnaire (IQ)

Class G

Class S

July

Student’s own choice by Final Questionnaire (FQ)

Class G

Class S

The 4th year (Bachelor)

The 6th year (Master)

Choice of Company (Business)

Choice of Job

(a)

(b)

Generalist

Specialist

58.1%

65.1%

Estimated Job...true

Fig. 4.5: Transition of students

"Job" : 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,

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.

"Business" : as the kind of ***company*** such as:

(a): trading, finance, banking, service, securities market, consultation, general construction, and so on,

(b): electric manufacturing, automobile manufacturing, precision instrument manufacturing, system integration, software development, and so on.

Estimation of the job

We know only the name of companies in which they joined, such as:

Canon Inc., IBM Japan Ltd., NEC, Toyota Motor Corp., Accenture, Nomura Research Institute Ltd., East Japan Railway Co., Kashima Corp., Sony Corp., Tokyo Mitsubishi UFJ Bank, and so on.

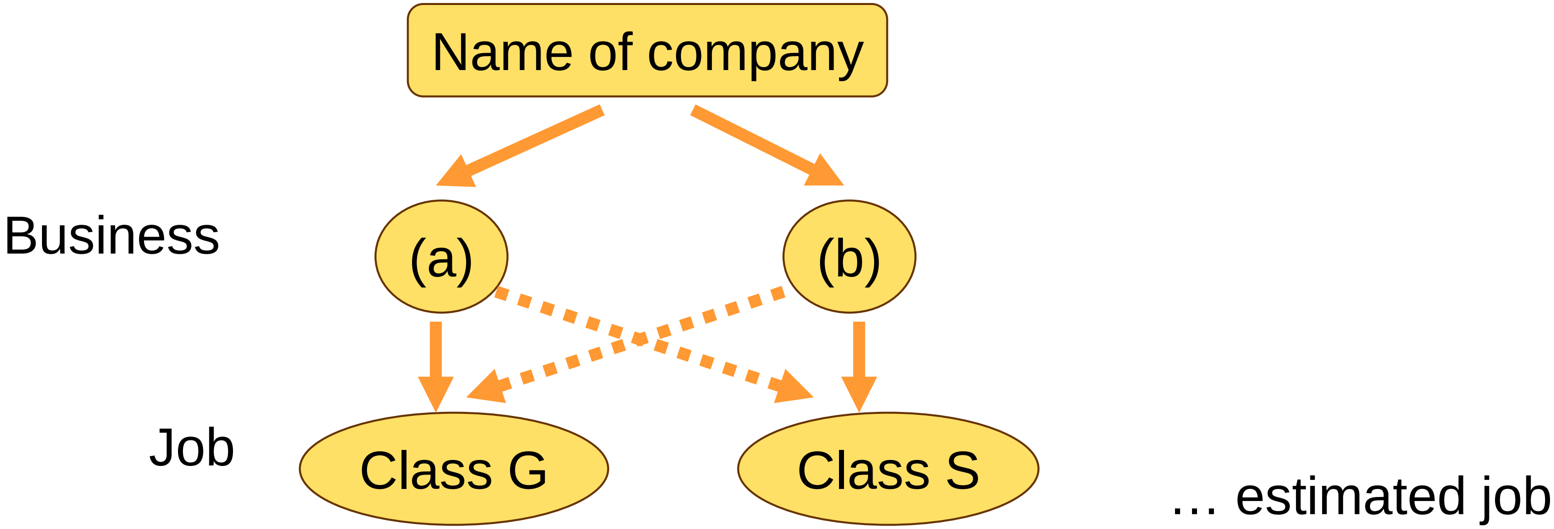


Fig. 4.6: Transition of students

Results of partition

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

AP: Automatic Partition
SEC: Students Estimated Choice

58.1%

Table 4.4: Numbers of partitioned students between AP and SEC

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

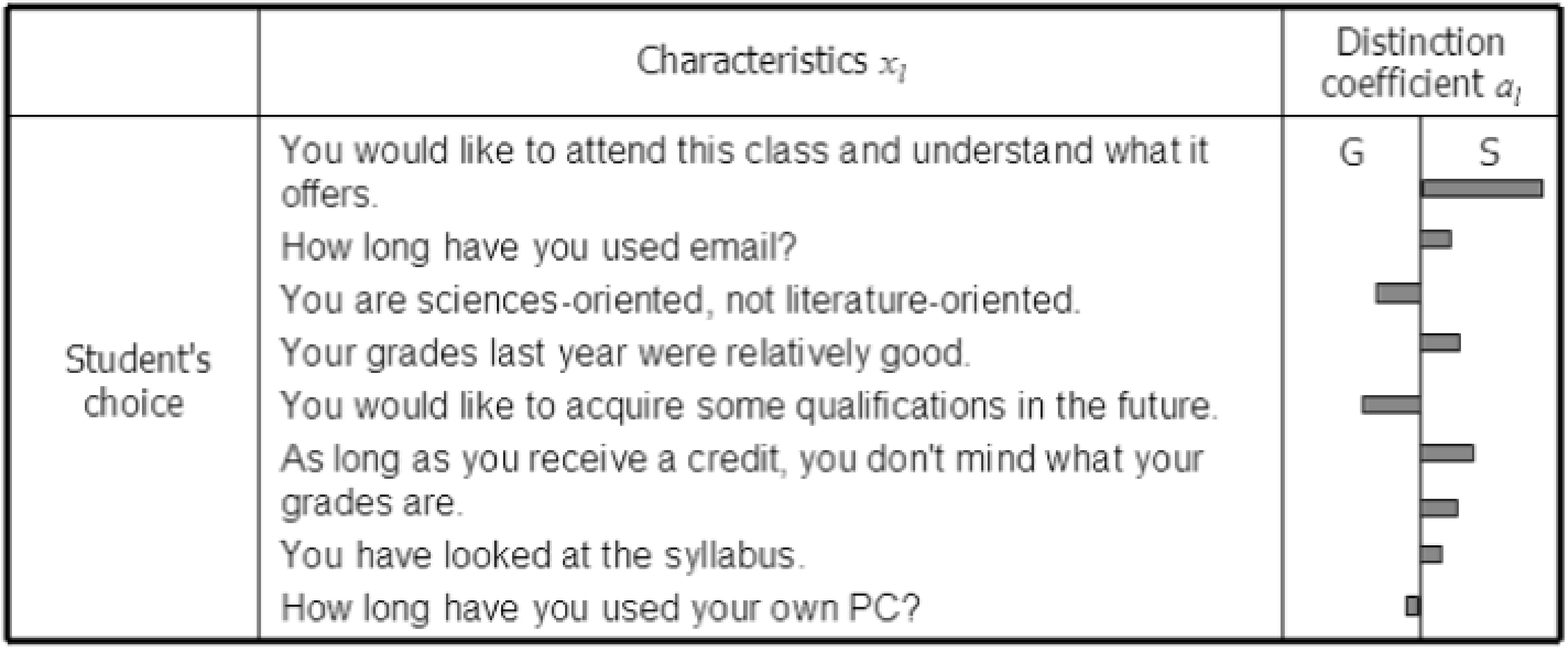
SOC: Sutudent's Own Choice

65.1%

Table 4.5: Numbers of partitioned students between SOC and SEC

Table 4.6(a) : Characteristics of Class G and Class S (by discriminant analysis)

(i) Students in Japan (Student's choice)

