Student Questionnaire Analyses Using the Clustering Method based on the PLSI Model

Takashi Ishida[1], Hisashi Hamada, Gendo Kumoi, Masayuki Goto, Shigeichi Hirasawa

Waseda University, Tokyo 169-8555, Japan
[1](tishida@fuji.waseda.jp)

ABSTRACT

This paper proposes a novel document clustering method based on Probabilistic Latent Semantic Indexing (PLSI) model. Recently, it is more strongly expected to improve quality of the education in universities. Student questionnaire is an effective way from which we can know the student's evaluation and make use it to improve the education quality. Free text format answer can be used to obtain in particular the opinion of the students directly. On the other hand, in the field of natural language processing, the various techniques and models for document classification or document clustering problems are developed. By the classification or the clustering for the questionnaire document, it is possible to make groups of students with the similar opinion. When a document set is large-scale, it is shown that good classification accuracy can be achieved in many studies. For a comparatively small-scale document set, the method using the PLSI model studied in recent years is remarkable, and it can achieve good classification accuracy to such small size of documents. In this study, our target documents for analysis are the student questionnaire with text free format for classrooms in university. Then, it is premised on the following.

(1) The document set is comparatively small-scale.
(2) The category which should be classified is not clear in advance.

By the clustering method using the PLSI model, the maximum likelihood estimator of the parameters of assumed probability distributions is calculated by using EM algorithm with a random initial value. From the estimated parameters, the probabilities belonging to clusters with each document are calculated, and a document is clustered with this value. In our proposal, the EM algorithm is run two times: In the first step, clustering method is applied to all the document sets. At this time, a pseudo learning document set is made of documents whose class can be estimated with high reliability. In the second step, the EM algorithm is again performed by using an initial value learned from the pseudo learning document set which is acquired in the first step, and clustering is performed. The evaluation experiment for the newspaper article and actual students' questionnaire document shows the efficiency of the proposed clustering method.

Keywords: student questionnaire, document clustering, PLSI model, faculty development

1. INTRODUCTION

Recently, with rapid development of information technology, various kind of information is stored up in WWW in the internet and computers by the day and the amount of accessible information is increasing quickly.

A large amount of information has come to be saved in the format of computerized text data. When investigating and analyzing these text documents, techniques to divide the documents into some groups with similar meaning by using the contents are very useful. The development of the methods of document classification or document clustering problems is being studied actively. The efficient document clustering method is required also in the field of university education. Since the effort for a class improvement has come to be needed, various types of questionnaires for students to evaluate the classrooms or lectures are performed at many universities. The answers of questionnaire are collected as a set of many documents. The trend of the students' opinion can be efficiently acquired by classifying and arranging these documents according to its content.

The studies on the questionnaire analysis using the method of the document classification for a class improvement are reported [2-5]. In these papers, the class model for explaining a student's score and the degree of satisfaction of a class is composed first, and the student's opinions for a class improvement are analyzed. In these studies, the Probabilistic Latent Semantic Indexing (PLSI) model which compresses a document-term matrix into low dimensionality based on a probabilistic model is used for classification or clustering of questionnaire documents.

The aim of this study is to improve the performance of document clustering method with the PLSI model [1]. The novel clustering algorithm using the initial-value dependability of the PLSI method is proposed. It uses the pseudo learning document set acquired from the first clustering step. To use the pseudo learning data which is constructed by documents whose class can be estimated with high reliability in the first step, the high accuracy of the clustering can be expected. The efficiency of the proposal method is shown by the evaluation experiment using the newspaper articles as benchmark data. Since it is premised on analyzing the student questionnaire for a

class improvement in this study, the target data is a comparatively small-scale document set.

2. PLSI MODEL

PLSI model compress a document-term matrix into low dimensionality one based on the probability model using the semantic hidden attribute classes [1]. The noise caused by many terms appeared in documents can be removed by reduction of the large dimensions of document models.

