

The 2007 IEEE International Conference on System, Man and Cybernetics  
October 7-10,2007, Montreal, Canada

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

**Shigeichi Hirasawa<sup>†</sup>, *Fellow, IEEE*,**

**Fu-Yih Shih<sup>‡</sup>, and**

**Wei-Tzen Yang<sup>+</sup>**

<sup>†</sup> Waseda University, Japan (hira@waseda.jp)

<sup>‡</sup> Leader University, Taiwan, R.O.C. (fuyih@mail.leader.edu.tw)

<sup>+</sup> Tamkang University, Taiwan, R.O.C. (018467@mail.tku.edu.tw)

---

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

# I . Introduction

## I . 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.

## I . Introduction

- 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) Technically professional engineer
- { (b) General and economical anaysist, sales engineer

## I . Introduction

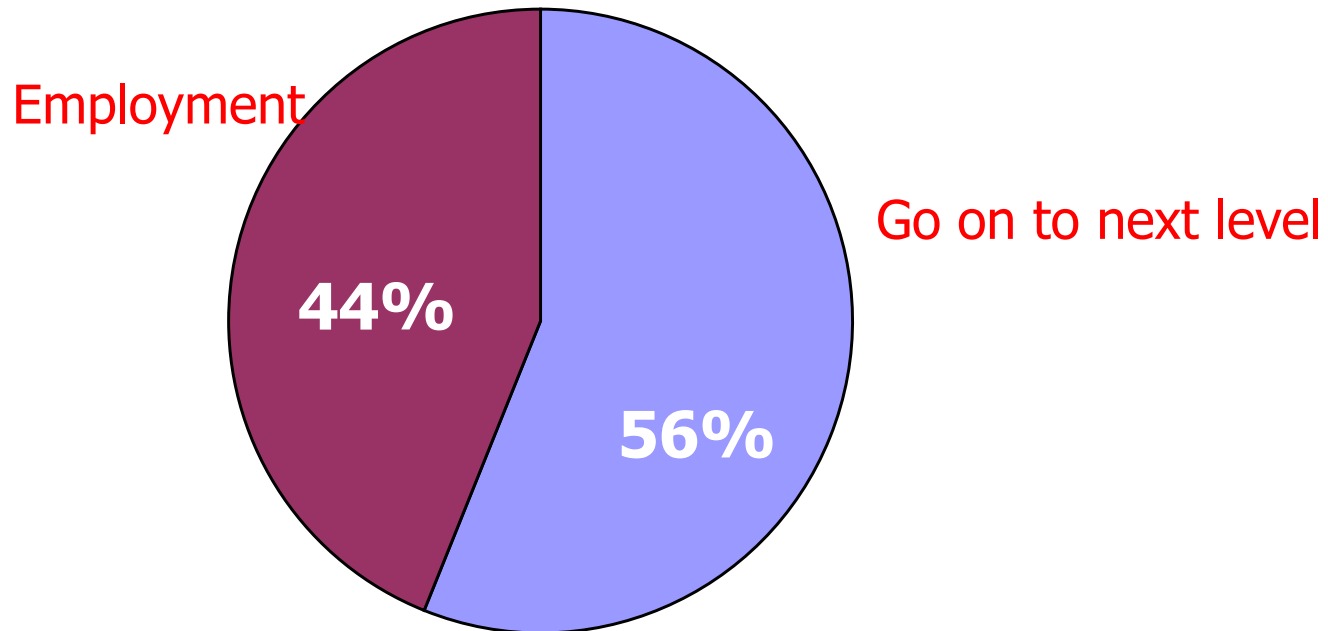


Fig.0-a: Example of future path of undergraduate students  
(Waseda University)

