TOPIC EXTRACTION FROM DIVIDED DOCUMENT SETS

Takeru Yokoi
Tokyo Metropolitan College of Industrial Technology, Shinagawa, Tokyo, Japan

Hidekazu Yanagimoto
Department of Engineering, Osaka Prefecture University, Sakai, Osaka, Japan

Keywords: Topic extraction, Sparse non-negative matrix factorization, Clustering.

Abstract: We propose here a method to extract topics from a large document set with the topics included in its divisions and the combination of them. In order to extract topics, the Sparse Non-negative Matrix Factorization that imposes sparse constrain only to a basis matrix, which we call SNMF/L, is applied to document sets. It is useful to combine the topics from some small document sets since if the number of documents is large, the procedure of topic extraction with the SNMF/L from a large corpus takes a long time. In this paper, we have shortened the procedure time for the topic extraction from a large document set with the combining topics that are extracted from respective divided document set. In addition, an evaluation of our proposed method has been carried out with the corresponding topics between the combined topics and the topics from the large document set by the SNMF/L directly, and the procedure times of the SNMF/L.

1 INTRODUCTION

A huge amount of information is published through networks such as the Internet due to rapid development of information technology. The electrical document information seemes especially to be popular. However, it is difficult for many people to deal with those information since the existing information is too much. To address this problem, it is necessary to organize those document information. Reserchers focused on topics included in the documents as one of the methods to organize the document information effectively (Cselle07). Though various methods have been proposed to extract topics from a document set such as clustering, matrix analysis, and so on, we focused on the Non-negative Matrix Factorization (NMF) (Hoyer04). The NMF is a method to factorize a matrix whose all elements are positive values into two matrices whose elements are also positive values approximately. It is reported that the column vector of one matrix of two factorized matrices, which is called basis matrix of the NMF, represents a topic in a document set. In this paper, we adopt the Sparse Non-negative Matrix Factorization Light (SNMF/L) (Kim07) which is one of the modified versions of the NMF in order to extract topics. The SNMF/L is the method that imposes the sparseness constrain to the basis matrix. The sparseness constrain lets the characteristic of a topic be more comprehensive.

When applying the SNMF/L to a document set for the topic extraction, a document is represented as a column vector, which is called a document vector, by the vector space model (Salton83) and a document set is represented by a term-document matrix. The document vector's dimension becomes larger as the number of document increases since the dimension depends on the number of index words in a document set. The index word is the characteristic or significant word for description of the document. The total number of index words becomes larger as the number of documents increase more. As a result, some problems such as the memory space and processing time raise. The problem on memory space is that the term-document matrix size becomes larger, memory is often insufficient for the various procedures for the matrix. Next, the problem on processing time is that the procedure time becomes long if the matrix size becomes large. In order to address these problems, we propose the method of topic extraction that divides a large document set into sub-document sets and combined the topics obtained each sub-document set.

In the following sections, we have presented

Yokoi T. and Yanagimoto H.
TOPIC EXTRACTION FROM DIVIDED DOCUMENT SETS.
DOI: 10.5220/0001822106540659
In *Proceedings of the Fifth International Conference on Web Information Systems and Technologies (WEBIST 2009)*, page
ISBN: 978-989-8111-81-4

overview of related works and an explanation of the method we have proposed. In sections 4 and 5, we have detailed our experimental procedures using the news articles and discussed our results. Lastly we present our conclusions and future work.

2 RELATED WORK

Traditionally, various methods have been applied to extract topics from a document set. In this paper, we especially focus on the methods based on the vector space model. The vector space model (Salton83) represents a document as a column vector whose element consists of the weight of an index word. The Euclidian distance, the cosine, and so on are used for the similarity. The one of the popular methods to extract topics from a document set is clustering (Yang99). After clustering documents, the centroid of each class is regarded as a topic.

Recently, the method such as factorization of a term-document matrix is focused for the topic extraction. At first, it was reported that the Independent Component Analysis (ICA) (Hyvarinen00) was applied for a term-document matrix so that its independent components represent topics by T.Kolenda (Kolenda00). E. Bingham extracted the topics from dynamical textual data such as chat lines with the ICA (Bingham03). In addition, we confirmed that it was possible for ICA to extract the topics from documents and proposed application of the ICA to the information filtering(Yokoi08). The independent component is possible to have negative elements so that it is difficult to comprehend the weight as a term weight directly.

