# A proposal of document recommendation based on topic model

**Yusei Yamamoto†**
Graduate School of Creative Science and Engineering,
Waseda University, Tokyo, Japan
Tel: (+81) 3-5286-3290, Email: yusei0809@ruri.waseda.jp

**Kenta Mikawa**
School of Information and Engineering,
Shonan Institute of Technology, Kanagawa, Japan
Tel: (+81) 466-30-0212, Email: mikawa@info.shonan-it.ac.jp

**Masayuki Goto**
School of Creative Science and Engineering,
Waseda University, Tokyo, Japan
Tel: (+81) 3-5286-3290, Email: masagoto@waseda.jp

**Abstract.** With rapid development on information society, the technology of document recommendation system which provides relevant information to user's interest from a large amount of text data has become even more important. For document recommendation, a method based on content-based filtering has been proposed, which selects documents for recommendation by calculating the similarity between documents which an active user has already read and making a list of candidate documents by using a data compression algorithm. However, this method needs a dictionary which includes all words appearing in the documents read by a user. Therefore, if the amount of documents read by a user is not enough, the precision of the similarity measure between documents can become worse. Moreover, it is usually natural to suppose several topics depending on categories in document data such as newspaper articles, whereas the conventional method cannot take account of these topic variations. To solve these problems, in this paper, we propose a new method for document recommendation which is based on the collaborative filtering with the topic model, thereby can achieve more generalization ability. We show that the proposed method provides more effective recommendation accuracy than the conventional method through the result of experiment.

**Keywords:** Document recommendation, Topic Model, Collaborative filtering, Content-based filtering

## 1. INTRODUCTION

With rapid development on information society, the technology of document recommendation system which provides relevant information according to user's interest from a large amount of document data has become even more important. For document recommendation, the method of content-based filtering using a data compression algorithm has been proposed (T. Watanabe et al. 2002). This method first needs a dictionary which represents all symbols appearing in browsed documents by using the LZ78 algorithm. The LZ78 algorithm is a source coding method which achieves the efficient compression by registering past matching sequences with the dictionary and coding a next sequence by using the number in the dictionary. Then, this method can calculate the similarity between documents which an active user has already browsed and each of candidate documents by referring above a dictionary. Finally, it selects a document with the highest similarity from a list of candidates to recommend.

However, if an amount of documents read by a user is larger, this method needs large computational cost for compressing all browsed documents, because this method needs to compress each of browsed documents individually referring all series of symbols appearing in each browsed document. Moreover, this method makes dictionaries from each of a set of browsed documents individually, therefore, it cannot take accounts of the total characteristics which a set of browsed documents have.

To solve these problem, Suzuki et al. proposed the method which unites all of various series of symbols which appeared in a set of browsed documents into a single larger

series of symbols, thereby can achieve more efficient calculation (T. Suzuki et al., 2011). Then, this creates a single dictionary for calculating the similarity. That is, this method calculates the similarity of each browsed document to a candidate by a single dictionary, although the previous method calculates the similarity by using multiple dictionaries.

However, this conventional method calculates the similarity using a single dictionary which is made by referring series of all symbols appearing in a united document. Therefore, if the different types of topics exist in a united document which a user has read, this method cannot achieve the stable recommendation. Moreover, it can be assumed that the tendencies of word frequencies vary depending on topics in documents, whereas the conventional method cannot take accounts of these variations. For example, in the case of newspaper articles, the words appearing in the topic "world cup soccer" are usually different from those in the topic "Major league baseball".

From the previous discussions, we propose a new method for document recommendation based on the topic model. Letting "topics" be various underlying latent characteristics of document in a category, e.g. "baseball" and "soccer" in a category "sports", the topic model can express these topics as the generative probability. In this study, we focus on one of the multi-topic models, Latent Dirichlet Allocation, which can express multiple topics of a document by assuming that each word in a document would has their own topic respectively (D. Blei 2001). Then, with this methods, the proposed method determines the documents for recommendation by calculating the similarities between a set of browsed documents and a candidate document.

And through the result of the experiment with Japanese news articles, we show the effectiveness of our proposed method.

## 2. PRELIMINARIES

In this section, we describe about the LZ78 algorithm which is used for calculating the similarity between a set of document which a user has read and a candidate document to recommend, and about the setting of document recommendation. Then, we describe about the previous study of document recommendation with the data compression algorithm.