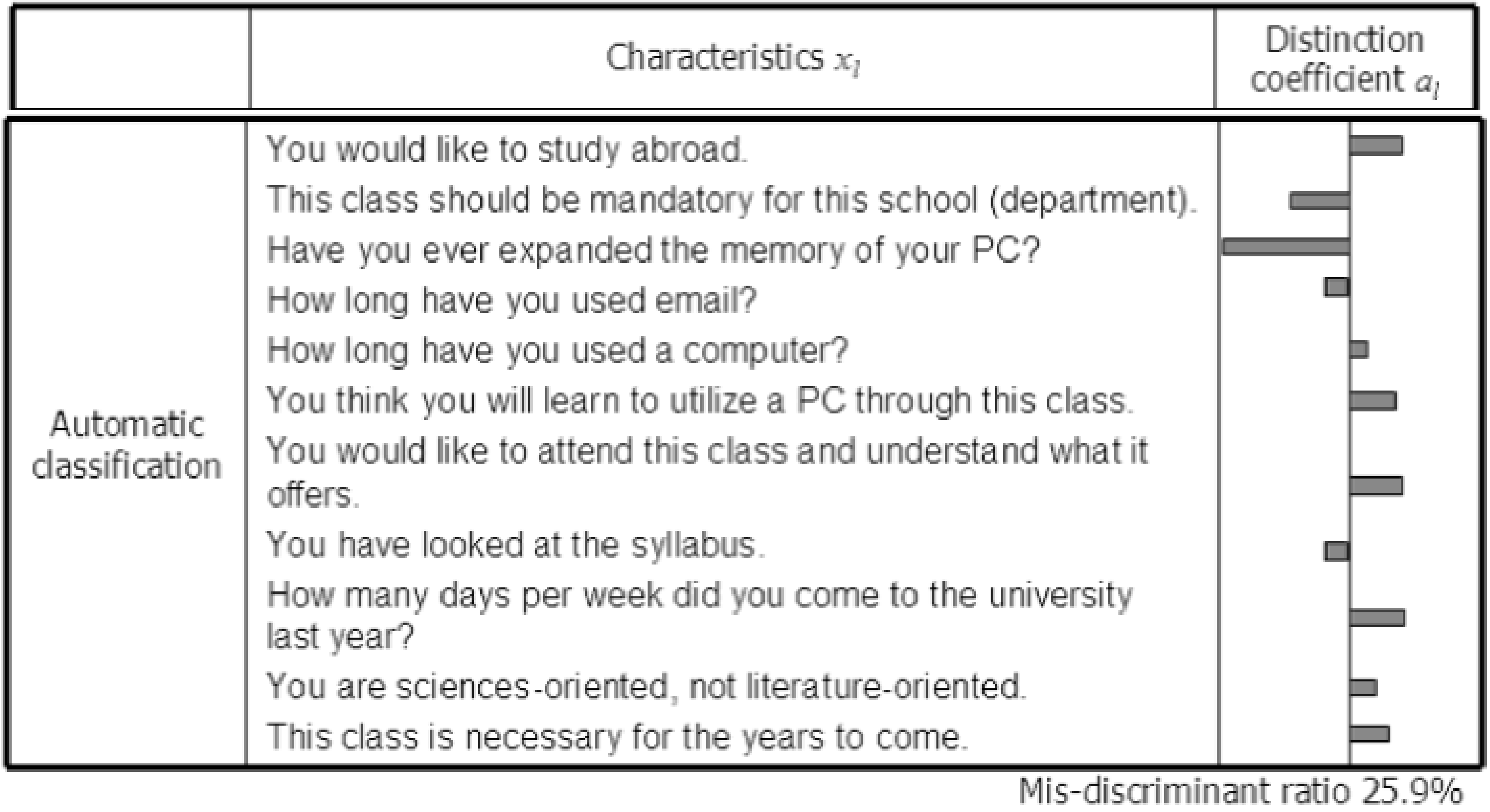


	Characteristics x_l	Distinction coefficient a_l	
		G	S
Student's choice	You would like to attend this class and understand what it offers.		
	How long have you used email?		
	You are sciences-oriented, not literature-oriented.		
	Your grades last year were relatively good.		
	You would like to acquire 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%

Table 4.6(b) : Characteristics of Class G and Class S (by discriminant analysis)

(i) Students in Japan (Automatic classfication)

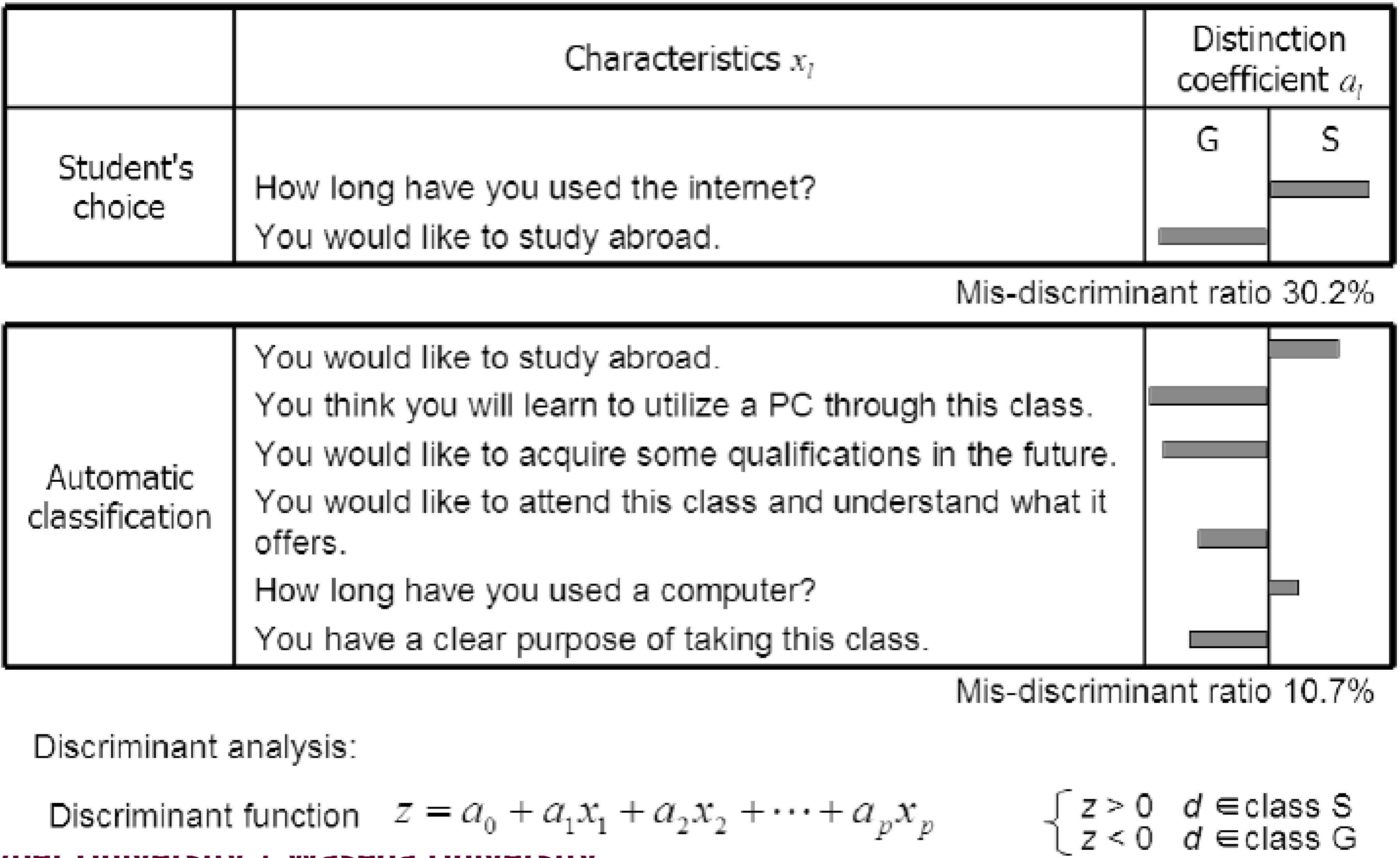


	Characteristics x_l	Distinction coefficient a_l
Automatic classification	You would like to study abroad.	
	This class should be mandatory for this school (department).	
	Have you ever expanded the memory of your PC?	
	How long have you used email?	
	How long have you used a computer?	
	You think you 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%

Table 4.7: Characteristics of Class G and Class S (by discriminant analysis)

(ii) Student’s in R.O.C.



	Characteristics x_l	Distinction coefficient a_l	
		G	S
Student's choice	How long have you used the internet?		▬
	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.	▬	

