Information Engineering Express
International Institute of Applied Informatics
2015, Vol.1, No.4, 65 – 74

Document Classification using Matrix Decomposition with Varied Viewpoints

Kaname Maruta [*] , Hidetoshi Nagai [*] , Teigo Nakamura [*]

Abstract

Classification results are not unique and can vary according to the user's viewpoint. If a document classification system ignores the user's viewpoints, classification will be different from the result desired by the user, and the difference between the user's desired result and the system's produced result can cause some inhibitions and oversights in information retrieval. Extracting the user's viewpoints from the classification examples performed preliminarily by the user allows us to configure classifications that reflect the user's desire. In this study, we propose four methods to extract viewpoints and three methods to classify documents using Nonnegative Matrix Factorization (NMF) matrix decomposition. We exhibit the results of comparative experiments with the original NMF, Semi-Supervised NMF (SSNMF) and our proposed methods.

Keywords: Document Classification, NMF, Matrix Decomposition,

1 Introduction

In recent years, the collection of information has become easier through the use of the Internet. Since there is an exhaustive collection of documents on the Internet, an efficient search method to find desired information is required. Search results clustering is one such efficient search method. It can locate the desired documents effortlessly by focusing on the target category. However, the resulting document classification is not unique to a particular target category, and there can be many returned results based on different category perspectives. In other words, the classification results can vary according to a user's viewpoints of classification. A document that may belong to more than one class is called a multi-label document. Similarities between two multi-label documents in a particular class are usually high and thus it is often difficult to classify them individually. If a document classification system ignores the user's viewpoints, the resulting classification will be different from the user's desired result, and the difference between the user's desired result and that of the system's can cause some inhibitions and oversights in information retrieval. Extracting the user's viewpoints from the classification examples that is performed preliminarily by the user allows us to configure classifications that reflect the user's desire.

[*] Kyushu Institute of Technology, Hukuoka, Japan

The user's viewpoint influences the classification result. We extract the viewpoint from the document vector of supervised document that is classified preliminarily by the user's viewpoint. The feature value of the document vector is term frequency.

We propose four methods of extracting the viewpoint information from the supervised document. Each method collects the viewpoint information for each class. We expect the extracted viewpoint information in all classes to be an approximation of the user's viewpoint. The first method of extracting viewpoint information uses the mean feature value of the correct class. The second method uses the ratio of the mean feature value in the correct and incorrect classes. The third method uses the max feature value of the correct class, and the fourth method uses the ratio of the max feature value in the correct and incorrect classes. Finally, we describe these methods in detail in Section 2.

After the viewpoint extraction, we use the viewpoint information for document classification. We propose three document classification methods using matrix decomposition. The classification methods decompose the document matrix into the basis matrix and reconstruction coefficient matrix. The reconstruction coefficient matrix represents the degree of relevance between each document and each class. Each document can then be classified into the class of maximum relevance value.

The first classification method uses simple matrix decomposition based on the pseudo-inverse matrix of the viewpoint matrix U_m. The second method of classification uses matrix U in NMF[1][2] in addition to method 1. The third method uses NMF-I[3]. We describe these methods in detail in Section 4.

We present the experimental results to classify the actual documents and compare the original NMF, Semi-Supervised NMF (SSNMF)[4] with our proposed method in Section 5.

2 User's viewpoint for classification

It is difficult to clearly describe the user's viewpoint of classification, because it largely depends on user's sensitivity. So, we approximate the user's viewpoints using the extracted information from the supervised documents. Therefore, we extract the viewpoint information from the document vector of supervised documents that are classified by the user's viewpoint.

We propose four methods of extracting the viewpoint information from the document vector. In the description of each extraction method, we describe how to calculate the degree of contribution of a term t to a class A.

$D = \{d_1, d_2, \ldots, d_n\}$ is a set of documents, and n is the total number of documents. $D_A = \{d_{A_1}, d_{A_2}, \ldots, d_{A_m}\} \subseteq D$ is a set of documents of class A. $m = \#(D_A)$ is the number of documents of class A. $\overline{D}_A$ is a complementary set of D_A. $f(D,d,t)$ is TF·IDF value of term t in the document d in D. A viewpoint vector of class A is a vector whose values are the degree of contribution of each term and a viewpoint matrix is a collection of viewpoint vectors for all classes.

