A Short Course at Tamkang University
Taipei, Taiwan, R.O.C., March 7-9, 2006

Classification and clustering methods for documents by probabilistic latent semantic indexing model

Shigeichi Hirasawa

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

Related area: data base, information retrieval, knowledge acquisition, text mining, classification, clustering

Abstract — Based on information retrieval model especially probabilistic latent semantic indexing (PLSI) model, we discuss methods for classification and clustering of a set of documents. A method for classification is presented and is demonstrated its good performance by applying to a set of benchmark documents with free format (text only). Then the classification method is modified to a clustering method and the clustering method is applied to a set of experimental documents with fixed and free formats to partition into two clusters, where the experimental documents are obtained from student questionnaires. Since the experimental documents are already categorized, the clustering method can be clearly evaluated its performance. The method has better performance compared to the conventional one based on the vector space model.

1. Introduction

Recent development in information retrieval techniques enables us to process a large amount of text data. The information retrieval techniques include classification and clustering techniques for a set of documents [BYRN99]. These techniques are used for not only classical information retrieval systems such as for technical paper archives, but also customer relationship management (CRM), knowledge management system (KMS), or questionnaire analysis (QA) at enterprises for the purpose of market research, or personal management.

In this paper, we discuss classification and clustering techniques based on the probabilistic latent semantic indexing (PLSI) model [Hoffman99], A new classification method [IIGSH03-a] [HC04] for a set of documents is presented using the maximum a posteriori probability criterion similar to the naive Bayesian technique. Using Japanese benchmark documents with free format [Mainichi95], the method is evaluated, and is demonstrated its good performance compared to conventional techniques for relatively small sets of documents, where a document with free format implies the texts. The method successfully uses a characteristic such that the EM algorithm converges a value dependent on an initial value. Then the classification method is modified into a new clustering method. The clustering method is applied to a set of real documents, where real documents are those obtained by student questionnaires which are composed of both fixed and free formats [HIIGS03] [IGH05]. A document with fixed format implies items such as those of selecting one from sentences, words, symbols, or numbers. While a document with free format implies the usual texts. We can find such documents in technical paper archives, questionnaires, or knowledge collaboration. In the case of paper archives, the documents with fixed format (called items in this paper) 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, i.e., discrete concepts.

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 fixed format are represented by an item-document matrix $G=[g_{mj}]$, where g_{mj} is the selected result of the item m (i_m) in the document j (d_j). The documents with free format are also represented by a term-document matrix $H=[h_{ij}]$, where h_{ij} is the frequency of the term i (t_i) in the document j (d_j). The dimensions of matrices G and H are $I\times D$, and $T\times D$, respectively. Both matrices are compressed into those with smaller dimensions by the probabilistic decomposition in PLSI [Hofmann99] similar to the single valued decomposition (SVD) in LSI (latent semantic indexing) [BYRN99]. The unobserved states are z_k ($k=1,2,\ldots,K$). Introducing a weight λ ($0 \le \lambda \le 1$), the log-likelihood function corresponding to matrix $[\lambda GT,\ (1-\lambda)HT]^{\mathrm{T}}$ is maximized by the EM algorithm [CH01], where A^T is the transposed matrix of A. Then we obtain the probabilities $\Pr(z_k)(k=1,2,\ldots,K)$, and the conditional probabilities $\Pr(t_i \mid z_k, i_m)$, and $\Pr(d_j \mid z_k)$. Using these probabilities, $\Pr(i_m, d_j)$ and $\Pr(t_i, d_j)$ are derived. We decide the state for d_j depending on $\Pr(z_k \mid d_j)$, and a similarity function between z_k and z_k' can be defined in the usual way, i.e., by cosine, or by inner product. By these preparations, we use the group average distance method with the similarity measure for agglomeratively clustering the state z_k's until the number of clusters becomes S, where $S \le K$ [HC03].

