Comparing LDA with pLSI as a Dimensionality Reduction Method in Document Clustering

Tomonari Masada, Senya Kiyasu, and Sueharu Miyahara

Nagasaki University, 1-14 Bunkyo-machi, Nagasaki 852-8521, Japan
{masada,kiyasu,miyahara}@cis.nagasaki-u.ac.jp

Abstract. In this paper, we compare latent Dirichlet allocation (LDA) with probabilistic latent semantic indexing (pLSI) as a dimensionality reduction method and investigate their effectiveness in document clustering by using real-world document sets. For clustering of documents, we use a method based on multinomial mixture, which is known as an efficient framework for text mining. Clustering results are evaluated by F-measure, i.e., harmonic mean of precision and recall. We use Japanese and Korean Web articles for evaluation and regard the category assigned to each Web article as the ground truth for the evaluation of clustering results. Our experiment shows that the dimensionality reduction via LDA and pLSI results in document clusters of almost the same quality as those obtained by using original feature vectors. Therefore, we can reduce the vector dimension without degrading cluster quality. Further, both LDA and pLSI are more effective than random projection, the baseline method in our experiment. However, our experiment provides no meaningful difference between LDA and pLSI. This result suggests that LDA does not replace pLSI at least for dimensionality reduction in document clustering.

1 Introduction

Document clustering is a classic problem of text mining. In recent years, clustering is proved to be effective in summarizing a search result or in distinguishing different topics latent in search results [29][7][5]. With respect to this type of application, clustering is expected to provide a result at query time. In contrast, enterprise documents stored in the intranet or the patent documents relating to a specific technical field form a document set which is not so small as a search result and, simultaneously, not so large as those targeted by open Web search services [12][15][18]. In this paper, we consider applications managing this type of document set, i.e., a document set of *middle-range* size and focus on *latent Dirichlet allocation (LDA)* [10] along with *probabilistic latent semantic indexing (pLSI)* [17], which are applicable to such document sets in realistic execution time. These two methods share the following special feature: topic multiplicity of each document is explicitly modeled. Therefore, we can consider topic mixture for each document. This feature makes LDA and pLSI differentiate from multinomial mixture model [24] and also from Dirichlet mixture model [21][19].

However, LDA employes a Bayesian inference framework, which makes LDA more theoretically attractive than pLSI.

In this paper, we use LDA and pLSI for dimensionality reduction of feature vectors in document clustering and check if LDA can replace pLSI for this task. Our original feature vectors have frequencies of words as their entries and thus are of dimension equal to the number of vocabularies. Both LDA and pLSI reduce the dimension of document vectors to the number of topics, which is far less than the number of vocabularies. Roughly speaking, we can regard each entry of the vectors of reduced dimension as a topic frequency, i.e., the number of words relating to each topic. We investigate the effectiveness of dimensionality reduction by conducting a clustering on feature vectors of reduced dimension.

Our experiment uses four different sets of Japanese and Korean Web articles. Each article set consists of tens of thousands of documents, i.e., a document set of middle-range size. We use a clustering method based on multinomial mixture with EM algorithm for parameter estimation. Multinomial mixture is well-known as an effective framework for text mining applications, e.g. junk e-mail filtering [26]. While we have also tested k-means clustering method, this does not give clusters of satisfying quality in comparison with multinomial mixture. Therefore, we do not include those results in this paper. In evaluating cluster quality, we compare the quality before and after dimensionality reduction via LDA and pLSI. Further, we compare these two methods with random projection [9], which we regard as the baseline method in this paper. We use the category assigned to each article as the ground truth for evaluating cluster quality. Therefore, we try to recover document categories based on the topic frequencies obtained by the two multi-topic document models, LDA and pLSI. While the inference of the correct number of clusters is important, this is beyond our scope. We have used the true number of clusters as an input.

The rest of the paper is organized as follows. Section 2 gives previous work concerning applications of LDA to real-world data. Section 3 includes a short description of LDA. Since pLSI has already become more widely accepted than LDA, we omit the details about pLSI from this paper and refer to the original paper [17]. The results of evaluation experiment is presented in Section 4. Section 5 draws conclusions and gives future work.

2 Previous Work

Recently, many applications of LDA to real-world problems are proposed, e.g. multimodal information integration [8][20], topic-author relationship analysis [16], expert finding [22] and subject extraction from digital library books [23]. However, these researches do not compare LDA with other probabilistic model. Sadamitsu et al. [28] conduct intensive experiments comparing LDA with pLSI and Dirichlet mixture. While we can learn important things about the applicability of LDA and other document models, their work compares these document models not from a practical viewpoint, but from a theoretical one, because the authors use *perplexity* as a measure for evaluation. Perplexity tells how well a

document model can generalize to test data, but does not tell how well a document model can solve text mining problems, e.g. information retrieval, document classification or document clustering.