I . Introduction

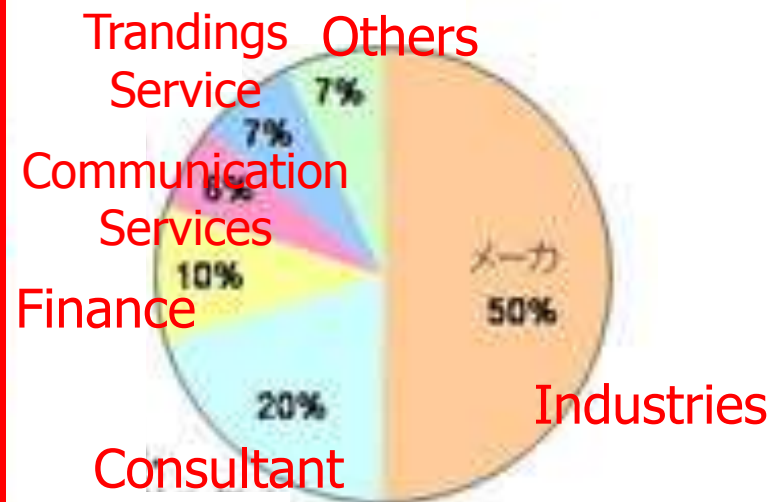


# 卒業生の進路

キヤノン	野村総合研究所
コンパックコンピュータ	ブライスウオータワースターパス
サントリー	三菱総合研究所
シャープ	ゴールドマンサックス証券
ソニー	三和銀行
東芝	JPモルガン証券
東レ	住友銀行
日本IBM	第一勧業銀行
日本電気	東海銀行
日産自動車	野村證券
富士通	富士銀行
本田技研	NTTデータ通信
松下電器産業	日本電信電話(NTT)
三菱電機	東海旅客鉄道
山之内製薬	情報堂
アクセンチュア	三井物産
CSK	鹿島建設
イトロイトーマクコンサルティング	日本経済新聞社
日本総合研究所	朝日新聞社

Major companies

就職業種(学部, 修士)



23

Fig.0-b: Example of jobs of undergraduate and graduate students (Waseda University)