To show the effectiveness of the methods, first we apply them into the benchmark test document set [Mainichi95] which has been already categorized. Then as an experiment, we apply the proposed method into a document set given by student questionnaires [SIGIH03], where the students are the members of a class (Introduction to computer science, in the second academic year,

undergraduate school) for the present author. The contents of the questionnaires consist of questions answered with fixed format: e.g., Are you interested in wearable computers? (Its answer is yes or no), and questions, with free format: e.g., write your image of computers. Merging the documents of students from two different classes, then the merged documents are partitioned into two categories. We show that each member of the partitioned classes coincides with that of the original classes at high rate. Its better performance is compared to the conventional method based on the vector space model. A final object of this experiment is to find helpful leads to the faculty development [HIASG04] [IIGSH03-b].

2. Information Retrieval Model

Early information retrieval systems adopted (1) Boolean model, and based on index terms (i.e., keywords) some of which are still in use for commercial purposes. To avoid over-simplification by this model, and to enable ranking the relevant document together with automatic indexing, (2) vector space (VS) model was proposed in early '70s [Salton71].

To improve the performance of the VS model, latent semantic indexing (LSI) model was studied by reducing the dimension of the vector space using single valued decomposition (SVD) [BYRN99].

As a similar approach, probabilistic latent semantic indexing (PLSI) model based on a statistical latent class model has recently been proposed by T. Hofmann [Hofmann99]. Information retrieval model are shown in Table 2.1.

Table 2.1: Information retrieval model

Base	Model
Set theory	(Classic) Boolian Model
	Fuzzy
	Extended Boolian Model
Algebraic	(Classical) Vector Space Model (VSM) [7]
	Generalized VSM
	Latent Semantic Indexing (LSI) Model [2]
	Probabilistic LSI (PLSI) Model [4]
	Neural Network Model
Probabilistic	(Classical) Probabilistic Model
	Extended Probabilistic Model
	Inference Network Model
	Bayesian Network Model

2.1 The Vector Space Model (VSM)

The VS model uses non-binary weights in the i-th (index) term (t_i) in the j-th document (d_j) for a given document set D and queries (q).

[Vector Space Model]

Let T be a term set used for representing a document set D. Let t_i (i=1,2,…,T) be the i-th term in T, where T is a subset of the all term set T_0 appeared in D, and d_j (j=1,2,…, D), the j-th document in 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 VS model, 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 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 VS model the weight w_{ij} is given by

$$w_{ij} = tf(i,j) \cdot idf(i) \qquad (2.1)$$

and is equal to a_{ij}.

The i-th row of the matrix A represents the frequency vector of the term t_i in D, and the j-th

column, that of d_j in T, we use the term vector $\boldsymbol{t}_i$ and the document vector $\boldsymbol{d}_j$ as

$$\boldsymbol{t}_i = (a_{i1}, a_{i2}, \ldots, a_{iD}) \quad (2.2)$$

$$\boldsymbol{d}_j = (a_{1j}, a_{2j}, \ldots, a_{Tj})^{\mathrm{T}} \quad (2.3)$$

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:

$$\boldsymbol{q} = (q_1, q_2, \ldots, q_T)^{\mathrm{T}} \quad (2.4)$$

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

$$s(q, d_j) = \frac{\boldsymbol{q}^{\mathrm{T}} \boldsymbol{d}_j}{|\boldsymbol{q}|\,|\boldsymbol{d}_j|}, \quad (2.5)$$

where $|x|$ is the norm of x.

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.