In this paper, we employ LDA as a dimensionality reduction method in document clustering and evaluate its effectiveness by inspecting the quality of document clusters. Although Blei et al. [10] use LDA for dimensionality reduction, the authors compare LDA with no other methods. Further, their evaluation task is a binary classification of Reuters-21578 corpus, a slightly artificial task. Elango et al. [14] also use LDA to reduce the dimension of feature vectors. However, their feature vectors are obtained from image data, and LDA is not compared with any other methods. In this paper, we use LDA as a dimensionality reduction method to clarify its applicability to the document clustering task by comparing it with pLSI. Futher, we compare LDA and pLSI with random projection [9], the baseline method in our experiment. Since LDA is proposed as a sophistication of pLSI, it is worthwhile to check if LDA can provide better results than pLSI.

Recently, LDA has been extented to enable an automatic determination of the number of clusters. This method is based on a probabilistic model, called Dirichlet process mixture (DPM) [11]. With DPM, we do not need to conduct dimensionality reduction first and then execute a clustering, because DPM can provide a far smaller number of probability distributions over topics than the number of documents. Each of these topic distributions, in turn, can be regarded as the feature of a cluster. In contrast, LDA gives as many topic distributions as documents, where we can observe no clustering effects. If LDA do not give better results than pLSI in our evaluation, we can conclude that LDA is not a good choice at least for dimensionality reduction in document clustering, because we can use pLSI when the efficiency in execution time is required, or can use DPM when high computational cost is allowed.

3 Latent Dirichlet Allocation

3.1 Details of Model

Latent Dirichlet Allocation (LDA) [10] is a document model which explicitly models topic multiplicity of each document. This feature differenciates LDA from multinomial mixture [24] and also from Dirichlet mixture [21]. Probabilistic latent semantic indexing (pLSI) [17] shares the feature of topic multiplicity modeling with LDA. However, pLSI requires heuristic computations for obtaining the probability of unknown documents and is also likely to result in overlearning.

We denote a document set by $D = \{d_1, \ldots, d_I\}$, the set of vocabularies (i.e., word types) appearing in D by $W = \{w_1, \ldots, w_J\}$ and the set of topics included in D by $T = \{t_1, \ldots, t_K\}$. Formally speaking, topics are the values of hidden variables of a document model. With respect to each topic, we have a multinomial distribution defined over W. Namely, the topic difference is represented by the difference of word probabilities.

In LDA, with respect to each document, we select a multinomial distribution defined over T according to the following Dirichlet distribution: $P(\theta; \alpha) =$

$\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_k^{\alpha_k}$. In generating documents, we regard each document d_i as an empty array of length equal to the document length n_i. We fill this array as follows. First, we select a multinomial over T from the Dirichlet distribution shown above. Second, we select a topic for each array element according to this multinomial over T. The topic assigned to the lth array element of document d_i is denoted by $\mathbf{z}_{il}$. The entire topic sequence of d_i is referred to by $\mathbf{z}_i$. Third, we select a word to fill this array element according to the multinomial over W which corresponds to the topic $\mathbf{z}_{il}$. The word filling the lth array element of document d_i is denoted by $\mathbf{x}_{il}$. The whole word sequence of d_i is referred to by $\mathbf{x}_i$. As we repeatedly select a topic for each word, we can explicitly model the topic multiplicity within the same document. Let β_{kj} be the probability of vocabulary w_j in the multinomial distribution corresponding to topic t_k. Note that $\sum_j \beta_{kj} = 1$ holds for all k. The model parameters of LDA are $\alpha_k (k = 1, \ldots, K)$ and $\beta_{kj} (k = 1, \ldots, K, j = 1, \ldots, J)$. The total number of parameters is $K + KJ$. The probability of the word sequence $\mathbf{x}_i$ of document d_i can be written as

$$P(\mathbf{x}_i; \alpha, \beta) = \int \sum_{\mathbf{z}_i} P(\theta; \alpha) P(\mathbf{z}_i|\theta) P(\mathbf{x}_i|\mathbf{z}_i, \beta) d\theta. \tag{1}$$

The probability of the word sequence of the whole document set D is equal to $\prod_i P(\mathbf{x}_i; \alpha, \beta)$. By maximizing the log of this probability, i.e., $\log \prod_i P(\mathbf{x}_i; \alpha, \beta) = \sum_i \log P(\mathbf{x}_i; \alpha, \beta)$, we can determine model parameter values.