The NMF(Hoyer04) has been applied to textual data and the column vectors of the basis matrix were reported to represent the topics in a document set. The basis matrix denotes one of the factorized matrices by the NMF. The NMF factorizes the non-negative matrix into two of non-negative matrices so that the element of the column vector in the basis matrix directly corresponds to a term weight. Xu. et al. proposed the bases are used for text clustering as one of the NMF applications for textual data (Xu03). In addition, the modified methods of the NMF have recently been paid attention (Berry07). We especially focus the NMF imposing the sparseness to one of the factorized matrices in those methods. Moreover, the conventional reports on the application of the NMF to documents targeted a statistic one document set. However, the size of document set becomes so large that it is difficult for the NMF to apply to it.

Our proposed method sequentially combines the topics based on the conventional reports that the NMF can extract topics from a document set.

3 TOPIC COMBINATION

In this section, a document vector, the SNMF/L for documents, and combination of topics are explained.

3.1 Document Vector

A document is represented with a vector with the vector space model(Salton83) and it is called a document vector. A document vector is a column vector of which the elements are the weights of the words in a document set. The ith document vector $\mathbf{d}_i$ is defined as:

$$\mathbf{d}_i = [w_{i1} \quad w_{i2} \quad \cdots \quad w_{iV}]^T \tag{1}$$

where w_{ij} signifies the weight for the jth word in the ith document, V signifies the number of words and $[\cdot]^T$ signifies the transposition. In this paper, w_{ij} is established by the tf-idf method and calculated as:

$$w_{ij} = tf_{ij} \log\left(\frac{n}{df_j}\right) \tag{2}$$

where tf_{ij} denotes the frequency of the jth word in the ith document, df_j denotes the number of documents including the jth word and n denotes the number of documents. The tf-idf method regards the words that appear frequently in a few documents as the characteristic features of the document. In addition, the n document vectors are denoted as $\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_n$ and the term-document matrix D is defined as follows:

$$D = [\mathbf{d}_1 \quad \mathbf{d}_2 \quad \cdots \quad \mathbf{d}_n]. \tag{3}$$

3.2 SNMF/L for Documents

The SNMF/L is one of the sparse NMF algorithms that can control the degree of sparseness in the basis matrix. The NMF approximately factorizes a matrix of which all the components have non-negative values into two matrices with components having non-negative values. When the NMF is applied to a document set, it has been reported that that bases represent the topics included in the document set. By using the SNMF/L in our proposal, the keywords of the topics are highlighted since only some words of each basis have weighted.

The NMF approximately factorizes a matrix into two matrices such as:

$$D = WH \tag{4}$$

where W is a $V \times r$ matrix containing the basis vectors $\mathbf{w}_j$ as its columns and H is a $r \times n$ matrix containing the coefficient vectors $\mathbf{h}_j$ as its rows. r is arbitrary determined as satisfying the following:

$$(n+V) \cdot r < n \cdot V. \tag{5}$$

In addition, the equation numbered 4 is also described as:

$$\mathbf{d}_j \approx W\mathbf{h}_j. \tag{6}$$

This means $\mathbf{d}_j$ is the linear combination of W weighted by the elements of $\mathbf{h}_j$.

Given a term-document matrix D, the optimal factors W and H are defined as the Frobenius norm between D and WH is minimized. The optimization problem is denoted as:

$$\min_{W,H} f(W,H) = \|D - WH\|_F^2, \ s.t. \ W,H > 0 \tag{7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $W,H > 0$ means that all elements of W and H are non-negative. In order to minimize $f(W,H)$, the following updates are iterated until $f(W,H)$ converges:

$$\bar{H}_{ij} = H_{ij} \frac{(W^T V)_{ij}}{(W^T WH)_{ij}} \tag{8}$$

$$\bar{W}_{ij} = W_{ij} \frac{(VH^T)_{ij}}{(WHH^T)_{ij}} \tag{9}$$

where X_{ij} denotes the ij element of matrix X, and $\bar{H}$ and $\bar{W}$ denote updated factors, respectively.