## I . Introduction

### 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 Securities Co., Ltd.

#### [Communication Services]

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

#### [Tradings and Services]

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

#### [Others]

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

## II . Questionnaire Analysis Model

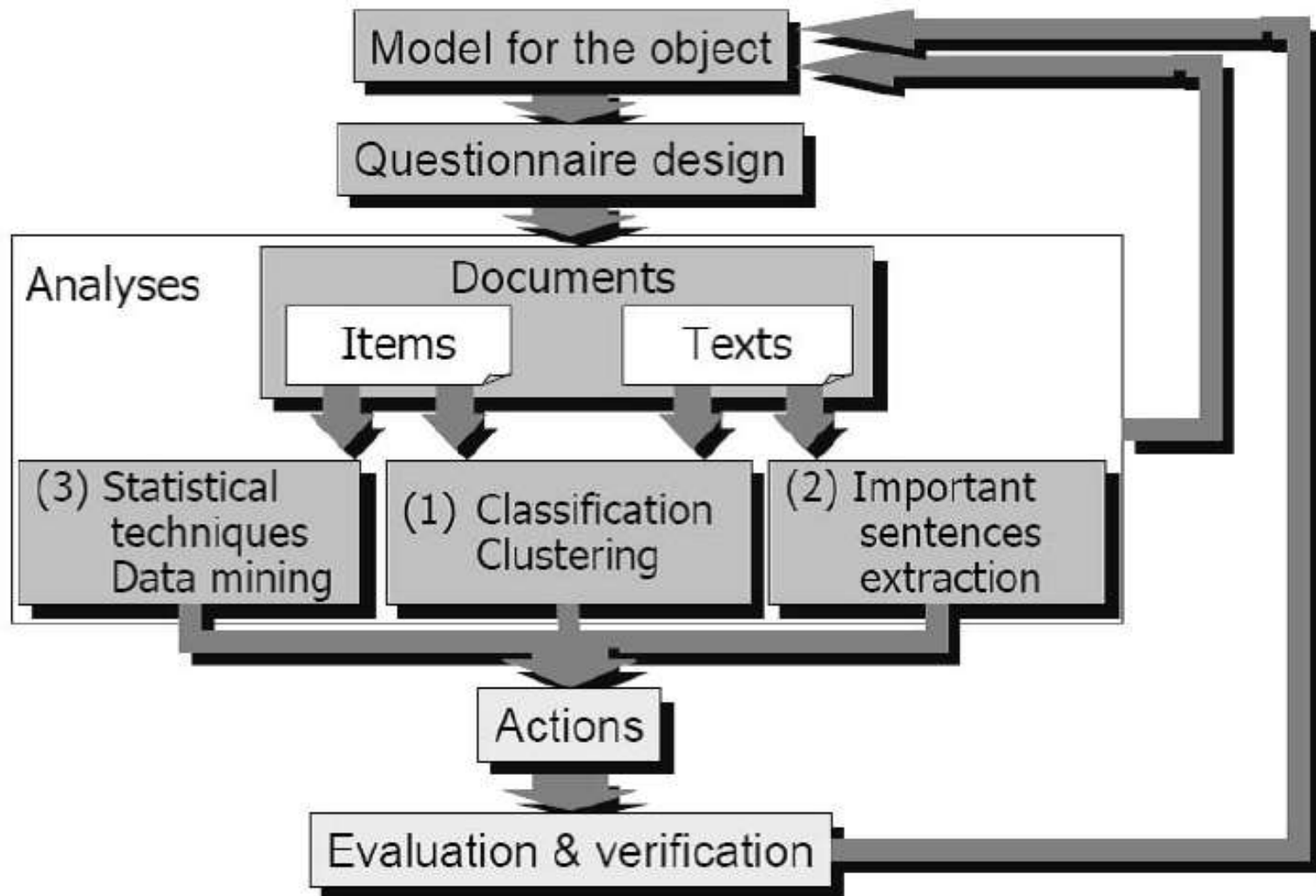


Fig. 2.1: Questionnaire analysis model



## II . 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:

## II . Questionnaire Analysis Model

- (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: II. Questionnaire Analysis Model

- 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.