3.2 Variational Inference

In this paper, we employ variational inference method [10], where two probability distributions, $Q(\theta; \gamma_i)$ and $Q(\mathbf{z}_i; \phi_i)$, are introduced with respect to each document as follows:

$$\begin{aligned}
&\log P(\mathbf{x}_i; \alpha, \beta) \\
&= \log \int \sum_{\mathbf{z}_i} P(\theta; \alpha) P(\mathbf{z}_i|\theta) P(\mathbf{x}_i|\mathbf{z}_i, \beta) d\theta \\
&= \log \int \sum_{\mathbf{z}_i} Q(\theta; \gamma_i) Q(\mathbf{z}_i; \phi_i) \frac{P(\theta; \alpha) P(\mathbf{z}_i|\theta) P(\mathbf{x}_i|\mathbf{z}_i, \beta)}{Q(\theta; \gamma_i) Q(\mathbf{z}_i; \phi_i)} d\theta \\
&\geq \int \sum_{\mathbf{z}_i} Q(\theta; \gamma_i) Q(\mathbf{z}_i; \phi_i) \log \frac{P(\theta; \alpha) P(\mathbf{z}_i|\theta) P(\mathbf{x}_i|\mathbf{z}_i, \beta)}{Q(\theta; \gamma_i) Q(\mathbf{z}_i; \phi_i)} d\theta
\end{aligned} \tag{2}$$

We can move from the third line to the fourth by applying Jensen's inequality and obtain a lower bound of $\log P(\mathbf{x}_i; \alpha, \beta)$ for each d_i. In variational inference, we maximize this lower bound in place of $\log P(\mathbf{x}_i; \alpha, \beta)$. $Q(\mathbf{z}_i; \phi_i)$ is equal to $\prod_{l=1}^{n_i} Q(\mathbf{z}_{il}; \phi_{il})$ where $\phi_{ilk}$is the probability of the assignment of topic t_k to the lth word of d_i. $\sum_{k=1}^{K} \phi_{ilk} = 1$ holds for all i and l. Further, $Q(\theta; \gamma_i)$ is a Dirichlet distribution defined over topic multinomials. While $Q(\theta; \gamma_i)$ plays a role similar to $P(\theta; \alpha)$, $Q(\theta; \gamma_i)$ is introduced separately for each document. The

details of variational inference for LDA is described in [10]. Here we only present the resulting update formulas.

$$\phi_{ilk} \propto \beta_{kj_{il}} \exp\{\Psi(\gamma_{ik}) - \Psi(\sum_{k'} \gamma_{ik'})\} \tag{3}$$

$$\gamma_{ik} = \alpha_k + \sum_l \phi_{ilk} \tag{4}$$

$$\beta_{kj} \propto \sum_i \sum_l \delta_{ilj} \phi_{ilk} \tag{5}$$

$$\alpha_k = \hat{\alpha}_k + f_k(\hat{\alpha}) + \sum_{k'} f_{k'}(\hat{\alpha}) \Big/ \Big\{ \frac{\Psi_1(\hat{\alpha}_k)}{\Psi_1(\sum_{k'} \hat{\alpha}_{k'})} - \sum_{k'} \frac{\Psi_1(\hat{\alpha}_k)}{\Psi_1(\hat{\alpha}_{k'})} \Big\}$$

where

$$f_k(\alpha) = \frac{\Psi(\sum_{k'} \alpha_{k'})}{\Psi_1(\alpha_k)} - \frac{\Psi(\alpha_k)}{\Psi_1(\alpha_k)} + \frac{\sum_i \{\Psi(\gamma_{ik}) - \Psi(\sum_{k'} \gamma_{ik'})\}}{N\Psi_1(\alpha_k)} \tag{6}$$

In Eq. 3, j_{il} is the index of the lth word of document d_i. Thus, the lth word of document d_i is $w_{j_{il}} \in W$. In Eq. 5, δ_{ilj} is equal to 1 if the lth word of d_i is w_j (i.e., $j_{il} = j$), and 0 otherwise. In Eq. 6, $\hat{\alpha}_k$ is a value for α_k obtained in the previous iteration. Ψ and Ψ_1 stand for digamma and trigamma functions, respectively. As Eq. 6 is an update formula only for α_k, we repeatedly use this formula until convergence. Our implementation in C language for this paper terminates this update iteration when $\sum_k \alpha_k$ changes by less than 0.000001 of the previous value. After executing Eq. 3, 4 and 5, we use Eq. 6 repeatedly and then return to Eq. 3. Our implementation terminates the entire iteration ranging from Eq. 3 to Eq. 6 when $\sum_k \alpha_k$ changes by less than 0.005 of the previous value.

In this paper, we regard γ_{ik} as a "pseudo-frequency" of topic t_k in document d_i. Roughly speaking, γ_{ik} is the "number" of words relating to topic t_k in document d_i. We have called γ_{ik} "pseudo-frequency," because this is not necessarily an integer. We use a K-dimensional vector $(\gamma_{i1}, \ldots, \gamma_{iK})$ as a feature vector of d_i after dimensionality reduction via LDA. This vector can be used as a feature vector by the following reason. By taking the sum of the both sides of Eq. 4 for all k, we have $\sum_k \gamma_{ik} = \sum_k \alpha_k + \sum_k \sum_{l=1}^{n_i} \phi_{ilk}$. Further, $\sum_k \sum_l \phi_{ilk}$ is equal to n_i, the document length of d_i, because $\sum_{k=1}^{K} \phi_{ilk} = 1$. Consequently, $\sum_k \gamma_{ik}$ is of the same order with document lengths.