2.1 Extraction method 1 (EM-1) - Mean

The degree of contribution of a term t to a class A is the mean value of $f(D,d,t)$ over $d \in D_A$. EM-1 can extract the average characteristics of each class. When a term t appears in many documents in the class, the degree of contribution of viewpoint becomes high.

$$\operatorname*{mean}_{d \in D_A} f(D,d,t) \tag{1}$$

2.2 Extraction method 2 (EM-2)- Raito of mean

The degree of contribution of a term t to a class A is the ratio of the mean value of $f(D,d,t)$ over $d \in D_A$ to the one over $d \in \overline{D}_A$. When a term t appears in many documents in $\overline{D}_A$, the degree of contribution of viewpoint becomes low.

$$\underset{d \in D_A}{\text{mean}} f(D,d,t) / \underset{d \in \overline{D}_A}{\text{mean}} f(D,d,t) \tag{2}$$

2.3 Extraction method 3 (EM-3) - Max

The degree of contribution of a term t to a class A is the maximum value of $f(D,d,t)$ over $d \in D_A$. For a term t, when the value of $f(D,d,t)$ for a document d is large and the value of $f(D,d',t)$ for the other documents d' in D_A is small, the value of EM-1 becomes low because of averaging, but the value of EM-3 is high because of maximization.

$$\max_{d \in D_A} f(D,d,t) \tag{3}$$

2.4 Extraction method 4 (EM-4) - Raito of max

The degree of contribution of a term t to a class A is the ratio of the maximum value of $f(D,d,t)$ over $d \in D_A$ to the one over $d \in \overline{D}_A$. We expect an effect of both EM-2 and 3.

$$\max_{d \in D_A} f(D,d,t) / \max_{d \in \overline{D}_A} f(D,d,t) \tag{4}$$

3 Background and Related Work

3.1 NMF

NMF decomposes a document matrix $X \in \mathbb{R}^{w \times n}$ into a basis matrix $U \in \mathbb{R}^{w \times k}$ and a reconstruction coefficient matrix $V \in \mathbb{R}^{n \times k}$, where w is the total number of feature, and k is the number of classes.

$$X \simeq UV^T \tag{5}$$

The basis matrix U and the reconstruction coefficient matrix V are constrained to non-negative. Decomposition by NMF reduces the dimension of the document data. In other words, the w-by-n matrix X is reduced to the k-by-n matrix V. NMF is suitable for clustering high dimensional and sparse data and it is often used for document classification. The matrix V is used for clustering. The j-th column of the matrix V represents the degree of relevance between the j-th class and each document. Therefore, the i-th document is classified by the formulas (6).

$$\arg\max_j \; v_{ij} \;\; , \tag{6}$$

where $v_{i,j}$ is the i-th row and the j-th column of the matrix V. The decomposed matrices U and V are estimated by solving the minimization problem for the optimal objective function J of NMF.

$$J = ||X - UV^T||^2 \tag{7}$$

Minimization optimal problem is solved using the method of Lagrange multipliers. Then update equation of the matrices U and V are obtained[1][2].

$$v_{ij} \leftarrow v_{ij} \frac{(X^T U)_{ij}}{(VU^T U)_{ij}} \quad , \quad u_{ij} \leftarrow u_{ij} \frac{(XV)_{ij}}{(UV^T V)_{ij}} \quad , \tag{8}$$

where $u_{i,j}$ and $v_{i,j}$ are the i-th row and the j-th column of the matrix U and the matrix V, respectively. $X_{i,j}$ is the i-th row and the j-th column of the matrix X. The initial value of the matrices U and V are usually composed of random values. And, there are a variety of extended version in NMF[3][4], a technique that combines another clustering method and NMF[5][6].

3.1.1 Initial value issues

Clustering results by NMF depends on the initial value of the matrices U and V. In other words, the clustering results vary when the different initial values are given. Therefore, it is necessary to select a good initial value[5][7], but it is usually a random value.

3.2 NMF-I

As described above, the clustering result depends on the initial values of the matrices U and V. Random values are usually used as initial values and the clustering results may converge as bad local solutions.