### III. 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.

# A. Class Model

## III. Student Questionnaire

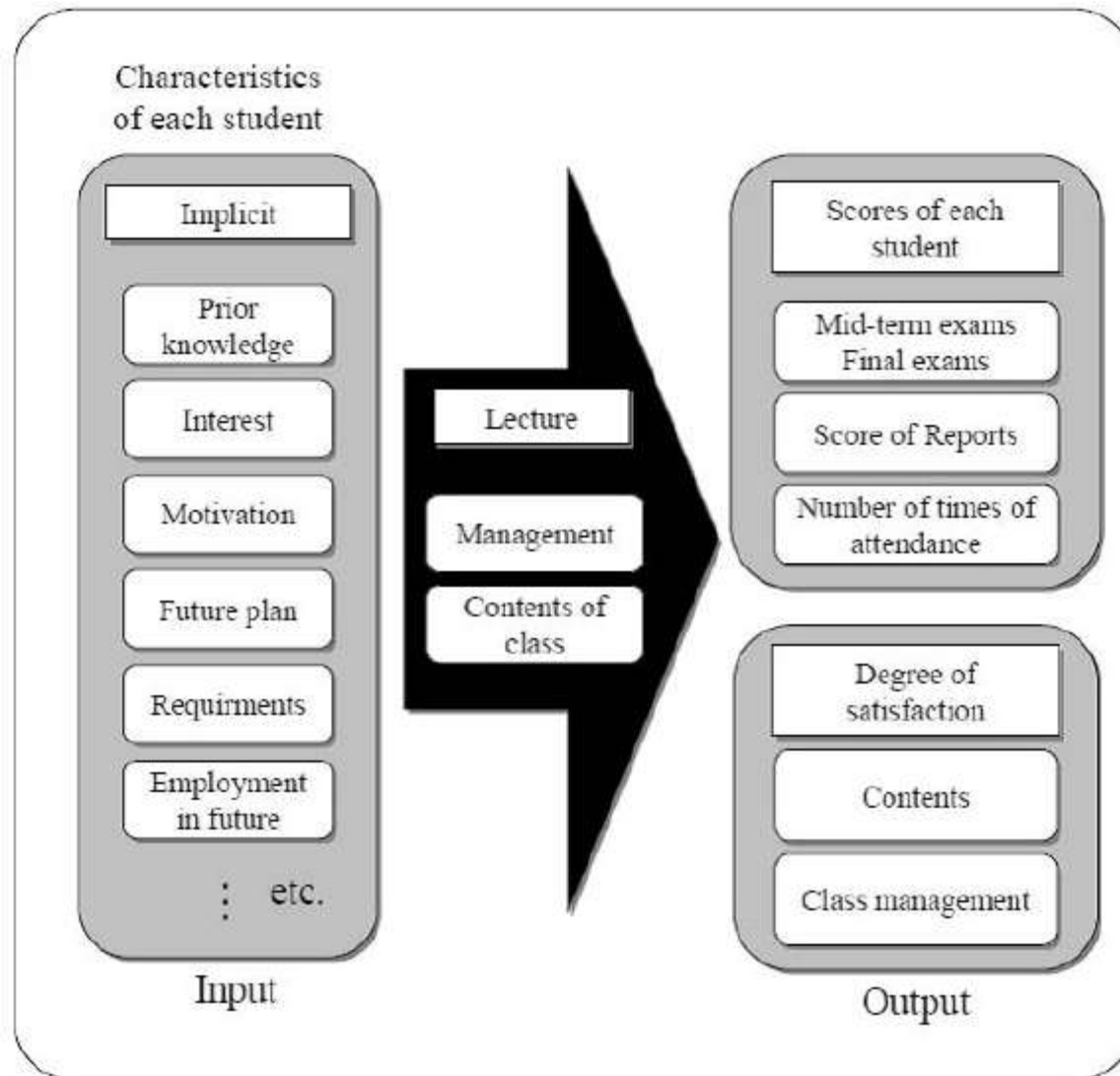


Fig. 2: Class model for the class "Introduction to Computer Engineering

## B. Design of Questionnaire

## III. Student Questionnaire

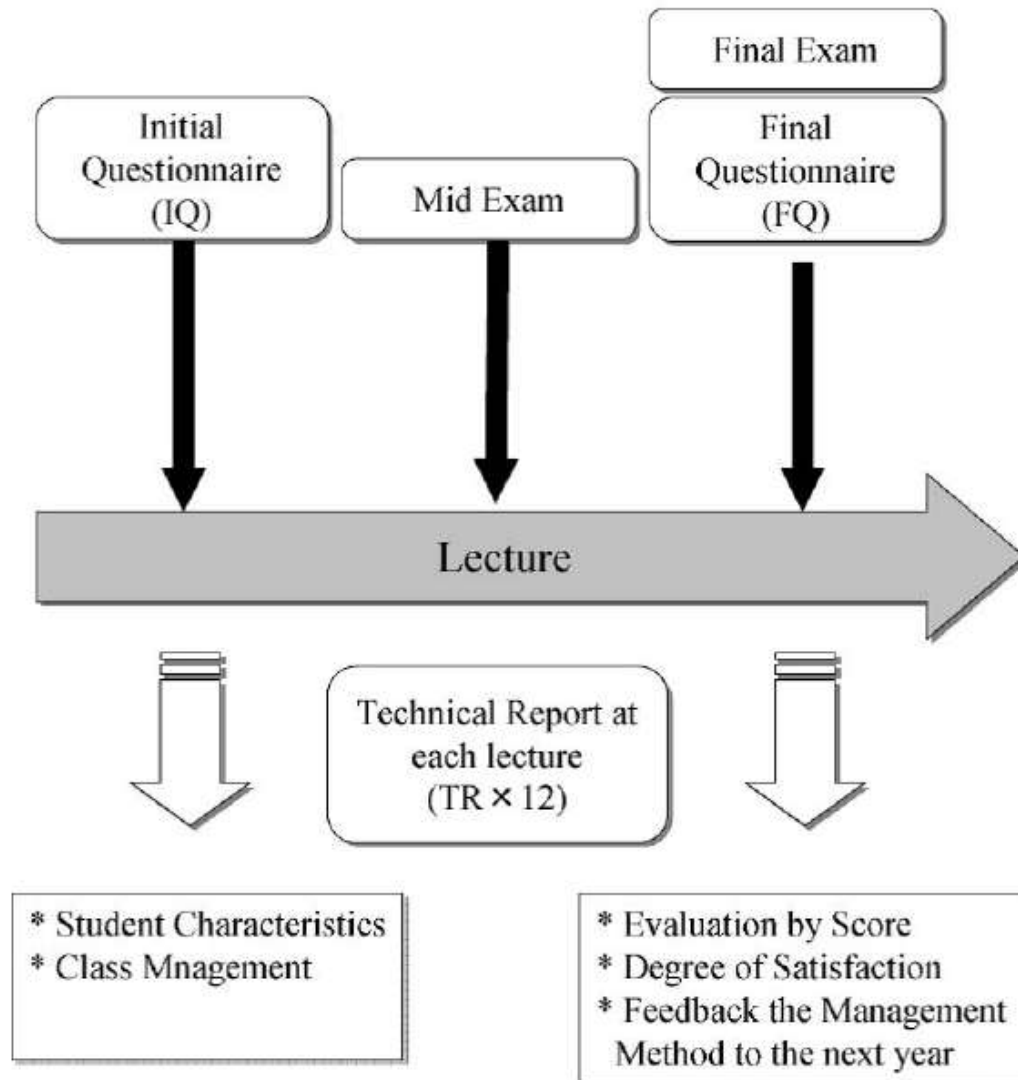


Fig. 3: Time schedule for class

## B. Design of Questionnaire

## III. Student Questionnaire

Table I : Data of class

Exercise	Contents
Initial Questionnaire (IQ) Item type Text type	7 questions (4-20 sub-questions each) 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 Text type	6 questions (6-21 sub-questions each) 5 questions (250-300 characters in Japanese each)

## B. Design of Questionnaire

## III. Student Questionnaire

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

Exercise		Examples (sub questions)
IQ	Item-type	<ul style="list-style-type: none"> <li>✓ For how many years have you used computers?</li> <li>✓ Do you have a plan to study abroad?</li> <li>✓ Can you assemble a PC?</li> <li>✓ Do you have a qualification related to information technology?</li> <li>✓ Write 10 technical terms in information technology which you know.</li> </ul>
	Text-type	<ul style="list-style-type: none"> <li>✓ Write about your knowledge and experience on computer.</li> <li>✓ What kind of work will you have after graduation?</li> <li>✓ What do you imagine from the name of this class subject name?</li> </ul>



## B. Design of Questionnaire

## III. Student Questionnaire

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

Exercise		Examples (sub questions)
