Industrial Engineering
& Management Systems
Vol 11, No 1, Mar 2012, pp.87-93
ISSN 1598-7248 | EISSN 2234-6473 |
http://dx.doi.org/10.7232/iems.2012.11.1.087

An Optimal Weighting Method in Supervised Learning of Linguistic Model for Text Classification

Kenta Mikawa*
Graduate School of Creative Science and Engineering, Waseda University
Takashi Ishida
Media Network Center, Waseda University
Masayuki Goto
Faculty of Science and Engineering, Waseda University

(Received: November 15, 2011 / Revised: January 28, 2012 / Accepted: February 3, 2012)

ABSTRACT

This paper discusses a new weighting method for text analyzing from the view point of supervised learning. The term frequency and inverse term frequency measure (*tf-idf* measure) is famous weighting method for information retrieval, and this method can be used for text analyzing either. However, it is an experimental weighting method for information retrieval whose effectiveness is not clarified from the theoretical viewpoints. Therefore, other effective weighting measure may be obtained for document classification problems. In this study, we propose the optimal weighting method for document classification problems from the view point of supervised learning. The proposed measure is more suitable for the text classification problem as used training data than the *tf-idf* measure. The effectiveness of our proposal is clarified by simulation experiments for the text classification problems of newspaper article and the customer review which is posted on the web site.

Keywords: Text Classification, Weighting Method, Vector Space Model, Cosine Similarity

* Corresponding Author, E-mail: mikawa@it.mgmt.waseda.ac.jp

1. INTRODUCTION

Due to development of information technology, the effectiveness of knowledge discovery from enormous document data is suggested in much of the literatures on this subject (Hearst. 1999). There are many web sites where customers can post their free comments about merchandise that they bought and used. On the internet, the number of customer reviews is increasing day by day. Therefore it has been easy to get a large amount of document data and analyze it for several purposes. Customer reviews consist of not only free comments but customer information and the degree of satisfaction about items as metadata. The analysis using the metadata is more helpful for knowledge discovery than using only text data. The techniques for text mining are developed for the purpose of getting information. Various methods have been proposed in this research field, for example, vector space model (Manning *et al*., 2008), (Mikawa *et al*., 2012), probabilistic model (Hofmann. 1999), (Bishop, 2006) and so on.

In this paper, a vector space model is the focus for document analysis. To construct a vector space model for document analysis, the documents are separated into terms or words (morphemes) by using the morphological analysis (Nagata. 1994). After that, each document is represented by a vector whose elements express the information of word frequency of appearance. Because the vector space is built by the information of word frequency, the characteristics of a document vector model should be remarkable: high dimension and sparseness. Generally speaking, untold thousands of words or more should be treated to represent a document vector using effective words appearing in all documents.

As mentioned above, there are enormous words which are appeared in whole documents. In addition, term frequency of each word varies widely in length. Therefore, the performance of text analyzing depends on term frequency of words which is appeared each documents. That is, it depends on the length of documents. To avoid this, several weighting approach for each word has been proposed. For instance, *tf-idf* weighting (Salton *et al.*, 1988), PWI (Probability-weighted amount of information) (Aizawa, 2000, 2003), mutual information (McCallum *et al*., 1998) and so on. And *tf-idf* weighting is one of the most famous method for weighting terms. However, it is proposed for information retrieval and the effectiveness is empirically shown. Therefore, the theoretical optimality is not proved. In addition, it doesn't use the metadata or side information for weighting each word. Nowadays, it can be easy to get or use those, and by using that information, it supposes to improve the performance of each analysis.

From above discussion, the purpose of this study is to propose a new weighting method for each word from the view point of supervised learning. We show the way of estimating an optimal word weighting by solving maximization problems. The effectiveness of this method is clarified by case experiments of applications to the customer review which are posted on web sites and newspaper articles which are used as a bench mark data.