[Truncated LSI Model]

Let an element of a term-document matrix $A \in R^{T \times D}$ be given by eq.(2.1). Then the matrix A is decomposed into A_K by the truncated SVD as follows:

$$A \to A_K = \begin{pmatrix} U_K & \hat{U} \end{pmatrix} \begin{pmatrix} \Sigma_K & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_K^{\mathrm{T}} \\ \hat{V} \end{pmatrix} = U_K \Sigma_K V_K^{\mathrm{T}} \quad (2.6)$$

where

$$U_K \in R^{T \times K}, \quad \Sigma_K \in R^{K \times K}, \quad V_K \in R^{D \times K}$$

and

$$K \le p \le \min\{T, D\}.$$

In eq.(2.6), $|A - A_K|_F$ 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 be given by the reduced rank matrix A_K by the truncated SVD, then a query vector $\boldsymbol{q} \in R^{T \times 1}$ in eq.(2.4) is represented by $\hat{\boldsymbol{q}} \in R^{K \times 1}$ in a space unit dimension K:

$$\hat{\boldsymbol{q}} = \Sigma_K^{-1} \boldsymbol{q} \in R^{K \times 1} \quad (2.7)$$

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

$$s(q, d_j) = \frac{\hat{\boldsymbol{q}}^{\mathrm{T}} \hat{\boldsymbol{d}}_j}{|\hat{\boldsymbol{q}}|\,|\hat{\boldsymbol{d}}_j|} \quad (2.8)$$

where

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

and

$$\boldsymbol{e}_j = (\overset{1}{0}, \overset{2}{0}, \cdots, \overset{j}{0}, 1, 0, \cdots, 0) \quad (2.9)$$

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 $A \in R^{T \times D}$ 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 $z_k \in Z$ (k=1,2,…,K), and Z, a set of states.

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. A graphical epresentation is depicted in Fig. 2.1.

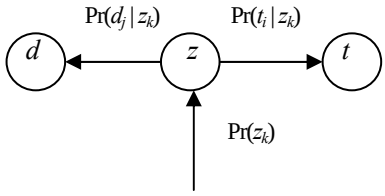

Fig. 2.1: A graphical model for the PLSI model

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

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

$$\Pr(t_i, d_j) = \sum_{z_k \in Z} \Pr(d_j) \Pr(t_i \mid z_k) \Pr(z_k \mid d_j) \tag{2.10}$$

$$= \sum_{z_k \in Z} \Pr(z_k) \Pr(t_i \mid z_k) \Pr(d_j \mid z_k) \tag{2.11}$$

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

$$K \le \max\{T, D\} \tag{2.12}$$

[PLSI Model]

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

$$L = \sum_{i,j} a_{ij} \log \Pr(t_i, d_j) \tag{2.13}$$

■

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.(2.11), the maximum value of eq.(2.13) is computed by alternating E-step and M-step until it converges.

E-step:

$$\Pr(z_k \mid t_i, d_j) = \frac{\Pr(z_k) \Pr(t_i \mid z_k) \Pr(d_j \mid z_k)}{\sum_{k'} \Pr(z_{k'}) \Pr(t_i \mid z_{k'}) \Pr(d_j \mid z_{k'})} \tag{2.14}$$

M-step:

$$\Pr(t_i \mid z_k) = \frac{\sum_j a_{ij} \Pr(z_k \mid t_i, d_j)}{\sum_{i',j} a_{i'j} \Pr(z_k \mid t_{i'}, d_j)} \tag{2.15}$$

$$\Pr(d_j \mid z_k) = \frac{\sum_i a_{ij} \Pr(z_k \mid t_i, d_j)}{\sum_{i,j'} a_{ij'} \Pr(z_k \mid t_i, d_{j'})} \tag{2.16}$$

$$\Pr(z_k) = \frac{\sum_{ij} a_{ij} \Pr(z_k \mid t_i, d_j)}{\sum_{i,j} a_{ij}} \tag{2.17}$$

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

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) [Hofmann99]. At the E-step for the TEM, the numerator and the each term of the denominator of eq.(2.14) are replaced by those to the power of β.

3. Proposed Methods

We propose new classification and clustering methods based on the PLSI model. The methods are strongly dependent on the fact and property that the EM algorithm usually converges to the local optimum solution from starting with an initial value. Hence we use a representative document as the initial value for the EM algorithm. Since the latent states are regarded as concepts in the PLSI model, the state corresponds to the category or the cluster. Consequently, we can state that the methods presented in this paper have good performance for a document set with relatively small size.