Therefore, we proposed the NMF-I[3] that solves the problem. NMF-I is suitable for supervised document classification. NMF-I is a method that uses the matrix U_s as the initial value of the matrix U . U_S is calculated from the supervised data. We expect U_s is close to the convergence value of the basis matrix in an ideal document classification.

$$U_s = X_{\text{train}} (V_{\text{train}}^T)^+ \tag{9}$$

In Eq.(9), $X_{\text{train}} \in \mathbb{R}^{w \times s}$ is a document matrix of only supervised data, $(V_{\text{train}})_{ij}$ is 1 if and only if the i-th training document is classified into the j-th class, otherwise 0. A^+ is the pseudo-inversion matrix of A.

3.3 SSNMF

SSNMF[4] is one of the semi-supervised NMFs that H.Lee proposed. A convergence direction of SSNMF is controlled by adding constraint to the objective function in original NMF. The objective function in SSNMF is Eq.(10).

$$J_{ss} = ||X - UV^T||^2 + \lambda ||L * (Y - WV^T)||^2 \tag{10}$$

The matrix $Y \in \mathbb{R}^{k \times n}$ is the supervised matrix representing the correct cluster of document data. Each supervised element of Y is 1 if it is a correct cluster element, otherwise 0. Each unsupervised element of Y is unknown. The marix $W \in \mathbb{R}^{k \times k}$ is a basis matrix of the constraint item. $L \in \mathbb{R}^{k \times n}$ is a weight matrix to consider only supervised data, and λ is a weight for the constraints.

$$L_{ij} = \begin{cases} 0.001 & \text{if } \mathrm{Y_{ij}} = 1 \\ 1 & \text{if } \mathrm{Y_{ij}} = 0 \\ 0 & \text{if } \mathrm{Y_{ij}} \text{ is unknown.} \end{cases} \tag{11}$$

This constraint controls the convergence direction of the matrix V to the direction in which the product of W and V is closer to Y. The constraint has the effect of increasing the degree of relevance between similar clusters in the matrix V.

In addition, SSNMF has applied K-means to V after the final update. Clustering result of SSNMF is the result of K-means.

4 Classification by Matrix Decomposition

After viewpoint extraction, we use the viewpoint information for document classification. Document classification uses the matrix decomposition. Therefore, we apply the viewpoint matrix $U_m \in \mathbb{R}^{w \times k}$ to the document classification as the viewpoint information. Matrix U_m is composed of viewpoint vector calculated by the extraction methods of section 2. We propose three document classification methods using a matrix decomposition. The classification methods decompose the document matrix into the basis matrix and the reconstruction coefficient matrix. The reconstruction coefficient matrix means the degree of relevance between each document and each class. They classify each document into the class whose relevance is maximum.

4.1 Classification method 1 (CM-1)

In CM-1, we use Eq.(12) that solves Eq.(5) for the reconstruction coefficient matrix V. Both the matrices U and V are not given in general. In that case, we estimate the matrices U and V using NMF. However, we can use a decomposed matrix other than NMF, by using the matrix U_m. In fact, it substitutes the matrix U_m for the matrix U in the Eq.(5), and calculate the matrix V in Eq.(12) we solved the substituted Eq.(5).

$$V^T = U_m^+ X \tag{12}$$

4.2 Classification method 2 (CM-2)

Classification results of CM-1 highly depend on the value of the matrix U_m, which is created from supervised document data. Therefore, there is a problem that the optimum viewpoint information and the classification results are not calculated, if a singular data is contained in the supervised document data. So, we add the matrix U calculated by NMF to U_m in Eq.(12), in order to make the basis matrix better. We expect better document classification than CM-1. In Eq.(13), μ is weight.

$$V^T = (\mu U_m + U)^+ X \tag{13}$$

4.3 Classification method 3 (CM-3)

CM-3 uses NMF-I. In fact, we use the matrix U_m as the initial value of the basis matrix in NMF-I. CM-3 is able to calculate the more flexible the basis matrix than CM-1 and 2. Therefore, we expect better document classification than CM-1 and 2.

5 Experiment

We verify the effectiveness of the proposed methods by comparing the classification results of the original NMF, SSNMF, the proposed methods, bag-of-words with Multinomial

Logistic Regression (MLR) [8] and Naive Bayes (NB).

The proposed methods to be verified are 12 types, that are a combination of four extraction methods and three classification methods.

5.1 Dataset

In the experiment, we use single label documents and mixtures of single and multi-label documents. Each single label document has just one correct class label. Each multi-label document has two or more correct class labels. We use the single label documents provided on the site of CLUTO[1]. k1a, k1b and wap consisted of a variety of web pages in Yahoo!. re0 is derived from Reuters text collection. tr31 and tr41 are test documents of TREC. fbis is from the Foreign Broadcast Information Service data of TREC-5.