In order to impose sparseness constraints on the basis matrix W, SNMF/L modifies the optimization function in the equation numbered 7 as following:

$$\min_{W,H} f(W,H) = \|D - WH\|_F^2 + \alpha \sum_{i=1}^{V} \|W(i,:)\|_1^2 \quad s.t. \ W,H > 0 \tag{10}$$

where $W(i,:)$ denotes the i-th row vector of W, and the parameter α is real non-negative value to control sparseness of W. The SNMF/L algorithm initializes a non-negative matrix W at first. Then, it iterates the following the Alternating Non-negativity Constrained Least Squares (ANLS) (Kim07) until convergence:

$$\min_{H} \|WH - D\|_F^2, \ s.t. \ H > 0 \tag{11}$$

$$\min_{W} \left\| \begin{pmatrix} H^T \\ \sqrt{\alpha}\mathbf{e}_{1\times r} \end{pmatrix} W^T - \begin{pmatrix} A^T \\ \mathbf{0}_{1\times V} \end{pmatrix} \right\|_F^2, s.t. \ W > 0$$

where $\mathbf{e}_{1\times r} \in R^{1\times r}$ is a row vector whose elements are all ones and $\mathbf{0}_{1\times V} \in R^{1\times V}$ is a zero vector whose elements are all zeros.

In this paper, we focus on the sparse row bases of W which represent the topics included in a document set. In addition, topic $\mathbf{t}$ obtained from a document set D_k is represented as follow vector:

$$\mathbf{t} = [t_1 \quad t_2 \quad \cdots \quad t_{V_k}]^T \tag{12}$$

where t_i is the weight of a word, and V_k denotes the number of index words included in the document set D_k.

3.3 Topic Combination

In order to extract topics from a large document set, we combine the topics obtained from some smaller sub-document sets of the large one. The combination is performed by bottom up hierarchical clustering such as dendrogram (Tou74). In this paper, we explain how to combine the topics obtained from two document sets, D_k and D_l.

At first, in order to perform the clustering of the topics obtained from different document sets, it is necessary to resolve a difference in the dimension of the topic vectors obtained from D_k and D_l. In order to resolve the difference, the extension of the index words for each document set is performed. The fixed number of index words V' is defined as:

$$V' = |V_k \cup V_l| \tag{13}$$

where $|\cdot|$ denotes the density of a set. Hence, the topic vector obtained from the document set D_k denoted by the equation numbered 12 is modified as follow:

$$\mathbf{t} = [t_1 \quad t_2 \quad \cdots \quad t_{V_k} \quad t_{V_{k+1}} \quad \cdots \quad t_{V'}]^T \tag{14}$$

where t_{V_k+1} to $t_{V'}$ are set to zeros. After extension of the index words, each topic vector is normalized and similar topics are combined. The similarity $s(\mathbf{t}_p, \mathbf{t}_q)$ between topics $\mathbf{t}_p$ and $\mathbf{t}_q$ is defined by Euclidian distance:

$$s(\mathbf{t}_p, \mathbf{t}_q) = \sqrt{\sum_{i=1}^{V'} (t_{pi} - t_{qi})^2}. \tag{15}$$

The topic $\mathbf{t}_p$ and $\mathbf{t}_q$ are put together into one topic if the Euclidian distance between them is nearer than the threshold. The novel topic vector constructed by combination of two topics $\mathbf{t}_p$ and $\mathbf{t}_q$ is defined as the median point of those topics. The novel topic vector $\mathbf{t}'$ is as follow:

$$\mathbf{t}' = \frac{\mathbf{t}_p + \mathbf{t}_q}{2}. \tag{16}$$

4 EXPERIMENTS AND RESULTS

In this section, we explain an evaluation experiment to confirm the effectiveness of the proposed method.

4.1 Experimental Environment and Procedures

In this paper, news articles of the day from 2006/11/13 to 2006/11/19 in "asahi.com" were used for an experimental data. The detail of the respective data is presented in Table 1.