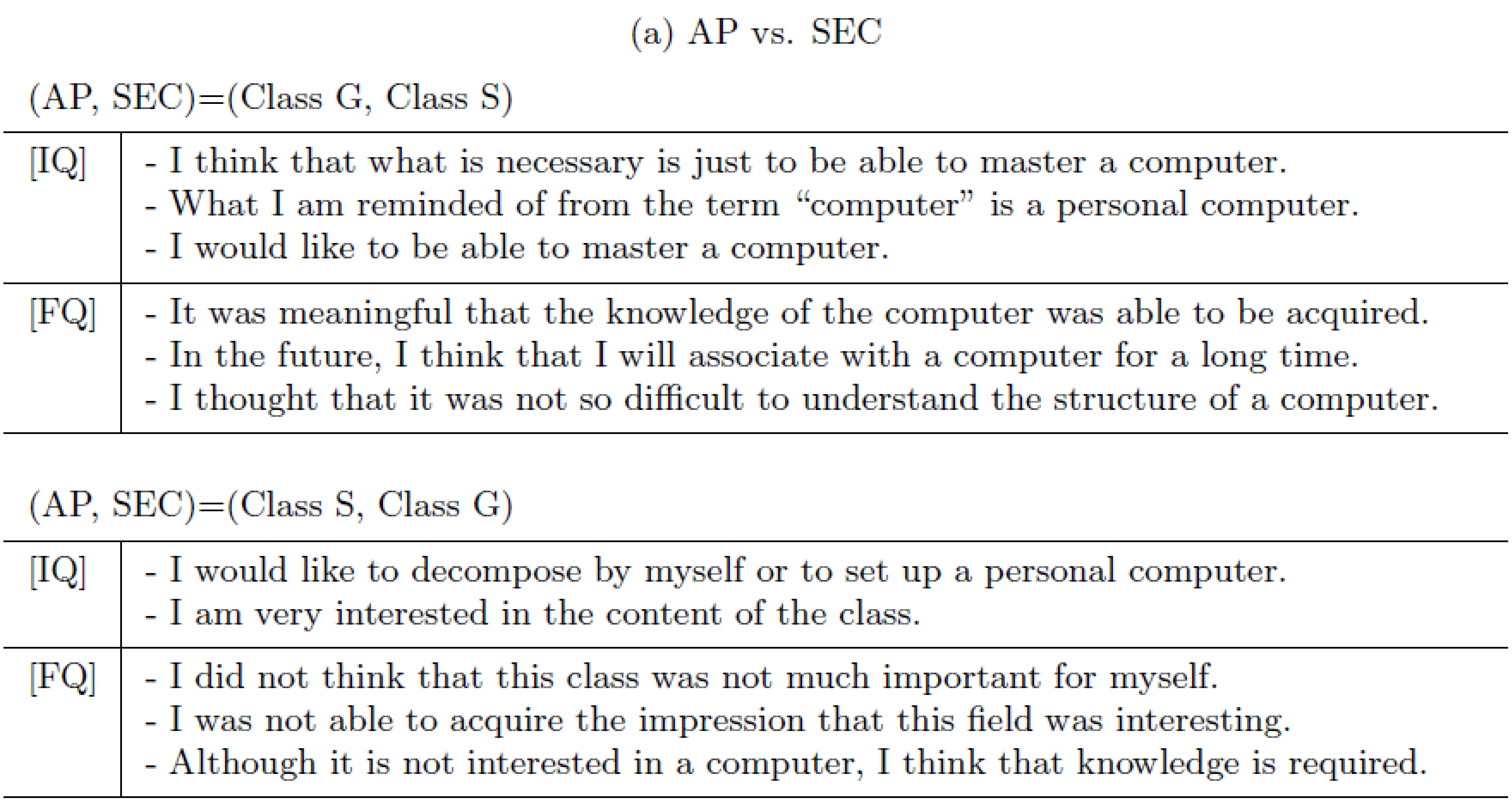
Mis-discriminant ratio 10.7%

Discriminant analysis:

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

Results of extracted important sentences

Table 4.8 : Extracted important sentences



(a) AP vs. SEC

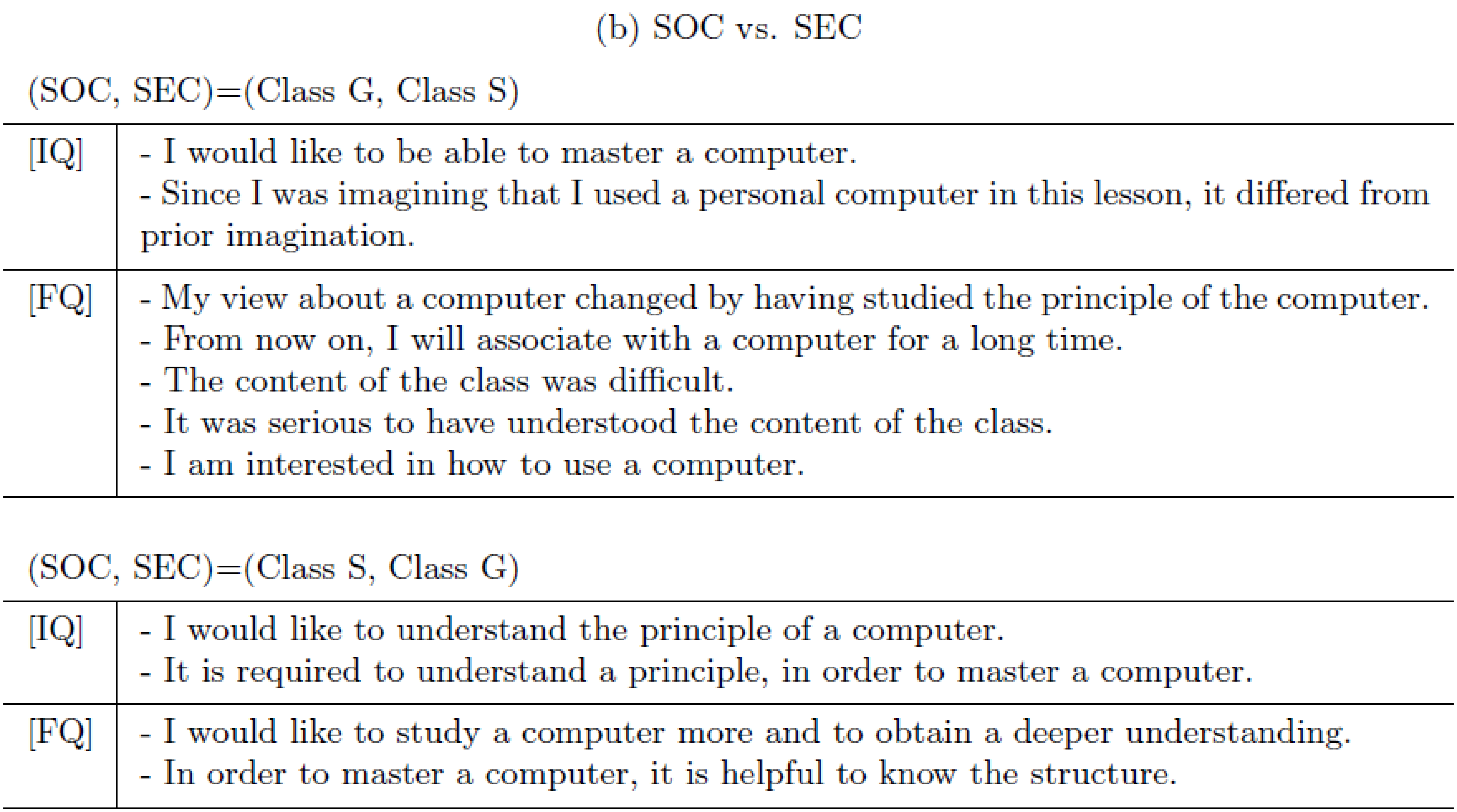
(AP, SEC)=(Class G, Class S)

[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.

Table 4.9 : Extracted important sentences



(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.

Discussion

(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 4.2 suggests us that the students at the 2nd academic year do not decide their future jobs. Hence they do not awake whether professional skill is required or not in their future.

(4) From the view-point of the hypothesis testing, under the hypothesis H_0: Two variables are independent, H_0 for Table 4.1 cannot be rejected, while H_0 for Table 4.5 can be rejected (See Appendix A).

Discussion

(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 who interested in computers, would chose Class S (Table 4.8 (a)).

(ii) There are many cases such that if the student wanted to learn only the method for using computers, he who graduated as a Master, will join an industry as a specialist (Table 4.8 (a)).

(iii) If the student who wanted to be a specialist, could not be interested in computers, he will become a generalist (Table 4.8 (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 4.8 (a)).

(v) If the student who chose Class G, changed his idea by learning the principle of computers, he becomes a specialist (Table 4.8 (b)).

Discussion

(vi) Even if the student felt that the lecture was difficult, he will become a specialist (Table 4.8 (b)).

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

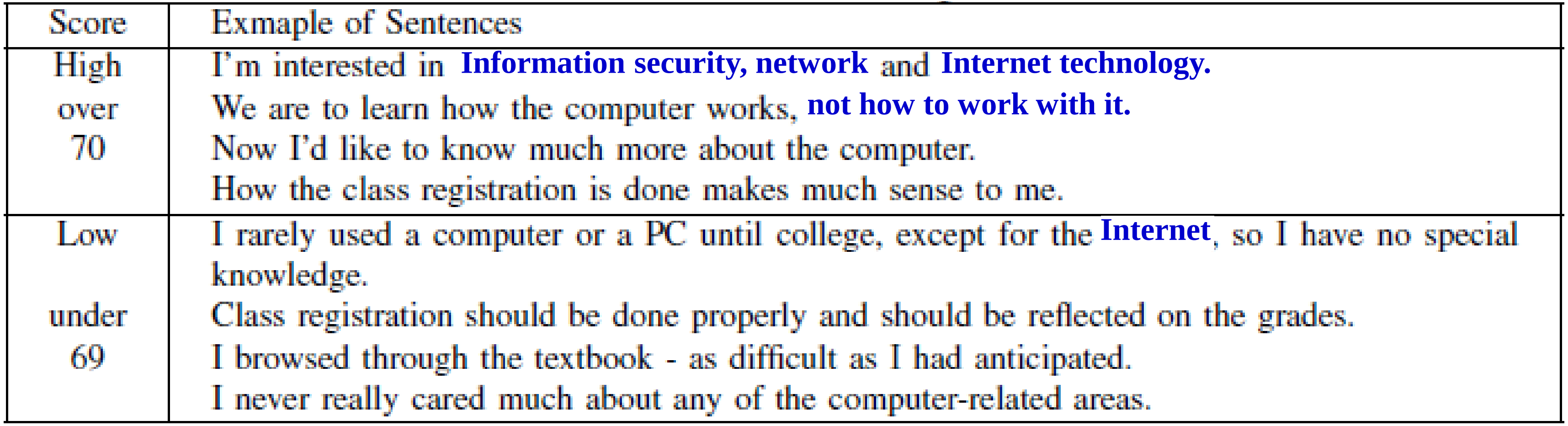
(6) Most of all students state that they will satisfy fruitful and interested contents of the lecture, and their choice of the Class S or Class G depends on the topics. Therefore, the contents of topics are very important.

