World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:3, No:2, 2009

Information Filtering using Index Word Selection based on the Topics

Takeru YOKOI, Hidekazu YANAGIMOTO and Sigeru OMATU

Abstract— We have proposed an information filtering system using index word selection from a document set based on the topics included in a set of documents. This method narrows down the particularly characteristic words in a document set and the topics are obtained by Sparse Non-negative Matrix Factorization. In information filtering, a document is often represented with the vector in which the elements correspond to the weight of the index words, and the dimension of the vector becomes larger as the number of documents is increased. Therefore, it is possible that useless words as index words for the information filtering are included. In order to address the problem, the dimension needs to be reduced. Our proposal reduces the dimension by selecting index words based on the topics included in a document set. We have applied the Sparse Non-negative Matrix Factorization to the document set to obtain these topics. The filtering is carried out based on a centroid of the learning document set. The centroid is regarded as the user's interest. In addition, the centroid is represented with a document vector whose elements consist of the weight of the selected index words. Using the English test collection MEDLINE, thus, we confirm the effectiveness of our proposal. Hence, our proposed selection can confirm the improvement of the recommendation accuracy from the other previous methods when selecting the appropriate number of index words. In addition, we discussed the selected index words by our proposal and we found our proposal was able to select the index words covered some minor topics included in the document set.

Keywords— Information Filtering, Sparse NMF, Index word Selection, User Profile, Chi-squared Measure

I. Introduction

Much information is published through networks such as the Internet due to the rapid development of information technology. People are able to read any number of documents easily and various information retrieval systems have been developed. Almost all of these systems search information based on the queries entered by users. However, it has recently become difficult to select appropriate queries or query combinations. As a result, there is a flood of information in which the most necessary information is buried in other information. To address this problem, a filtering system focusing on the user's interest is proposed.

In dealing with a document in information filtering, we often use a document vector [1]. Each element of such a vector corresponds to the weight of a word in a corpus. A word that indicates the features of a document is called an index word. In order to obtain the index words, it is necessary to divide a document into words and remove stop words such as articles, conjunctions, etc., which themselves do not indicate the features of the document. If we try to construct the document vectors from many documents using these index words, the number of remaining words will still be so large that the dimension of the document vector also becomes very large. It would, thus, be more efficient to select index words which would reduce the dimension of the document vector, especially since unnecessary words for identification of the documents are included even when the stop words are removed. These are considered noise in information filtering and index word selection would remove such noise. In addition, the accuracy of information filtering is higher using selected index words than the all words including unnecessary words.

In this paper, we propose a method to select the index words focusing on the topics included in a set of documents for the construction of a document vector from a document set which can be applied to the information filtering system based on the user's interest. Moreover, we verify the effectiveness of our index word selection for improvement of the information filtering accuracy. This index word selection uses the Non-negative Matrix Factorization with Sparseness Constraints (NMFSC) [2] and Chi-square value method. The NMFSC adds a sparseness constraint to the Non-negative Matrix Factorization (NMF) [3] so that it makes the characteristics of the basis and coefficient of NMF more comprehensive than with the NMF itself. Our proposal selects some words from each basis of the NMFSC, which are referred to as keywords, and the frequently co-occurring words with the keywords as index words. Keywords are significant in the representation of the documents and the frequently co-occurring words also have significant features in the representation of the topics. In addition, when the NMF is applied to a document set, it has been reported that the topics included in the document set can be obtained [4] [5].

In the following sections, we have presented an 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 test collection "MEDLINE" [6] and discussed our results. Lastly we present our conclusions

Takeru YOKOI is with the Tokyo Metropolitan College of Industrial Technology, Shinagawa, Tokyo, Japan (corresponding author to provide phone: +81-3-3471-6331; fax: +81-3-3471-6338; e-mail: takeru@s.metri-cit.ac.jp).

Hidekazu YANAGIMOTO and Sigeru OMATU are with the Department of Engineering, Osaka Prefecture University, Sakai, Osaka, Japan (e-mail: {hidekazu, omatu}@cs.osakafu-u.ac.jp).

and future work.