In estimating parameters of LDA, we have also tried a collapsed variational Bayesian inference [27] only for one of the four datasets in the evaluation experiment. As for the details, please refer to the original paper. This inference gives a probability that a topic is assigned to a vocabulary appearing in a document for all $K \times J \times I$ combinations of topics, vocabularies and documents. Therefore, by taking a summation of these probabilities over the vocabularies appearing in a document with respect to a fixed topic, we can have a value of the same meaning as γ_{ik} shown above, i.e., a "pseudo-frequency" of a topic in each document.

4 Evaluation Experiment

4.1 Document Sets

In the evaluation experiment, we use one document set of Japanese Web news articles, one document set of questions from a Japanese Q&A Web site and two sets of Korean Web news articles.

The first set consists of news articles published at Japan.internet.com [1] from 2001 to 2006. Every article is uniquely labeled by one of the following six categories: mobile phone, Web business, e-commerce, Web finance, Web technology and Web marketing. We use MeCab morphological analyzer [3] to split every document into a sequence of word tokens. Then we count the frequencies of all vocabularies and eliminate the vocabularies of low frequency and those of high frequency. The resulting document set, denoted by JIC, includes 28,329 articles. The sum of the lengths of all documents in JIC amounts to 4,108,245. The number of vocabularies is 12,376. As the number of categories is six, we subdivide this set into six disjoint subsets by clustering. The number of documents and the document length sum for each category are included in Table 1.

The second set includes the queries submitted to a Japanese Q & A Web site, called "OKWave" [4]. In this experiment, we have not used the answers to each query, because we think that some of them explicitly introduce noisy information for document clustering. Every question is uniquely labeled by one of the following 11 categories: information for computer engineers, regional information, entertainment, digital life, business and career, money, daily life, education, society, hobbies, health and beauty. Here we also use MeCab morphological analyzer and eliminate the vocabularies of low frequency and those of high frequency. This document set, denoted by OKWAVE, includes 70,555 articles. The sum of the lengths of all documents is 2,511,221, and the number of vocabularies is 13,341. We split this set into 11 disjoint subsets by clustering. Note that the average document length of this document set is far shorter than the other three sets. Table 2 provides the number of documents and the document length sum for each category. We have used a collapsed variational Bayesian inference only for this set, because this inference method shows an advantage in computational cost when the number of documents is large.

Table 1. JIC dataset

category	# of docs	sum of doc lengths
mobile phone	3,049	499,368
Web business	9,059	1,214,335
e-commerce	2,522	327,264
Web finance	2,994	398,995
Web technology	6,109	922,164
Web marketing	4,596	746,119
total	28,329	4,108,245

Table 2. OKWAVE dataset

category	# of docs	sum of doc lengths
info. for comput. engineers	7,055	333,624
regional info.	4,103	128,800
entertainment	8,241	206,832
digital life	11,909	410,648
business and career	5,985	236,157
money	4,150	180,271
daily life	7,672	342,287
education	7,149	232,377
society	4,589	170,617
hobbies	6,725	159,666
health and beauty	3,030	109,942
total	70,555	2,511,221

Table 3. S2005 dataset

category	# of docs	sum of doc lengths
economy	6,172	461,592
international	3,048	216,462
politics	3,608	286,375
society	9,221	590,190
total	22,049	1,554,619

Table 4. S2006 dataset

category	# of docs	sum of doc lengths
administration	1,503	124,657
culture	4,870	347,438
economy	6,745	549,081
entertainment	1,710	125,787
international	2,498	186,753
policitics	3,806	324,076
region	3,923	280,676
society	8,946	607,158
sport	3,016	185,054
total	37,017	2,730,680

Two Korean document sets are obtained by gathering articles published at Seoul newspaper Web site [6] from 2005 to 2006. One set consists of the articles published in 2005. The articles of this set belong to one of the following four categories: economy, international, politics and society. Another set includes the articles published in 2006. The articles of this set belong to one of the following nine categories: administration, culture, economy, entertainment, international, politics, region, society and sports. We use KLT version 2.10b [2] for Korean morphological analysis. For each of the two sets, we eliminate the vocabularies of low frequency and those of high frequency. We denote the resulting document sets S2005 and S2006, respectively. S2005 includes 22,049 articles and 14,563 vocabularies. The sum of the lengths of all documents in S2005 amounts to 1,554,619. We conduct a clustering on this set and obtain four disjoint clusters, because the number of categories is four. S2006 includes 37,017 documents and 25,584 vocabularies. The document length sum is 2,730,680. We split this set into nine disjoint subsets by clustering. The number of documents and the document length sum for each category are given in Table 3 and Table 4 for S2005 and S2006, respectively.

