May 2004
A Short Course at Tamkang University,

Knowledge Discovery from Questionnaires — A Case of Improvements for Quality of Education —

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 viewpoints of both data mining and text mining, where the answers by the fixed format is called "items", and those by the free format, simply "texts". In this paper, an algorithm for simultaneously processing answers with both the items and the texts is developed. An algorithm for extracting important sentences from documents is also developed. Using these algorithms combined with statistical techniques, a method for analyzing the questionnaires is presented. The method is applied to a case of improvements for the quality of education for which the student questionnaire is executed to a class, and we obtain useful knowledge which leads to faculty developments.

Keywords—questionnaire, information retrieval, PLSI model, classification, clustering, data mining, text mining, statistical analysis, faculty development, improvement of education

1 Introduction

Recently, many enterprises introduce questionnaires for the purpose of extracting minds, intentions or opinions of customers. It is necessary to develop an efficient way for knowledge discovery from the mass by means of electronic documents with the standpoints of market research, knowledge management and customer relationship management.

Usual questionnaires consist of two types: one is questions answered by a fixed format such as multiple choice, and the other, by a free format which implies text. The former is called "items", and the latter, simply "texts". The questionnaires that consist of both types are also used.

In this paper, we discuss knowledge discovery from the questionnaire consisting of both the items and the texts. (1) An algorithm for simultaneously processing answers with both the items and the texts and (2) An algorithm for extracting important sentences from the text-parts of the documents are developed. By combining (3) Statistical techniques applied to the item-parts, characteristics of each class or cluster are clarified. The results obtained in these analyses give us useful knowledge to manage the object.

The method of knowledge discovery discussed in this paper is applied to a case of improvements for the quality of education. A student questionnaire of the class: Introduction to Computer Science, second academic year of our department which the author teaches, is executed. Problems of partitioning students of the class into a few subclasses by their characteristics are evaluated. The purposes of these problems are to improve the degree of satisfaction of the students and increase the effectiveness of education.

In this paper, we show a questionnaire analysis model in section 2. The applied case to student questionnaire is mainly discussed in detail in section 3. Section 4 concludes this paper.

2 Questionnaire Analysis Model

The method for analyzing the questionnaire is shown in Fig. 2.1 as a questionnaire analysis model.

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

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

Next, analyses are executed as follows:

(1) The set of documents is classified or clustered by the proposed algorithm for classification or clustering [HC03] [HIASG04] [IIGSH03-a]. Note that both the items and the texts are processed together, not separately.

(2) For the texts only, important sentences, or the parts of them are extracted from the documents by the proposed algorithm for extracting important information [IIGH02] [SIGIH03] [ISH03] [SIGH04]. These results are helpful to easily understand the opinions and directly give useful information of the classes or clusters.

(3) For the items [1] only, statistical techniques such as multiple linear regression analysis, discriminant analysis and so on, 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 [Hofmann99] [CH01] which is known to be one of the most powerful model in information retrieval systems [BYRN99]. The proposed algorithm based on PLSI

* Department of Industrial and Management Systems Engineering, School of Science and Engineering, Waseda University, 3-4-1, Ohkubo, Shinjuku, Tokyo 169-8555 Japan. Phone:+81-3-5286-3290. Fax:+81-3-5273-7215. E-mail:hirasawa@hirasa.mgmt.waseda.ac.jp

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

model exhibits good performance in classification or clustering especially for a small size of the document set. In (2), we have also proposed the algorithm to select important sentences by extracting representative words or sentences based on Japanese language processing.

The results obtained by combining (1) and (3) give the profile of each class 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 of 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 objective is generated by the feedback loop if necessary.