In section 2, basic formulation of vector space model and weighting methods which have already proposed are explained. In section 3, the proposed method of weighting each word and the way of its estimation is explained. The illustration of simulation experiments in order to clarify the effectiveness of our proposal and the results acquired from the experiments are explained in section 4. Finally, the conclusion of this study is stated in section 5.

2. BASIC INFORMATION FOR ANALYSIS OF TEXT DOCUMENTS

In this paper, the vector space model is adopted to represent the document data. In this section, the premises and the notations of this research are defined and relative methods are explained.

2.1 Similarity among Documents

Let the set of documents be $\Delta = \{d_1, d_2, \cdots, d_D\}$, and the set of categories to which each document in training data belongs be $C = \{c_1, c_2, \cdots, c_N\}$. Here, D and N are the numbers of documents and categories respectively.

All documents in Δ are separated into morphologies by the morphological analysis. Selecting valid words from the given morphologies by using frequencies, mutual information, or other methods, a word set is constructed to represent documents in the vector space. Let the word set from the all documents in Δ be $\Sigma = \{w_1, w_2, \cdots, w_W\}$.

Then each document can be expressed as a point in the vector space. Here, W is the number of different valid words which appear in Δ and it is equivalent to the dimension of the vector space. Here, let the frequency of a word w_j in the document $\boldsymbol{d}_i$ be v_j^i, and it can be expressed by W-dimensional vector as follows:

$$\boldsymbol{d}_i = (v_1^i, v_2^i, \cdots, v_W^i,)^T, \tag{1}$$

where, T means transposition of a vector. That is, the document space can be constructed by regarding each component of vectors as the information about the frequency of each word.

By expressing document as a vector, a similarity or distance metrics between documents can be measured in the vector space. Here, the distance metric of document vectors $\boldsymbol{d}_i$ and $\boldsymbol{d}_j$ can be expressed by using the Euclid distance which is traditionally used as follows:

$$\begin{aligned} \mathrm{dis}_e(\boldsymbol{d}_i, \boldsymbol{d}_j) &= \sqrt{(\boldsymbol{d}_i - \boldsymbol{d}_j)_T(\boldsymbol{d}_i - \boldsymbol{d}_j)}, \\ &= \sqrt{\textstyle\sum_{k=1}^{W}(v_k^i - v_k^j)^2}. \end{aligned} \tag{2}$$

However, the Euclid distance sometimes cannot work effectively to treat the document data. For example, many elements in a document vector are almost 0 and this property is called “sparseness.” If there are only a few different words between two documents, these documents are judged as having high similarity.

On the other hand, the cosine similarity measure between document vectors has asymptotically good performance in a sparse vector space (Goto *et al*., 2007), (Goto *et al*., 2008).

$$\begin{aligned} \mathrm{sim}_c(\boldsymbol{d}_i, \boldsymbol{d}_j) &= \frac{\boldsymbol{d}_i^T \boldsymbol{d}_j}{\|\boldsymbol{d}_i\| \|\boldsymbol{d}_j\|}, \\ &= \frac{\sum_{k=1}^{W} v_k^i v_k^j}{\sqrt{\sum_{k=1}^{W}(v_k^i)^2}\sqrt{\sum_{k=1}^{W}(v_k^j)^2}} \end{aligned} \tag{3}$$

2.2 Weighting Method for Documents

The way of weighting can be expressed the Hadamard product of document vector $\boldsymbol{d}_i$ and weighting vector $\boldsymbol{f}$. Here, let weighting vector $\boldsymbol{f}$ be

$$\boldsymbol{f} = (f_1, f_2, \cdots, f_W)^T \tag{4}$$

Here, f_k is weighting function for the word w_k. By using that, let the weighted document vector be $\boldsymbol{d}_i^*$, and it is given by

$$\begin{aligned} \boldsymbol{d}_i^* &= \boldsymbol{f} \odot \boldsymbol{d}_i \\ &= (f_1 v_1^i, f_2 v_2^i, \cdots, f_W v_W^i)^T \end{aligned} \tag{5}$$

where, let $\boldsymbol{d}_i^* = \boldsymbol{f} \odot \boldsymbol{d}_i$ is Hadamard product of $\boldsymbol{f}$ and $\boldsymbol{d}_i$.