2.1 Hidden attribute class

It can be considered that words and documents have latent semantics. For example, the same word may be used in a different meaning for each document, or different words may be used in the same senses. In the PLSI model, the latent sense or topic which exists behind words or documents is assumed as a concept of hidden attribute class.

2.2 Maximum likelihood estimation on PLSI model

The probabilities of a document $d_i \in \mathcal{D}$ $(i = 1,2,\cdots I)$ and a word $w_j \in \mathcal{W}$ $(j = 1,2,\cdots J)$ are expressed as a conditional probability of a hidden attribute class $z_k \in \mathcal{Z}$ $(k = 1,2,\cdots K)$. In the PLSI model, it assumes that occurrences of document d_i and the word w_j are independent each other under the conditions in which hidden class z_k is given. Then the joint probability of d_i and w_j is given as follows:

$$P(d_i, w_j) = \sum_{k=1}^{K} P(d_i \mid z_k) \cdot P(w_j \mid z_k) \cdot P(z_k). \quad (1)$$

Let $n(d_i, w_j)$ be a frequency of appearance of the word w_j which actually occurs in the document d_i. Then the EM algorithm calculates the probabilities $P(d_i \mid z_k)$, $P(w_j \mid z_k)$ and $P(z_k)$ which maximize the log likelihood function *L*

$$L = \sum_{i=1}^{I} \sum_{j=1}^{J} n(d_i, w_j) \cdot \log P(d_i, w_j). \quad (2)$$

2.3 EM Algorithm

The EM algorithm is an iterative method for calculating a maximum likelihood estimator (MLE), when the hidden parameter which cannot be observed exists. It can calculate a local optimal solution by repeated calculation. Arbitrary initial values are given to each $P(d_i \mid z_k)$, $P(w_j \mid z_k)$ and $P(z_k)$ respectively, and the following equations are iteratively calculated until the values converge.

[EM Algorithm]

E-Step

$$P(z_k \mid d_i, w_j) = \frac{P(z_k) \cdot P(d_i \mid z_k) \cdot P(w_j \mid z_k)}{\sum_{k'} P(z_{k'}) \cdot P(d_i \mid z_{k'}) \cdot P(w_j \mid z_{k'})}. \quad (3)$$

M-Step

$$P(w_j \mid z_k) = \frac{\sum_i n(d_i, w_j) \cdot P(z_k \mid d_i, w_j)}{\sum_i \sum_{j'} n(d_i, w_{j'}) \cdot P(z_k \mid d_i, w_{j'})}, \quad (4)$$

$$P(d_i \mid z_k) = \frac{\sum_j n(d_i, w_j) \cdot P(z_k \mid d_i, w_j)}{\sum_{i'} \sum_j n(d_{i'}, w_j) \cdot P(z_k \mid d_{i'}, w_j)}, \quad (5)$$

$$P(z_k) = \frac{\sum_i \sum_j n(d_i, w_j) \cdot P(z_k \mid d_i, w_j)}{\sum_i \sum_j n(d_i, w_j)}. \quad (6)$$

2.4 Tempered EM Algorithm

In practice, the Tempered EM (TEM) algorithm is applied instead of the conventional EM algorithm to the PLSI model. In this algorithm, a hyper-parameter β (<1) with a small positive value is introduced, and an Eq. (3) is replaced by the following equation:

$$P(z_k \mid d_i, w_j) = \frac{P(z_k) \cdot \left[P(d_i \mid z_k) \cdot P(w_j \mid z_k)\right]^{\beta}}{\sum_{k'} P(z_{k'}) \cdot \left[P(d_i \mid z_k) \cdot P(w_j \mid z_k)\right]^{\beta}}. \quad (7)$$

3. DOCUMENT CLUSTERING USING PLSI MODEL

3.1 Clustering Algorithm

Since the hidden classes $z_k \in \mathcal{Z}$ $(k = 1,2,\cdots K)$ in the PLSI model can be regarded as respectively indicating some kinds of concepts, documents with the same concept can be clustered into the same class by using PLSI model. Now, let a document set be clustered into *S* groups. And the number of the hidden class *K* is set to *S* (*K* = *S*).