4.3 Verification of class model by IQ and FQ

(1) Scores of students

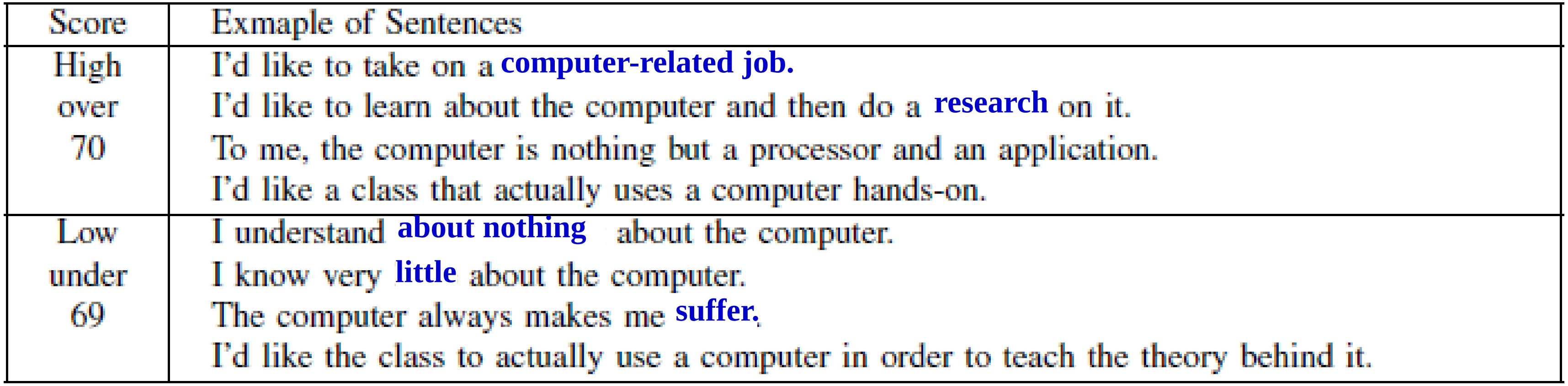
Table 4.10: Sentences extracted from text-type questionnaire for scores of students

(i) Students in Japan



Score	Exmaple 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 college, 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	Exmaple 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 about 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.

Discussions

From Table 4.10:

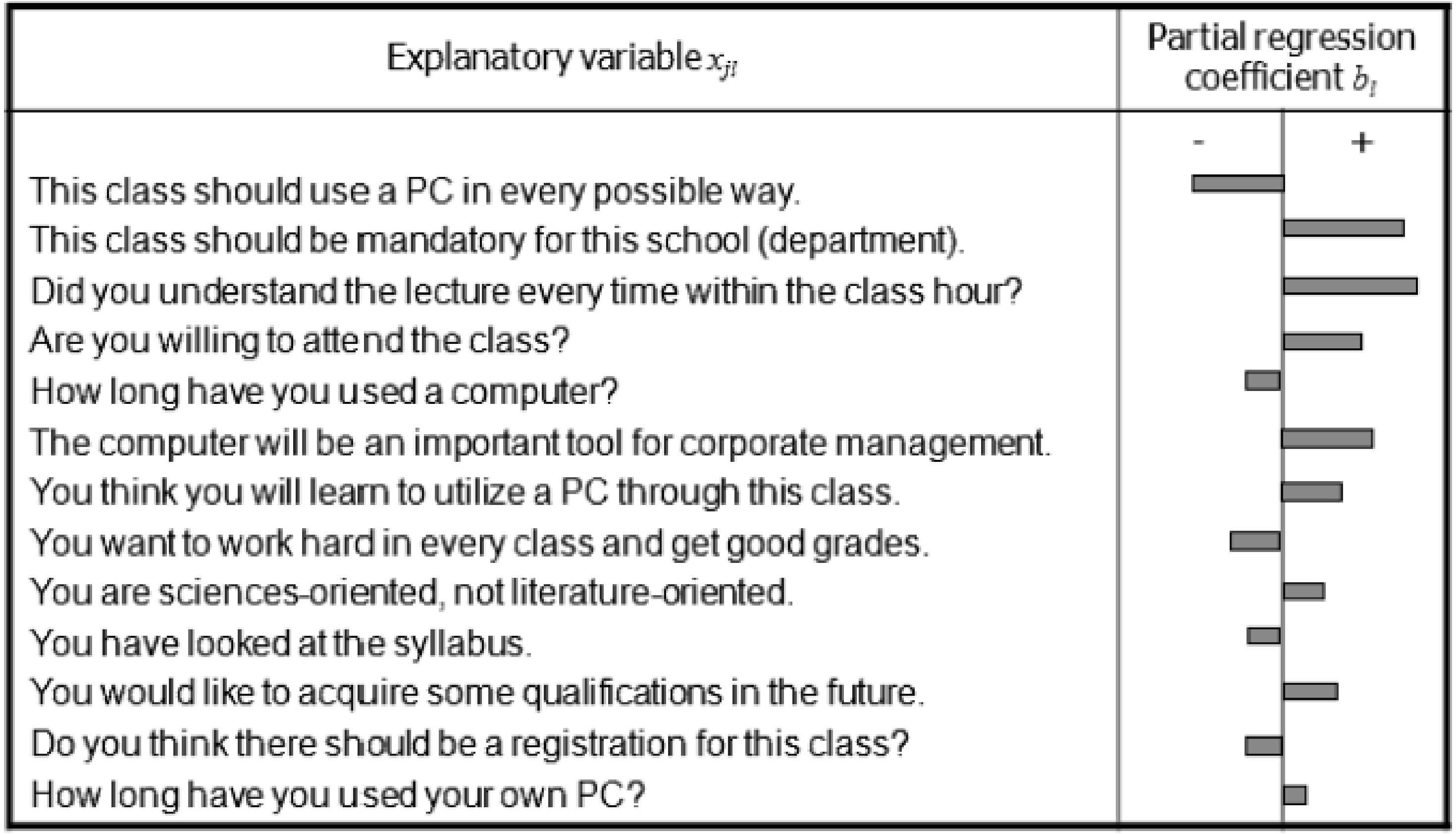
- Students in higher level both in Japan and in R.O.C. are interested in computer. This would be quite natural.
- Students in lower level do not have prior knowledge in computer.

(2) Degree of satisfaction

Table 4.11: Interpretation of degree of satisfaction by item-type questionnaire (by multiple regression analysis)

(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?	