There are several weighting methods for text data analyzing. As mentioned above, the most famous method is *tf-idf* weighting. It can be calculated the product of *tf*(term frequency, that is v_j^i) and *idf*(inverse document frequency). Here, the *tf* is representing the frequencies of each term. And *idf* is a monotonically decrease function of an appearance rate of term in documents. Let $df(w_k)$ be a number of documents in which the word w_k appears. Then the *idf* weighting for the word w_k is given by

$$idf(w_k) = \log \frac{D}{df(w_k)}. \tag{6}$$

Here, let the weight function $\boldsymbol{f}_k$ be $idf(w_k)$, *tf-idf* weighting can be expressed as follows:

$$\boldsymbol{d}_i^* = (idf(w_1)v_1^i,\ idf(w_2)v_1^i, \cdots, idf(w_W)v_1^i)^T. \tag{7}$$

3. THE METHOD OF SUPERVISED WEIGHTING FOR TEXT CLASSIFICATION

As mentioned above, *tf-idf* weighting is not made use of side information or metadata when calculating each word weight. In the basic text classification problem, training data has information which category training data belongs to. Therefore, by making the most use of the information for weighting each word can be improved its performance.

In this section, we propose the optimal weighting method and its estimation for text classification. And we derive the expression of optimal weight is given by the solution of maximization problem.

To formulate the learning algorithm to estimate the optimal weighting from training data set, the centroid of each category is defined as equation (8). The centroid can be calculated by using the training data with known category. Let the centroid vector of category c_n ($n = 1, 2, \cdots, N$) be $\boldsymbol{g}_n = (g_1, g_2, \cdots, g_{nW})^T$. Then the centroid vector $\boldsymbol{g}_n$ is given by

$$\begin{aligned} \boldsymbol{g}_n &= \frac{1}{|c_n|} \sum_{\boldsymbol{d}_i \in c_n} \boldsymbol{d}_i \\ &= \frac{1}{|c_n|} \sum_{\boldsymbol{d}_i \in c_n} (v_1^i, v_2^i, \cdots, v_W^i)^T \end{aligned} \tag{8}$$

Here, the $|c_n|$ is a number of documents contained category c_n. And the same as equation (5), weighted centroid of category c_n is given by

$$\begin{aligned} \boldsymbol{g}_n^* &= \boldsymbol{f} \odot \boldsymbol{d}_n \\ &= (f_1 g_{n1}, f_2 g_{n2}, \cdots, f_W g_{nW})^T \end{aligned} \tag{9}$$

Normally, training data should be allocated by near the centroid of its category because a new document data is classified by the distance from the centroids. Due to that, the optimization of the weight should maximize the similarity between training data and its centroid.

From the above discussion, the estimation of optimal weighting vector $\boldsymbol{f}$ is to maximize cosine similarity between weighted document vector $\boldsymbol{d}_i^*$ and weighted centroid vector $\boldsymbol{g}_n^*$. If the weighting vector can be obtained by the maximization of the similarity between data and its category's centroid, the proposed weighting method is expected to have a good performance to classify the data into proper categories. Then, our optimal weighting vector $\hat{\boldsymbol{f}}$ is given by

$$\begin{aligned} \hat{\boldsymbol{f}} &= \arg\max_{\boldsymbol{f}} \sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in C_n} \mathrm{sim}_c(\boldsymbol{d}_i^*, \boldsymbol{g}_n^*), \\ &= \arg\max_{\boldsymbol{f}} \sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in C_n} \frac{\sum_{k=1}^{W} f_k v_k^i f_k g_{nk}}{\sqrt{\sum_{k=1}^{W} (f_k v_k^i)^2} \sqrt{\sum_{k=1}^{W} (f_k g_{nk})^2}}, \\ &= \arg\max_{\boldsymbol{f}} \sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in C_n} \frac{\sum_{k=1}^{W} v_k^i (f_k)^2 g_{nk}}{\sqrt{\sum_{k=1}^{W} (f_k v_k^i)^2} \sqrt{\sum_{k=1}^{W} (f_k g_{nk})^2}}. \end{aligned} \tag{10}$$

Here, $\boldsymbol{f} = (f_1, f_2, \cdots, f_W)^T$ is weighting vector, and it satisfies $\sum_{k=1}^{W} f_x = 1$.