II. Related works

Traditionally, the focus has been on changing the space constructed by words into one constructed by latent semantics to reduce the dimension of a document vector. The Latent Semantic Analysis (LSA) [7] is a popular method to analyze and extract the latent semantics of the documents and reduce the dimension focusing on the variance of the words' weight. The Independent Component Analysis (ICA) has also been reported to enable the retrieval of the latent semantics of the documents by evaluating the independence of a basis which describes the features of a document set [8]. We have reported on a method which reduces the dimension by projecting the space into the one structured by independent components [9].

It has also been reported that the NMF can extract the topics included in a document set [4]. Xu. W. et al. proposed the document clustering using the topics obtained by the NMF [10]. It is the one of the application examples that the NMF is applied for documents. M.W. Berry et al. introduced some NMF's applications and they applied the NMF for text mining to extract topics as one of the NMF's applications [11]. Tsuge et al. have proposed a method for dimension reduction by projecting the space into the one structured from the bases obtained by the NMF [5]. Above researches reported that the bases of the NMF represented the topics included in the document set if the NMF is applied to a document set.

Concerning with the NMF, other variations of the NMF were also studied. Attention was focused on the sparseness of the variations for the elements of the bases and coefficients. P. O. Hoyer proposed Non-negative Sparse Coding (NNSC) [12]. This method adds a small reconstruction error with a sparseness criterion to the objective function defined as the Euclidian least-square function, realizing the addition of the properties of sparseness for the bases and coefficients. Liu et al. proposed Sparse Nonnegative Matrix Factorization (SNMF) [13]. This method used a divergence term instead of the Euclidian objective function used in NNSC. Moreover, the NMFSC, which is used in this paper, is a more recent work related to the addition of sparseness constraints to the conventional NMF.

Various methods to extract the most significant words from a document or document set have been proposed. Here, we describe two of the works most closely related to ours. First, Matsuo and et al. [14] have used co-occurrence to extract important words from a document. In this research, the bias of the probabilistic distributions between the co-occurrence and the appearance of the most frequent words in the document were measured. Such frequent words are referred to as keywords. They have evaluated the bias by chi-squared measure and selected the most important words. However, if we extract the keywords from a document set only by term frequency, some words included in a few documents which identify minor topics in a corpus tend to be ignored. Since our goal is to construct document vectors of all the documents in a corpus, words included in the minor topics cannot be ignored.

Next, Osawa and et al. [15] have proposed a method to select the most important words from WWW using the Key Graph based on the most frequent words. They have evaluated the significance of the words by their co-occurrence using a graph structure between the most frequent and other words.

Our proposal selects some words which represent the features of the topics. The words have maximum weight within each basis of the NMFSC and are referred to as the keywords instead of the most frequent words used in a previous report [12]. In addition, we selected other words related to the topics. The selection is performed according to the chi-squared measure between each word and the key words. We finally regarded the keywords and words selected by chi-squared measure as the index words.

III. Topic-based index words selection

In this section, a document vector, a user profile, the NMFSC for the documents and selection of the index words using the chi-squared measure are explained.

A. Document vector and User profile

A document vector is a column vector of which the elements are the weights of the words in a corpus. The *i*th document vector $\mathbf{d}_i$ denotes

$$\mathbf{d}_i = [\omega_{i1} \quad \omega_{i2} \quad \cdots \quad \omega_{iV}]^{\mathrm{T}} \cdots (1)$$

where ω_{ij} signifies the weight for the *j*th word in the *i*th document, V signifies the number of words and $[\cdot]^{\mathrm{T}}$ signifies transposition. In this paper, ω_{ij} is determined by the tf-idf method and calculated as

$$\omega_{ij} = tf_{ij} \log\left(\frac{N}{df_j}\right) \cdots (2)$$

where tf_{ij} denotes the frequency of the *j*th word in the *i*th document, df_j denotes the number of documents including the *j*th word and N denotes the number of documents. The tf-idf method regards the words which appear frequently in a few documents as the characteristic features of the documents. 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:

$$\mathrm{D} = [\mathbf{d}_1 \quad \mathbf{d}_2 \quad \cdots \quad \mathbf{d}_N] \cdots (3)$$

A user profile describes a user's interest. We define $\mathbf{u}$ as a column vector of which the element consists of the weight of word, as is expressed in Eq. (2).