4.2 Clustering Method

To obtain document clusters, we use a clustering based on multinomial mixture, an effective framework for text mining [26][24]. While we have also tested k-means method, no better results are obtained. Therefore, we do not include those results in this paper. In conducting a clustering with multinomial mixture, we randomly initialize model parameter values and execute EM algorithm [13] 20 times for each document set. We also use a smoothing techinque [24] and an annealing technique [25]. In applying a smoothing, we linearly mix the background word probability (i.e., the word probability in the entire document set) to the cluster-wise word probability. We test the following four mixture ratios for smoothing: 0.0, 0.01, 0.1 and 0.3. The ratio of 0.0 corresponds to the case where we have no smoothing effect. When the ratio is 0.3, for example, we use $(1-0.3)\times(\text{cluster-wise word probability})+0.3\times(\text{background word probability})$ in place of the cluster-wise word probability when updating parameter values in EM algorithm. Only for OKWAVE dataset, which includes many short articles, we use the mixture ratio 0.5 instead of 0.01 and consequently use the following four mixture ratios: 0.0, 0.1, 0.3 and 0.5. This is because large smoothing is likely to give clusters of good quality for a set of short documents.

4.3 Evaluation of Cluster Quality

We evaluate the quality of clusters by F-measure, the harmonic mean of precision and recall. Precision and recall are computed as follows. We call the category assigned to the largest number of articles in a given cluster *dominating category* of the cluster. The precision of a cluster is defined to be the ratio of the number of the articles of dominating category to the size of the cluster. The recall of a

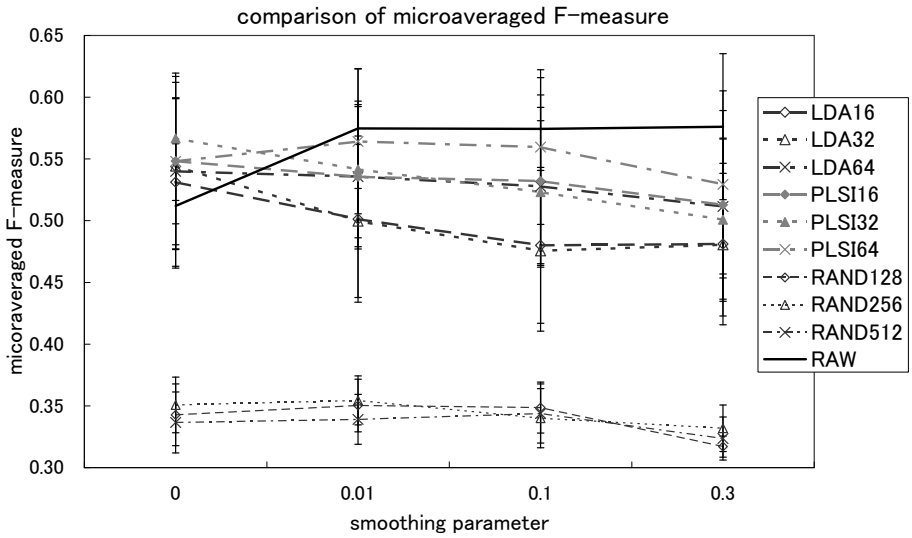

Fig. 1. Comparison of microaveraged F-measure for S2005 dataset.

cluster is the ratio of the number of the articles of dominating category to the number of articles of that category from the entire document set.

To obtain a precison and recall for each clustering result, we compute the sum of the numerators and the sum of the denominators used for computing precisions and recalls for different clusters included in the clustering result, and divide the former sum by the latter sum. Consequently, we have two evaluation measures called *microaveraged precision* and *microaveraged recall*. For example, when we have a clustering result consisting of three clusters whose precisions are 2/3, 5/8 and 4/7, microaveraged precision of this clustering result is (2+5+4)/(3+8+7). Microaveraged recall is also computed in the same manner. From definition, when there are at least one categories which do not dominate any clusters, microaveraged precision can be different from microaveraged recall. Therefore, we use the harmonic mean of microaveraged precision and microaveraged recall as an integrated evaluation for a clustering result. In this paper, we simply call this harmonic mean *F-measure* in the rest of the paper.

In our experiment, we run a clustering algorithm 20 times on a document set from randomly initialized parameter values. Consequently, we obtain 20 F-measures for each document set. We use the average and the standard deviation of these 20 F-measures for evaluating the performances of different dimensionality reduction methods. Since we use four mixture ratios (i.e., 0.0, 0.01, 0.1 and 0.3, or, 0.0, 0.1, 0.3 and 0.5 only for OKWAVE set) for smoothing as is described in Section 4.2, we have four evaluation results for each document set with respect to each dimentionality reduction method.