FQ	Item-type	<ul style="list-style-type: none"> <li>✓ Could you understand the contents of this lecture?</li> <li>✓ Was the midterm test difficult?</li> <li>✓ Was it easy to read the handwritings on the white-board?</li> <li>✓ Do you think the contents of this lecture to be useful to yourself?</li> <li>✓ Do you want to finish this course even if it is optional?</li> <li>✓ Which are you interested in applied technology or the fundamentals of computers?</li> <li>✓ Which do you choose class (S) or class (G)?</li> </ul>
	Text-type	<ul style="list-style-type: none"> <li>✓ Do you want to be a member of laboratories related to the information technology?</li> <li>✓ In the future, will you get a job in industries related to the information technology?</li> <li>✓ Did your image on computers change after taking this lecture?</li> </ul>

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/](http://www.hirasa.mgmt.waseda.ac.jp/users/comp-eng/)

## IV. Algorithm used for Analyses

### Information Retrieval Model

Text Mining:

- Information Retrieval including
- Clustering
- Classification

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 <b>Probabilistic LSI (PLSI) Model [Hofmann99]</b> Inference Network Model Bayesian Network Model

# Document

## IV. Algorithm used for Analyses

Format		Example in paper archives		matrix
Fixed format	Items	<ul style="list-style-type: none"> <li>- The name of authors</li> <li>- The name of journals</li> <li>- The year of publication</li> <li>- The name of publishers</li> </ul>	<ul style="list-style-type: none"> <li>- The name of countries</li> <li>- The year of publication</li> <li>- The citation link</li> </ul>	$G \in \{0,1\}^{I \times D}$
Free format	Texts	The text of a paper <ul style="list-style-type: none"> <li>- Introduction      - Preliminaries</li> <li>.....</li> <li>- Conclusion</li> </ul>		$H \in \{0,1,2,\dots\}^{T \times D}$

