Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 11(4): 241-246, Dec. 2013 Regular paper

Enhancing Text Document Clustering Using Non-negative Matrix Factorization and WordNet

Chul-Won Kim[1] and Sun Park[2*], *Member*, *KIICE*

[1]Department of Computer Engineering, Honam University, Gwangju 506-714, Korea
[2]Networked Computing System Lab., Gwangju Institute of Science and Technology, Gwangju 500-712, Korea

Abstract

A classic document clustering technique may incorrectly classify documents into different clusters when documents that should belong to the same cluster do not have any shared terms. Recently, to overcome this problem, internal and external knowledge-based approaches have been used for text document clustering. However, the clustering results of these approaches are influenced by the inherent structure and the topical composition of the documents. Further, the organization of knowledge into an ontology is expensive. In this paper, we propose a new enhanced text document clustering method using non-negative matrix factorization (NMF) and WordNet. The semantic terms extracted as cluster labels by NMF can represent the inherent structure of a document cluster well. The proposed method can also improve the quality of document clustering that uses cluster labels and term weights based on term mutual information of WordNet. The experimental results demonstrate that the proposed method achieves better performance than the other text clustering methods.

Index Terms: Non-negative matrix factorization, Semantic features, Term mutual information, Text document clustering, WordNet

I. INTRODUCTION

Traditional document clustering methods are based on the bag of words (BOW) model, which represents documents with features, such as weighted term frequencies. However, these methods ignore semantic relationships between terms within a document set. The clustering performance of the BOW model is dependent on a distance measure of document pairs. However, this distance measure cannot reflect the real distance between two documents because the documents are composed of high-dimension terms with respect to complicated document topics. Further, the results of clustering documents are influenced by the document properties or the cluster forms desired by the user. Recently, internal and external knowledge-based approaches have been used for overcoming the problems of the vector model-based document clustering method [1-3].

Internal knowledge-based document clustering determines the inherent structure of a document set by using a factorization technique [4-9]. These methods have been studied intensively, and although they have many advantages, the successful construction of semantic features from the original document set remains limited with respect to the organization of very different documents or the composition of similar documents [10].

External knowledge-based document clustering exploits the term ontology constructed using an external knowledge database with respect to Wikipedia or WordNet [11-14]. The

Received 28 December 2012, **Revised** 13 March 2013, **Accepted** 22 March 2013
***Corresponding Author** Sun Park (E-mail: sunpark@nm.gist.ac.kr, Tel: +82-62-715-3136)
Networked Computing System Lab., Gwangju Institute of Science and Technology, 123 Cheomdangwagi-ro, Buk-gu, Gwangju 500-712, Korea.

Open Access http://dx.doi.org/10.6109/jicce.2013.11.4.241 print ISSN: 2234-8255 online ISSN: 2234-8883

term ontology techniques can improve the BOW term representation of document clustering. However, it is often difficult to locate a comprehensive ontology that covers all the concepts mentioned in the document collection, which leads to a loss of information [1, 12]. Moreover, the ontology-based method incurs a relatively high cost as the ontology has to be constructed manually by knowledge engineers and domain experts.

In order to resolve the limitations of knowledge-based approaches, in this paper, we propose a text document clustering method that uses non-negative matrix factorization (NMF) and WordNet. The proposed method combines the advantages of the internal and external knowledge-based methods. In the proposed method, first, meaningful terms of a cluster for describing the cluster topics of documents are extracted using NMF. The extracted terms well represent the document clusters through semantic features (i.e., internal knowledge) that have the inherent structure of the documents. Second, the term weights of documents are calculated using the term mutual information (TMI) of the synonyms of documents terms obtained from WordNet (i.e., external knowledge). The term weights can easily classify documents into an appropriate cluster by extending the coverage of a document with respect to a cluster label.

The rest of this paper is organized as follows: Section II reviews related works on text document clustering. Section III describes the NMF algorithm. Section IV presents the proposed text document clustering method. Finally, Section V presents the evaluation and experimental results, and Section VI concludes this paper.

II. RELATED WORKS

Recently, a knowledge-based document clustering method, which is used for increasing the efficiency of document clustering, has been proposed; the techniques used in the method can be divided into internal and external knowledge-based techniques.

As an internal knowledge-based approach, Li et al. [8] pro-posed a document clustering algorithm, called the adaptive subspace iteration (ASI), which explicitly models the subspace structure and works well for high-dimensional data. This is influenced by the composition of the document set for document clustering. To overcome the orthogonal problem of latent semantic indexing (LSI), Xu et al. [4] proposed a document partitioning method based on NMF in the given document corpora. The results from the abovementioned method have a stronger semantic interpretation than those from LSI, and the clustering result can be derived easily using the semantic features of NMF. However, this method cannot be kernelized because the NMF must be performed in the original feature space of the data points. Wang et al. [9] used clustering with local and global regularization (CLGR), which uses local label predictors and global label smoothing regularizers. They achieved satisfactory results because the CLGR algorithm uses fixed neighborhood sizes. However, the different neighborhood sizes cause the final clustering results to deteriorate [9].