### 2.1 DOCUMENT RECOMMENDATION

Letting $N$ numbers of all document data be $D = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \dots, \boldsymbol{d}_n, \dots, \boldsymbol{d}_N\}$, we assume that a user has already read $J$ documents $D_G = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \dots, \boldsymbol{d}_j, \dots, \boldsymbol{d}_J\}$. And we designate a list of $U$ numbers of candidate documents for recommendation as $D_U = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \dots, \boldsymbol{d}_u, \dots, \boldsymbol{d}_U\}$, where both of browsed documents $D_G$ and candidate documents $D_U$ are subset of all documents data $D$, that is, $D_G \subset D$, $D_U \subset D$. In document recommendation, given a set of documents $D_G$, we calculate the similarity between each browsed document in $D_G$ and each candidate document in $D_U$, then, determine $\boldsymbol{d}_u$ with the highest similarity from a list of candidate documents in $D_U$.

### 2.2 LZ78 ALGORITHM

We explain about the LZ78 algorithm which is used in the conventional method for calculating the similarity between a set of documents which a user has read and a list of candidates to recommend. The LZ78 algorithm is the compression technique proposed by Ziv and Lempel (J. Ziv and A. Lempel ,1978). Here, the $j$-th document $\boldsymbol{d}_j$ is expressed as a series of symbols appearing in order $\boldsymbol{d}_j = (d_{j1}, d_{j2}, \dots, d_{jR_j})$, where $R_j$ is the length of a sequence of symbols in $j$-th document $\boldsymbol{d}_j$. Here, a input document $\boldsymbol{d}_j$ is compressed as double $< i, c >$, where we let $i$ be the dictionary number with which the longest matching is registered and $c$ be the symbols following the longest matching.

For example, if the document data $\boldsymbol{d} = \{"a", "b", "a", "b", "a", "a"\}$ is given, this document compressed as $< 0, a >, < 0, b >, < 1, b >, < 1, a >$ , and a dictionary as below is created (Table1). There, the compression rate is calculated as 0.6667.

Table1: An example of a dictionary

| Dictionary Number | Symbol |
|---|---|
| 0 | NULL |
| 1 | a |
| 2 | b |
| 3 | ab |
| 4 | aa |

### 2.3 Pattern Representation scheme using Data Compression

*Pattern Representation scheme using Data Compression* (PRDC) is the method which calculates the similarity between a set of browsed document $D_G$ and each candidate $\boldsymbol{d}_u$ with the LZ78 algorithm, then determine the candidate documents for recommendation. After $J$ numbers of dictionaries $e = \{e_1, e_2, \dots, e_J\}$ are created from each in the browsed documents set $D_G =$

$\{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_j, \ldots, \boldsymbol{d}_J\}$ with the LZ78 algorithm, the similarity between a candidate document $\boldsymbol{d}_u$ and a browsed document $\boldsymbol{d}_j$, $\mathrm{sim}(\boldsymbol{d}_u, \boldsymbol{d}_j)$, is calculated as follows:

$$\mathrm{sim}(\boldsymbol{d}_u, \boldsymbol{d}_j) = \frac{l_{com}(\boldsymbol{d}_u, e_j)}{l_{in}(\boldsymbol{d}_u)}, \quad (1)$$

where $l_{com}(\boldsymbol{d}_u, e_j)$ is the length of sequence of the document $\boldsymbol{d}_u$ after the compression by the $j$-th dictionary $e_j$, and $l_{in}(\boldsymbol{d}_u)$ is the original length of the document $\boldsymbol{d}_u$ before the compression. For recommendation, after calculating the similarity between document $\boldsymbol{d}_u$ and a set of browsed documents $D_G$ by the equation (2), the document with the highest similarity is determined to be recommended.

$$\mathrm{sim}(\boldsymbol{d}_u, D_G) = \min_{\boldsymbol{d}_j \in D_G} \left( \frac{l_{com}(\boldsymbol{d}_u, e_j)}{l_{in}(\boldsymbol{d}_u)} \right). \quad (2)$$

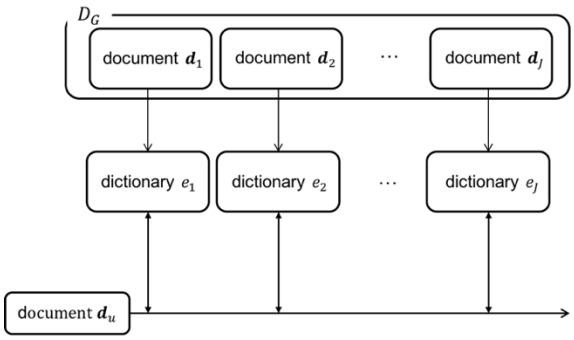


Figure1. The graphical image of PRDC.