4.4 Evaluation Results

The evaluation results for S2005 and S2006 are provided in Fig. 1 and Fig. 2, respectively. For JIC, we obtain the results shown in Fig. 3. Finally, Fig. 4 shows the results for OKWAVE dataset. The horizontal axis represents mixture ratio of smoothing. The vertical axis represents F-measure. Each graph shows the average of 20 F-measures obtained from 20 executions of clustering. The width of each marker indicates plus/minus one standard deviation of the 20 microaveraged F-measures. In all figures, the graph labeled with RAW shows the average of 20 F-measures when we use no dimensionality reduction. Without dimensionality reduction, the quality of clusters gets better when we apply smoothing by choosing a non-zero value for mixture ratio.

The graphs labeled with LDA16, LDA32 and LDA64 present the averages of 20 microaveraged F-measures obtained when we reduce the vector dimension to 16, 32 and 64 via LDA, respectively. For any of these three cases, smoothing does not improve the quality of clusters. This seems because the dimensionality reduction implies a smoothing effect. Further, LDA provides F-measures comparable with RAW. We can say that LDA can reduce the dimension of feature vectors without degrading the cluster quality.

The graphs labeled with PLSI16, PLSI32 and PLSI64 indicate the results when we use pLSI for dimensionality reduction by setting the number of topics 16, 32 and 64, respectively. The standard deviation markers for pLSI intersect with those of LDA. Namely, LDA is not superior to pLSI as a dimensionality reduction method in document clustering.

However, both LDA and pLSI provide clusters of far better quality than random projection. The graphs having labels RAND128, RAND256 and RAND512 give the averages of 20 F-measures obtained by conducting a clustering on the feature vectors of dimension 128, 256 and 512, respectively, where dimensionality reduction is realized by random projection. When we reduce the dimension to 16, 32 or 64 by random projection, the cluster quality gets disastrous. Hence, only for random projection, we provide the evaluation results of clustering by using the vectors of dimension 128, 256 and 512. Further, for OKWAVE dataset, the reduced vectors obtained by random projection always give quite small F-measures (0.1181˜0.1331). Therefore, we do not include these results in Fig. 4. This fact suggests that random projection is not applicable to the feature vectors of short documents, i.e., document vectors with many zero entries.

The cluster quality obtained for S2005 is better than that for S2006, because the number of categories of S2005 is far less than that of S2006. Although the number of categories of JIC is a little larger than that of S2005, clusters of better quality are obtained for JIC than for S2005. This seems due to the fact that the average document length of JIC ($\approx$ 145.0) is far larger than that of S2005 ($\approx$ 70.5). Longer documents may result in document clusters of higher quality. The results of OKWAVE dataset are comparable with those of S2006 due to the similar number of categories.

As for LDA, we have an issue of computational resources. Our experiment is conducted on a desktop PC equipped with Intel Core2 6600 2.40GHz CPU and

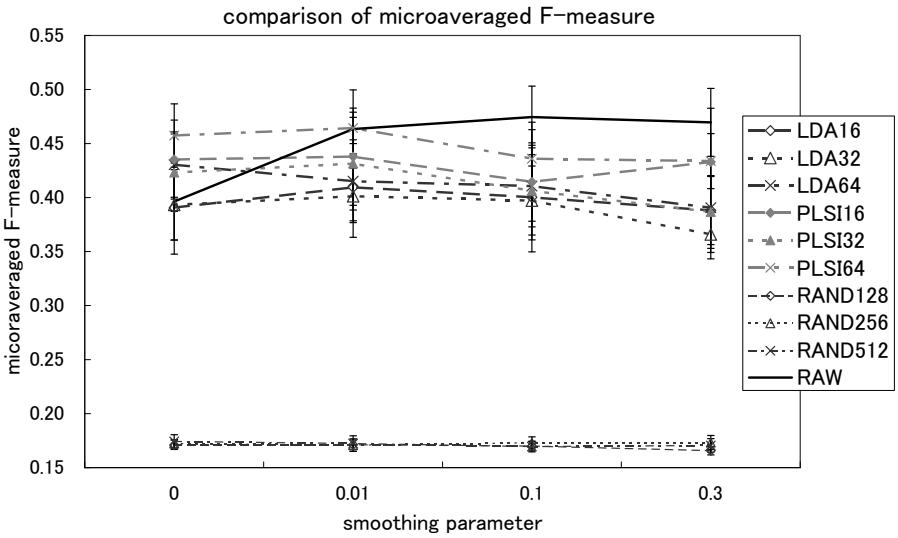

Fig. 2. Comparison of microaveraged F-measure for S2006 dataset.

with 2G byte main memory. For the dataset S2006 including 37,017 articles and 2,730,680 word tokens, the variational inference has required nearly 40 minutes (resp. 90 minutes) for the case of 16 topics (resp. 32 topics). When the number of topics is 64, the execution time has amounted to nearly five hours due to swapping. This issue can be addressed by splitting a document set into several subsets and parallelizing the computation as is described in [23]. However, our results show that pLSI is more favorable when computing resource is a severe problem. Further, even when the resource problem is not severe, we can use DPM [11], a more sophisticated version of LDA, at least for document clustering.