$$\mathbf{u} = [u_1 \quad u_2 \quad \cdots \quad u_V]^{\mathrm{T}} \cdots (4)$$

If the *i*th word is included in the interesting documents, the value of u_i takes high value. On the other hand, if the *i*th word exists in the uninteresting ones, the value of u_i dose low value. In addition, the user profile is constructed using a centroid of the documents as follows:

$$\mathbf{u} = \alpha \sum_{\mathbf{d}_k \in \mathrm{D_I}} \mathbf{d}_k - \beta \sum_{\mathbf{d}_l \in \mathrm{D_U}} \mathbf{d}_l \cdots (5)$$

where α and β are coefficients for each document, D_I and D_U denotes the document set including the interesting documents and uninteresting ones respectively. This formula is referred to the Rocchio's formula [16].

B. NMFSC for the documents and keyword extraction

The NMFSC adds the sparseness constraint for the bases and coefficients to the NMF. The NMF approximately factorizes a matrix of which all the elements have non-negative values into two matrices with elements having non-negative values. If the NMF is applied to a document set, it has been reported that the bases represent the topics included in the document set [4]. By using the NMFSC and not the NMF in our proposal, the keywords of the topics are considered to be highlighted since only some words of each basis have weight.

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

$$\mathrm{D} \approx \mathrm{WH} \cdots (6)$$

where W is an $V \times r$ matrix containing the basis vectors $\mathbf{w}_k$ as its columns and H is an $r \times N$ matrix containing the coefficient vectors $\mathbf{h}_i$ as its rows. r is determined as satisfying the following:

$$(N + V) \cdot r < N \cdot V \cdots (7).$$

In addition, equation (6) is also denoted as:

$$\mathbf{d}_k \approx \mathrm{W}\mathbf{h}_k \cdots (8).$$

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

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

$$\mathrm{F} = \|\mathrm{D} - \mathrm{WH}\|_{\mathrm{F}}^2 \cdots (9)$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. In order to minimize F, the following updates are iterated until F converges:

$$\overline{\mathrm{H}}_{\mathrm{ij}} = \mathrm{H}_{\mathrm{ij}} \frac{(\mathrm{W}^{\mathrm{T}}\mathrm{V})_{\mathrm{ij}}}{(\mathrm{W}^{\mathrm{T}}\mathrm{WH})_{\mathrm{ij}}} \cdots (10)$$

$$\overline{\mathrm{W}}_{\mathrm{ij}} = \mathrm{W}_{\mathrm{ij}} \frac{(\mathrm{VH}^{\mathrm{T}})_{\mathrm{ij}}}{(\mathrm{WHH}^{\mathrm{T}})_{\mathrm{ij}}} \cdots (11)$$

where $\overline{\mathrm{H}}$ and $\overline{\mathrm{W}}$ denote updated factors, and X_{ij} denotes the ij element of matrix X. The NMFSC adds the sparse constraint to the NMF. The sparseness of each basis can be evaluated by

$$\mathrm{sparseness}(\mathbf{w}_{\mathrm{i}}) = \frac{\sqrt{V} - \left(\sum_{\mathrm{j}} |w_{kj}| / \sqrt{\sum_{\mathrm{j}} w_{kj}^2}\right)}{\sqrt{V} - 1} \cdots (12).$$

This function evaluates to unity if and only if $\mathbf{w}_k$ contains a single non-zero element, and takes a value of zero if and only if all the elements are equal. The sparseness measure of $\mathbf{h}_k$ can be also defined. However, since our proposal focuses on the bases, we applied the sparseness constraint only to the bases. The NMFSC devises the update equation to fill the sparseness using a projection operator which enforces sparseness.

Here, we represent the algorithm of the NMFSC, as follows:

1. Initialize W and H to random positive matrices.
2. Project each column of W to be non-negative with an unchanged L2 norm and L1 norm set to achieve the desired sparseness.
3. Iterate following steps until equation (9) converges.
   i. Set $\overline{\mathrm{W}} := \mathrm{W} - \mu_{\mathrm{W}}(\mathrm{WH} - \mathrm{V})\mathrm{H}^{\mathrm{T}}$, where μ_{W} is a small positive constant.
   ii. Project each column of $\overline{\mathrm{W}}$ to be non-negative with an unchanged L2 norm and L1 norm set to achieve the desired sparseness. L1 norm mentioned above is determined by substituting the L2 norm of each column of $\overline{\mathrm{W}}$ for Eq. (12).
   iii. $\overline{\mathrm{H}}_{\mathrm{ij}}$ is updated by Eq. (10).