3.1 Classification method [IIGSH03-a]

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

[Proposed classification method]

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

$$\boldsymbol{d}^*_k = \frac{1}{n_k} \sum_{\boldsymbol{d}_j \in C_k} \boldsymbol{d}_j \tag{3.1}$$

where n_k is the number of selected documents to compute the representative document vector from C_k.

(2) Compute the probabilities $\Pr(z_k)$, $\Pr(d_j \mid z_k)$

and $\Pr(t_i|z_k)$ which maximizes eq.(2.13) by the TEM algorithm, where $|Z|=K$.

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

$$\max_k \Pr(z_k \mid d_j) = \Pr(z_{\hat{k}} \mid d_j) \Rightarrow d_j \in z_{\hat{k}} \tag{3.2}$$

■

By the algorithm described above, a set of documents is classified into K categories. If we can obtain the K representative documents prior to classification, they are used for $\boldsymbol{d}^*_k$ in eq.(3.1).

3.2 Clustering method [HC03][HIASG04]

Suppose a set of documents to be clustered into S clusters, where the S clusters are denoted by $c_1, c_2, \ldots, c_S$.

[Proposed clustering method]

(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 eq.(2.13) by the TEM algorithm, where $|Z|=K$.

(2) Decide the state $z_{\hat{k}}$ for d_j as

$$\max_k \Pr(z_k \mid d_j) = \Pr(z_{\hat{k}} \mid d_j) \Rightarrow d_j \in z_{\hat{k}} \tag{3.3}$$

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

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

$$s(z_k, z_{k'}) = \frac{\boldsymbol{z}_k^{\mathrm{T}} \boldsymbol{z}_{k'}}{|\boldsymbol{z}_k||\boldsymbol{z}_{k'}|} \tag{3.4}$$

$$\boldsymbol{z}_k = (\Pr(t_1 \mid z_k), \Pr(t_2 \mid z_k), \cdots, \Pr(t_T \mid z_k))^{\mathrm{T}} \tag{3.5}$$

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. ■

By the above algorithm, a set of documents is clustered into S clusters.

4. Experimental Results

We first apply the classification method to the set of benchmark documents [Sakai99], and verify its effectiveness. We then apply the clustering method to the set of student questionnaires as real documents to be analyzed whose answers are written in both fixed and free formats. All documents applied in this paper are written in Japanese.

4.1 Document sets

The document sets which we use as experimental data are shown in Table 4.1.

Table 4.1: Document sets

	contents	format	# words T	amount	categorize	# selected document D_L+D_T
(a)	Articles of Mainichi news paper in '94 [Mainichi95]	free (texts only)	107,835	101,058 (see Table 4.2)	yes (9+1 categories)	300(S=3)
(b)						200～300 (S=2～8)
(c)	questionnaire (see Table 4.3 in detail)	fixed and free (see Table 4.3)	3,993	135+35	Yes (2 categories)	135+35

Table 4.2: Selected categories of newspaper

category	contents	# articles D_L+D_T	# used for training D_L	# used for test D_T
C_1	business	100	50	50
C_2	local	100	50	50
C_3	sports	100	50	50
Total		300	150	150

Table 4.3: 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.

Table 4.4: 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

As shown in Table 4.1, the benchmark data in Japan [Mainichi95],[Sakai99] is a document set composed of 101,058 articles of Mainichi newspaper in '94, which is prepared for. The articles are categorized into 9 categories and the others, dependent on their contents (edited location in newspaper) such as economics, business, sports, or local. (a) is a case of 3 categories as shown in Table 4.2, and (b) are cases of S (=2～8) categories, each of which contains 100～450 articles.

While (c) in Table 4.1 is actual data i.e., student questionnaires for which the present authors want to analyze for obtaining useful knowledge from the data in order to manage the classes. Effective clustering gives a proper class partition depending on students' interests, their levels, their experiences and so on.