The procedure of experiment for our proposed method is as follows:

1. A term-document matrix was constructed for each document set. The column vector of the matrix was the document vector defined by the equation numbered 1. In this experiment, nouns in the document set were used as the index words of a document vector. In addition, those nouns were obtained by morphological analysis using MECAB.
2. The SNMF/L was applied to each document set, i.e. the set of news articles for the respective day presented in Table 1. The SNMF/L procedure was continued until iteration times reached the max iteration one. We set the parameter of α to 0.7 in the equation numbered 12, and the max iteration times to $20,000$. In addition, the numbers of topics that we extracted from each document set are also shown in Table 1.
3. The clustering was performed for the topics obtained from two document sets. In this paper, we set the parameter of the distance in order to combine the topics to 0.8.
4. Those topics were evaluated.

Table 1: The details of experimental data. "Data No." denotes the day of articles, "# of Article" denotes the number of articles included in each set of the articles, and "# of Word" denotes the number of index words included in each document set. The instance of xx in Data No. denotes "2006/11/xx". In addition, "# of Topic" denotes the number of extracted topics from each document set.

Data No.	# of Article	# of Word	# of Topic
13	67	2,488	20
14	95	2,988	30
15	87	2,621	30
16	88	2,672	30
17	101	3,206	30
18	84	2,844	30
19	52	2,186	20

As a comparable method, we applied SNMF/L to all document sets combining two document sets presented in Table 1, and compared the obtained topics. This comparable method corresponded with the conventional application of the NMF to a document set, i.e. applied to a statistic and one document set. According to this experiment, we discuss the difference between our proposal and the conventional one. Table 2 presents the details of combined document sets. In addition, the parameter of α in the equation numbered 12 in the comparable experiment was established to 0.7, that was the same value to the experiment of our proposed method. In this paper, we regarded the topic extracted from combined data as an original topic. We evaluate how many topics extracted by our proposed method covered with the original topics. After selecting the most similar topic from the topics that our proposed method extracts in the perspective of the cosine similarity, we manually judged the correspondence between those two topics.

Table 2: The details of combined document sets. "Comb No." denotes the numbers of combined document sets, "# of Article" denotes the number of articles in each combined document set, "# of Word" denotes the number of index words of the combined document set, that is calculated by the equation numbered 13, and "# of Topic" denotes the number of topics that we extracted in the comparable method. In addition, the instance, xx-yy, in "Comb No." denotes the data combining the document sets numbered xx and yy.

Comb No.	# of Articles	# of Words	# of Topics
13-14	162	4,235	50
13-15	154	3,933	50
13-16	155	4,013	50
13-17	168	4,482	50
13-18	151	4,195	50
13-19	119	3,741	40
14-15	182	4,236	60
14-16	183	4,323	60
14-17	196	4,709	60
14-18	179	4,500	60
14-19	147	4,107	50
15-16	175	3,951	60
15-17	188	4,423	60
15-18	171	4,191	60
15-19	139	3,798	50
16-17	189	4,469	60
16-18	172	4,230	60
16-19	140	3,835	50
17-18	185	4,596	60
17-19	153	4,252	50
18-19	136	3,839	50

Besides, for each experiment, we evaluated the procedure time to extract topics by SNMF/L. These experimental environments such as the machine specification, the operation system and used software are presented in Table 3.

Table 3: The specification and software used in this experiment.

CPU	Intel Core2 CPU 2.66GHz
Memory	4GB
OS	Windows Vista SP1
Software	Matlab 6.1 (SNMF/L), Java (Clustering)

4.2 Experimental Results

In this section, the results of the experiments are presented. Table 4 presents the procedure time of SNMF/L for the document set of each day.

Table 4: The procedure times and the errors of SNMF/L for the document set of each day. "Time" denotes the procedure time. In addition, [m] denotes a minutes.

Data No.	Time [m]
13	22
14	56
15	52
16	52
17	70
18	53
19	18

Table 5 presents the percentage of the number of corresponding topics between the original and our proposal, the procedure time for the combined document sets.