Next, the projection operator which enforces sparseness by setting the L1 norm is defined as follows. Here, for the given vector $\mathbf{x}$, the closest non-negative vector $\mathbf{s}$ can be determined with a given L1 norm and L2 norm.

1. Set $\mathrm{s}_{\mathrm{i}} := \mathrm{x}_{\mathrm{i}} + \frac{(\mathrm{L1} - \sum_{\mathrm{i}} \mathrm{x}_{\mathrm{i}})}{\dim(\mathbf{x})}, \forall \mathrm{i}$, where $\dim(\mathbf{x})$ denotes the number of dimension of $\mathbf{x}$.
2. Set $\mathrm{Z} := \{\emptyset\}$
3. Iterate the following steps:
   i. Set $\mathrm{m}_{\mathrm{i}} := \begin{cases} 0 & if\ i \in Z \\ \frac{\mathrm{L1}}{(\dim(\mathbf{x}) - \mathrm{size(Z)})} & if\ i \notin \mathrm{Z} \end{cases}$, where size(Z) denotes the element count of Z.
   ii. Set $\mathbf{s} := \mathbf{m} + \alpha(\mathbf{s} - \mathbf{m})$, where $\alpha \geq 0$ is selected such that the resulting $\mathbf{s}$ satisfies the L2 norm constraint.
   iii. If all elements of $\mathbf{s}$ are non-negative, return $\mathbf{s}$, end
   iv. Set $\mathrm{Z} := \mathrm{Z} \cup \{\mathrm{i}; \mathrm{s}_{\mathrm{i}} < 0\}$
   v. Set $\mathrm{s}_{\mathrm{i}} := 0, \forall \mathrm{i} \in \mathrm{Z}$
   vi. Calculate $\mathrm{c} := (\sum_{\mathrm{i}} \mathrm{s}_{\mathrm{i}} - \mathrm{L1}) / \big(\dim(\mathbf{x}) - \mathrm{size(Z)}\big)$
   vii. Set $\mathrm{s}_{\mathrm{i}} := \mathrm{s}_{\mathrm{i}} - \mathrm{c}, \forall \mathrm{i} \notin \mathrm{Z}$
   viii. Go to i

Following the above steps, we obtained the sparse bases which represent the topics included in the document set. In addition, our proposal picks up some words which have maximum weight from each basis as the keywords of the respective topic. These keywords are part of the reconstructed index words. Here, each keyword is denoted by *g* and a set of those keywords by G.

C. Selection of the index words as related to the topics

In order to select index words related to the topics based on the keywords, we used the chi-squared measure. In this work, the chi-squared measure evaluates the distribution bias between the co-occurrence of the keywords and each word as well as the appearance of the keywords. Here, *t* denotes the words other than the keywords in a corpus. If there is a large bias between the co-occurrence probability of the word *t* and the keywords G and the appearance probability of G, the chi-squared measure is deemed high. If the word *t* is used generally throughout the corpus, i.e., if the word *t* occurs with all the keywords evenly across the text, the co-occurrence probability distribution is not biased. Therefore, the chi-squared measure of the word *t* is low. Since important words are considered to occur with some

specified keywords, we can judge whether the word t is important or not by evaluating the chi-squared measure. In other words, if we select a word whose chi-squared measure is high, the word is considered to be closely concerned with the keywords and expresses a feature of the topic.