## 3. CONVENTIONAL METHODS

### 3.1 Pattern Representation scheme using Data Compression with a United Document

In the PRDC, the similarity between a candidate document $\boldsymbol{d}_u$ and a browsed document $\boldsymbol{d}_j$ is calculated by each of $J$ numbers dictionary $e_1, e_2, \ldots, e_J$. Therefore, if an amount of documents read by a user is larger, this method costs huge amounts of calculation for compressing each browsed document $\boldsymbol{d}_j$. Moreover, because this method compresses a set of browsed documents $D_G$ individually, it cannot take account of the total characteristics of user's preference.

To solve these problem, Suzuki et al. proposed the method, *Pattern Representation scheme using Data Compression with a United Document* (PRDCUD), which creates a single larger dictionary $E$ by compressing a single united document $\overline{D}_G$ which is made by uniting all browsed documents $D_G = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_J\}$ into a single sequence, $\overline{D}_G = \left(d_{11}, \ldots, d_{1R_1}, d_{21}, \ldots, d_{2R_2}, \ldots, d_{J1}, \ldots, d_{JR_J}\right)$ instead of creating $J$ numbers of dictionaries $e_1, e_2, \ldots, e_J$ from each document in a set of browed documents $D_G$, Consequently, the similarity between a candidate document $\boldsymbol{d}_u$ and a set of browsed documents $D_G$, $\mathrm{sim}(\boldsymbol{d}_u, D_G)$, is calculated as follows:

$$\mathrm{sim}(\boldsymbol{d}_u, D_G) = \mathrm{sim}(\boldsymbol{d}_u, \overline{D}_G) = \frac{l_{com}(\boldsymbol{d}_u, E)}{l_{in}(\boldsymbol{d}_u)}, \quad (3)$$

where, $l_{com}(\boldsymbol{d}_u, E)$ is the length of sequence of the document $\boldsymbol{d}_u$ after the compression by a single dictionary $E$.

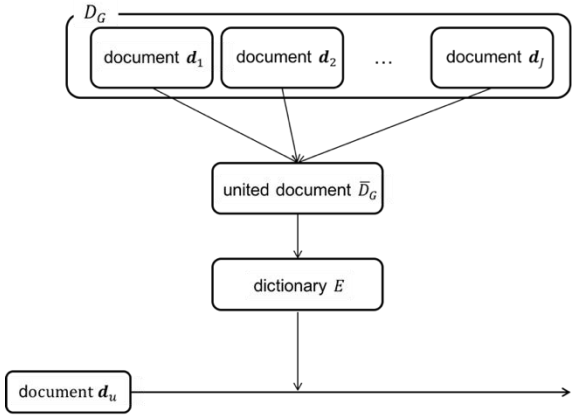


Figure2. The graphical image of PRDCUD.

### 3.2 COMBINED METHOD

In the combined method, the similarity between a candidate document $\boldsymbol{d}_u$ and a set of browsed documents $D_G$ is calculated also from the viewpoint of bag-of-words, in addition to PRDCUD. Here, we complementary designated the vector of the $j$-th document $\boldsymbol{d}_j$ as $\boldsymbol{v}_{\boldsymbol{d}_j} = (x_{j1}, x_{j2}, \ldots, x_{jV})$, where $V$ is the size of vocabulary (the number of words) that appear in all documents, and $x_{jv}$ is the number of frequencies of $v$-th word $w_v$ in the $j$-th document $\boldsymbol{d}_j$. The word set is denoted by $W = \{w_1, w_2, \ldots, w_V\}$. Then, letting $\boldsymbol{v}_{\overline{D}_G}$ be the vector that represents the bag-of-words of a united document $\overline{D}_G$, and letting $\boldsymbol{v}_{\boldsymbol{d}_u}$ be the vector of the bag-of-words of a candidate document $\boldsymbol{d}_u$, the similarity between a candidate document $\boldsymbol{d}_u$ and a set of browsed documents $D_G$ is calculated by

the equation (4).

$$sim(\boldsymbol{d}_u, \bar{D}_G) = \frac{\boldsymbol{v}_{\boldsymbol{d}_u} \cdot \boldsymbol{v}_{\bar{D}_G}}{|\boldsymbol{v}_{\boldsymbol{d}_u}| \cdot |\boldsymbol{v}_{\bar{D}_G}|} \log\left(\frac{l_{com}(\boldsymbol{d}_u, E)}{l_{in}(\boldsymbol{d}_u)}\right) \quad . \tag{4}$$