4.2 Classification problem: (a)

4.2.1 Experiment conditions of (a)

As shown in Table 4.2, we choose three categories. 100 articles are randomly chosen from each category. The half of them, D_L, is used for training, and the rest of them, D_T, for test.

As baseline classification methods to be compared to the proposed method, the following conventional methods are evaluated, where we call the classification method by the VS model simply as the VS method, that by the LSI model as the LSI method, and that by the PLSI model as the PLSI method.

The VS method:

classified by the cosine similarity measure between the representative document vector and a given document vector in the VS model where a_{ij} is given by eq.(2.1).

The LSI method:

the same as the VS method except that the term-document matrix $[a_{ij}]$ is compressed by SVD in the LSI model, where K=81 which corresponds to the condition that the cumulative distribution rate=70[%].

PLSI method:

the same as VS the method except that the matrix $[\Pr(d_j|z_k)]$ is compressed by the PLSI model, where K=10.

The proposed method uses K=3.

4.2.2 Results of (a)

Table 4.5 for each method shows the classified number of articles $n_{k\hat{k}}$ from category C_k to $C_{\hat{k}}$, hence the number of the diagonal element $n_{k\hat{k}}$ implies that of correct classification. Table 4.6 indicates the classification error rate for each method.

Table 4.5: 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 4.6: Classification error rate

method	classification error [%]
VS method	42.7
LSI method	38.7
PLSI method	20.7
Proposed method	6.0

Classification process by the EM algorithm is shown in Fig. 4.1 for step 1, 4, and 4096. At step 1, almost all document vectors are located in the center of the triangle. Then they move to each state z_k (k=1,2,3) depending on the probability $\Pr(z_k|d_j)$ (k=1,2,3) at step L as L

increases. Finally, each document vector is on the lines with satisfying $\Pr(z_1 \mid d_j)+\Pr(z_2 \mid d_j)+\Pr(z_3 \mid d_j)=1$.

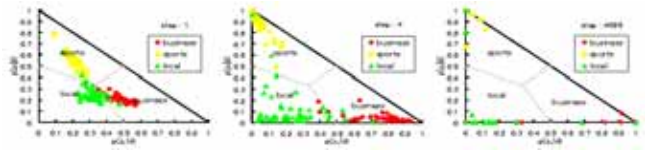

Fig. 4.1: Classification process by EM algorithm

We see that the proposed classification method is clearly superior compared to the conventional methods.

4.3 Classification problem (b)

4.3.1 Experiment conditions of (b)

We choose S = 2, 3, …, 8 categories, each of which contains D_L=100～450 articles randomly chosen. The half of them D_L is used for training, and the rest of them D_T, for test.

4.3.2 Results of (b)

The number of documents used for training D_L vs. the classification error rate P_C with parameters S is shown in Fig. 4.2. The number of categories S vs. the classification error rate P_C with parameters D_L, is also shown in Fig. 4.3, where we always choose $D_L = D_T$.

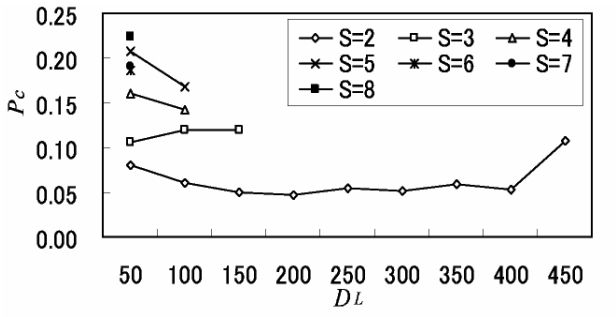

Fig. 4.2 Classification error rate for D_L

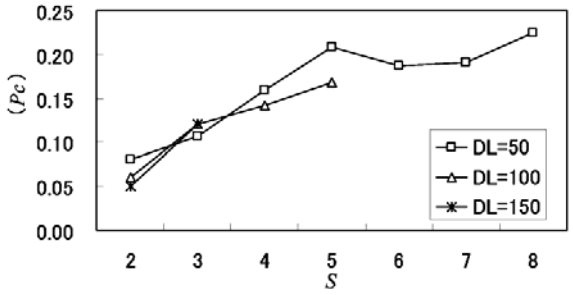

Figure 4.3 Classification error rate for S

From Figs.4.2 and 4.3, we see that the proposed classification method has good performance for small size of document sets, and the classification error P_C increases as the number of categories S increases.