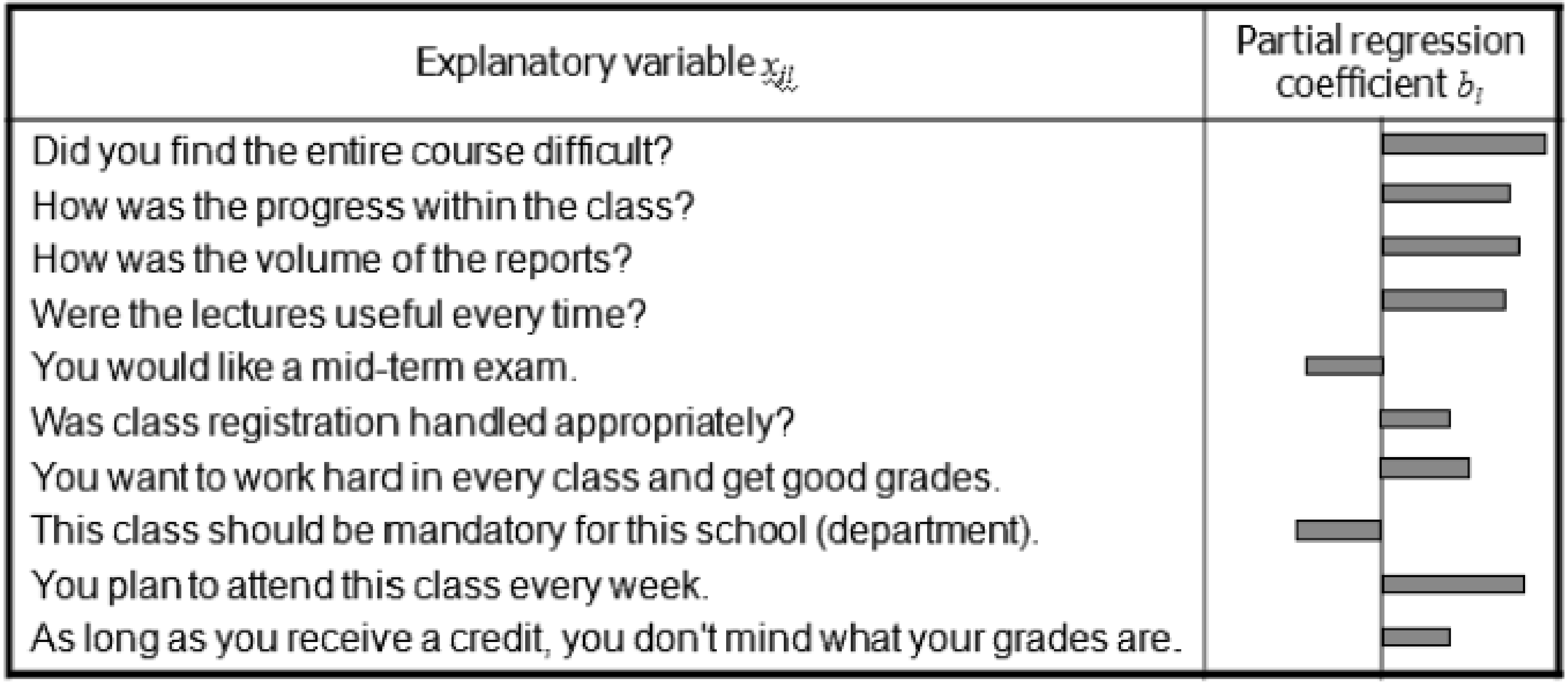
Contribution ratio = 0.766

(2) Degree of satisfaction

Table 4.12: Interpretation of degree of satisfaction by item-type questionnaire (by multiple regression analysis)

(i) Students in Japan

Satisfaction in terms of **Class management**



Explanatory variable x_{ji}	Partial regression coefficient b_j
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.	

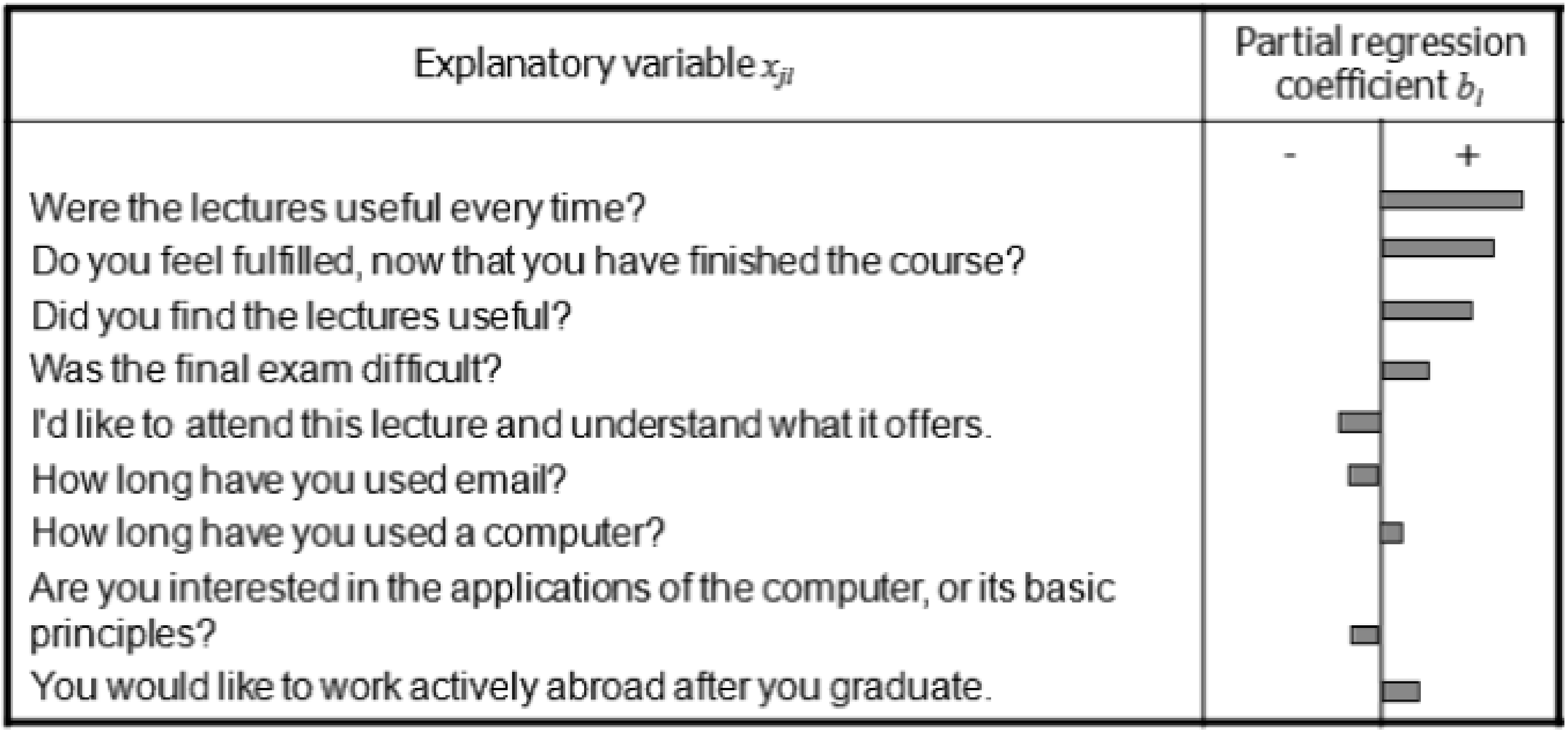
Contribution ratio＝0.782

(2) Degree of satisfaction

Table 4.13: Interpretation of degree of satisfaction by item-type questionnaire (by multiple regression analysis)

(ii) Students in R.O.C

Satisfaction in terms of **Contents of the lecture**



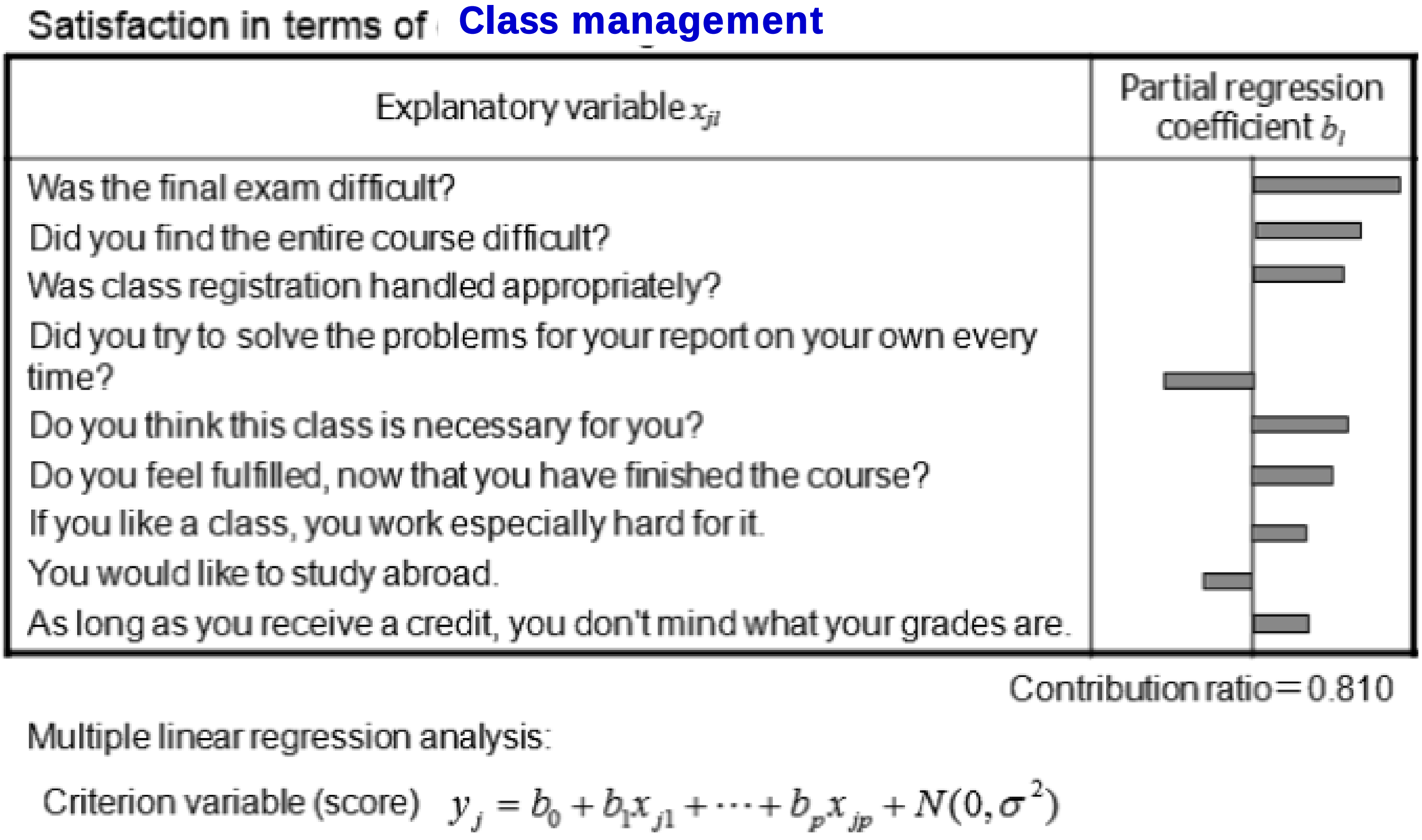
Explanatory variable x_{jl}	Partial regression coefficient b_l
	- +
Were the lectures useful every time?	
Do you feel fulfilled, now that you have finished the course?	
Did you find the lectures useful?	
Was the final exam difficult?	
I'd like to attend this lecture and understand what it offers.	
How 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