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

$d_j$ : The  $j$ -th document

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

$t_i$ : 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

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

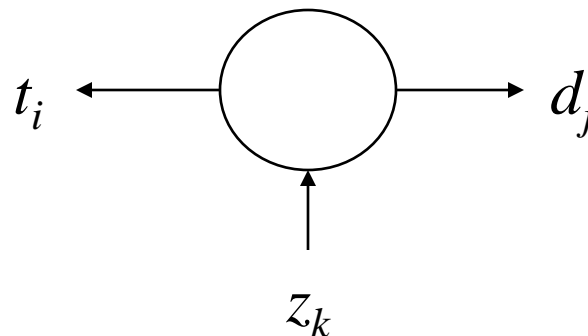
the number of term  $t_i$  in document  $d_j$

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  $d_j$



# The Probabilistic LSI (PLSI) Model

E) Similarity function:

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

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.(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:

$$L = \sum_{i,j} a_{ij} \log \Pr(t_i, d_j) \quad (4.1)$$

# EM Algorithm

[EM algorithm]

According to eq.(1), the maximum value of eq.(4.1) is computed by alternating E-step and M-step until it converges.

E-step:

$$\Pr(z_k | t_i, d_j) = \frac{\Pr(z_k) \Pr(t_i | z_k) \Pr(d_j | z_k)}{\sum_{k'} \Pr(z_{k'}) \Pr(t_i | z_{k'}) \Pr(d_j | z_{k'})} \quad (4.2)$$

M-step:

$$\Pr(t_i | z_k) = \frac{\sum_j a_{ij} \Pr(z_k | t_i, d_j)}{\sum_{i',j} a_{i'j} \Pr(z_k | t_{i'}, d_j)} \quad (4.3)$$

$$\Pr(d_j | z_k) = \frac{\sum_i a_{ij} \Pr(z_k | t_i, d_j)}{\sum_{i,j'} a_{ij'} \Pr(z_k | t_i, d_{j'})} \quad (4.4)$$

$$\Pr(z_k) = \frac{\sum_{i,j} a_{ij} \Pr(z_k | t_i, d_j)}{\sum_{i,j} a_{ij}} \quad (4.5)$$

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

## 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 ( $C_1, C_2, \dots, C_K$ )

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

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

where  $n_k$  is the number of selected documents to compute the representative document vector from  $C_k$  and  $\vec{d}_j = (a_{1j}, a_{2j}, \dots, a_{Dj})^T$ , 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  $|\mathcal{Z}| = K$
- (3) Decide the state  $z_{\hat{k}} (= C_{\hat{k}})$  for  $\vec{d}_j$  as

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

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



## B. Clustering Algorithm [10]

$S$  : The number of clusters ( $C_1, C_2, \dots, C_S$ )

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

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

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

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

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

## 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.

# V. Questionnaire Analyses

## A. Verification of class model by IQ

Class G (generalist): wide and shallow technical topics

Class S (specialist): technical and professional topics

Table III : Contents of topics

Class	Contents
Class G	<ul style="list-style-type: none"><li>- History of computers, fundamental concepts in computer</li><li>- Basics of architecture</li><li>- Basics of hardware</li><li>- Basics of software</li><li>- Applications of information technology etc.</li></ul>
Class S	<ul style="list-style-type: none"><li>- Architecture(stack machine, binary system, processor architecture)</li><li>- Hardware(logic design, logical circuit, automaton)</li><li>- Software(operating system, UNIX, language processor) etc.</li></ul>

Table IV : Partition of Class G and Class S

## (i) Students in Japan

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

## (ii) Students in R.O.C.

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

Table V : Characteristics of Class G and Class S (by **discriminant analysis**)  
 (i) Students in **Japan** (Student's choice)

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

Mis-discriminant ratio 30.5%


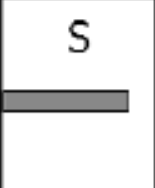
Table V : Characteristics of Class G and Class S (by discriminant analysis)  
 (i) Students in Japan (Automatic classification)

	Characteristics $x_j$	Distinction coefficient $a_j$
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 V : Characteristics of Class G and Class S (by discriminant analysis)

(ii) Student's in R.O.C.

