Knowledge Acquisition from Documents with both Fixed and Free Formats*

Shigeichi Hirasawa
Department of Industrial and
Management Systems Engineering
School of Science and Engineering
Waseda University, Japan
hirasawa@hirasa.mgmt.waseda.ac.jp

Wesley W. Chu
Computer Science Department
School of Engineering and Applied Science
University of California, Los Angeles, U.S.A.
wwc@cs.ucla.edu

Abstract – *Based on techniques in information retrieval, we discuss methods for knowledge acquisition from documents composed of both fixed and free formats. The documents with the fixed format imply items such as those of selecting one from sentences, words, symbols, or numbers. While the documents with the free format are the usual texts. In this paper, starting with an item-document matrix and a term-document matrix used for the representation of a document set, we propose a new method for knowledge acquisition taking simultaneously into account of both fixed and free formats. A method based on the Probabilistic Latent Semantic Indexing (PLSI) model is used for clustering a set of documents. The proposed method is applied to a document set given by questionnaires of students which is taken for the purpose of faculty development. We show the effectiveness of the proposed method compared to the conventional method.*

Keywords: Information Retrieval, PLSI, Clustering.

1 Introduction

Recent developments in text mining techniques enable us to handle a large amount of documents. Based on techniques in information retrieval, methods for knowledge acquisition from documents composed of both fixed and free formats are discussed. The documents with the fixed format are answers to multiple choice questions (such as those of selecting one from sentences, words, symbols, or numbers) called items in this paper. While the documents with the free format are the usual texts. We can find such documents in technical paper archives, questionnaires, or knowledge sharing systems. In the case of the paper archives, the documents with the fixed format (items) correspond to the name of authors, the name of journals, the year of publication, the name of publishers, the name of countries, and so on.

In this paper, as is found in the traditional vector space model of information retrieval systems, a co-occurrence matrix is used for the representation of a document set. The documents with the fixed format are represented by an item-document matrix $G = [g_{mj}]$, where g_{mj} is the selected result of the m-th item (i_m) in the j-th document (d_j). The documents with the free format are also represented by a term-document matrix $H = [h_{ij}]$, where h_{ij} is the frequency of the i-th term (t_i) in the j-th document (d_j). The dimensions of matrices G and H are $I * D$, and $T * D$, respectively. Both matrices are compressed into those with smaller dimensions by the probabilistic decomposition in PLSI (Probabilistic Latent Semantic Indexing) model [4] similar to Single Valued Decomposition (SVD) in LSI (Latent Semantic Indexing) [2]. The unobserved states are denoted by z_k, where $k = 1, 2, \cdots, K$, and $K \leq \max\{T, D\}$. Introducing a weight λ ($0 \leq \lambda \leq 1$), the log-likelihood function corresponding to the sum of λG and $(1-\lambda)H$ is maximized by the EM algorithm [1]. Then we obtain the probabilities $\Pr(z_k)$ ($k = 1, 2, \cdots, K$), and the conditional probabilities $\Pr(t_i|z_k)$, and $\Pr(d_j|z_k)$. Using these probabilities, $\Pr(i_m, d_j)$ and $\Pr(t_i, d_j)$ are derived. A similarity function between d_j and $d_{j'}$ (or query q) can be defined in the usual way, i.e., by cosine, or by inner product. Then we can construct a clustering system, where the latent state z_k plays an important role in clustering systems based on PLSI model. Furthermore, by these preparations, we can also construct ranking systems or classification systems.

As an experiment, we apply the proposed method to a document set given by questionnaires of students [6], where the students are the members of classes for one of the present authors. The contents of the questionnaires consist of questions as the fixed format: e.g., Do you know the meanings of the term "ubiquitous"? Its answer is given by selecting the number $1, 2, \cdots, 5$, where the number is larger as knowing it better, and questions as the free format: e.g., What kind of area in computer

*A part of the work leading to this paper was done at UCLA during a sabbatical year of S.H. as a visiting faculty in 2002.

technology are you interested in ? (within 250-300 Chinese and Japanese character). By the proposed method, we can reasonably divide students into two classes prior to starting the classes. The characteristics of each class of students are clarified by statistical analysis, and it helps to manage the class by selecting objects, topics, or examples for each class. We show the effectiveness of the proposed method and its better performance compared to the conventional methods. A final object of this experiment is to find helpful leads to the faculty development.