The expected probability p_g denotes an unconditional probability of a keyword g in the set of keywords G. Here, n_t denotes the frequency of the co-occurrence of the word t and the set of the key words G. The frequency of the co-occurrence of the word t and the keyword g ($g \in$ G) is denoted as $freq(t, g)$. In this paper, we have defined the co-occurrence as being when each key word g and the word t are included in the same document. In addition, since the corpus is constructed by documents, we have defined p_g as follows:

$$p_g = \frac{\text{the number of documents including } g}{\text{total number of documents including G}} \cdots (13).$$

Thus, $\chi^2(t)$, which is the chi-squared measure for a word t, is defined as:

$$\chi^2(t) = \sum_{g \in \mathrm{G}} \frac{\left(freq(t,g) - n_t p_g\right)^2}{n_t p_g} \cdots (14).$$

After calculating the $\chi^2(t)$ for each word t, we selected the words whose chi-squared measure are added to the keywords G as the reconstructed index words. This method expects to cover the minor topics obtaining not only keywords of the topics but the words spread in the neighbor of the keywords.

IV. Experiments and Results

In this section, the experiments for our proposal and other comparable methods are explained. One comparable method selects the keywords by term frequency. Another comparable method is LSA, which is popular to reduce the dimension of a document vector. Another method uses all the words without index word selection.

A. *Experimental environment and procedures*

A test collection was used for the evaluation of a similar document retrieval system, MEDLINE, for the experimental data. MEDLINE consists of 1,033 English documents in medical science and bio science. In addition, the test collection has 30 retrieval queries. Associations or connections, which are represented by binary labels, are given for each query to confirm if they are appropriate. We used 5 queries. Our experiments used the associations for the construction of a user profile and an evaluation of the filtering was carried out with the user profile.

Table 1 shows the five associations used as well as the number of documents related to each query and its content. In Table 1, "# of related Doc." denotes the number of documents which we would like to select.

For the first step of the experiment, each document was represented as a document vector with the vector space model. The words of weight were all stemmed words without the stop words in the document set and the total number of words was 7,014. We used the SMART stop list [17] when removing the stop words.

Table 1. Organization of the experimental data

No.	# of related Doc.	Query content
1	37	The crystalline lens in vertebrates
8	10	The effects of pesticide on the bone marrow
12	8	Effect of azathioprine on LE, particularly in regard to renal lesions.
16	12	Separation anxiety in infancy and in preschool children
29	38	Hereditary implications of prolonged neonatal obstructive jaundice associated with liver pathology.

The NMFSC was applied to these document vectors and hundreds of bases which were considered to characterize the topics included in the document set were obtained. We tried to obtain 100 bases, 300 bases and 500 bases. The numbers of extracted bases denote about 10%, 30% and 50% of the number of documents, respectively. The keywords which represented a topic were selected depending on the weight of the word in each basis. We tried to select as many keywords from each basis and the total number was about 1,000 words. In addition, these keywords were set to a part of the index words. Then, using these keywords, we calculated the chi-squared measure for each word and added the words whose chi-squared measures were high to the index words until the number of index words reached 30% or 50% of the total number of words. In addition, we tried the experiment when the index words consisted of only the keywords. Finally, we reconstructed the document vectors with these index words in order to make the user profile and evaluated the filtering accuracy of the profile. A user profile was constructed using Eq. (5), and coefficients α and β in Eq. (5) were determined as:

$$\alpha = 1, \beta = \frac{\mathrm{N_I}}{\mathrm{N_U}} \cdots (15)$$

where $\mathrm{N_I}$ and $\mathrm{N_U}$ denote the number of documents related and unrelated to the query respectively. Similarities as defined by the inner product between a user profile and documents were adopted in order to classify the documents. The similarity *Sim* between the user profile $\mathbf{u}$ and the document $\mathbf{d}_i$ was defined as:

$$Sim = \mathbf{u}^{\mathrm{T}}\mathbf{d}_i.$$

If the similarity *Sim* was more than 0, the document $\mathbf{d}_i$ was regarded to have association. In addition, in constructing the user profile, we performed a leave-one-out method and evaluated by the percentage of the correct determination of associations. This method is referred to as "NMFSC". The experimental process is described in the following steps:

Step1. Construct the document vectors with the vector space model.
Step2. Apply the NMFSC to the document vectors and obtain hundreds of bases.
Step3. Extract the keywords whose weight is the highest in each basis and set those words to index words.

Step4. Calculate the chi-squared measure by (14).
Step5. Assign the words whose chi-squared measure is high to the index words.
Step6. Reconstruct the document vectors with the selected index words.
Step7. Construct the user profile.
Step8. Perform the recommendation from documents for evaluation using the user profile.