	Characteristics $x_i$	Distinction coefficient $a_i$	
		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:

$$\text{Discriminant function } z = a_0 + a_1x_1 + a_2x_2 + \dots + a_px_p$$

$$\begin{cases} z > 0 & d \in \text{class S} \\ z < 0 & d \in \text{class G} \end{cases}$$



# Discussions (for A)

## From Table IV:

- It is shown that the degree of agreement between the student's own choice and automatic classification are **approximately 60% by IQ only**.
- Although our method is probably not accurate enough to use automatic classification, but it would be still useful **to assist and to guide students**.
- We know that most of all students **do not decide their future jobs yet** in their second academic year.
- It is worth noting from our experience that **the student's own choice is not always true**.
- For example, it would be interesting whether a graduated student who is a member of staff at industry chose Class S or not. Past data by graduated students in their second year can be effectively used to this analysis.

## From Table V:

- Automatic classification gives interesting tendency such that the students in Class S **like to learn actively** and **wish to go to study abroad**.
- There is almost no difference between students in Japan and in R.O.C.

## B. Verification of class model by IQ and FQ

### (1) Scores of students

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

#### (i) Students in Japan

Score	Exmample of Sentences
High over 70	I'm interested in <b>Information security, network</b> and <b>Internet technology</b> . We are to learn how the computer works, <b>not how to work with it</b> . 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 <b>Internet</b> , 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	Exmample of Sentences
High over 70	I'd like to take on a <b>computer-related job</b> . I'd like to learn about the computer and then do a <b>research</b> 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 <b>about nothing</b> about the computer. I know very <b>little</b> about the computer. The computer always makes me <b>suffer</b> . I'd like the class to actually use a computer in order to teach the theory behind it.

# Discussions (for B-(1))

## From Table VI:

- 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.

# B. Verification of class model by IQ and FQ

## (2) Degree of satisfaction

Table VII : 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 $x_{ji}$	Partial regression coefficient $b_j$
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

## B. Verification of class model by IQ and FQ

### (2) Degree of satisfaction

Table VII : 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 $x_{jt}$	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

## B. Verification of class model by IQ and FQ

### (2) Degree of satisfaction

Table VII : 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 $x_{ji}$	Partial regression coefficient $b_i$	
	-	+
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

## B. Verification of class model by IQ and FQ

### (2) Degree of satisfaction

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

Contribution ratio = 0.810

Multiple linear regression analysis:

$$\text{Criterion variable (score)} \quad y_j = b_0 + b_1x_{j1} + \dots + b_px_{jp} + N(0, \sigma^2)$$

# Discussions (for B-(2))

## From Table VII:

- 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.).



## B. Verification of class model by IQ and FQ

### (3) Partition by Class G and Class S

Table VIII : Interpretation of partition for Class G or Class S (by discriminant analysis)

(i) Students in [Japan](#) 1

Characteristics $x_j$	Distinction coefficient $a_j$	
	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**

## B. Verification of class model by IQ and FQ

### (3) Partition by Class G and Class S

Table VIII: Interpretation of partition for Class G or Class S (by discriminant analysis)  
(ii) Students in R.O.C

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

Mis-discriminant ratio 10.7%

Discriminant analysis:

Discriminant function  $z = a_0 + a_1x_1 + a_2x_2 + \dots + a_px_p$

$$\begin{cases} z > 0 & d \in \text{class S} \\ z < 0 & d \in \text{class G} \end{cases}$$

## Discussions for (B-(3))

### From Table VIII:

- 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**.

## C. Clustering of students

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

Table IX : Results of clustering

$$K = 2$$

$\lambda$	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

$$K = 3$$

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

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

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

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

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

# Discussions (for C)

## From Table IX:

- 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.

# Discussions (for C)

## From Table X:

- 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 XI:

- In the case of  $K = 3$ ,  $\lambda = 0.5$ , extracted feature words represent that the cluster **z3** 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 XII : Important sentences extracted from text-type questionnaire (IQ only)  
for scores of students in R.O.C.

(i) By translating Chinese into [Japanese](#) ;

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

# Additional experiments

Table XII : 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 (for AE)

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 XII :

- There are little differences between Table VI (ii), Table XII (i) and (ii).
- Directly Chinese text processing for students in low scores extracts positive sentences.

## VI. Conclusions and future works

- 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.