And every component comprising both of the vectors, $\boldsymbol{v}_{\boldsymbol{d}_u}$ and $\boldsymbol{v}_{\bar{D}_G}$ is modified by weighting with the tf-idf measure as follows:

$$x_{iv} = \frac{\log(tf_v + 1)}{\log(V_G)} \log\left(\frac{N}{idf_v}\right) \tag{5}$$

where, $tf_v$ is the number of frequency of the $v$-th word $w_v$, $idf_v$ is the number of documents which include the $v$-th word $w_v$, and $V_G$ is the total summed amount of frequencies of all words in a set of browsed documents $D_G$ and a candidate document $D_U$.

## 4. PROPOSED METHOD

In this section, we describe about the document recommendation based on the topic model. Especially, we focus on the Latent Dirichlet Allocation which is one of the multi-topic models. Then, we describe about how to calculate the similarity between a set of browsed documents and a candidate document for recommendation.

### 4.1 LATENT DIRICHLET ALLOCATION

We express "topics" as the various underlying latent characteristics splitting a category, e.g. "baseball", "soccer" and "rugby" in the category "sports". Generally speaking, the tendency of word frequencies is mainly dependent on these topics in each document.

For modeling these topics of each document, we focus on the Latent Dirichlet Allocation (LDA) which is one of multi-topic models. The LDA represents a document with multiple topics by assuming that each word in a document has their own topic respectively.

Here, we redefine the characteristic of $j$-th document $\boldsymbol{d}_j$ as $\boldsymbol{w}_{\boldsymbol{d}_j} = \left(w_{\boldsymbol{d}_j,1}, w_{\boldsymbol{d}_j,2}, \dots, w_{\boldsymbol{d}_j,n_{\boldsymbol{d}_j}}\right)$, where $w_{\boldsymbol{d}_j,l}$ is the $l$-th word in document $\boldsymbol{d}_j$ and $n_{\boldsymbol{d}_j}$ is the total summed amounts of word frequencies in the document $\boldsymbol{d}_j$. Then, each latent variable $\boldsymbol{z}_{\boldsymbol{d}_j} = \left(z_{\boldsymbol{d}_j,1}, z_{\boldsymbol{d}_j,2}, \dots, z_{\boldsymbol{d}_j,n_{\boldsymbol{d}_j}}\right)$ is assigned to each word in the document, which designates a latent topic to which each word belongs. Now, letting $\boldsymbol{\theta}_{\boldsymbol{d}_j} = \left(\theta_{\boldsymbol{d}_j 1}, \theta_{\boldsymbol{d}_j 2}, \dots, \theta_{\boldsymbol{d}_j M}\right)$ be the belonging probability of document to each latent topic, where $M$ is the number of latent topics, the total probability $p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta} | \boldsymbol{\alpha}, \boldsymbol{\beta})$ is expressed as follows:

$$p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta} | \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{n=1}^{N} p(\boldsymbol{\theta}_{\boldsymbol{d}_n} | \boldsymbol{\alpha}) \prod_{v=1}^{n_{\boldsymbol{d}_j}} p(z_{\boldsymbol{d}_n\ v} | \boldsymbol{\theta}_{\boldsymbol{d}_n}) p(w_{\boldsymbol{d}_n, v} | z_{\boldsymbol{d}_n\ v}, \boldsymbol{\beta}) \tag{6}$$

where, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the hyper parameters, and the Latent Dirichlet Allocation can model the various types of topics by adjusting these hyper parameters, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

### 4.2 RECOMMENDATION ON TOPIC MODEL

The parameters $\boldsymbol{\theta}_{\boldsymbol{d}_n} = \left(\theta_{\boldsymbol{d}_n 1}, \theta_{\boldsymbol{d}_n 2}, \dots, \theta_{\boldsymbol{d}_n M}\right)$ of all documents $D$ are estimated by the LDA, which designates the belonging probability of a document to each of $M$ topics. After that, the proposed method first calculates the similarity among all documents by adopting the clustering method to these parameters. Among them, all of a set of browsed documents $\boldsymbol{d}_j$ and a set of candidate documents $\boldsymbol{d}_n$ are assigned to one of $K$ clusters $S = \{S_1, S_2, \dots, S_K\}$, where we designated $S$ as a set of clusters and $S_k$ as the $k$-th cluster.