5 Conclusion

This paper provides the results of an evaluation experiment concerning dimensionality reduction in document clustering. We use LDA and pLSI to reduce the dimension of document feature vectors which are originally of dimension equal to the number of vocabularies. We conduct a clustering based on multinomial mixture for the set of the feature vectors of original dimension and for the set of the vectors of reduced dimension. We also compare LDA and pLSI with random projection. The results show that LDA can reduce the dimension of document feature vectors without degrading the quality of document clusters. Further, LDA is far superior to random projection. However, our experiment tells no significant difference between LDA and pLSI. When we consider the issue of computational cost, we have no positive reason to promote LDA beside pLSI for dimensionality reduction in document clustering.

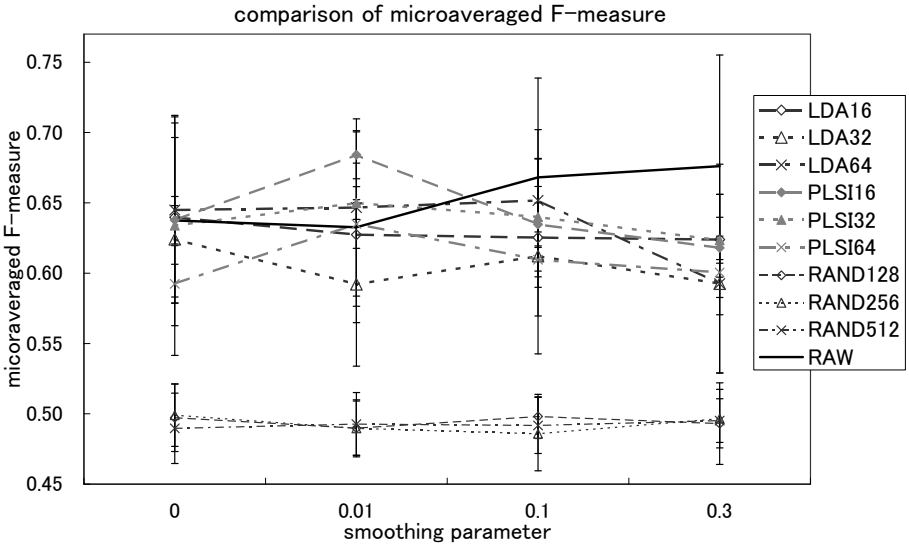

Fig. 3. Comparison of microaveraged F-measure for JIC dataset.

The variational inference for LDA, however, gives a wide variety of results. In this paper, we only use a part of the results, i.e., "pseudo-frequencies" of topics with respect to each document (γ_{ik} in Eq. 4). In addition to this, we can obtain topic probabilities with respect to each word token (ϕ_{ilk} in Eq. 3), word probabilities with respect to each topic (β_{kj} in Eq. 5) and α_k in Eq. 6 which can be regarded as an importance of each topic in the entire document set. These information cannot directly be obtained by Gibbs sampling, an alternative inference framework for LDA [16][22][23]. Our future work is to propose better applications of LDA to various text mining problems by utilizing the above parameters effectively.

References

1. http://japan.internet.com/
2. http://nlp.kookmin.ac.kr/HAM/kor/
3. http://mecab.sourceforge.net/
4. http://okwave.jp/
5. http://www.quintura.com/
6. http://www.seoul.co.kr/
7. http://vivisimo.com/
8. K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei and M. Jordan. Matching Words and Pictures. *Journal of Machine Learning Research*, Vol. 3, pp. 1107-1135, 2003.
9. E. Bingham and H. Mannila. Random Projection in Dimensionality Reduction: Applications to Image and Text Data. in *Proc. of KDD'01*, pp. 245-250, 2001.