Conventional clustering algorithm using the PLSI model [2, 3] is shown as follows:

[Conventional Clustering Method (CM)]

(Step1) Set random initial values to the parameters $P(d_i \mid z_k)$, $P(w_j \mid z_k)$ and $P(z_k)$, then calculate MLE of each one by TEM algorithm.

(Step2) Each document $d_i \in \mathcal{D}\,(i = 1,2,\cdots I)$ is assigned to the hidden class $z_k \in \mathcal{Z}\,(k = 1,2,\cdots K)$ with maximum value of $P(z_k \mid d_i)$.

In the PLSI model, the EM or TEM algorithms are used in order to compute a MLE. However, the different partial solutions are usually derived for the same document data depending on the set of initial values of the algorithm. Therefore, since the correct solution is not always obtained when an initial value is set up at random, it is necessary to repeat calculation using the many different initial values. In addition, the problem caused by the value of probabilities of the hidden classes can be pointed out. In (Step 2) of the conventional algorithm, the class of document d_i is assigned to z_k whose values of $P(z_k \mid d_i)$ is a maximum in all hidden classes. However, the value of $P(z_k \mid d_i)$ is sometimes relatively small and two or more hidden classes have the almost same value of probability. In these cases, it is known that the clustering cannot be performed well in many cases.

4. PROPOSED METHOD

It can be regarded that the value of $P(z_k \mid d_i)$ means a degree of assignment to the class z_k of the document d_i. When the maximum value of $P(z_k \mid d_i)$ over the hidden classes is small and there is small difference in the values mutually, the document d_i had a high possibility of being assigned to any class z_k. Therefore, the good clustering performance cannot be expected in such case of the conventional method.

On the contrary, when the maximum $P(z_k \mid d_i)$ has a clearly large value compared with other z_k, it was shown that the good clustering performance can be obtained by some experiments.

Then, we may judge that the document d_i has a high reliability of the clustering when $P(z_k \mid d_i)$ has sufficiently larger value than other $z_{k'}$. In this way, by feeding back the information with high reliability, it is expected to estimate more pertinent probability and to improve a clustering precision.

When we can previously obtain the document set whose components have the known classes to which each document belongs (it's called "learning document set"), the effective document classification method can be constructed. In these setting, a document classification method which set an initial value of EM algorithm by the occurrence probability calculated from the learning document set has been proposed [4, 6]. It has been shown that these classification methods achieve a good classification precision to the document set with small number of documents.

Using the concept of learning technique of above classification methods, we propose a novel document clustering methods. That is, at first, the MLE of probability $P(z_k \mid d_i)$ is calculated by performing EM algorithm. And a threshold R is set up. We consider that if the maximum value of $P(z_k \mid d_i)$ for a document d_i is larger than a threshold R, then d_i is regard as a document which can be assigned to a class z_k with high reliability. Such documents are fed back as the pseudo learning documents. Once a pseudo learning document set is obtained and the classes of the documents in the set are regarded as the correct values, the initial value of a probability can be calculated from these document set as same as the document classification method described above. Again the EM algorithm is performed setting the initial values. And using the estimated MLE of the second EM algorithm, the documents will be clustered.

In a pseudo learning document set, let a frequency of the word w_j which appears in a class z_k be denoted by $f_{k,j}$ (k=1,2,…K, j=1,2,…J). The initial value of EM algorithm is calculated as follows from a pseudo learning document set.

$$P(w_j \mid z_k) = \frac{f_{k,j} + \alpha}{\sum_{j'}(f_{k,j'} + \alpha)}, \tag{8}$$

$$P(d_i \mid z_k) = \frac{1}{I}, \tag{9}$$

$$P(z_k) = \frac{1}{K}. \tag{10}$$

Here, α (>0) is a offset parameter.