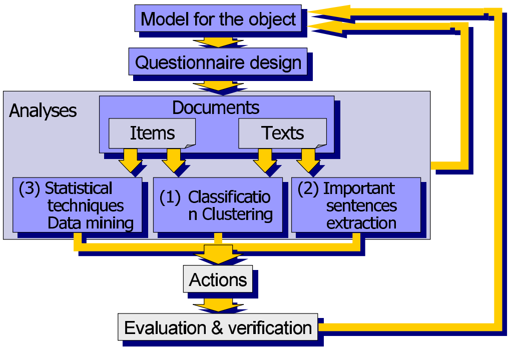

Model for the object
Questionnaire design
Analyses
Documents
Items
Texts
(3) Statistical techniques Data mining
(1) Classificatio n Clustering
(2) Important sentences extraction
Actions
Evaluation & verification

Figure 2.1: Questionnaire analysis model

3 Case of Student Questionnaire

A class model in this object is shown in the Fig.3.1.

A technique to extract requests of the students from the questionnaire is discussed by applying the questionnaire analysis model. First, relationships between the degree of satisfaction, score and the characteristics of the students and so on 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 analysis 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 universities in Japan, recently attentions to the quality assurance of the education program by Japan Accreditation Board for Engineering Education (JABEE) becomes important as for improvements of a class.

3.1 Class model

We have proposed a class model for the class “Introduction to Computer Science” as shown in Fig. 3.1 [IIGSH03-b].

The implicit characteristics of each student are essentially measured by questionnaire. While explicit characteristics are objectively given by numerical data of the class. These characteristics generate explanatory variables. Then each student yields his or her final score and the degree of satisfaction as the result of the class. The degree of satisfaction is also measured by questionnaire. The final score and the degree of satisfaction play as criterion variables in this model. Besides these variables, we can expect to get some information regarding to classroom management such as partition of the class. Usually there exist many differences between each student in level, interested area, experiences, and motivation before beginning the class, since the class is opened in Department of Industrial and Management Systems Engineering. Hence a proper partition of the class depending on features such as the future plan of each student is desirable. The partitions shown in Table 1 can be considered, where G stands for a generalist course, S, for a specialist course by estimating his or her future work. According to this model, we can effectually design the questionnaire.

Characteristics of each student		
(1)	Implicit characteristics	
	prior knowledge	
	interested area	⇒ Final score
	motivation	
	future plan	
	intention	
	profession	
	employment in future	⇒ Degree of satisfaction
	etc.	
(2)	Explicit characteristics	
	number of attendances	⇒ Partition of the class
		Class G
	scores of	Class S
	reports	
	mid-term exam	
	final exam	
	etc.	
	Explanatory variables	Objective variables

Figure 3.1: Class model

3.2 Contents of Questionnaire

A questionnaire was applied to the class: “Introduction to computer science”. 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 test (MT) and final test (FT) are explicit characteristics of each student. We analyze them by using statistics, data mining, and information retrieval techniques that include classification and clustering. The example of contents of a questionnaire is shown in Table 3.2 and in Table 3.3.

3.3 Questionnaire analyses

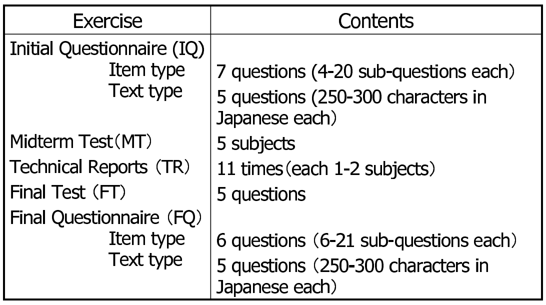
3.3.1 Verification of class model by IQ

Before starting the class, we discuss problems on the class management and the lecture plan. By using

Table 3.1: 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.

Table 3.2: 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)

only IQ, the following problems are considered:

(1) Prediction of the scores
(2) Partition of the students of the class

(1) Prediction of scores

For explaining the score by only IQ, it is difficult to distinguish even whether he get a high score or not (The rate of contribution of the multiple linear regression analysis was 51% .)