For comparison, we tried three other methods. One method was based on the term frequency in a corpus. The method, which was similar to [9], extracted the words whose total term frequency in the corpus was the highest as the keywords. We extracted the same number of keywords from the various bases to obtain a total of 1,000 words. This method is referred to as "TF". The next method was LSA which is a popular method to reduce the dimension of the document vectors by projecting them in the word space into spaces structured by latent semantics. Singular Value Decomposition (SVD) is often used to obtain the latent semantics and the singular vectors are considered to represent the latent semantics. The singular vectors are obtained by applying SVD to a term-document matrix as follows:

$$\mathrm{D} = \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}} \cdots (16)$$

where row vectors U, V are singular vectors. Σ is a diagonal matrix whose elements are singular values σ_i, *i=1, 2, ..., p* where *p* means the number of singular vectors, which correspond with $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_{\mathrm{p}}$. We selected the *k* singular vectors with the largest singular values. In addition, *k* was determined by a cumulative contribution ratio r_{k} defined as:

$$r_k = \sum_{i}^{\mathrm{k}} \sigma_i^2 / \sum_{j}^{p} \sigma_j^2 \cdots (17)$$

Now r_k was set to 0.7, 0.8, and 1.0. Then, the document vectors were projected into the space structured by the latent semantics. This method is referred to as "LSA". The other method uses the total words as the index words. This method is referred to as "ORIG". The numbers of document vector's dimension for each method are presented in Table 2. "NMFSC500" denotes the results of experiments in which 500 bases were extracted by the NMFSC, as are "NMFSC300" and "NMFSC100". Moreover, "Key" means that the index words consist only of keywords, and "30%" and "50%" means that the number of index words is about 30% and 50%, respectively, of the total number of words included in the document set. The figures "0.7", "0.8" and "1.0" of LSA denote the cumulative contribution ratios. In addition, Table 2 presents the number of keywords extracted from each topic, i.e. each basis of the NMFSC, for NMFSC500, NMFSC300 and NMFSC100. The differences of the number of the dimension for the same percentage cause why the words were removed depending on the frequency or chi-squared measure respectively. Hence, we selected the words which have the same frequency or chi-squared measure even though the number of the selected words exceeded the determined percentage. Moreover, if the exceedance went over 3%, the number of the words was less than the determined percentage.

Table 2. The numbers of document vector's dimension for each method. "DIM" denotes the number of document vector's dimension and "# for each" denotes the number of the keywords extracted from each topic.

Methods		DIM	# for each
NMFSC500	Key	965	2
	30%	2,087	
	50%	3,466	
NMFSC300	Key	880	3
	30%	2,091	
	50%	3,571	
NMFSC100	Key	993	10
	30%	2,100	
	50%	3,511	
TF	Key	1,000	-
	30%	2,159	
	50%	3,529	
LSA	0.7	422	-
	0.8	554	
	1.0	1,033	
ORIG		7,014	-

B. Experimental results

The percentages of an accurate determination of our method for each query are presented in Tables 3. The description in the table follows the ones in Table 2. In addition, the keywords of NMFSC300 and TF are presented in Table 4 for discussion on the differences between NMFSC and TF. The examples in the Table 4 are the 20 keywords which has the most document frequency. Fig 1 shows the number of keywords for each document frequency with respective methods.

V. Discussion

Comparing the results of our proposal shown in Tables 3, our proposal could obtain more accurate associations in all of the experiments. The reason may be that this method enables selection of the necessary words to represent a document appropriately. Focusing on the results of No.1, the accuracy rises about 10% above the ORIG accuracy when the number of bases and index words are set to 300 and 30% respectively. As with previous results, the accuracies improve about 5% from ORIG for the result of No.8 when setting the number of bases and index words as 500 bases and 30% of the total index words,

Table 4. Comparison of the document frequency of the keywords for NMFSC300 and TF. "df" denotes the document frequency of each word.