Table 1: Document data set of single label

Data	docs	terms	class	similarity
K1a	2340	21839	20	0.219
K1b	2340	21839	6	0.220
Re0	1504	2886	13	0.276
Wap	1560	6460	20	0.212
Tr31	927	10128	7	0.191
Tr41	878	7454	10	0.171
Fbis	2463	2000	17	0.252

We use the a mixture of documents provided on the Asahi newspaper[2] site. In this paper, we use the mixture documents from two classes in the experiment, in order to easily compare the difference in viewpoints. The mixture of documents consisted of single label documents of label-α, single label documents of label-β, and multi-label documents of label-$\alpha\beta$. In Tables 1 and 3, "similarity" means the similarity between clusters.

Table 2: Label configuration of mixture documents

Data(label-α , label-β)	label-α	label-β	multi label-$\alpha\beta$	total docs
ps(politics , sport)	40	40	19	99
se(sport , economy)	60	60	29	149
et(economy , technology)	200	200	100	500
it(incident , technology)	200	200	100	500

5.2 Setting of viewpoint and the correct label for multi-label

In the experiment, we decide that each multi-label document is classified into only one class to simplify the evaluation of the classification results. Class-α and classβ denote the sets of documents whose labels are label-α and label-β, respectively. Therefore, each multi-label document belongs to either class-α or class-β. Although it is desirable that we use some complex viewpoints, which have certain consistency like real viewpoints, it is difficult to set up such a viewpoint. To simplify the setting of the viewpoint, we introduce the lopsided

[1] http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download/

[2] http://asahi.com

Table 3: Data set of mixture documents

Data(viewpoint)	docs	terms	class	similarity
ps(politics)	99	3149	2	0.236
ps(sport)	99	3149	2	0.283
se(sport)	149	3317	2	0.249
se(economy)	149	3317	2	0.268
et(economy)	500	7048	2	0.513
et(technology)	500	7048	2	0.424
it(incident)	500	7374	2	0.477
it(technology)	500	7374	2	0.460

viewpoints of viewpoint-α and viewpoint-β which represent the viewpoints that classify all the multi-label documents into class-α and class-β, respectively. In other words, when we use viewpoint-α, the correct label of all the multi-label documents is label-α. When we use viewpoint-β, the correct label of all the multi-label documents is label-β.

In the single label documents, we used five documents for each class as supervised documents. In the mixed label documents, we used 10 documents of label-α and label-β each, either the set of label-α or the set of label-β containing five documents of multi-label as supervised documents. In Table 3, "(viewpoint)" represents the specified correct label of multi-label documents of label-$\alpha\beta$.

5.3 Evaluation

We used the Entropy, Purity, RandIndex, Precision and Recall[9] as a measure for accuracy of classification results. We used $\mathrm{H_m}$, the harmonic mean of these five scores as an overall measure. Because the smaller the Entropy value the better the clustering, we use $(1-\text{Entropy})$ instead of Entropy in calculating the harmonic mean. The N is the total number of documents.

$$\text{Entropy} = \sum_{i=1}^{k} \frac{|C_i|}{N} \times \left(-\sum_{h=1}^{k} P(A_h|C_i)\log P(A_h|C_i)\right) \tag{14}$$

$$\text{Purity} = \frac{1}{N}\sum_{i=1}^{k} \max_h |C_i \cap A_h| \tag{15}$$

In Eqs.(14) and (15), k is the number of clusters, C_i is the i-th resulting cluster, and A_h is the h-th correct cluster. The RandIndex is a measure of the similarity between the correct and resulting clusters.

$$\text{RandIndex} = \frac{TP+TN}{TP+FP+FN+TN} \tag{16}$$

TP is the number of document pairs which belong to the same correct cluster and also to the same resulting cluster. TN is the number of document pairs which belong to the different correct cluster and also to the different resulting cluster. FP is the number of document pairs which belong to the different correct cluster but belong to the same resulting cluster. FN is the number of document pairs which belong to the same correct cluster but belong to the different resulting cluster.

$$\text{Precision} = \frac{TP}{TP+FP}\ ,\quad \text{Recall} = \frac{TP}{TP+FN} \tag{17}$$

5.4 Result

In the experiment, we classified 20 times by changing the supervised sets and the initial value of NMF to random values and calculated the average. We set $\mu = 1$ for all experiments from preliminary experiments. NMF was updated 100 times in the experiment. We show the single label results in Fig.1 and the mixture documents results in Fig.2 that is the harmonic mean of the results in two viewpoint.