The algorithm of our proposal method is shown below.

[Proposal Clustering Method (PM)]

(Step1) Set random initial values to the parameters $P(d_i \mid z_k)$, $P(w_j \mid z_k)$ and $P(z_k)$, then calculate MLE of each one by TEM algorithm.

(Step2) Form a pseudo learning document set is composed by collection of the documents d_i which satisfy $\max_{z_k} P(z_k \mid d_i) > R$.

(Step3) Set the initial value of EM algorithm by computation from a pseudo learning document set.

(Step4) Assigne each document $d_i \in \mathcal{D}\,(i = 1,2,\cdots I)$ to the hidden class $z_k \in \mathcal{Z}\,(k = 1,2,\cdots K)$ with maximum value of recalculated $P(z_k \mid d_i)$.

5. EVALUETION EXPERIMENT

5.1 Test Data

Our proposed clustering method is applied to the '94 Mainichi newspaper article [7] and the answer of the actual student questionnaire written by Japanese language. The correct label of the class is beforehand given to the newspaper article data.

In our study, the small size document set is targeted. For the newspaper data, three classes, that is, economy, sport, and social, are used. 50 documents are prepared for each class, and let them be a document set. The clustering algorithm is performed to the document set.

A student questionnaire data is the students' answers for the questionnaire carried out in 2003 in the subject of "computer engineering" which is a class of a department of science and engineering. The reply of a student questionnaire is composed of the mixture of the item-selection and the free-text types.

5.2 Evaluation Criteria

Accuracy rate (AR) is used as an evaluation criterion for the clustering performance. Accuracy rate is calculated as a ratio of the number of the documents which are correctly clustered to that of the all documents.

5.3 Setting of Parameters

From the result of preliminary experiments, β is set to 0.9 for the newspaper article, 0.7 for student questionnaire. A threshold R in the proposed method is set to 0.99 for the newspaper data, 0.90 for the questionnaire data. Also α in Eq. (8) is set to 0.5 and λ in Eq. (11) is set to 0.5.

For both the conventional method (CM) and the proposed method (PM), EM algorithm is performed using the 300 initial values generated at random.

5.4 Results of Experiment

(1) For newspaper data

The scatter diagram showing the relationship between the accuracy rate (AR) and the logarithm likelihood (LL) of the CM and the PM to the 300 initial values is shown in a Fig.1. In addition, the result of the document classification method (LM) using the learning document set for calculating the initial value is also shown in Fig.1 for comparison. In this method, for each class, 50 documents are prepared respectively.

Table 1 shows the average of the LL and AR for the 300 initial values on CM and PM.

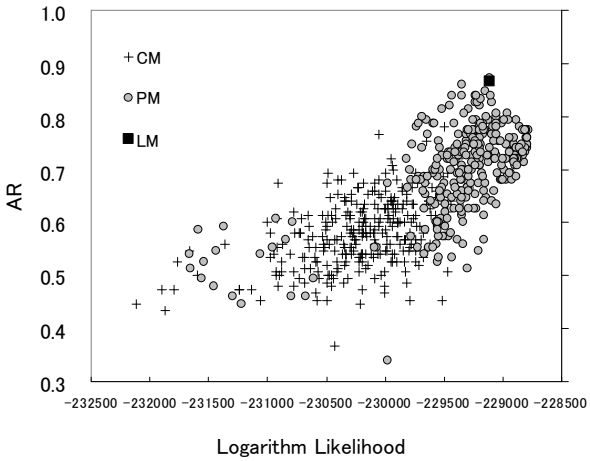

Fig.1. Relationship between AR and log likelihood of CM and PM (Newspaper)

Table 1. Average of LL and AR of CM and PM

	Logarithm Likelihood	Accuracy Rate
CM	-230214	0.580
PM	-229373	0.697

In addition, the values of LL and AR of the CM, PM and LM in the case that AR takes the highest value are shown in Table 2.