2 Information Retrieval Model

In early information retrieval systems, some of which are still in use for commercial purposes adopt index terms (keywords) which is refereed to as (1) Boolean model. To avoid over-simplification by this model, and to enable ranking the relevant document together with automatic indexing, (2) Vector Space Model (VSM) was proposed in early '70s [7].

To improve the performance of VSM, Latent Semantic Indexing (LSI) model was studied by reducing the dimension of the vector space using Single Valued Decomposition (SVD) [2].

As a similar approach, Probabilistic Latent Semantic Indexing (PLSI) model based on a statistical latent class model has recently been proposed by T. Hofmann [4].

2.1 The Vector Space Model (VSM)

The VSM uses non-binary weights in the i-th (index) term (t_i) in the j-th document (d_j) for a given document set $\mathcal{D}$ and queries (q).

[Vector Space Model]

Let $\mathcal{T}$ be a term set used for representing a document set $\mathcal{D}$. Let t_i $(i = 1, 2, \cdots, T)$ be the i-th term in $\mathcal{T}$, where $\mathcal{T}$ is a subset of the all term set $\mathcal{T}_0$ appeared in $\mathcal{D}$, and d_j $(j = 1, 2, \cdots, D)$, the j-th document in $\mathcal{D}$. Then a term-document matrix $A = [a_{ij}]$ is given by the weight $w_{ij} \geq 0$ associated with a pair (t_i, d_j). □

In the VSM, the weight w_{ij} is usually given by so-called the tf-idf value, where tf stands for the term frequency, and idf, the inverse document frequency. When the number of the i-th term (t_i) in the j-th document (d_j) is $f_{i,j}$, then $tf(i,j) = f_{i,j}$. When the number of documents in $\mathcal{D}$ for which the term t_i appears is $df(i)$, then $idf(i) = \log(D/df(i))$. The tf-idf value is calculated by their product. As the result, for the VSM the weight w_{ij} is given by

$$w_{ij} = tf(i,j) \cdot idf(i) \tag{1}$$

and is equal to a_{ij}.

The i-th row of the matrix A represents the frequency vector of the term t_i in $\mathcal{D}$, and the j-th column, that of d_j in $\mathcal{T}$, we use the term vector $\boldsymbol{t}_i$ and the document vector $\boldsymbol{d}_j$ as

$$\boldsymbol{t}_i = (a_{i1}, a_{i2}, \cdots, a_{iD}) \tag{2}$$

$$\boldsymbol{d}_j = (a_{1j}, a_{2j}, \cdots, a_{Tj})^{\mathrm{T}} \tag{3}$$

where $\boldsymbol{x}^{\mathrm{T}}$ is the transposed vector of $\boldsymbol{x}$. Similar to the vector $\boldsymbol{d}_j$, we also use a query vector $\boldsymbol{q}$ for a query q by the weight associated with the pair (t_i, q) as follows:

$$\boldsymbol{q} = (q_1, q_2 \cdots, q_T)^{\mathrm{T}} \tag{4}$$

Then we can define the similarity $s(q, d_j)$ between q and d_j. In the case of measuring it by cosine of the angle between the vectors $\boldsymbol{q}$ and $\boldsymbol{d}_j$, we have

$$s(q, d_j) = \frac{\boldsymbol{q}^{\mathrm{T}} \boldsymbol{d}_j}{|\boldsymbol{q}||\boldsymbol{d}_j|} \tag{5}$$

2.2 The Latent Semantic Indexing (LSI) Model

The LSI model is accomplished by mapping each document and query vector into a lower dimensional space by using SVD [2].

[Truncated LSI Model]

Let a term-document matrix $A \in \mathcal{R}^{T*D}$ is given by eq.(1). Then the matrix A is decomposed into A_K by the truncated SVD as follows:

$$\begin{aligned} A \to A_K &= \left(U_K \hat{U}\right) \begin{pmatrix} \Sigma_K & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_K^{\mathrm{T}} \\ \hat{V} \end{pmatrix} \\ &= U_K \Sigma_K V_K^{\mathrm{T}} \end{aligned} \tag{6}$$

where

$$U_K \in \mathcal{R}^{D*K}$$
$$\Sigma_K \in \mathcal{R}^{K*K}$$
$$V_K \in \mathcal{R}^{T*K}$$

and

$$K \leq p \leq \max\{T, D\}.$$

In eq.(6), $|A - A_K|_F$ is minimized for any K, where p is the rank of A, and $|\cdot|_F$ is the Frobenius matrix norm. □