Equation (10) can be expressed by matrix operation by using the diagonal matrix which is constructed its elements by $(f_k)^2$. To use that, the optimal weight of each word can be estimated to maximize cosine similarity between training data $\boldsymbol{d}_i$ and each category's centroid $\boldsymbol{g}_n$ by using the matrix.

In order to solve the above problem, we introduce the diagonal metric matrix $M = [m_k]$ ($k = 1, 2, \cdots, W$) with the elements $(f_k)^2$, that is

$$M = \begin{pmatrix} m_1 & & 0 \\ & m_2 & \\ & \ddots & \\ 0 & & m_W \end{pmatrix} = \begin{pmatrix} (f_1)^2 & & 0 \\ & (f_2)^2 & \\ & \ddots & \\ 0 & & (f_W)^2 \end{pmatrix}.$$

Here, $M = [m_k]$ is satisfying $|M| = 1$, and $|M|$ is determinant of M.

From above discussion, the equation (10) can be rewritten as follows:

$$\begin{aligned} \hat{M} &= \arg\max_{M} \sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in c_n} \mathrm{sim}_c(\boldsymbol{d}_i^*, \boldsymbol{g}_n^*), \\ &= \arg\max_{M} \sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in c_n} \frac{\boldsymbol{d}_i^T M \boldsymbol{g}_n}{\|\boldsymbol{d}_i^*\| \|\boldsymbol{g}_n^*\|}, \end{aligned}$$

$$= \arg\max_{M} \sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in c_n} \frac{\sum_{k=1}^{W} v_k^i m_k g_{nk}}{\sqrt{\sum_{k=1}^{W} m_k (v_k^i)^2} \sqrt{\sum_{k=1}^{W} m_k (g_{nk})^2}}. \quad (11)$$

Supject to $|M| = 1$

From the above formulation, the following theorem is obtained.

Theorem 1: The metric matrix $\hat{M} = [\hat{m}_k] (k = 1, 2, \cdots, W)$ which is satisfies the equation (11) is given by

$$\hat{m}_k = \frac{\left[\prod_{k=1}^{W} \left\{ \sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in c_n} v_k^i g_{nk} \right\}\right]^{\frac{1}{W}}}{\sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in c_n} v_k^i g_{nk}}. \quad (12)$$

(For the proof, see APPENDIX.)

From above discussion, the similarity between document vector and centroid can be calculated by using the metric matrix $\hat{M}$ given by the equation (12) as follows:

$$\mathrm{sim}_M(\boldsymbol{d}_i, g_n) = \frac{\boldsymbol{d}_i^T \hat{M} \boldsymbol{g}_n}{\|\boldsymbol{d}_i^*\| \|\boldsymbol{g}_n^*\|}, = \frac{\sum_{k=1}^{W} v_k^i \hat{m}_k g_{nk}}{\sqrt{\sum_{k=1}^{W} \hat{m}_k (v_k^i)^2} \sqrt{\sum_{k=1}^{W} \hat{m}_k (g_{nk})^2}} \quad (13)$$

After calculating similarity among test data and all centroids by equation (13), test data is classified to the most similar category.

4. EXPERIMENTS

4.1 Experimental Condition

In this section, simulation experiments are conducted to verify the effectiveness of our method by using Japanese document data in practice. The experiments are performed for the two data sets, i.e, customer reviews which are posted on a web site and newspaper articles with pre-assigned categories. The suitable weight is estimated by learning through these data. And the effectiveness is confirmed by the performance of classification of test data into the correct categories.

The basic process of experiments is as follows.

Step1: Calculation of all centroids from training data.