(2) Partition of students of class

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 have the following three partitions. As shown in Table 3.1, we provide two courses depending on the contents of topics of the lecture:

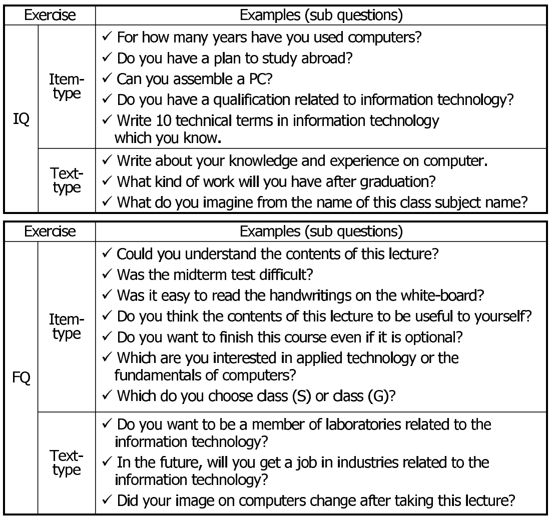
(a) Partition by the contents of topics

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

We can also provide two classes depending on the level of the lecture and on the management of the class:

(b) Partition by the student's level

Table 3.3: Contents of questionnaire



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

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

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

Class H: a higher level
Class L: a lower level

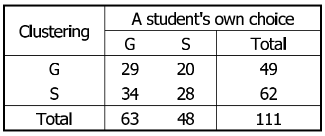
(c) Partition by the class managing method:

Class E: exercise- and practice-based lecture
Class T: test-based lecture

A partition (a) of Class S and Class G is examined by a clustering algorithm using both the item-type and the text-type questionnaire [HW03] of only IQ. Since this algorithm requires representative vectors (pseudo documents), they are obtained from the same questionnaire by graduate students (or the senior students whose jobs in future were decided). Then clusters are automatically generated.

The result of this clustering compared with student's own choice are shown in Table 3.4. The characteristics extracted by discriminant analysis are shown in Table 3.5.

Table 3.4: Classification of Class G and Class S



Clustering	A student's own choice		
	G	S	Total
G	29	20	49
S	34	28	62
Total	63	48	111

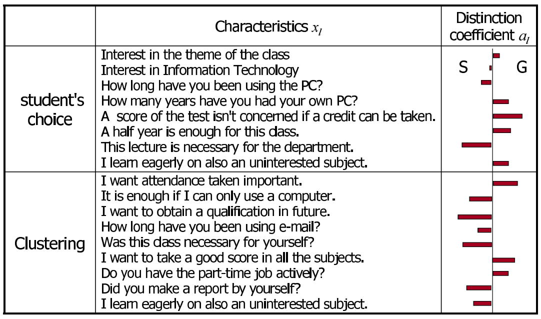
The prediction error for classification is summarized in Table 3.6, where the prediction error is calculated by the difference between the automatic clustering and student's choice given by the item-type questionnaire of FQ.

3.3.2 Verification of class model by IQ and FQ

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

(1) Scores of students We expect to explain the scores of the midterm test (MT) and of final test (FT)(as

Table 3.5: Characteristics of Class G and Class S



	Characteristics x_l	Distinction coefficient a_l
student's choice	Interest in the theme of the class	
	Interest in Information Technology	
	How long have you been using the PC?	
	How many years have you had your own PC?	
	A score of the test isn't concerned if a credit can be taken.	
	A half year is enough for this dass.	
	This lecture is necessary for the department.	
	I leam eagerly on also an uninterested subject.	
Clustering	I want attendance taken important.	
	It is enough if I can only use a computer.	
	I want to obtain a qualification in future.	
	How long have you been using e-mail?	
	Was this dass necessary for yourself?	
	I want to take a good score in all the subjects.	
	Do you have the part-time job actively?	
	Did you make a report by yourself?	
	I leam eagerly on also an uninterested subject.	