Table 2. LL and AR in the case of the highest AR in each method

	Logarithm Likelihood	Accuracy Rate
CM	-229493	0.780
PM	-229108	0.867
LM	-229116	0.873

Next, AR by each class (economy, sports, and society) in each method of Table 2 is shown in Table 3.

Table 3. AR by each class

	economy	sports	Society	all
CM	0.88	0.68	0.78	0.78
PM	0.86	0.84	0.90	0.87
LM	0.80	0.90	0.92	0.87

When a feedback is performed in the PM, the number of the documents whose tentative classes are correct in a pseudo learning document set is shown in Table 4 about the following cases.

(a) AR rises by feedback.
(b) AR descends by feedback.

Each row of Table 4 means the result in a different initial value. Only the results of upper rank of (a) or (b) are shown.

Table 4. Relationship of the AR and the number of the pseudo learning document assigned to a correct class

(a) Case that AR rises by feedback

Accuracy Rate (AR)		Pseudo learning documents	
CM	PM	Rate of correct class	Number of documents
0.473	0.780	0.76	25
0.473	0.773	0.79	19
0.513	0.693	0.91	32
0.547	0.833	0.92	24
0.620	0.733	0.73	49

(b) Case that AR descends by feedback

Accuracy Rate (AR)		Pseudo learning documents	
CM	PM	Rate of correct class	Number of documents
0.600	0.460	0.71	42
0.467	0.340	0.64	42
0.627	0.527	0.74	46
0.580	0.480	0.70	43
0.613	0.540	0.74	42

The AR of the CM and the PM for the top 10 of the large LL at the first step of the EM algorithm is shown in Table 5.

Table 5. AR of the CM and PM for the top 10 of the large logarithm likelihood at first step of EM algorithm

(a) Case that AR rises by feedback.

Rank	CM		PM	
	LL	AR	LL	AR
1	-229251	0.660	-228829	0.733
2	-229294	0.673	-229096	0.767
3	-229337	0.627	-228794	0.740
4	-229376	0.700	-229007	0.787
5	-229378	0.720	-229151	0.773
6	-229438	0.640	-228823	0.713
7	-229466	0.507	-228787	0.773
8	-229468	0.633	-228796	0.760
9	-229472	0.620	-228858	0.733
10	-229474	0.693	-228958	0.807

(2) For questionnaire data

The proposed clustering method (CM) is applied to the questionnaire answer documents of 86 students who take the class of "computer engineering." The aim of this questionnaire is to divide the students into two classes according to the content of a class shown in Table 6.

Table 6. Contents of the each class

Class	Contents
A	For general consideration - History of computer - Fundamental concepts in computer - Basic of architecture, hardware, software - Applied technology etc.
B	For specialist - Computer architecture - Hardware - Software etc.

Here, we assume that the actual answers of the students for a question "when the class is divided into two, which do you choose Class A or B?" are their correct categories of the clustering. The average of AR and LL of the clustering result for 300 initial values by CM and PM is shown in Table 7.

Table 7. AR for students' questionnaire

	LL	AR
CM	-781831	0.556
PM	-781718	0.566

The scatter diagram showing the relationship of AR and LL of the CM and the PM for the 300 initial values is shown in Fig. 2.

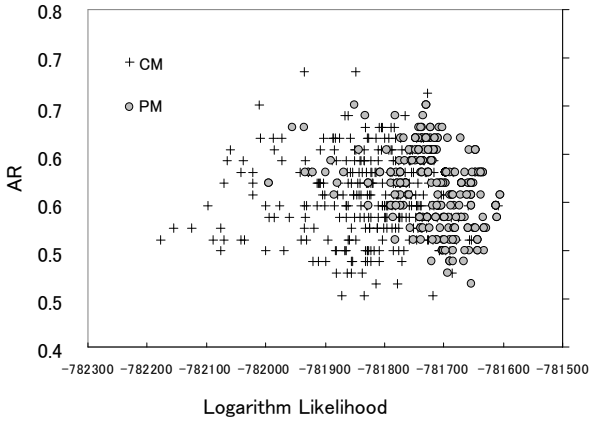

Fig.2. Relationship between AR and log likelihood of CM and PM (Students' questionnaire)

5.5 Discussions

(1) Fig. 1 and Table 1 show that the LL and the AR are overall improved by the PM.