6 Discussion

6.1 Result of single label documents

We discuss the best combination of the classification methods and the extraction methods for the document set of single label. From Fig.1, a good classification method combined with all the extraction methods is CM-3. The combination of EM-2+CM-3 and EM4+CM-3 better than the other methods in the classification experiment of single label.

We compare EM-2+CM-3 to NMF, NMF-I, and SSNMF. EM-2+CM-3 shows 14.1% better average than NMF, 18.3% better average than MLR, 28.8% better average than NB, 2.40% better average than NMF-I, 19.5% better average than SSNMF in Hm.

EM-2+CM-3 shows the best classification performance in the single label document data used in the experiment.

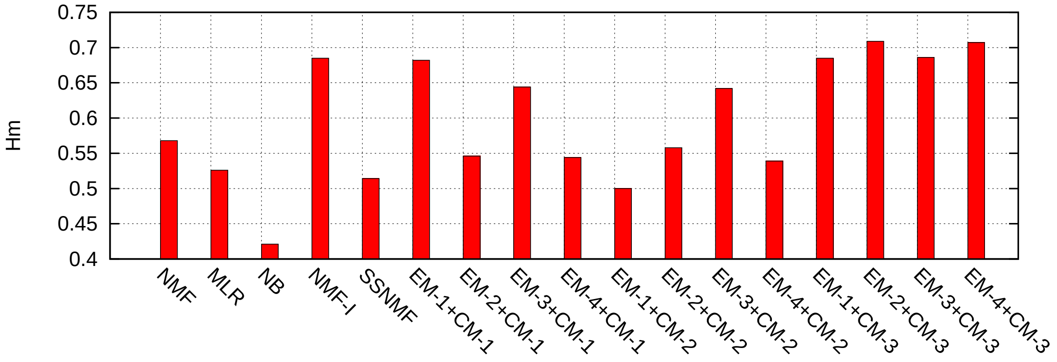

Figure 1: Single-label - Macro-average of seven data

6.2 Result of mixed label documents

We discuss the best combination of the extraction and classification methods for the document set of mixed label. From Fig.2, a good classification method combined with all the extraction methods is CM-3.

The combination of EM-4+CM-3 is best in the classification experiment of mixed label. In particular, EM-2+CM-3, EM-3+CM-3, and EM-4+CM-3 show better performance than the other methods. The result of data “et” is bad. It is probably because the similarity between the clusters of the data “et” is higher than that of the other data in the Table 3.

We compare the method EM-4+CM-3 to NMF, NMF-I and SSNMF. EM-4+CM-3 shows 34.4% better average than NMF, 15.3% better average than MLR, 38.8% better average than NB, 5.70% better average than NMF-I, 33.4% better average than SSNMF in Hm.

EM-4+CM-3 shows the best classification performance in the mixture documents data used in the experiment.

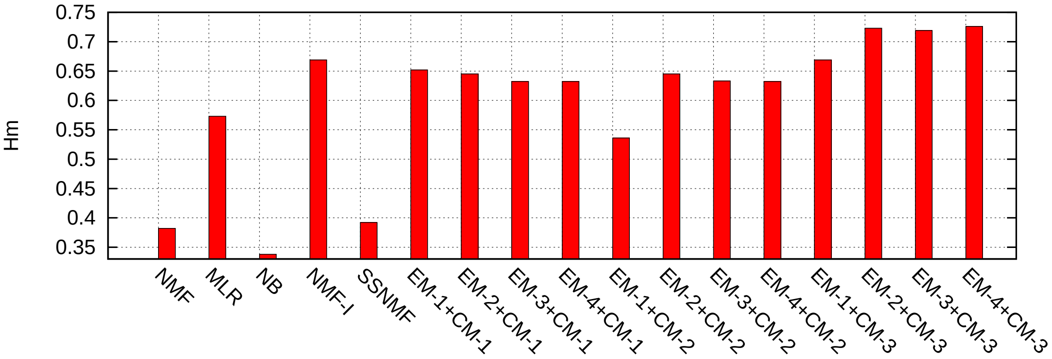

Figure 2: Mixed-label - Macro-average of four data and two viewpoint

6.3 Difference of Viewpoint

We discuss the effect of the difference of viewpoint. In particular, we focus on CM-3, because it shows the best classification performance for the mixture document. When we use viewpoint-α, we regard only label-α as the correct label of multi-label documents. However, the original multi-label documents also have label-β and even if we use viewpoint-α, the multi-label documents are easy to be classified into class-β. So, when the extraction methods can't extract the viewpoint well, classification performance will be low. From Fig.2, even if we use viewpoint-α or use viewpoint-β, classification performance is good. Therefore, the extraction methods would have been able to extract the viewpoint well.