Table 5: The numbers of correspondence topics and the procedure times for combined document sets. "% of Acc." denotes the percentage of the number of the corresponding topics.

Comb No.	% of Acc. [%]	Time [m]
13-14	44	233
13-15	52	196
13-16	54	189
13-17	52	201
13-18	54	170
13-19	53	93
14-15	60	230
14-16	63	246
14-17	58	280
14-18	60	323
14-19	70	163
15-16	65	284
15-17	67	316
15-18	60	371
15-19	70	159
16-17	65	350
16-18	60	317
16-19	72	172
17-18	58	337
17-19	74	191
18-19	72	149

5 DISCUSSION

At first, we focused on the percentage of the number of the corresponding topics in Table 5. The average of the percentage is 61%. While the accurracy is especially low around 50% with the combined document set including the document set numbered 13, it is high around 70% with the combined document set including the one numbered 19. The one of the reasons is why the numbered of topics that truly exist in a document set was not miss matched. For example, focusing on the document set numbered 13-14, that results the lowest corresponding percentage, some topics on "Matsuzaka" were obtained. Such topics should be put together into a few topics or one topic. In addition, with SNMF/L, we can avoid it even though we performed clustering with Euclidian distance. Only few topics were integrated.

Next, we discuss the procedure time for the SNMF/L. The difference is remarkable between the SNMF/L for the single document set and combined one. The procedure times of the combined document sets are about 10 times longer than that of the single ones. The cause of the differences is certainly due to the size of the term-document matrix, however, mainly due to the rank of basis matrix. Focusing on the differences of the result for document set combining the document sets numbered 13 and 19, and the others, there is the remarkable difference of the procedure time despite an equal number of index words. In addition, when applying the SNMF/L to the document set combining three document sets, the SNMF/L process has not finished for two days in our experimental environment. If the size of a document set becomes larger, it has to extend the number of topics to extract. Therefore, our proposed method, i.e. applying SNMF/L to sub-document sets respectively, can contribute to shorten the procedure time.

Finally, we remark on the divisions of a document set. In our experiment, we have treated a document set as the one that collects up the articles by their date. Our goal is, in fact, the division should be performed by any criteria, but this experiment was useful for the topic extraction, since the set of news articles divided by the date has included various topics.

6 CONCLUSIONS

In this paper, we have proposed the combination of the topics extracted from sub-document sets and discussed the difference between the combined document set and the sub-document sets with the topic extraction using SNMF/L. As a result, the 60% over of topics are same ones between the two methods. In addition, our proposed method has advantage in the view point of the procedure time.

We will have to discuss how many topics should be extracted and the distance function used in combining the topics, that we used Euclidian distance in this paper, in the future work.

ACKNOWLEDGEMENTS

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in Aid for Young Scientist (Start Up), 20860085, 2008.

REFERENCES

A. Hyvarinen, E. O. (2000). Independent component analysis: A tutorial. *Neural Network*, 13:411–430.

E.Bingham, A.Kaban, M. (2003). Topic identification in dynamical text by complexity pursuit. *Neural Processing Letters*, 17(1):69–83.

G. Cselle, K. Albrecht, R. Wattenhofer (2007). Buzztrack: Topic detection and tracking in email. In *IUI2007*.

G.Salton, M.J.McGill (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company.

H. Kim, H. Park (2007). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23:1495–1502.

M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, R. J. Plemmons (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173.

P.O.Hoyer (2004). Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469.

T. Yokoi, H. Yanagimoto, S. Omatu (2008). Improvement of information filtering by independent components selection. volume 163, pages 49–56. Wiley.

T. Kolenda, L. K. Hansen (2000). Independent components in text. In *Advances in Independent Component Analysis*. Springer-Verlag.

Tou J. T.,Gonzalez R. C. (1974). *Pattern Recognition Principles*. Addison-Wesley, Reading.

Xu. W., Liu. X., Gong. Y. (2003). Document clustering based on non-negative matrix factorization.

Y. Yang, J. Carbonell, R. Brown, T. Pierce, B. T. Archibald, X. Liu (1999). Learning approaches for detecting and tracking news events. *IEEE Inteligent Systems*, 14(4):32–43.