4.4 Clustering problem: (c)

As stated above, we demonstrated the effectiveness of the proposed classification method. Based on this verification, we extend it to a clustering method.

We assume that characteristics of the students in Class CS are different from those in Class IS, because their majors are obviously distinct.

First, the documents of students in Class CS and those in Class IS are merged. Then the merged documents are partitioned into two clusters by the clustering method stated in 3.2, as shown in Fig. 4.4. We can expect that one cluster contains only the documents in Class CS and the other cluster, in Class IS. Since we know whether the document comes from Class CS or Class IS, the experiment is regarded as a classification problem, hence we can easily evaluate the performance of the clustering method by clustering error $\Pr(\{k \neq \hat{k}\}) = C(e)$.

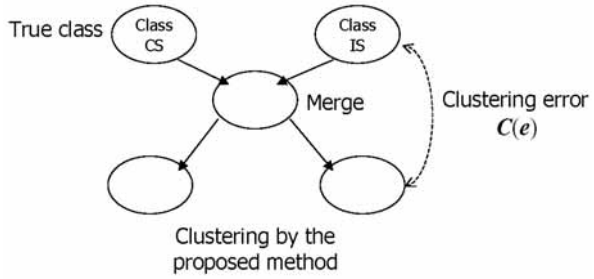

Fig. 4.4: Class partition problem by clustering method

4.4.1 Experiments conditions of (c)

Substituting $[\lambda G^T, (1-\lambda)H^T]^T$ into $A=[a_{ij}]$ in eq.(2.13), the log-likelihood function L is computed [HC03].

This condition is added to the clustering method stated in 3.2 before step (1). Then documents given by student questionnaire in two classes, Class CS and Class IS are applied. As shown in Table 4.4, the total number of students is 170.

As another clustering method to be compared to that developed in this paper, the VS clustering method is evaluated, where the VS (clustering) method uses *tf-idf* value given by eq.(2.1) for matrix A in the VS model.

4.4.2 Results of (c)

Since S=2, clustering error occurs when d_j in Class CS is classified into Class IS, and vice versa. Fig. 4.5 shows the clustering error rate $C(e)$ vs. λ, where λ is the weight for matrices G and H.

If λ=0, then only the matrix H is used which implies the case of use of text (free format) only.

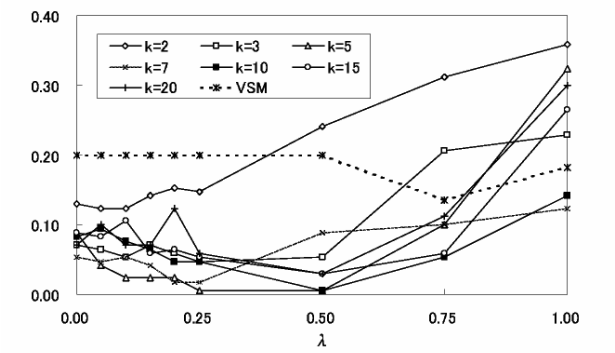

Fig. 4.5: Clustering error rate $C(e)$ vs. λ

The result shows the superiority of the clustering method discussed in this paper. Choosing λ=0.5 will be favorable to minimize the clustering error. We see that $C(e)$ decreases as K increases.

If K becomes large, however, the performance will go down because of overfitting. Fig. 4.6 shows that there is the optimum value of K to minimize $C(e)$, although it is difficult to find it out. We also show clustering process for the EM algorithm at step 1, 4, and 1024 for K=2 and K=3 in Fig. 4.7. We see that the EM algorithm works well for clustering document sets.