Table 3. Comparison of accuracies for applying methods

No.	NMFSC500			NMFSC300			NMFSC100			TF			LSA			ORIG
	Key	30%	50%	Key	30%	50%	Key	30%	50%	Key	30%	50%	0.7	0.8	1.0	
No.1	0.756	0.837	0.830	0.792	**0.891**	0.811	0.819	0.868	0.819	0.837	0.836	0.833	0.816	0.816	0.813	0.813
No.8	0.715	**0.719**	0.716	0.630	0.652	0.710	0.678	0.642	0.695	0.716	0.691	0.672	0.669	0.659	0.674	0.674
No.12	0.864	**0.884**	0.782	0.726	0.772	0.697	0.671	0.760	0.709	0.641	0.691	0.690	0.715	0.720	0.726	0.726
No.16	0.640	0.613	0.711	0.656	0.678	0.713	0.701	0.698	**0.744**	0.733	0.712	0.701	0.689	0.690	0.700	0.700
No.29	0.674	0.792	**0.817**	0.686	0.716	0.754	0.785	0.766	0.799	0.797	0.785	0.770	0.766	0.768	0.766	0.766

NMFSC300		TF	
Index Word	df	Index Word	df
present	299	Study	357
develop	182	Patient	302
relat	176	present	299
tissu	147	result	282
compar	144	Case	254
level	142	increas	252
consid	125	effect	246
form	122	Treat	225
human	114	Cell	215
measure	112	found	210
rate	111	Norm	204
typ	111	observ	202
mean	92	develop	182
numb	92	High	179
therap	92	Relat	176
remain	85	Show	168
total	82	Time	162
mechan	80	suggest	157
direct	79	Active	156
coltur	72	produc	153

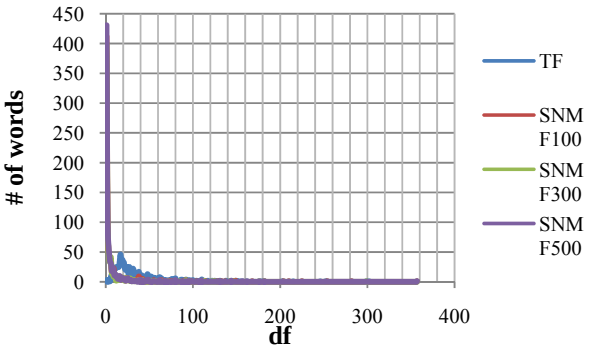

Fig 1. The number of keywords for each document frequency. "df" denotes the document frequency and "# of words" denotes the number

respectively. In the result of No12, the accuracy for the 500bases and 30% index words rises about 16% from ORIG. The experiment for No.16 results that our proposal improved by about 5% from ORIG when using 100 bases and 50% index words. Our proposal's result of No.29 improved about 5% when setting the 500 bases and 50% index words. These describe that the filtering accuracy improves dramatically by setting the appropriate number of bases and index words, the filtering accuracy improves dramatically. Especially, in these experiments, our proposal achieved to improve about from 5% to 16% compared with the accuracy of ORIG and LSA when the appropriate number of the bases and index words can be determined. In addition, the accuracies are able to be improved when using the detailed topics, i.e. a lot of bases.

Moreover, our method has an advantage that the documents can be represented with the index terms. This is considered to benefit actual filtering in that the query or other added documents do not require transforming the words into latent semantics.

Next, comparisons of our proposal to the results of "TF" were evaluated. Focusing on the results of No.1, the highest accuracy with this method is obtained when the number of bases and index words are set to 300 and 30%, respectively. Compared with the accuracy of "TF" with 30% of the total words, our method shows more effectiveness. However, when the index words are established by only the keywords, the accuracy of "TF" is better than our proposal. The same tendency appeared in other data, especially, queries No. 16 and 29. When focusing on the results of Nos.16 and 29 of "TF", the accuracies lessen with the addition of the index words obtained by chi-squared measure. It is considered that almost all necessary index words that separate the related documents are included in the keywords. The other results using "TF" show that the accuracies became also worse by adding the index words. With our proposal, the necessary index words can be appropriated by adding the index words obtained with the chi-squared measure. Moreover, our proposal obtained the highest accuracy in the results of the 4 methods for each query.