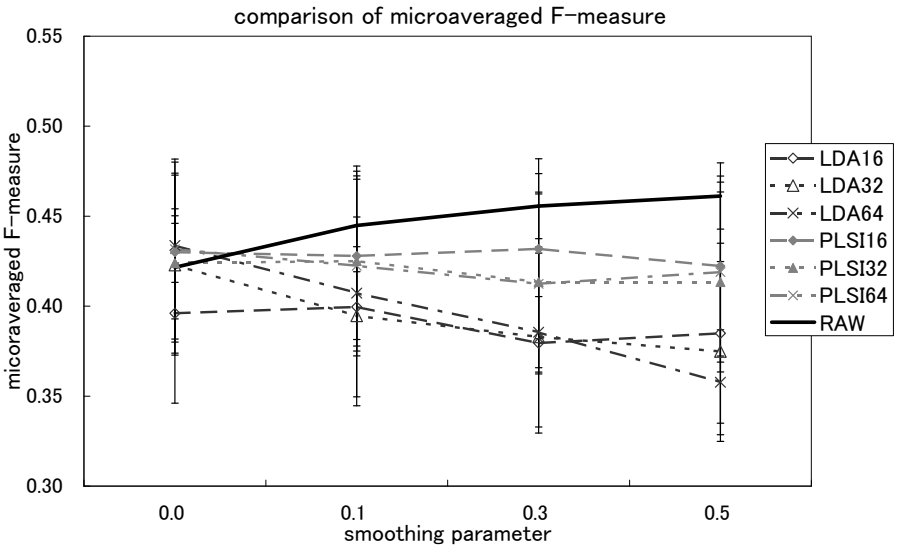

Fig. 4. Comparison of microaveraged F-measure for OKWAVE dataset. We do not include the cluster quality for the case of random projection, because this case almost always gives heavily degenerated clustering results where all input vectors are put into one or two clusters.

10. D. Blei, A. Y. Ng and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022, 2003.
11. D. Blei and M. I. Jordan. Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis*, Vol. 1, No. 1, pp. 121-144, 2005.
12. J. G. Conrad, K. Al-Kofahi, Y. Zhao and G. Karypis. Effective Document Clustering for Large Heterogeneous Law Firm Collections. in *Proc. of ICAIL'05*, pp. 177-187, 2005.
13. A. P. Dempster, N. M. Laird and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B, Vol. 39, No. 1, pp. 1-38, 1977.
14. P. K. Elango and K. Jayaraman. Clustering Images Using the Latent Dirichlet Allocation Model. 2005. (available at `http://www.cs.wisc.edu/~pradheep/`)
15. M. Fattori, Giorgio Pedrazzi and Roberta Turra. Text Mining Applied to Patent Mapping: a Practical Business Case. *World Patent Information*, Vol. 25, pp. 335-342, 2003.
16. T. Griffiths and M. Steyvers. Finding Scientific Topics. in *Proc. of the National Academy of Sciences*, 101 (suppl. 1), pp. 5228-5235, 2004.
17. T. Hofmann. Probabilistic Latent Semantic Indexing. in *Proc. of SIGIR'99*, pp. 50-57, 1999.
18. F.-C. Hsu, A. J.C. Trappey, C. V. Trappey, J.-L. Hou and S.-J. Liu. Technology and Knowledge Document Cluster Analysis for Enterprise R&D Strategic Planning. *International Journal of Technology Management*, Vol. 36, No.4 pp. 336-353, 2006.
19. R. E. Madsen, D. Kauchak and C. Elkan. Modeling Word Burstiness Using the Dirichlet Distribution. in *Proc. of ICML'05*, pp. 545-552, 2005.

20. T. J. Malisiewicz, J. C. Huang and A. A. Efros. Detecting Objects via Multiple Segmentations and Latent Topic Models. 2006. (available at `http://www.cs.cmu.edu/~tmalisie/`)
21. T. Minka. Estimating a Dirichlet distribution. 2000. (available at `http://research.microsoft.com/~minka/papers/`)
22. D. Mimno and A. McCallum. Expertise Modeling for Matching Papers with Reviewers. in *Proc. of KDD'07*, pp. 500-509, 2007.
23. D. Mimno and A. McCallum Organizing the OCA: Learning Faceted Subjects from a library of digital books. in *Proc. of JCDL'07*, pp. 376-385, 2007.
24. K. Nigam, A. McCallum, S. Thrun and T. Mitchell. Text Classification from Labeled and Unlabeled Documents Using EM. *Machine Learning*, Vol. 39, No. 2/3, pp. 103-134, 2000.
25. K. Rose, E. Gurewitz, and G. Fox, A Deterministic Annealing Approach to Clustering. *Pattern Recognition Letters*, Vol. 11, pp. 589-594, 1990.
26. M. Sahami, S. Dumais, D. Heckerman and E. Horvitz. A Bayesian Approach to Filtering Junk Email. AAAI Technical Report WS-98-05, 1998.
27. Y. W. Teh, D. Newman and M. Welling. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. in *Proc. of NIPS'06*, pp. 1353-1360, 2006.
28. M. Yamamoto and K. Sadamitsu. Dirichlet Mixtures in Text Modeling. CS Technical report CS-TR-05-1, University of Tsukuba, 2005.
29. H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma. Learning to Cluster Web Search Results. in *Proc. of SIGIR'04*, pp. 210-217, 2004.