Classifying error rate23.6%

Discriminant analysis

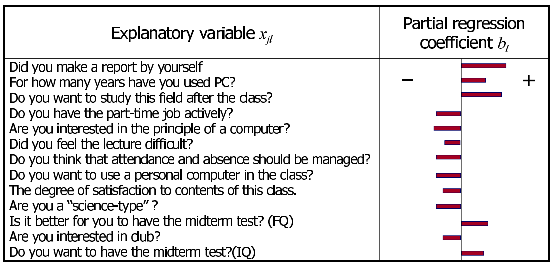
Discriminant function $z = a_0 + a_1x_1 + a_2x_2 + \cdots + a_px_p$ $\begin{cases} z \geqq 0 & d \in \text{class G} \\ z < 0 & d \in \text{class S} \end{cases}$

Table 3.6: Prediction error for partition

(a) S or G	(b) H or L	(c) E or T
0.40	0.40	0.33

intermediate criterion variables) by the item-type questionnaire (as explanatory variables) of IQ and FQ. The result by the multiple linear regression analysis is shown in Table 3.7. Summarized sentences [IIGH02] extracted from the text-type questionnaire of IQ and FQ based on the scores are shown in Table 3.8.

Table 3.7: Explanation of scores by item-type questionnaire



Explanatory variable x_{jl}	Partial regression coefficient b_l
Did you make a report by yourself	
For how many years have you used PC?	
Do you want to study this field after the dass?	
Do you have the part-time job actively?	
Are you interested in the principle of a computer?	
Did you feel the lecture difficult?	
Do you think that attendance and absence should be managed?	
Do you want to use a personal computer in the class?	
The degree of satisfaction to contents of this dass.	
Are you a "science-type" ?	
Is it better for you to have the midterm test? (FQ)	
Are you interested in dub?	
Do you want to have the midterm test?(IQ)	

Multiple linear regression analysis

Contribution ratio=0.742

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

(2) Degree of satisfaction Similar to the above experiment, the item-type questionnaire (as explanatory variables) of IQ and FQ can explain the degree of satisfaction (as criterion variables) for the contents of topics and for the class management as shown in Table 3.6.

(3) Favorite partition The reasons why the students choose Class G or S, and Class E or T as their favorite partition (as criterion variables) are shown in Table 3.10 and 3.11, respectively.

The important sentences extracted from the text-type questionnaire are shown in Appendix A depending on partition of the students.

Table 3.8: Summarized sentences extracted from text-type questionnaire

Score	Example of Sentences
High Over 70	- I think that this class is the one for giving interest to a computer. - Since I was interested in the structure of a computer, I wants to participate the lecture eagerly. However, since I have almost no prior knowledge, I am worried in the ability to catch up with a class. - Since I think that deep understanding in the field of information technology cannot be obtained without knowledge of computer structure, I want to learn firmly at this opportunity.
Low Under 69	- I can only use a few functions of computers. For example, Internet, Excel or Word etc. - I cannot effectively use a personal computer. Therefore, I think that I want to know various things related to computers. - I cannot imagine the contents of the class from the name of subject "Introduction to computer science" well. And this subject would be uninterested for me.

Table 3.9: Explanation of degree of satisfaction by item-type questionnaire

Satisfaction with Contents of the class

Explanatory variable	Partial regression coefficient	t - value
FQ: I want to study this field after the class.	1.5	5.6
FQ: The reports are necessary for every week.	1.0	5.4
FQ: I actively attended the dass.	0.9	4.2
FQ: This lecture is necessary for our department.	0.8	4.1
FQ: I had been interested in contents of a lecture.	0.9	3.9
IQ : I want to have a qualification related to information technologies.	-1.0	-3.6
IQ : Attendance should be taken.	-0.5	-3.5
IQ : How many days in a week in average do you come to university?	-1.3	-3.2
IQ : I don't care if I lost a credit.	0.9	3.0
IQ : I prefer science to literature.	0.4	2.8

Contribution ratio=0.85

Satisfaction for the Class Management

Explanatory variable	Partial regression coefficient	t - value
FQ: The reports are necessary for every week.	1.6	5.1
FQ: The degree of interest of the contents of this dass.	0.2	4.1
IQ : I clearly have an object to learn in this class.	-1.4	-3.8
FQ: I am interested in the contents of this class.	1.5	3.7
IQ : I finished this dass even if it is optioned.	1.3	3.3
IQ : This class is sufficient for a half year.	1.2	3.1
IQ : I don't care even if I lost a credit.	1.5	3.0
IQ : I checked a syllabus.	0.6	2.7
FQ: I finished this class even if it is optioned.	0.9	2.6
FQ: I want to have a qualification related to information technologies.	-0.9	-2.6
FQ: I can use the most of a PC by this dass.	-0.9	-2.2
FQ: I made the reports by myself.	-0.8	-2.0

Contribution ratio= 0.60

3.4 Discussions

(1) It is difficult to predict the student's final score by only IQ, and there is no explanation capability by the linear regression analysis. This can be, however, thought probably to be a natural result.