Let the term-document matrix A is given by the reduced rank matrix A_K by the truncated SVD, then a query vector $\boldsymbol{q} \in \mathcal{R}^{T*1}$ in eq.(4) is represented by $\hat{\boldsymbol{q}} \in \mathcal{R}^{K*1}$ in a space unit dimension K:

$$\hat{\boldsymbol{q}} = \Sigma_K^{-1} \boldsymbol{q} \ \in \mathcal{R}^{K*1} \tag{7}$$

[1] then $s(q, d_j)$ is also computed by

$$s(q, d_j) = \frac{\hat{\boldsymbol{q}}^{\mathrm{T}} \hat{\boldsymbol{d}}_j}{|\hat{\boldsymbol{q}}||\hat{\boldsymbol{d}}_j|} \tag{8}$$

[1] In the other case, $\hat{\boldsymbol{q}} = \Sigma_K^{-1} U_K^{\mathrm{T}} \boldsymbol{q} \in \mathcal{R}^{K*1}$.

where

$$\hat{\boldsymbol{d}}_j = \Sigma_K V_K^{\mathrm{T}} \boldsymbol{e}_j \ \in \mathcal{R}^{K*1}$$

and

$$\boldsymbol{e}_j = (\overset{1}{0}, \overset{2}{0}, \overset{\cdots}{\cdots}, 0, \overset{j}{1}, 0, \cdots, 0) \tag{9}$$

is the j-th canonical vector.

2.3 The Probabilistic Latent Semantic Indexing (PLSI) Model

In contrast to the LSI model, the PLSI model is based on mixture decomposition derived from a latent state model. A term-document matrix $A = [a_{ij}]$ is directly given by term frequency $tf(i,j) = f_{i,j}$, i.e., a_{ij} is the number of a term t_i in a document d_j. In the LSI model, the matrix $A \in \mathcal{R}^{T*D}$ is decomposed into A_K with smaller dimension by SVD, using principal eigenvectors. While in the PLSI model, the matrix A is probabilistically decomposed into K unobserved states, where the k-th state is denoted by $z_k \in \mathcal{Z}$, and $\mathcal{Z}$, a set of satates.

First, we assume both (i) an independence between pairs (t_i, d_j) , and (ii) a conditional independence between t_i and d_j, i.e., the term t_i and the document d_j are independent conditioned on the latent state z_k.

The joint probability of t_i and d_j, $\Pr(t_i, d_j)$ is given by

$$\begin{aligned} \Pr(t_i, d_j) &= \sum_{z_k \in \mathcal{Z}} \Pr(d_j) \Pr(t_i|z_k) \Pr(z_k|d_j) \quad (10) \\ &= \sum_{z_k \in \mathcal{Z}} \Pr(z_k) \Pr(t_i|z_k) \Pr(d_j|z_k) \quad (11) \end{aligned}$$

The number of the set of the states, or the cardinality of $\mathcal{Z}$, $||\mathcal{Z}|| = K$ satisfies

$$K \leq \max\{T, D\} \tag{12}$$

[PLSI Model]

Let a term-document matrix $A = [a_{ij}]$ is 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:

$$L = \sum_{i,j} a_{ij} \log \Pr(t_i, d_j) \tag{13}$$

□

The maximization technique usually used for the likelihood function is the Expectation Maximization (EM) algorithm. The EM algorithm performs iteratively E-step and M-step as follows:

[EM algorithm]

According to eq.(11), the maximum value of eq.(13) is computed by alternating E-step and M-step until it converges.