Here, we denote $\boldsymbol{\mu}_k$ as the centroid of the $k$-th cluster $S_k$, that is calculated by the equation (7),

$$\boldsymbol{\mu}_k = \frac{1}{|S_k|} \sum_{\boldsymbol{d}_n \in S_k} \boldsymbol{\theta}_{\boldsymbol{d}_n} \quad , \tag{7}$$

where, $|S_k|$ is the number of documents belonging to the $k$-th cluster $S_k$.

Given the centroid of each cluster $\boldsymbol{\mu}_k$, the similarity between these centroids and each parameter $\boldsymbol{\theta}_{\boldsymbol{d}_n}$ is calculated in Euclidian distance. Then, each document $\boldsymbol{d}_n$ is assigned to $\hat{k}$-th cluster $S_{\hat{k}}$ with the minimum Euclidian distance between the centroid $\boldsymbol{\mu}_k$ and $\boldsymbol{\theta}_{\boldsymbol{d}_n}$.

$$\hat{k} = \min_{\boldsymbol{\theta}_{\boldsymbol{d}_n}} \left\| \boldsymbol{\mu}_k - \boldsymbol{\theta}_{\boldsymbol{d}_n} \right\|^2 \quad , \tag{8}$$

These calculations, the equations (6) and (7), are iterated until the belonging clusters of all documents are unchanged.

After above procedure of clustering is converged, the similarity between a candidate document $\boldsymbol{d}_u$ and a set of browsed documents $\bar{D}_G$ is calculated. And this calculation is applied only to the sets of a candidate document $\boldsymbol{d}_u$ and

a browsed document $\boldsymbol{d}_j$ belonging to the same cluster. Then, the document with the highest similarity is recommended.

$$\mathrm{sim}(\boldsymbol{d}_u, \bar{D}_G) = \min_{\boldsymbol{d}_u, \boldsymbol{d}_j \in S_k} \mathrm{sim}(\boldsymbol{d}_u, \boldsymbol{d}_j) = \frac{\Sigma m \theta_{\boldsymbol{d}_u m} \cdot \theta_{\boldsymbol{d}_j m}}{|\boldsymbol{\theta}_{\boldsymbol{d}_u}| \cdot |\boldsymbol{\theta}_{\boldsymbol{d}_j}|}. \quad (9)$$

Here, we show the algorithm of the proposed method as follows.

Table2: The algorithm of the proposed method.

| | |
|---|---|
| **Step1** | Learn each belonging probability $\boldsymbol{\theta}_{\boldsymbol{d}_j} = \left(\theta_{\boldsymbol{d}_j 1}, \theta_{\boldsymbol{d}_j 2}, \ldots, \theta_{\boldsymbol{d}_j M}\right)$ of all documents $D$ to each of $M$ latent topics by adopting LDA. |
| **Step2** | Assign all documents data $D$ to one of K clusters. |
| **Step3** | Calculate the similarity between a candidate document $\boldsymbol{d}_u$ and a browsed document $\boldsymbol{d}_j$ $\mathrm{sim}(\boldsymbol{d}_u, \boldsymbol{d}_j)$, where this calculation is applied only to the sets belonging to the same cluster. |

## 5. EXPERIMENTS

In this section, we describe the experiments of recommendation with newspaper articles, and we show the effectiveness of our proposed method in comparison with the conventional method.

### 5.1 EXPERIMENTAL CONDITION

For the experiment, we used Mainichi newspaper articles published in 2010, where the number of categories is 8. And we extract 300 numbers of documents from each category so that the total number of documents is 2,400. Then, we select 30 documents as the browsed documents from arbitrary $T(\leq 8)$ numbers of categories. And we select 400 documents as the candidate documents. Among them, 50 are relevant documents, i.e., documents extracted from $T$ numbers of categories. And the rest 350 are irrelevant documents, i.e., documents extracted from other categories. And we change $T$ from 1 to 5.

For example, in the case $T = 2$, relevant documents are extracted as 15 documents each (total is 30) from two categories, and from 6 categories irrelevant documents are extracted.

In the proposed method, we empirically set the number of latent topics as $M = 100$, and hyper parameter $\boldsymbol{\alpha}$ as 0.5 and hyper parameter $\boldsymbol{\beta}$ as 0.01 respectively. And we adopt Gibbs Sampling to estimate the belonging probability $\boldsymbol{\theta}_{\boldsymbol{d}_n}$ of each document. And we updated the topics 1,000 times with Gibbs Sampling. For clustering, we set the number of clusters as $K = 15$.