(2) Degree of satisfaction

Table 4.14: Interpretation of degree of satisfaction by item-type questionnaire (by multiple regression analysis)

(ii) Students in R.O.C



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

Discussions

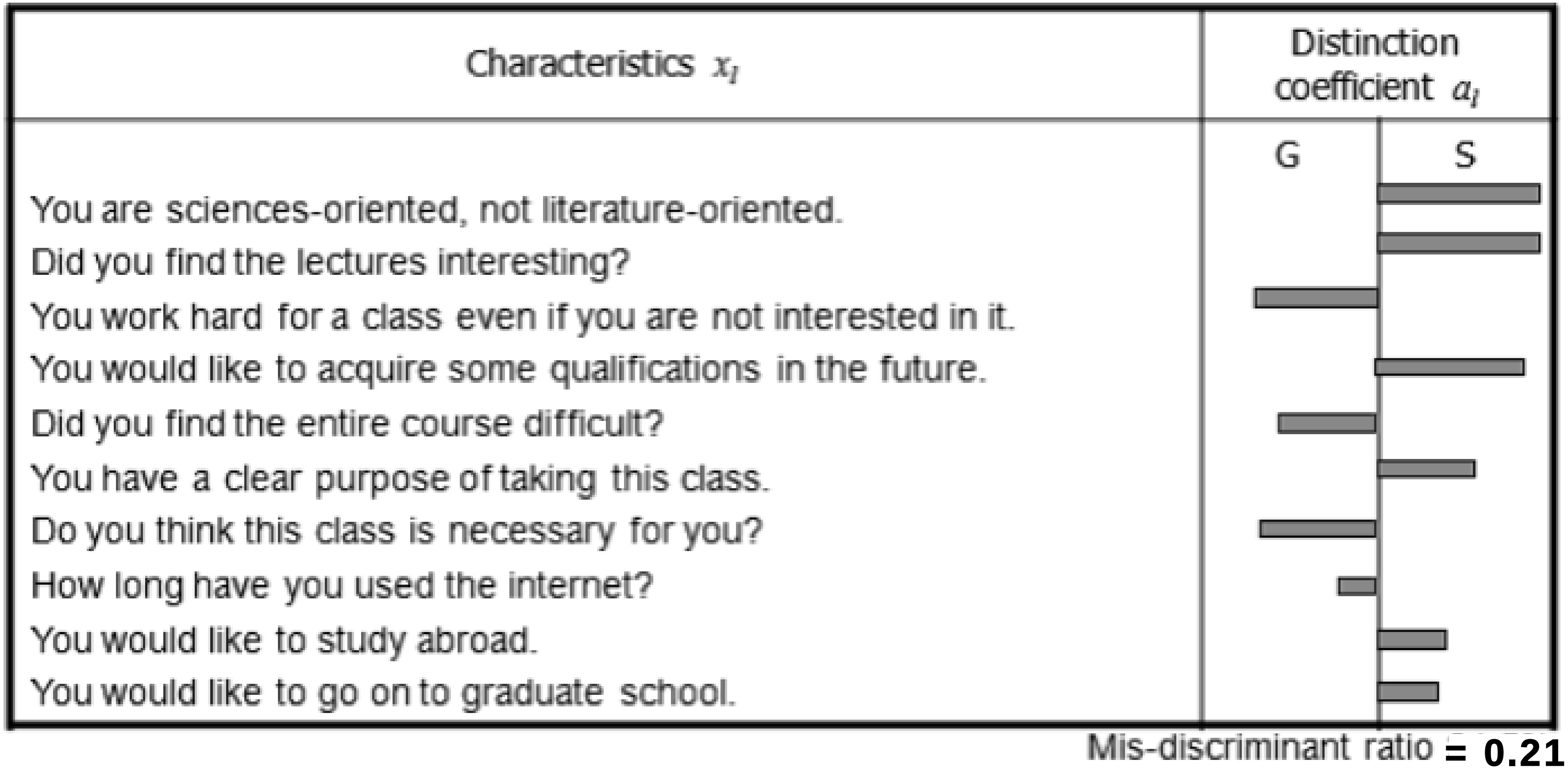
From Table 4.11-4.14:

- 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.
- 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.).

(3) Partition by Class G and Class S

Table 4.15: Interpretation of partion for Class G or Class S (by discriminant analysis)

(i) Students in Japan

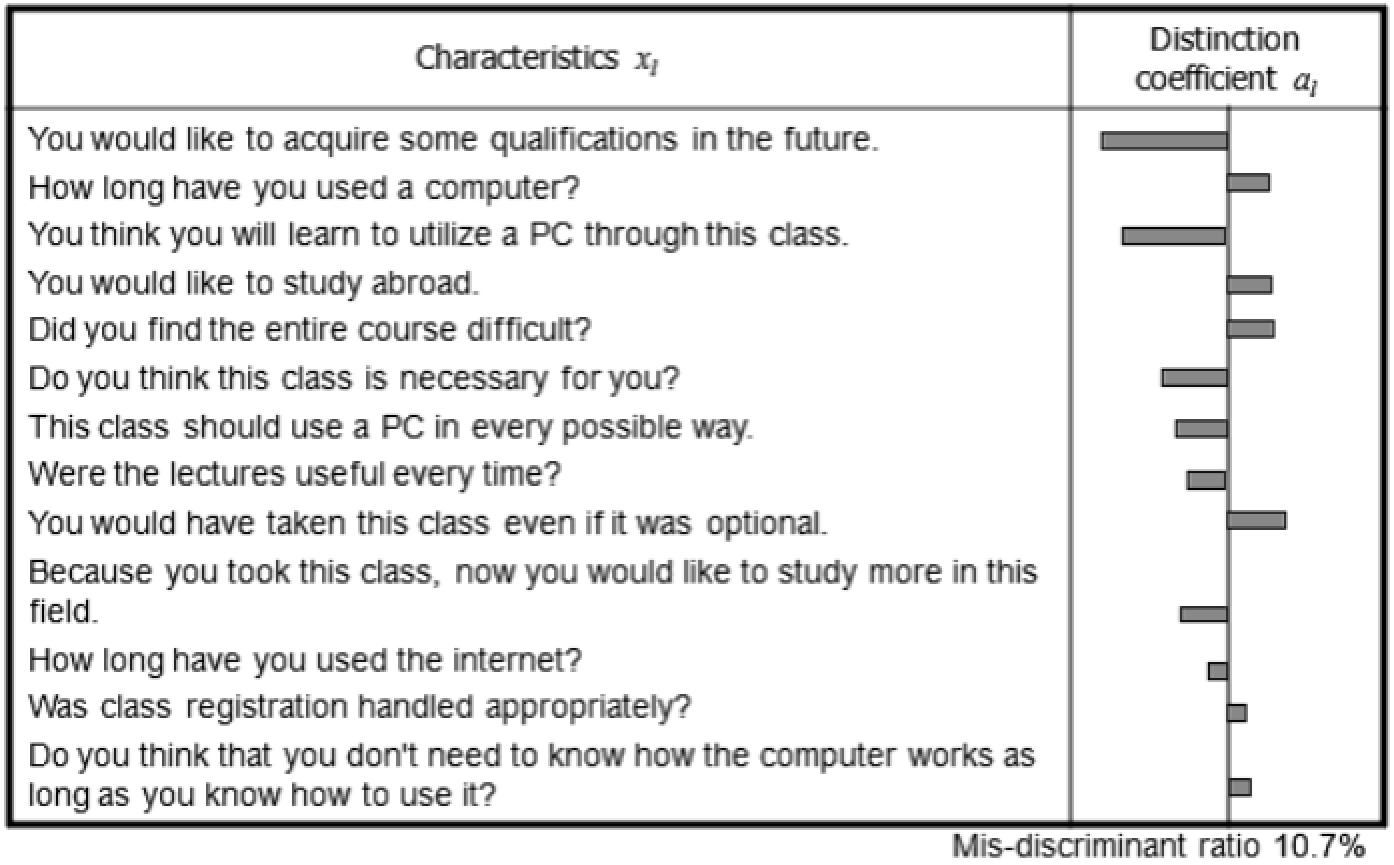


Characteristics x_l	Distinction coefficient a_l	
	G	S
You are sciences-oriented, not literature-oriented.		▬
Did you find the lectures interesting?		▬
You work hard for a class even if you are not interested in it.	▬	
You would like to acquire some qualifications in the future.		▬
Did you find the entire 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 = 0.215