(2) Although the prediction error rates of the partition problems are 30 - 40[%] for which only the item-type questionnaire of IQ are used, combining with the important sentences extracted from the text-type questionnaire gives useful information for managing the class. According to the characteristics of each class, we can improve the quality of education.

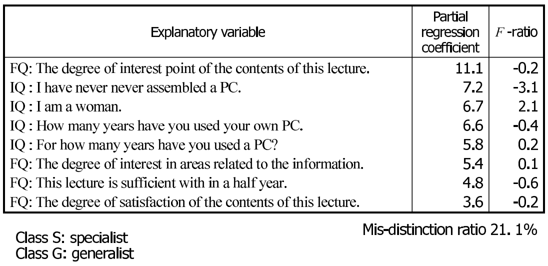
(3) The student's own choice is insufficient for partition of Class G and Class S.

(4) It is possible to explain the student's final score by IQ and FQ. However, it is difficult to get the suggestion to improve the student's score, although we can get a student's tendency.

(5) It is a little difficult to explain the degree of satisfaction regarding the class management, but easy to explain that regarding the contents of topics by IQ and FQ.

(6) It is possible to explain the favorite partition to the students by IQ and FQ.This suggest us a proper partition to the next year by taking into

Table 3.10: Explanation of favorite partition Class G or Class S

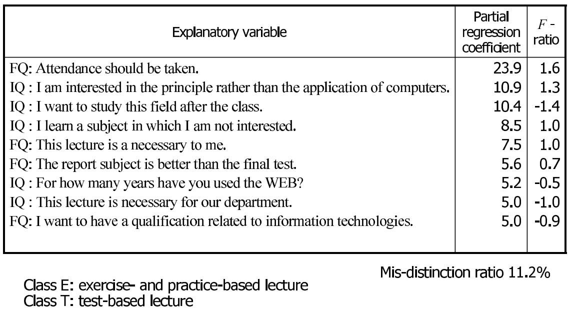


Explanatory variable	Partial regression coefficient	F -ratio
FQ: The degree of interest point of the contents of this lecture.	11.1	-0.2
IQ : I have never never assembled a PC.	7.2	-3.1
IQ : I am a woman.	6.7	2.1
IQ : How many years have you used your own PC.	6.6	-0.4
IQ : For how many years have you used a PC?	5.8	0.2
FQ: The degree of interest in areas related to the information.	5.4	0.1
FQ: This lecture is sufficient with in a half year.	4.8	-0.6
FQ: The degree of satisfaction of the contents of this lecture.	3.6	-0.2

Mis-distinction ratio 21. 1%

Class S: specialist
Class G: generalist

Table 3.11: Explanation of favorite partition Class E or Class T



Explanatory variable	Partial regression coefficient	F -ratio
FQ: Attendance should be taken.	23.9	1.6
IQ : I am interested in the principle rather than the application of computers.	10.9	1.3
IQ : I want to study this field after the class.	10.4	-1.4
IQ : I learn a subject in which I am not interested.	8.5	1.0
FQ: This lecture is a necessary to me.	7.5	1.0
FQ: The report subject is better than the final test.	5.6	0.7
IQ : For how many years have you used the WEB?	5.2	-0.5
IQ : This lecture is necessary for our department.	5.0	-1.0
FQ: I want to have a qualification related to information technologies.	5.0	-0.9

Mis-distinction ratio 11.2%

Class E: exercise- and practice-based lecture
Class T: test-based lecture

account causal relations obtained in this year.