Furthermore, we use dissimilarity of the viewpoint matrix U'_m for viewpoint-α and the viewpoint matrix U''_m for viewpoint-β, as an indicator to measure the effect of viewpoint. The viewpoint matrix is dependent on the viewpoint. Therefore, the matrix varies when we use a different viewpoint. In other words, when each extraction method is able to extract the viewpoint as expected, U_m will probably vary greatly. So, we calculated the dissimilarity between U'_m and U''_m. We compare the four extraction methods to examine which method extracts the viewpoint best. From Fig.2, C-2, C-4 show better classification performance than C-1, C-3. And from Table 4, it was found that the dissimilarity was high in EM-2, EM-4. That is, when the dissimilarity is high, the viewpoint matrix U_m affects the good classification performance. And EM-2, EM-4 use the ratio of each cluster. In other words, it was found that the influence of viewpoint is large using the ratio of each cluster.

Table 4: Dissimilarity of the viewpoint matrix

Data	EM-1	EM-2	EM-3	EM-4
ps	0.009	**1.733**	0.389	**1.744**
se	0.010	**1.781**	0.391	**1.764**
et	0.010	**1.742**	0.385	**1.719**
it	0.009	**1.735**	0.389	**1.732**

7 Conclusion

To classify documents according to a user's requirements, we have to extract the user's viewpoint information for classification. We proposed four methods of extracting viewpoints and three methods of classifying documents based on matrix decomposition. In both single label documents and mixed label documents, the classification result of EM-2+CM-3 and EM-4+CM-3 were best in the experiments. In classification methods, CM-3 using NMF with appropriate initial value was best. Since the number of supervised data is small, we thought that CM-3 that estimates the optimal value in iterative calculation became better results than CM-1 and CM-2 that directly calculate the value. EM-2 and EM-4 using the ratio between a class and all other classes increase weight of words that are emphasized only in each class. Since the result of EM-2 and EM-4 were good, in the classification according to the viewpoint, we thought that the classfication method that emphasizes the words representing the unique features of the each class is effective. Classifying documents into more than two classes and classifying documents using complex viewpoints remains problematic for the future.

References

[1] D.D.Lee, H.S.Seung, "Algorithms for Non-negative Matrix Factorization", NIPS , pp.556-562 , (2000).

[2] W.Xu, X.Liu, Y.Gong, "Document clustering based on non-negative matrix factorization", in Proc.ACM SIGIR Conf.Research and Development in Information Retrieval , Toronto, ON, Canada, (2003).

[3] K.Maruta, H.Nagai, T.Nakamura, "NMF with Supervised Constraints for Document Classification (in Japanese)", SIG-IFAT, IAS, pp14-21, (2013).

[4] H.Lee , J.Yoo , S.Choi "Semi-Supervised Nonnegative Matrix Factorization", IEEE SIGNAL PROCESSING LETTERS , Vol.17 No.1 , pp.4-7 , JANUARY (2010).

[5] H.Shinnou, M.Sasaki, "Ping-Pong Document Clustering by using NMF and Linkage Based Refinement" , IPS japan, NL Technical Reports, Vol.2007, No.47, pp.7-12, (2007).

[6] C.Ding,T.Li,W.Peng : "On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing", Computational Statistics and Data Analysis 52 , 3913 - 3927, (2008).

[7] S.Hotta, S.Miyahara, "An Initialization Method for Non-negative Matrix Factorization and Its Applications", IEICE Technical Reports, PRMU, Vol.102, No.652, pp.19-24, (2003).

[8] A.Agresti, "An Introduction to Categorical Data Analysis", John Wiley & Sons, (2006).

[9] C.D.Manning et al., "Introduction to Information Retrieval", trans.K.Iwano et al., Kyoritu Shuppan, (2012).