The external knowledge-based techniques for document clustering include TMI with conceptual knowledge by WordNet [11], concept mapping schemes from Wikipedia [12], concept weighting from domain ontology [13], and fuzzy associations with condensing cluster terms by WordNet [14].

III. NON-NEGATIVE MATRIX FACTORIZATION

This section reviews the NMF theory along with the corresponding algorithm. In this paper, we define the matrix notation as follows: Let X_{*j} be the j-th column vector of matrix X, X_{i*} be the ith row vector, and X_{ij} be the element of the ith row and the j-th column. NMF decomposes a given $m \times n$ matrix A into a non-negative semantic feature matrix W and a non-negative semantic variable matrix H, as shown in Eq. (1) [10].

$$A \approx WH , \tag{1}$$

where W denotes an $m \times r$ non-negative matrix and H represents an $r \times n$ non-negative matrix. Usually, r is chosen to be smaller than m or n; hence, the total sizes of W and H are smaller than the size of the original matrix A.

Further, an objective function is used for minimizing the Euclidean distance between each column of A and its approximation $\widetilde{A} = WH$; this function was proposed by Lee and Seung [10]. As the objective function, the following Frobenius norm is used:

$$\Theta_E(W,H) \equiv \left\|A - WH\right\|_F^2 \equiv \sum_{i=1}^{m}\sum_{j=1}^{n}\left(A_{ij} - \sum_{l=1}^{r} W_{il}H_{lj}\right)^2 . \tag{2}$$

W and H are continuously updated until $\Theta_E(W,H)$ converges under the predefined threshold or exceeds the set number of repetitions. The update rules are as follows:

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu}\frac{(W^T A)_{\alpha\mu}}{(W^T WH)_{\alpha\mu}} . \tag{3}$$

$$W_{i\alpha} \leftarrow W_{i\alpha}\frac{(AH^T)_{i\alpha}}{(WHH^T)_{i\alpha}} . \tag{4}$$

IV. PROPOSED DOCUMENT CLUSTERING METHOD

This paper proposes a document clustering method that uses cluster label terms generated by NMF and term weights based on the TMI of WordNet. The proposed method consists of three phases: preprocessing, extraction of cluster terms and term weights, and document clustering. In the subsections below, each phase is explained in full.

A. Preprocessing

In the preprocessing phase, Rijsbergen's stop words list is used for removing all stop words, and word stemming is removed using Porter stemming algorithm [15]. Then, the term document frequency matrix A is constructed from the document set.

B. Cluster Term Extraction and Term Weight Calculation

This section consists of two phases: cluster term extraction and term weight calculation. The cluster terms corresponding to the properties of the document clusters are extracted by using the semantic features of NMF; these terms can explain the topic of the document cluster well.

The extraction method can be described as follows: First, the term document frequency matrix A is constructed by executing the preprocessing phase. Second, the number of clusters (i.e., the number of semantic features r) is set, and NMF is performed on matrix A to decompose the two sematic feature matrices W and H. Finally, matrix W and Eq. (5) are used for extracting the cluster terms. The column vector of matrix W corresponds to the cluster, and the row vector of matrix W refers to the terms of the document; that is, an element of matrix W (i.e., the semantic feature value) indicates the extent to which the term reflects the cluster. The equation for extracting cluster terms is as follows:

$$C^p \leftarrow A_{ij} \ if \ \ p = \arg\max_{1 \le j \le r} (W_{ij} \ge asf), \tag{5}$$

where C^p denotes the term set of the pth cluster and A_{ij} represents the term corresponding to the semantic feature of the i-th row and the j-th column in matrix W. The average semantic feature value, *asf*, is as follows:

$$asf = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} W_{ij}}{m \times n}, \tag{6}$$

where m denotes the total number of rows (i.e., the number of terms) and n represents the total number of columns (i.e., the number of clusters).

Example 1 illustrates the cluster term extraction.

Example 1. Table 1 shows the six documents (i.e., the extracted the [2]'s Figure 4.10). Table 2 shows the term document frequency matrix generated in the preprocessing phase, described in Table 1. Table 3 presents the sematic feature matrix W obtained through NMF from Table 2, and the result of the average of non-zero elements of the semantic features vector *asf* calculated using Eq. (6). Table 4 shows the results of the extracted cluster terms from Table 3, which match the semantic feature values greater than the average semantic feature value *asf*.

The term weights are calculated using TMI based on the synonyms of WordNet. WordNet is a lexical database for the English language where words (i.e., terms) are grouped in synsets consisting of synonyms and thus representing a specific meaning of a given term [16].