Step2: From equation (12), learning metric matrix $\hat{M}$ from training data.

Step3: From equation (13), to calculate similarity among test data and all category's centroid. And it classifies that with most similar category.

As mentioned above, two different types of data are used on this experiment. The first one is articles from the Mainichi newspaper in 2005 which are used as a benchmark data for document classification. The second one is customer review that is posted on a web site in order to recognize the performance of our method for real data. News articles in the Mainichi newspaper consist of several categories (economics, sports, politics and so on) and this time, three and five categories are extracted at random. Customer review consists of not only text data but degree of customer's satisfaction as mentioned above. In this experiment, we use two categories of that which are highest degree of satisfaction and lowest one.

A condition of experiments is shown in Table 1.

Table 1. A Condition of Experiments.

	The number of category	The number of training data	The number of test data
Mainichi newspaper	3 and 5 categories	100 to 500 data in each category	500 data in each category
Customer review	2 categories	500 data in each category	500 data in each category

For the sake of comparison, the experiment of the common cosine measure with only term frequency (*tf* measure) and *tf-idf* measure between different two data points was also performed. The criteria of evaluation are classification accuracy rate. It's the ratio of the number of documents which are classified into correct categories to the total number of documents.

4.2 Result of Experiments

The results of experiments are shown in Figure 1, Figure 2 and Figure 3. Figure 1 and Figure 2 show the cases of the newspaper article. The former is the case of three categories and the latter is that of five.

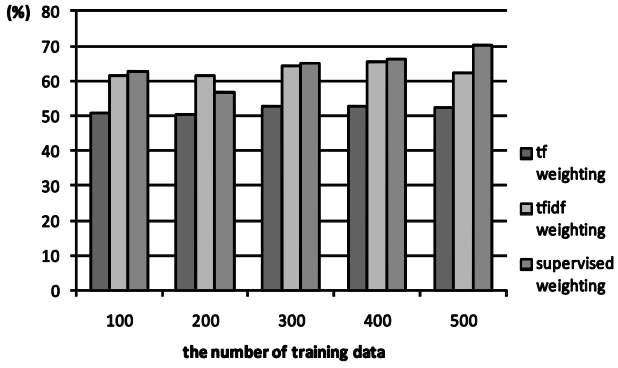

Figure 1. Result for Classification Accuracy Using Newspaper (3 Categories).

In addition, Figure 3 shows the case of customer review. It shows supervised weighting (proposed method),

tf-idf weighting and only use term frequency (*tf* weighting).

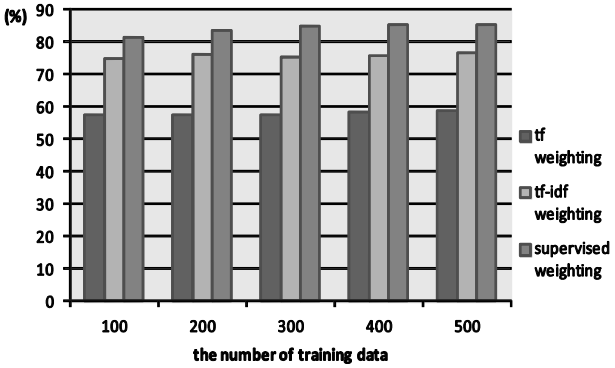

Figure 2. Result for Classification Accuracy Using Newspaper (5 Categories).

From the Figure 1 and Figure 2, the proposed method is basically superior to *tf-idf* weighting and *tf* weighting methods. However, from the Figure 3, the proposed method and the other two methods are almost the same performance. That is, the performance of proposed method is more suitable when newspaper article is used.

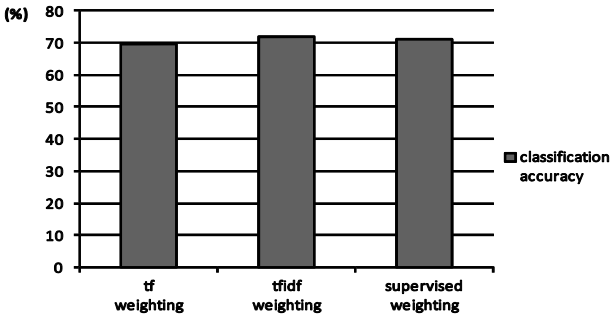

Figure 3. Result for Classification Accuracy Using Customer Review.