(3) Partition by Class G and Class S

Table 4.16: Interpretation of partion for Class G or Class S (by discriminant analysis)
(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 $z = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_p x_p$ $\begin{cases} z > 0 & d \in \text{class S} \\ z < 0 & d \in \text{class G} \end{cases}$

Discussions

From Table 4.15-4.16:

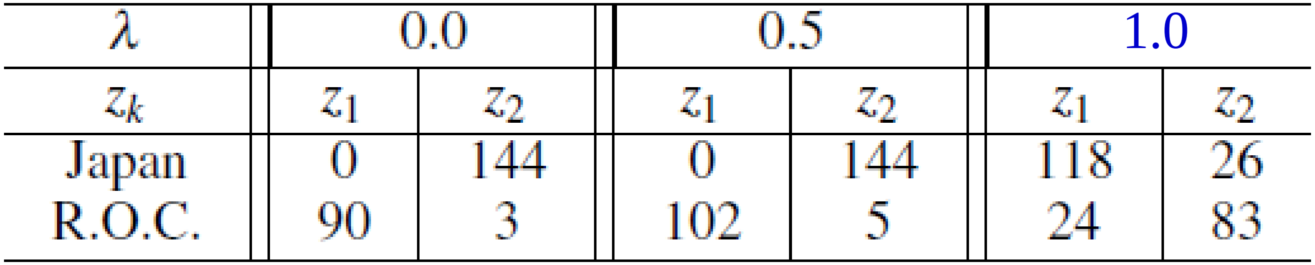
- Comparing to IQ only (Table V), it is more clear to interpret better partition to students by IQ and FQ. This suggests that proper partition to the next year should take causal relations obtained in this year into account.
- The students who are classified to Class S like sciences rather than literature,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

4.4 Clustering of students in Japan and R.O.C.

The clustering algorithm is applied to intentionally merged documents of both students in Japan and those in R.O.C.

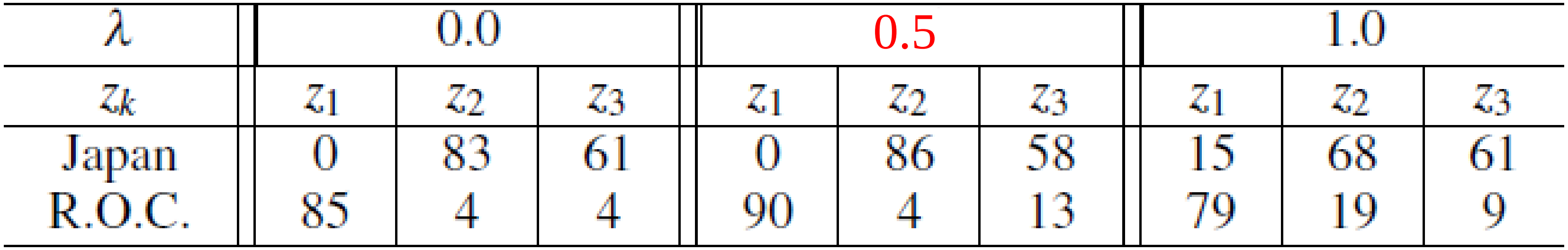
Table 4.17: Results of clustering

$K = 2$



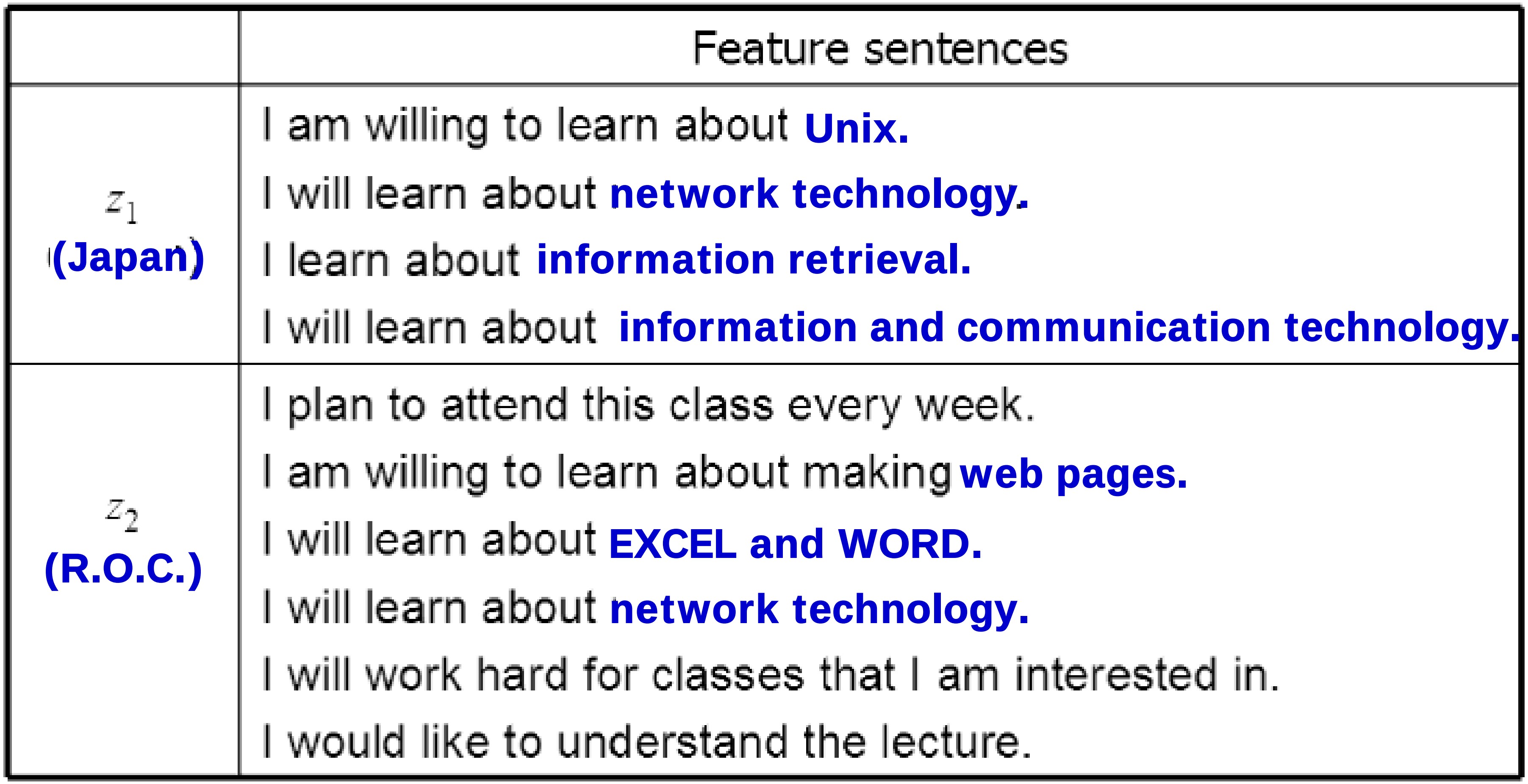
λ	0.0		0.5		1.0	
z_k	z_1	z_2	z_1	z_2	z_1	z_2
Japan	0	144	0	144	118	26
R.O.C.	90	3	102	5	24	83