In the conventional method, after the similarities of all candidate documents to a set of browsed documents are sorted, the top $r$ documents are recommended. On the other hand, in the proposed method, after each similarity of all candidate documents to a set of browsed documents which belong to the same cluster as the browsed documents are sorted, the top $r$ documents are recommended.

And we evaluate both of the conventional and the proposed method by $F$-measure which is calculated as follows:

$$accuracy = \frac{RD_{\leq r}}{r}, \quad (10)$$

$$recall\_rates = \frac{RD_{\leq r}}{RD}, \quad (11)$$

$$F = \frac{2 \times accuracy \times recall\_rates}{accuracy + recall\_rates}, \quad (12)$$

where $RD$ is the number of all relevant documents, $RD_{\leq r}$ is the number of relevant documents in the top $r$ documents. And we change the rank $r$ from 25 to 75, then, in each rank, we calculated $F$-measure ten times, and calculated the averages of all $F$-measures as the result.

### 5.2 RESULT OF EXPERIMENTS

We show the result of the experiments in Table3.

Table3: The result of conventional and proposed method.

| Number of Categories | PRDCUD | Combine Method | Proposed Method |
|---|---|---|---|
| $T = 1$ | 0.3806 | 0.3942 | **0.6222** |
| $T = 2$ | 0.3093 | 0.2984 | **0.5755** |
| $T = 3$ | 0.2230 | 0.2382 | **0.4598** |
| $T = 4$ | 0.1960 | 0.2272 | **0.3647** |
| $T = 5$ | 0.1677 | 0.2034 | **0.2531** |

From Table3, we can show the effectiveness of the proposed method in all extracted numbers of categories $T$. For document data where the variation of words appearing in the document changes according to their topics such as newspaper article, we can conclude that it is more effective to refer the word frequencies in the

document than the sequence of words. That is why the combined method can improve $F$-measure a little in comparison with the PRDCUD method. However, this method cannot grasp the characteristics of all documents, thus, if an amount of documents browsed by a user is not enough, this method cannot achieve stable recommendation. On the other hand, the proposed method can take account of topics in each document by training from all documents. That is why our proposed method shows better $F$-measure consistently.

However, as the number of extracted categories $T$ becomes larger, it can be thought that it becomes more difficult to grasp the variation of word frequencies. That is why in the case that $T$ becomes larger, $F$-measure of the proposed method becomes worse gradually.

## 6. CONCLUSION AND FUTURE WORK

In this paper, we propose the new method for document recommendation using the Latent Dirichlet Allocation. While the conventional method refers only the sequence of words appearing in a document, the proposed method assumes the latent characteristic of a document underlying a category. Through the result of the experiment with Japanese newspaper articles, we show the effectiveness of the proposed method in comparison with the conventional methods. For recommending the document data whose word frequencies depends mainly on their topics such as newspaper articles, the proposed method can be more suitable. It is the future work to investigate the hybrid method for document recommendation combining data compression and the topic model.


### ACKNOWLEDGEMENT

The authors would like to express thanks to Mr. Yang Tianxiang, Mr. Gendo Kumoi, Dr. Haruka Yamashita, and all members of Goto laboratory for their support for our research. This study was partly supported by JSPS KAKENHI Grant Numbers, 26282090 and 26560167.


## REFERENCES


Bishop, C. M. (2006) Pattern Recognition and Machine Learning, Springer.

T. Suzuki, S. Hasegawa, T. Hamamoto, A. Aizawa (2011) Document Recommendation Using Data Compression, *Procedia – Social and Behavioral Sciences*, **27**, 150-159.

J. Ziv and A. Lempel (1978) Compression of individual sequences via variable-rate coding, *IEEE Trans. Information Theory*, **24**, 530-6

T. Watanabe, K, Sugawara, and H. Sugihara (2002) A new pattern representation scheme using data compression, *IEEE Trans. Pattern Analysis and Machine Intelligence*, **24**, 570-90

T. Hoffman (1999) Probabilistic Latent Semantic Analysis, *In Proc. of Uncertainty in Artificial Intelligence*, UAI99, 289–296

D. Blei, A.Y. Ng and M. Jordan, (2001) Latent Dirichlet Allocation, *Journal of Machine Learning Research*, **3**, 993-1022

T. Griffiths and M. Steyvers (2004) Finding Scientific Topics, *Proceedings of the National academy of Sciences of the United States of America*, 1, 5228-5235