4.3 Discussion

From the result, the proposed method is more suitable than conventional methods when newspaper article is used. In the proposed method, category information of training data as weighting parameter is taken into consideration. This property can work well for text classification. On the other hand, *tf-idf* weighting is not used each category's characteristics. As the result, the performance is improved.

However, the performance of the proposed method which can be solved regarded as the optimal solution is not improved drastically in the customer review classification. Because customer write their comment freely for merchandise which they bought and used. Due to the fact, it may appear many invalid words and the tendency of words is not different between each category. The customer reviews are too high of freedom of text data for the proposed method to estimate such complex statistical structure. By these reason, the proposed method doesn't work effectively. It is necessary to improve the method as getting better performance for customer review (real data) as the future work.

5. CONCLUSION

In this paper, the suitable weighting method by estimating optimal weight from a training data set is proposed. Our proposal is based on supervised learning with optimal metric matrix $\hat{M}$ which can be calculated by a training data set.

From the simulation experiment, the proposed method is superior than conventional method when newspaper article is used.

A future work is to calculate the contribution ratio in each word which is weighted by the proposed method for the text classification in order to improve performance.

ACKNOWLEDGEMENT

The authors would like to acknowledge Prof. Shigeichi Hirasawa, Cyber University, and Mr. Gendo Kumoi, Waseda University, for their useful suggestions and cooperation for our research.

REFERENCES

Aizawa, A. (2000), The Feature Quantity: An Information Theoretic Perspective of Tfidf-like Measures, *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 104-111.

Aizawa, A. (2003), An Information-theoric perspective tf-idf Measure, *Information Processing and Management*, **39**, 45-65.

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

Goto, M., Ishida, T., and Hirasawa, S. (2007), Statistical Evaluation of Measure and Distance on Document Classification Problems in Text Mining, *IEEE International Conference on Computer and Information Technology*, 674-679.

Goto, M., Ishida, T., Suzuki, M., and Hirasawa, S. (2008), Asymptotic Evaluation of Distance Measure on High Dimensional Vector Space in Text Mining, *International Symposium on Information Theory and its Applications*.

Hearst, M. A. (1999), Untangling text data mining, *ACL '99 Proceedings*, 3-10.

Hofmann, T. (1999), Probabilistic Latent Semantic Indexing, *Proceeding of the 22nd International Con-*

ference on Research and Development in Information Retrieval, 50-57.

Manning, C. D., Raghavan, P., and Schuetze, H. (2008), Introduction to Information Retrieval, *Cambridge University Press*.

McCallum, A. and Nigam, K. (1998), A Comparison of Event Models for Naive Bayes Text Classification, *Proceeding of AAAI-98 Workshop on Learning for Text Categorization*, 41-48.

Mikawa, K., Ishida, T., and Goto, M. (2012), A Proposal of Extended Cosine Measure for Distance Metric Learning in Text Classification, *Proceeding of 2011 IEEE International Conference on the Systems, Man, Cybernetics (SMC)*, 1741-1746.

Nagata, M. (1994), A Stochastic Japanese morphological analyzer using a forward-DP backward-A^* best search algorithm, *Proceeding of the 15th International Conference on Computational Linguistics*, 201-207.

Salton, G. and Buckley, C. (1988), Term-Weighting Approaches in Automatic Text Retrieval, *Information Processing and Management*, **24**(5), 513-523.

APPENDIX: Proof of Theorem 1

We show the proof of Theorem 1 as follows.