In order to discuss our proposal's character, we focus the selected index words, especially the difference from ones of TF. Focusing the Table 4, the keywords of "TF" are the words which have larger document frequency than the one of "NMFSC300". This notes that "TF" collects the keywords covering documents more widely than the NMFSC. Moreover, considering with the results presented in Fig 1, "NMFSC" extracts a lot of keywords which have less document frequency than "TF". In the figure, the index words obtained by TF include the most words whose document frequency is about 20. In contrast, the number of index words obtained by the NMFSC peaks around the ones whose document frequency less than 5. Especially, the keywords that are appeared in only a document are extracted a lot. This notes our method can cover the minor topics uncovered by term frequency in the document set. The aim of selection of the necessary words which previous methods could not obtain in order to construct a document vector was achieved when considered with the various topics included in a document set.

VI. Conclusion

In this paper, we have proposed a method of selecting index words in order to construct a document vector based on the topics by using the NMFSC and then applying these selected index words to a filtering system based on the user's interest. As a result, the filtering accuracies could be improved when setting the appropriate number of topics and index words respectively. This work showed that our method could obtain the minority topics and cover the index words included in the documents related to these minority topics. Covering the index words to characterize the document for separation was considered to be significant.

We will have to evaluate the advantage of our proposal's machine time for other methods and apply to the bigger corpus than MEDLINE. In addition, determination of the appropriate number of bases and index words automatically will also be the subject of our future work.

Acknowledgment

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

[1] G.Salton, M.J.McGill: "Introduction to Modern Information Retrieval", McGraw-Hill Book Company, 1983.

[2] P.O.Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints"', Journal of Machine Learning Research, Vol. 5, pp. 1457-1469, 2004.

[3] D.Lee and H.Seung, "Algorithms for non-negative matrix factorization", NIPS 2000, 2000.

[4] D.Lee and H.Seung, "Learning the parts of objects by non-negative matrix factorization", Nature, Vol. 401, pp.788-791

[5] S.Tsuge, M.Shishibori, S.Kuroiwa *and* K.Kita: "Dimensionality Reduction Using Non-negative Matrix Factorization for Information Retrieval", Natural Language Processing and Knowledge Engineering Mini Symposium, IEEE SYSTEMS, MAN, AND CYBERNETICS 2001 (NLPKE), pp.960-965, 2001

[6] E. P. Jiang: "Information Retrieval and Filtering Using the Riemannian SVD", Ph.D. Thesis, Dept. of Computer Science, The University of Tennessee, Knoxville, TN, 1988.

[7] S.Deerwester, T.Dumais, T.Landauer, W.Furnas and A.Harshman: "Indexing by Latent Semantic Analysis", Journal of the Society for Information Science, Vol.41, No.6, pp.391-497

[8] T.Kolenda and L.K.Hansen: "Independent Components in Text", Advances in Independent Component Analysis, Springer-Verlag, 2000.

[9] T.Yokoi, H.Yanagimoto and S.Omatu: "The Proposal for the Way to Recommend Information with ICA", The Ninth Int. Synp. on Artificial Life and Robotics(AROB 9th '04), Proc. pp. 694-697, 2004

[10] Xu. W., Liu. X., Gong. Y.:"Document Clustering Based On Non-negative Matrix Factorization", Proceedings of SIGIR'03, pp.267-273, 2003.

[11] M.W. Berry, M. Browne, A.N. Langville, "Algorithms and Applications for Approximate Nonnegative Matrix Factorization", V.P. Pauca, and R.J. Plemmons, Computational Statistics & Data Analysis 52(1), pp. 155-173, 2007.

[12] P.O.Hoyer, "Nonnegative Sparse Coding", Proc. IEEE Workshop Neural Networks for Signal Processing, 2002

[13] Xu.W., Liu. X., Gong. Y., "Nonnegative Matrix Factorization for Visual Coding", Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing(ICASSP2003), 2003

[14] Y.Matsuo and M.Ishizuka, "Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information", Int'l Journal on Artificial Intelligence Tools, Vol.13, No.1, pp.157-169, 2004

[15] Yukio Ohsawa, Nels E. Benson and Masahiko Yachida, "KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor", Proc. Advanced Digital Library Conference (IEEE ADL'98), pp.12-18 (1998)

[16] J. Rocchio: "Relevance Feedback in Information Retrieval", The SMART Retrieval System Experiments in Automatic Document Processing, pp313-323, 1971.

[17] SMART stop-list
ftp://ftp.cs.cornell.edu/pub/smart/english.stop