(2) In the CM and the PM, the AR is nearly proportional to LL. From this result, it is considered to be justified to perform clustering using the MLE calculated by EM algorithm. However, Tables 2 and 5 show that the AR of the clustering does not

necessarily become the maximum when the LL is the maximum.

(3) From Tables 2 and 3, when a good initial value is chosen, it is thought that the PM has achieved the higher AM than the document classification method using the learning document (LM). The higher AR can be achieved by clustering using various initial values. However, it is necessary to clarify which initial value is adopted in the case of utilization. Table 5 shows that the performance of clustering becomes good when the LL takes higher value. Therefore, if the initial values with the higher LL are adopted, it can be said that the good performance of clustering is achieved to some extent.

(4) Table 4 shows that the clustering performance in the PM becomes good when the correct answer rate of a pseudo learning document takes higher value. Furthermore, we found that it affects the AR that less incorrect document is contained in pseudo learning document set rather than that many correct answer documents are contained in it. The above result suggests that (1) if the class of document in a pseudo document set is correct, the good clustering performance can be achieved with a small number of the learning documents, (2) few errors of a class of the documents fed back in pseudo learning document set are desirable.

(5) From the results of the clustering for the students' questionnaire by Fig. 2 and Table 6, we find that the AR is not proportional to the LL for both CM and PM. Moreover, the AR does not take so higher value. However, it can be said that this is because the CM and PM are not necessarily performed so that the students may be divided into the class which we set to the correct answer hypothetically. Therefore, it does not suggest that the performance of the clustering is poor.

6. CONCLUDING REMARKS

This paper proposed a novel clustering method which uses the pseudo learning document based on a PLSI model. The aim of this study is a proposal of the new clustering method for students' questionnaire analysis, and the application to a small-scale document set has been focused.

The result of the evaluation experiment shows that our proposed method has a better performance of clustering rather than the conventional one.

As future works, the performance of clustering when increasing the number of hidden attribute class should be evaluated. In addition, it is necessary to analyze the students' questionnaire document which is our final target. The efficiency of our clustering method was already shown by the experiment using a newspaper article with a correct class of each document in this study. It is a future work to extract the features of each cluster (class) and to give semantic index to each class of students clustered by our method.

REFERENCES

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

[2] T. Ishida, M. Goto, and S. Hirasawa, "Analysys of student questionnaire in the lecture of computer science," (in Japanese) Computer Education, CIEC, vol.18, pp.152-157, July 2005.

[3] S. Hirasawa, T. Ishida, H. Adachi, M. Goto, and T. Sakai, "A Document Classification Method and its application to Questionnaire Analyses," (in Japanese) Proc. of 2005 Spring Conference on Information Management, JASMIN, pp.83-84, 2005.

[4] J. Ito, T. ishida, M. Goto, T. Sakai, and S. Hirasawa, "A method for extracting important sentences using co-occurrence similarities between words," (in Japanese) Forum on Information Technology 2003, pp.83-84, Tokyo, Sept. 2003.

[5] S. Hirasawa, F. Shih, and W. Yang, "Student questionnaire analyses for class management by text mining both in Japanese and in Chinese," Proc. 2007 IEEE International Conference on System, Man and Cybernetics, pp. 398-403, Montreal, Canada, Oct. 2007.

[6] H. Hamada, G. Kumoi, T. Ishida, Y. Tsai, and S. Hirasawa, "Chinese text categorization based on PLSI model," (in Japanese) Proc. of 2008 Spring Conference on Information Management, JASMIN, 2008.

[7] Mainichi Newspaper CD '94, Naigai Associates, 1995.