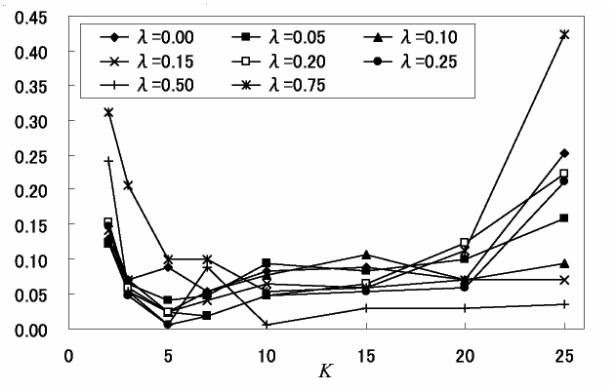

Fig. 4.6: Clustering error rate $C(e)$ vs. K

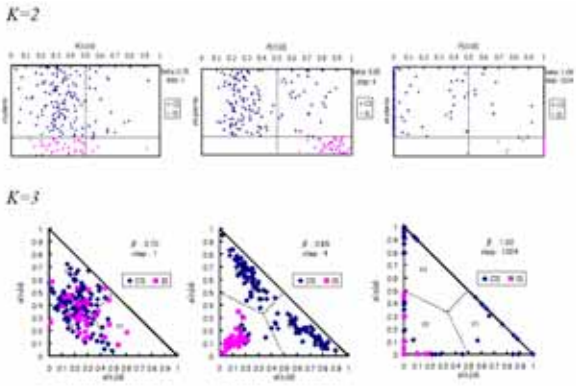

Fig. 4.7: Clustering process by EM algorithm

5. Concluding Remarks

We have proposed a classification method for a set of documents and extend it to a clustering method. The classification method exhibits its better performance for a document set with comparatively small size by using the idea of the PLSI model.

The clustering method also has good performance. We show that it is applicable to documents with both fixed and free formats by

introducing a weighting parameter λ.

In the case of clustering problem (c), we tried to apply the MDL principle [Rissanen83] to decide the optimum value of K. We see that the negative log-likelihood function, $-L$, decreases slowly as K increases. While the penalty term, $\frac{K}{2}\log D$, increases rapidly as K increases. If the number of the clusters S and that of the states K are small, then the optimum value of K will be small. This suggests us that there is a possibility to apply Bayesian probabilistic latent semantic indexing (BPLSI) model [GITSH03], [GIH03] into the clustering problems. Although the optimum value of the number of the states K is still hard to decide, the method is robust in choosing K for small K and S.

As an important related study, it is necessary to develop a method for abstracting the characteristics of each cluster [HIIGS03], [IIGSH03-b]. An extension to a method for a set of documents with comparatively large size also remains as a further investigation.

Acknowledgement

The author would like to thank Mr. T. Ishida for his valuable support to write this paper.

This work was partially supported by Waseda University Grant for Special Research Project no:2005B-189.

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.

[HC04] S. Hirasawa, and W.W. Chu "Classification methods for documents with both fixed and free format by PLSI model," Proc. of 2004 *Int. Conf. on Management Science and Decision Making*, pp.427-444, Taipei, Taiwan, R.O.C., May 2004.

[HIAGS05] S. Hirasawa, T. Ishida, H. Adachi, M. Goto, and T. Sakai, " A Document classification and its applicaion to questionnaire analyses," (in Japanese), *Proc. of 2005 Spring Conference in Information Management, JASMIN*, pp.54-57, Tokyo, June 2005.

[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.

[IGH05] T. Ishida, M. Gotoh, and S. Hirasawa, "Analysys of student questionnaire in the lecture of computer science," (in Japanese) *Computer Education, CIEC*, vol.18, pp.152-159,

July 2005.

[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.

[Mainichi95] Mainichi Newspaper Co., Ltd., '94 Mainichi Newspaper Article, CD, Naigai Associate, 1995.

[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.

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