E-step:

$$\Pr(z_k|t_i, d_j) = \frac{\Pr(z_k) \Pr(t_i|z_k) \Pr(d_j|z_k)}{\sum_{k'} \Pr(z_{k'}) \Pr(t_i|z_{k'}) \Pr(d_j|z_{k'})} \tag{14}$$

M-step:

$$\begin{aligned} \Pr(t_i|z_k) &= \frac{\sum_j a_{ij} \Pr(z_k|t_i, d_j)}{\sum_{i',j} a_{i'j} \Pr(z_k|t_{i'}, d_j)} \quad (15) \\ \Pr(d_j|z_k) &= \frac{\sum_i a_{ij} \Pr(z_k|t_i, d_j)}{\sum_{i,j'} a_{ij'} \Pr(z_k|t_i, d_{j'})} \quad (16) \\ \Pr(z_k) &= \frac{\sum_{i,j} a_{ij} \Pr(z_k|t_i, d_j)}{\sum_{i,j} a_{ij}} \quad (17) \end{aligned}$$

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

To avoid overtraining to the data in the EM algorithm, a temperature variable β ($\beta > 0$) is used, that is called a Tempered EM (TEM) [4]. At the E-step for the TEM, the numerator and the each term of the denominator of eq.(14) are replaced by those to the power of β.

3 Formats of Documents

We assume that the documents discussed in this paper are consist of the following two formats.

(1) The fixed format

The fixed format documents are represented by a set of items. An item-document matrix $G = [g_{mj}]$, $g_{mj} \in \{0,1\}^{I*T}$, where g_{mj} is the value of the i-th item (i_m) in the j-th document (d_j), and the number of the items is I. The value of g_{mj} is given by

$$g_{mj} = \begin{cases} 0, & i_m \text{ is absent in } d_j; \\ 1, & i_m \text{ is present in } d_j \end{cases} \tag{18}$$

(2) The free format

The free format documents are represented by a set of terms. A term-document matrix $H = [h_{ij}]$, $h_{ij} \in \{0, 1, \cdots\}$ is a $T * D$ matrix whose elements are non-negative integers. The value of h_{ij} is given by

$$h_{ij} = tf(i,j) \tag{19}$$

4 Proposed Methods

We propose a new clustering method based on PLSI model. For the purpose of simultaneously maximizing the log-likelihood function corresponding to G and H, the idea of the joint probabilistic model [1] is applied.

Let the number of the latent states $||\mathcal{Z}|| = K$, we represent the method for obtaining S clusters, where $K \geq S$.

(1) Choosing a proper integer K, compute the log-likelihood function :

$$L = \sum_j \left[\lambda \sum_m \frac{g_{mj}}{\sum_{m'} g_{m'j}} \log \sum_k \Pr(i_m|z_k) \Pr(z_k|d_j) + (1-\lambda) \sum_i \frac{h_{ij}}{\sum_{i'} h_{i'j}} \log \sum_k \Pr(t_i|z_k) \Pr(z_k|d_j) \right] \quad (20)$$

where λ is a weight such that $0 \leq \lambda \leq 1$. To maximize it, the TEM algorithm is used. Then we have the probabilities $\Pr(i_m|z_k)$, $\Pr(t_i|z_k)$, and $\Pr(d_j|z_k)$.

(2) By the probability $\Pr(z_k|d_j)$, decide the state z_{k_0} for d_j as

$$\max_k \Pr(z_k|d_j) = \Pr(z_{k_0}|d_j) \rightarrow d_j \in z_{k_0} \quad (21)$$

If $S = K$, then let d_j be a member of z_{k_0}.

(3) If $S < K$, then compute a similarity measure $s(z_k, z_{k'})$:

$$s(z_k, z_{k'}) = \frac{\boldsymbol{z}_k^{\mathrm{T}} \boldsymbol{z}_{k'}}{|\boldsymbol{z}_k||\boldsymbol{z}_{k'}|} \quad (22)$$

$$\boldsymbol{z_k} = (\Pr(t_1|z_k), \Pr(t_2|z_k), \cdots, \Pr(t_T|z_k))^{\mathrm{T}} \quad (23)$$

and use the group average distance method with the similarity measure $s(z_k, z_{k'})$ for agglomeratively clustering the states z_k's until the number of clusters becomes S. Then we have S clusters, and the members of each cluster are those of a cluster of states.

(4) Analyze the characteristics of the members of each cluster by statistical techniques. Then we have knowledge acquisition.

5 Experimental Results

We shall demonstrate experimental results to show the effectiveness of the proposed method.

5.1 Student Questionnaires for Class

(a) Class data

The objects of this experiment are students of two classes as shown in Table 1 in which one of the present authors teaches. The students of the former are the second year of Undergraduate School, Department of Industrial and Management Systems Engineering. The name of the subject is "Introduction to Computer Science", which is called Class CS. The class data consist of Initial Questionnaires (IQ), Final Questionnaires (FQ), Mid-term Test (MT), Final Test (FT), and Technical Report (TR). The students of the latter are from the second through the fourth year of all undergraduate schools. The name of the subject is "Introduction to Information Society", which is called Class IS. The class data also consist of IQ, FQ, and of First Report(R1), Second Report(R2), Third Report(R3) and Fourth Report(R4).