To solve for M under the $|M| = 1$ is equivalent to solve following maximization:

$$\hat{M} = \arg\max_{M} \sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in c_n} \mathrm{sim}_c(\boldsymbol{d}_i^*, \boldsymbol{g}_n^*), \\ = \arg\max_{M} \sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in c_n} \frac{\boldsymbol{d}_i^T M \boldsymbol{g}_n}{\|\boldsymbol{d}_i^*\| \, \|\boldsymbol{g}_n^*\|} \tag{14}$$

In the following, to simplify the calculation, document $\boldsymbol{d}_i^*$ and centroid $\boldsymbol{g}_n^*$ is normalized by its norm. In other words, it can be said $\|\boldsymbol{d}_i^*\| = \|\boldsymbol{g}_n^*\| = 1$.[1)] So equation (14) becomes

$$\hat{m}_k = \arg\max_{M} \sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in c_n} \left\{ \sum_{k=1}^{W} v_k^i m_k g_{nk} \right\} \tag{15}$$

Here, M is diagonal matrix and $|M| = 1$. Therefore,

$$|M| = \prod_{k=1}^{W} m_k = 1.$$

Therefore,

$$\sum_{k=1}^{W} \log m_k = 0 \tag{16}$$

is derived.

To maximize the equation (15) under the equation (16), Lagrange multipliers method is used. Here, let Lagrange multiplier be λ, and Lagrange function can be defined

$$L = \sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in c_n} \left\{ \sum_{k=1}^{W} v_k^i m_k g_{nk} \right\} - \lambda \left\{ \sum_{k=1}^{W} \log m_k \right\} \tag{17}$$

Here, L is partially differentiated with respect to m_k. And put them 0,

$$\frac{\partial L}{\partial m_k} = \sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in c_n} v_k^i g_{nk} - \lambda \left\{ \frac{1}{m_k} \right\} = 0, \tag{18}$$

is derived. To solve the equation (18) for m_k,

$$m_k = \frac{\lambda}{\sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in c_n} v_k^i g_{nk}}. \tag{19}$$

is derived.

Here, let M^{-1} be $[m_m^{-1}]$. Then form the equation (19), m_m^{-1}. is

$$m_m^{-1} = \frac{\sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in c_n} v_k^i g_{nk}}{\lambda}. \tag{20}$$

Here, let set A be $A = [a_{kl}]$ and set a_k be

$$a_k = \sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in c_n} v_k^i g_{nk}. \tag{21}$$

Then $\boldsymbol{A}$ is

$$m_k^{-1} = \lambda^{-1} a_k. \tag{22}$$

Therefore,

$$a_k = \lambda m_k^{-1}, \tag{23}$$

is derived. From the equation (22), $A = \lambda M^{-1}$, is acquired. By the characteristics of M,

$$|A| = \lambda^W \left| M^{-1} \right| = \lambda^W, \tag{24}$$

is derived. Therefore, λ becomes

$$\lambda = |A|^{1/W}, \tag{25}$$

Accordingly, form equations (19) and (25), $\hat{m}_k$ becomes

$$\hat{m}_k = \lambda a_k^{-1} = |A|^{1/W} a_k^{-1}. \tag{26}$$

1) In case of $\|\boldsymbol{d}_i^*\| \neq 1$, $\|\boldsymbol{g}_n^*\| \neq 1$, it can be easy to extend.

Here, because A is diagonal matrix, $|A|$ is given by

$$|A| = \prod_{k=1}^{W} \left\{ \sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in c_n} v_k^i g_{nk} \right\}. \quad (27)$$

Consequently,

$$|A|^{1/W} = \left[\prod_{k=1}^{W} \left\{ \sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in c_n} v_k^i g_{nk} \right\} \right]^{1/W}, \quad (28)$$

is derived.

Accordingly, from equations (26) and (28), each $\hat{m}_k$ becomes

$$\hat{m}_k = \frac{\left[\prod_{k=1}^{W} \left\{ \sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in c_n} v_k^i g_{nk} \right\} \right]^{1/W}}{\sum_{n=1}^{N} \sum_{\boldsymbol{d}_i \in c_n} v_k^i g_{nk}}. \quad (29)$$

The proof is complete. □