$K = 3$



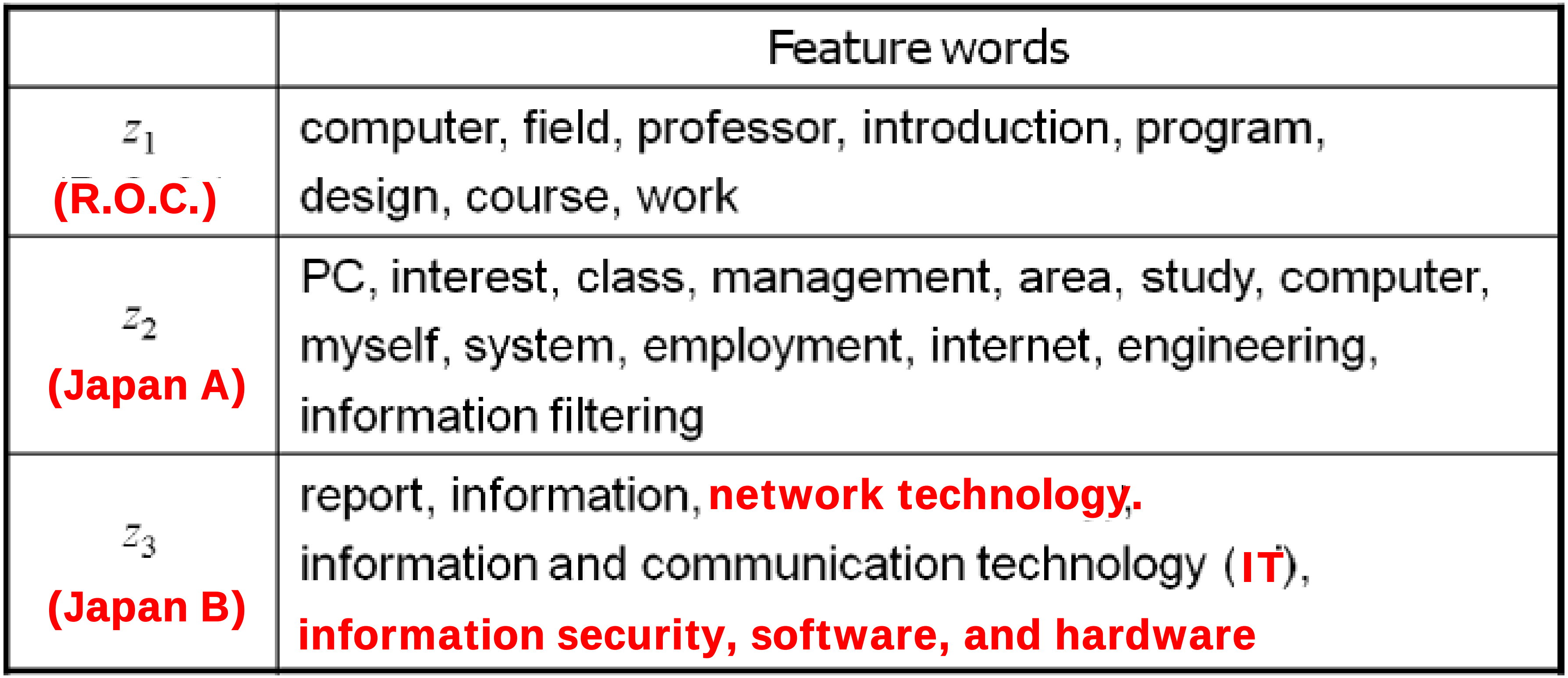
λ	0.0			0.5			1.0		
z_k	z_1	z_2	z_3	z_1	z_2	z_3	z_1	z_2	z_3
Japan	0	83	61	0	86	58	15	68	61
R.O.C.	85	4	4	90	4	13	79	19	9

Table 4.18: Extracted feature sentences in the case $K = 2, \lambda = 1.0$



	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 4.19: Extracted feature words in the case $K = 3, \lambda = 0.5$



	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**

Discussions

From Table 4.17:

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

From Table 4.18:

- 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 4.19:

- In the case of K = 3, λ = 0.5, extracted feature words represent that the cluster z_3 contains more professional students.

Additional experiments

Difference of text processing methods between by automatic translating Chinese and by directly Chinese:

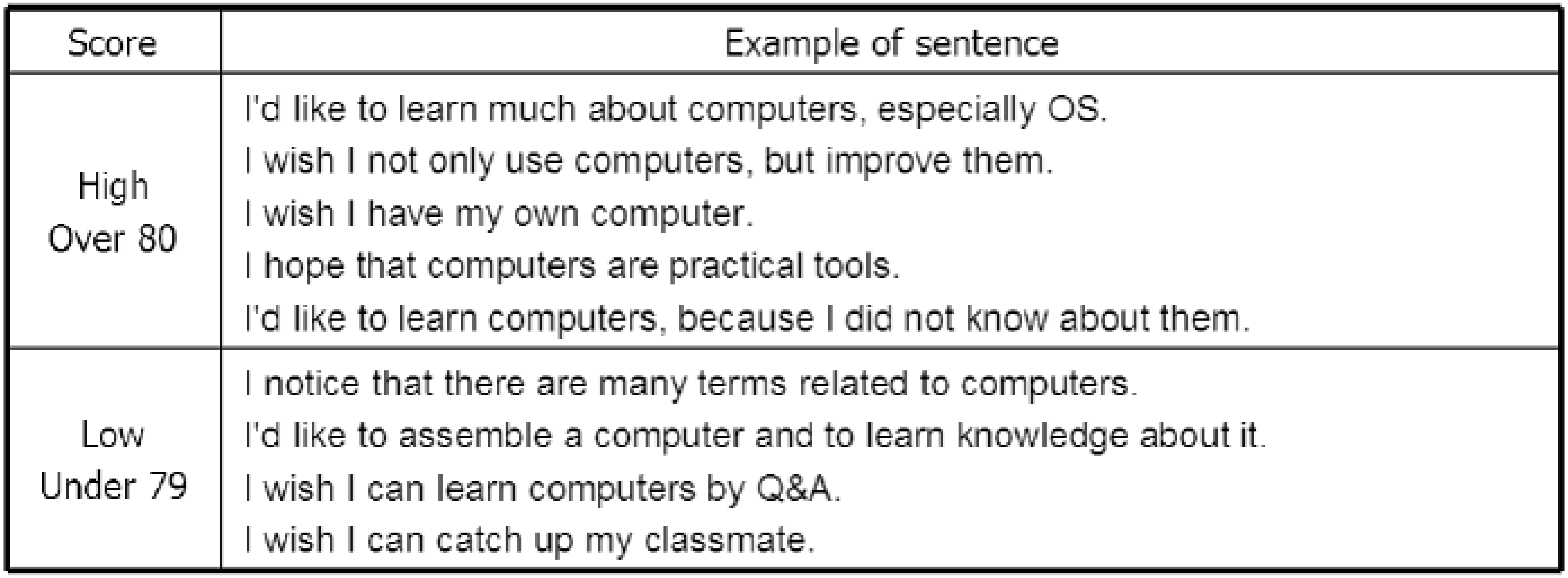
Table XII shows important sentences extracted from text-type questionnaire (IQ only) for high or low scores of students in R.O.C.

The (i) in this table corresponds to (ii) of Table VI.

Additional experiments

Table 4.20: Important sentences extracted from text-type questionnaire (IQ only) for scores of students in R.O.C.

(i) By translating Chinese into Japanese

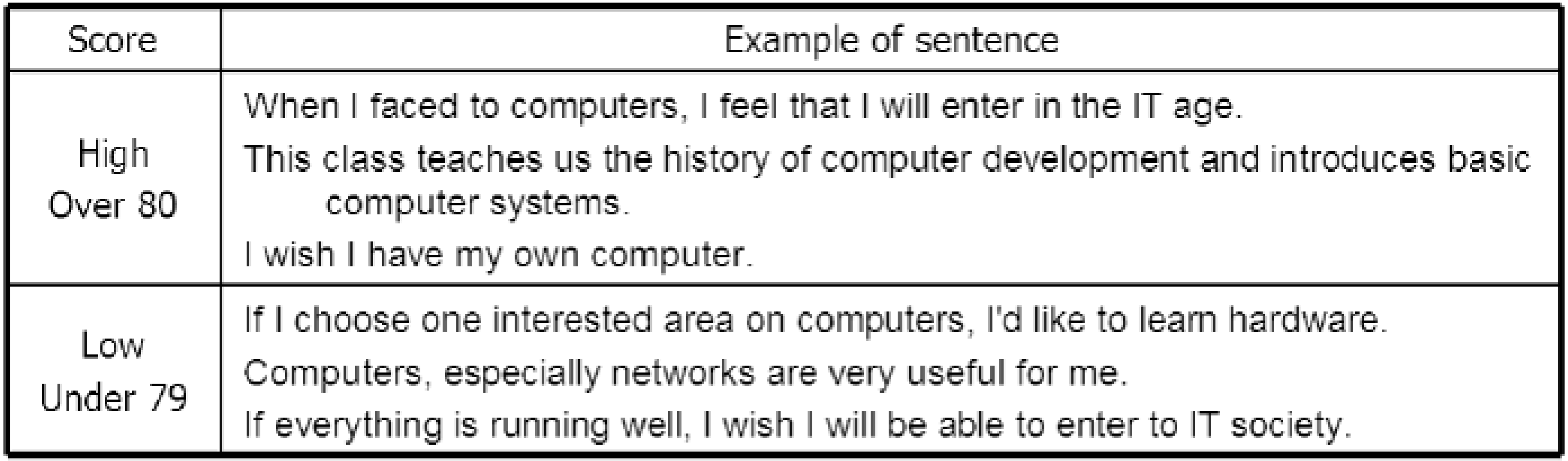


Score	Example of sentence
High Over 80	I'd like to learn much about computers, especially OS. I wish I not only use computers, but improve them. I wish I have my own computer. I hope that computers are practical tools. I'd like to learn computers, because I did not know about them.
Low Under 79	I notice 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.

Additional experiments

Table 4.21: Important sentences extracted from text-type questionnaire (IQ only) for scores of students in R.O.C.

(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 IT society.

Discussions

It is possible to realize the system for Chinese language, where we can use

- automatic indexing by N-gram or
- morpheme in Chinese (ii).

From Table 4.20-4.21:

- There are little differences between Table 4.10, Table 4.20 and 4.21.
- Directly Chinese text processing for students in low scores extracts positive sentences.

5. Concluding Remarks

- 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.

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 plan, in a sense that they are going to learn not only their future job but their unsophisticated thirst for knowledge.