Table 1: 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

The data of the j-th student includes the items and texts in IQ and FQ, and the numerical data of scores (normalized in [0,100]) in MT, FT, and TR for Class CS, and in R1, R2, R3, and R4 for Class IS.

(b) Contents of questionnaires

The IQ is questionnaires composed of items with fixed format and of texts with free format. These contents of IQ are briefly shown in Table 2.

The same Initial Questionnaires(IQ) are applied to both classes. The IQ of the j-th student is refereed to as the j-th document.

Table 2: Contents of IQ

Format	Number of questions	Examples
fixed (item)	7 major questions[2]	- For how many years do you use 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.

The number of items $||\mathcal{I}|| = I = 3293$, and that of terms $||\mathcal{T}|| = T = 3993$ in this experiment.

5.2 Experiment 1 (E1)

First, the documents of the students in Class CS and those in Class IS are merged. Then the merged documents are divided into two classes ($S = 2$) by the proposed method. Assume that the characteristics of

the students in Class CS are clearly different from that in Class IS, since their majors are obviously distinct. Then we evaluate whether each member of the divided two classes coincides with that of the original classes , i.e., Class CS and Class IS (assuming to be the true classes).

5.3 Results of E1

Applying the proposed method to the student questionnaires (IQ) of both Class CS and Class IS, results obtained are shown in Figures 1 and 2, and Table 3.

(1) As shown in Fig.1, there is the optimum value of λ for given K, where x-axis is λ and y-axis is the rate of clustering error $C(e)$, which is 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. We see that $C(e)$ decreases as K increases.

(2) The dendrogram of clustering states z_k's by the group average distance method is depicted in Fig.2, which shows the members of each cluster are almost the same for any K, i.e., there always exists one state (the right most one in Fig.2). This figure tells us the process at (3) of Section 4.

(3) The performance of the VSM is inferior compared to the proposed method as shown in Fig. 1. The proposed method has well enough performance such that it performs $C(e) < 0.1$. If the K-means method is used at the clustering step (at (3) of Section 4), it cannot improve the performance of clustering. For example, for the case of $S = K = 2$, it gives $C(e) \simeq 0.411$.

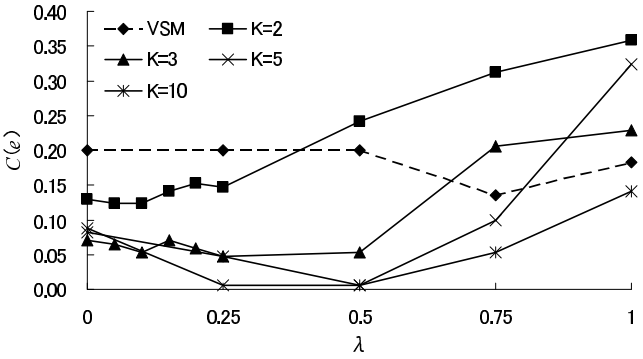

Figure 1: The rate of clustering error $C(e)$ for K

(4) Differences of characteristics of the students between Class CS and Class IS should be evaluated for knowledge acquisition. One of the approaches to clarify them is to apply statistical techniques such as multivariate analysis[3]. As an example, applying discriminant analysis techniques, results are shown in Table 3.

5.4 Experiment 2 (E2)

According to the validity of E1, we divide the students of Class CS into two classes prior to starting the class