3.5 Conclusions and future works

It can be concluded that we obtain useful information to improve the class management by student questionnaire with both the item-type and the text-type. The result shows verification of the class model for "Introduction to computer science". The degree of satisfaction for the students should be investigated in detail as a future work. Questionnaire 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.

4 Concluding Remarks

We have shown the effective way for knowledge discovery from questionnaire by combining data mining and text mining techniques. One of the most remarkable points is to construct a questionnaire with both fixed format and free format and to simultaneously process them. The effective algorithm to extract the important sentences from questionnaire with free format (text) is also provided. A model which simply exhibits the real problem and the design of the questionnaire based on this model are important to successfully apply this method to actual problems. We have applied the method to improve the quality of education. Although there are many problems, we obtain effective information which leads to faculty developments. The developments of other algorithms such as data mining technique for classification and clustering which should be added to this method is a further research. ,

Acknowledgement

The author would like to thank Mr. T. Ishida, Mr. J. Itoh, Dr. T. Sakai, and Dr. M. Gotoh for their helpful and fruitful discussions.

References

[BYRN99] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, ACM Press, New York, 1999.

[CH01] 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.

[GIH03] 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.

[GITSH03] M. Gotoh, J. Itoh, T. Ishida, T. Sakai, and S. Hirasawa, "A method to analyze a set of documents based on Bayesian statistics," (in Japanese) *Proc. of 2003 Fall Conference on Information Management*, JASMIN, pp.28-31, Hakodate, Nov. 2003.

[GSIIH03] 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.

[HC03] 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.

[HIASG04] S. Hirasawa, T. Ishida, H. Adachi, T. Sakai, and M. Goto, "Classification and clustering methods for documents and their application to analyses of student questionnaires," *submitted to ER2004*.

[HIIGS03] S. Hirasawa, T. Ishida, J. Ito, M. Goto, and T. Sakai, "Analyses on student questionnaires with fixed and free formats," (in Japanese) *Proc. of Comp. Edu. JUCE*, pp.144-145, Sept. 2003.

[Hofmann99] T. Hofmann, "Probabilistic latent semantic indexing," *Proc. of SIGIR'99*, ACM Press, pp.50-57, 1999.

[IIGH02] J. Itoh, T. Ishida, M. Gotoh, and S. Hirasawa, "A method for extracting important sentences using co-occurrence similarities between words ," (in Japanese) *IEICE 2002 FIT*, pp.83-84, Tokyo, Sept. 2002.

[IIGSH03-a] J. Itoh, T. Ishida, M. Gotoh, T. Sakai, and S. Hirasawa, "Knowledge discovery in documents based on PLSI," (in Japanese) *IEICE 2003 FIT*, pp.83-84, Ebetsu, Sept. 2003.

[IIGSH03-b] T. Ishida, J. Itoh, M. Gotoh, T. Sakai, and S. Hirasawa, "A model of class and its verification," (in Japanese) *Proc. of 2003 Fall Conference* *on Information Management*, JASMIN, pp.226-229, Hakodate, Nov. 2003.

[ISH03] 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.

[Sakai99] T. Sakai, et al., “BMIR-J2: A Test Collection for Evaluation of Japanese Information Retrieval Systems,” *ACM SIGIR Forum*, Vol.33, No.1, pp.13-17, 1999.

[Salton71] G. Salton, The SMART Retrieval System - Experiments in Automatic Documents Processing, Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

[SIGIH03] T. Sakai, J. Itoh, M. Gotoh, T. Ishida, and S. Hirasawa, “Efficient analysis of student questionnaires using information retrieval techniques,” (in Japanese) *Proc. of 2003 Spring Conference on Information Management*, JASMIN, pp.182-185, Tokyo, June 2003.

[SIGH04] T. Sakai, T. Ishida, M. Gotoh, and S. Hirasawa, “A student questionnaires analysis system based on natural language expressions,” (in Japanese) *submitted to IEICE 2004 FIT.*

[Rissanen83] J. Rissanen, “A universal prior for integers and estimation by minimum description length,” *Ann. Statist.* vol.11, no.22, pp416-431, 1983.