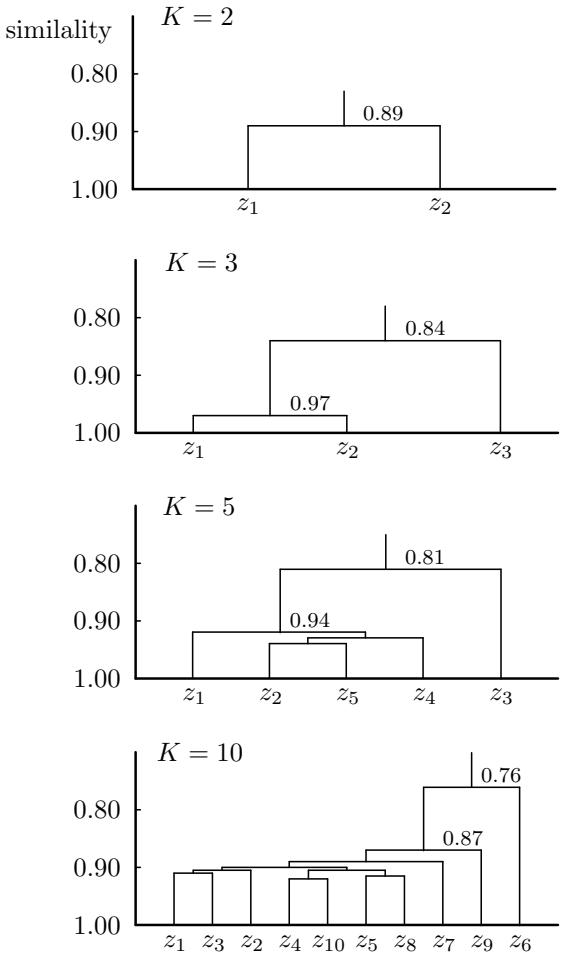

Figure 2: The dendrogram of clustering states for E1

by only taking IQ from students. It can be expected that students in the same class have similar characteristics. If we know the characteristics of the students of each class, we can effectively use this knowledge for the management of the class.

5.5 Results of E2

Applying the proposed method to the student questionnaires (IQ) for only Class CS, results obtained are shown in Tables 4 and 5.

(1) As a special requirement for this problem, it is desirable to divide students into two classes with almost the same number. It is very hard to derive the optimum solution, because it requires a large amount of computational work. For this problem, we choose empirically better classes. The rate $C(e)$ between the cases K and K' is shown in Table 4, which implies that the members of a class vary for different K.

(2) Differences of characteristics of students between divided two classes are evaluated for each division which are interpreted in Table 5. The most convenient case for characteristics of students should be chosen.

Table 3: 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.

Table 4: The rate $C(e)$ between the cases K and K'

	$K=2$	$K=3$	$K=5$	$K=10$
$K'=2$	0.00%	20.74%	41.48%	40.00%
$K'=3$		0.00%	41.48%	34.07%
$K'=5$			0.00%	29.63%
$K'=10$				0.00%

5.6 Discussions on Experiments

(1) The present contents of Initial Questionnaires (IQ) are proper for E1, they should, however, be improved for E2.

(2) Performance of the proposed method is dependent on the structure of characteristics of the students.

(3) If we derive multiple solutions for dividing students into two classes, it is possible to choose better division from a viewpoint of class management.

6 Concluding Remarks

We proposed a method for clustering a set of documents with fixed and free formats. By weighting the matrix for items G and that for texts H by λ, and $(1-\lambda)$, respectively, and by agglomeratively clustering latent states with a similarity measure, we have shown that it is possible to reasonably divide a set of documents with any number of clusters, although the performance is dependent on the structure of given documents.

As a further study, we should clarify how to determine the number of states K and the value of a weight λ properly. A method for abstracting the characteristics of each cluster from the viewpoint of a probabilistic model is also an important further investigation[5].

Table 5: 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.

Acknowledgement

One of the present authors S.H. wishes to thank Mr.J. Itoh and Mr.T. Ishida for their fruitful discussions and helpful works especially for processing experimental data. He also likes to thank Drs.T. Sakai and M. Gotoh for their valuable comments.

References

[1] 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.

[2] S. Deerwester, S.T. Dumais, G.W.Furnas, T.K. Landauer, and R. Harshman, Indexing by Latent Semantic Analysis, J. of the Society for Information Science, 41, pp.391-407, 1990.

[3] M. Gotoh, T. Sakai, J. Itoh, T. Ishida, and S. Hirasawa, "Knowledge discovery form questionnaires with items and texts," (in Japanese) to appear in Proc. of 2003 PC Conference, Kagoshima, Japan, Aug. 2003.

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

[5] J. Itoh, T. Ishida, M. Gotoh, T. Sakai, and S. Hirasawa, "Knowledge discovery in documents based on PLSI," (in Japanese) to appear in Proc. of FIT, Ebetsu, Japan, Sept. 2003.

[6] 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, June 2003.

[7] G. Salton, The SMART Retrieval System, Prentice Hall, 1971.