Table 1. Document set of composition of six documents

Document	Document content
d1	A course on integral equations
d2	Tractors for semi-groups and evolution equations
d3	Automatic differentiation of algorithms: theory, implementation, and application
d4	Geometrical aspects of partial differential equations
d5	Ideals, varieties, and algorithms—An introduction to computational algebraic geometry and commutative algebra
d6	Oscillation theory for neutral differential equations with delay

Table 2. Term document frequency matrix from Table 1

Term	Document					
	d1	*d2*	*d3*	*d4*	*d5*	*d6*
course	1	0	0	0	0	0
integral	1	0	0	0	0	0
equations	1	1	0	1	0	1
tractors	0	1	0	0	0	0
semi-groups	0	1	0	0	0	0
evolution	0	1	0	0	0	0
automatic	0	0	1	0	0	0
different	0	0	1	1	0	1
algorithms	0	0	1	0	1	0
theory	0	0	1	0	0	1
implementation	0	0	1	0	0	0
application	0	0	1	0	0	0
geometric	0	0	0	1	1	0
aspects	0	0	0	1	0	0
partial	0	0	0	1	0	0
ideals	0	0	0	0	1	0
varieties	0	0	0	0	1	0
introduction	0	0	0	0	1	0
computational	0	0	0	0	1	0
algebra	0	0	0	0	2	0
commutative	0	0	0	0	1	0
oscillation	0	0	0	0	0	1
neutral	0	0	0	0	0	1
delay	0	0	0	0	0	1

Table 3. The *asf* and semantic feature matrix *W* by NMF from Table 2

Term	*r1* (cluster 1)	*r2* (cluster 2)	*r3* (cluster 3)
course	0	0.384	0
integral	0	0.384	0
equations	0	**1.976**	0
tractors	0	0.476	0
semi-groups	0	0.476	0
evolution	0	0.476	0
automatic	0.007	0	**0.979**
different	0.029	**0.952**	**1.100**
algorithms	**1.007**	0	**0.976**
theory	0	0.342	**1.203**
implementation	0.007	0	**0.979**
application	0.007	0	**0.979**
geometric	**1.044**	0.570	0
aspects	0.045	0.610	0
partial	0.045	0.610	0
ideals	**0.999**	0	0
varieties	**0.999**	0	0
introduction	**0.999**	0	0
computational	**0.999**	0	0
algebra	**1.999**	0	0
commutative	**0.999**	0	0
oscillation	0	0.506	0.224
neutral	0	0.506	0.224
delay	0	0.506	0.224
as	0.656	0.627	0.689

NMF: non-negative matrix factorization.

Table 4. Extracted cluster terms

Title	*r1*	*r2*	*r3*
Class label term	algorithms, geometric, ideals, varieties, introduction, computational algebra, commutative	equations, different	automatic, different, algorithms, theory, implementation, application

Class label terms may be restricted by the properties of a document cluster and the document composition. To resolve this problem, in this study, we use the term weight of documents by using the TMI on synonyms obtained from WordNet. Term weights of the document are calculated by using Jing's TMI as in Eq. (7) [11]. In the equation for Jing's TMI, δ_{il} indicates the semantic information between two terms. If term A_{lj} appears in the synonyms of A_{ij} obtained from WordNet, δ_{il} will be treated in the same level for different A_{ij} and A_{lj}, otherwise, δ_{il} will be set to zero.

$$\widetilde{A}_{iij} = A_{ij} + \sum_{\substack{l=1 \\ i \neq l}}^{m} \delta_{il} A_{lj} \,. \tag{7}$$

C. Document Clustering

This section explains document clustering using cosine similarity between the cluster terms and the term weights of the documents. The proposed method is described as follows: First, the cosine similarity between the cluster terms and the term weights is calculated using Eq. (8). Then, the document having the highest similarity value with respect to the class label is added to a document cluster [3, 15].

The cosine similarity function between the sentence vectors and the query is computed as follows [15]:

$$sim(A_{*a}, A_{*b}) = \frac{A_{*a} \cdot A_{*b}}{|A_{*a}| \times |A_{*b}|} = \frac{\sum_{i=1}^{m} A_{ia} \times A_{ib}}{\sqrt{\sum_{i=1}^{m} A_{ia}^2} \times \sqrt{\sum_{i=1}^{m} A_{ib}^2}}, \tag{8}$$

where A_{*a} and A_{*b} denote the ath document and the bth document, respectively. Further, m denotes the number of terms.

V. EXPERIMENTS AND EVALUATION

In this study, we use the dataset of 20 newsgroups for the performance evaluation [17]. To evaluate the proposed method, mixed documents were randomly chosen from the abovementioned dataset. A normalized mutual information metric related to Eqs. (9) and (10) was used for measuring the document clustering performance [2-4, 7-9]. The cluster numbers for the evaluation method were set in the range of 2 to 10, as shown in Fig. 1. For each given cluster number k, 50 experiments were performed on different randomly selected clusters, and the final performance values were the average of the values obtained from these experiments.

The normalized mutual information metric $\overline{MI}$ was used for measuring the document clustering performance [2-4, 7-9]. To measure the similarity between the two sets of document clusters $C = \{c_1, c_2, ..., c_k\}$ and $C' = \{c'_1, c'_2, ..., c'_k\}$, the following mutual information metric $MI(C,C')$ was used:

$$MI(C,C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)}, \tag{9}$$

where $p(c_i)$ and $p(c'_j)$ denote the probabilities that a document arbitrarily selected from the corpus belongs to c_i and c'_j, respectively, and $p(c_i, c'_j)$ denotes the joint probability that the selected document simultaneously belongs to c_i and c'_j. $MI(C, C')$ takes values between zero and $max(H(C), H(C'))$, where $H(C)$ and $H(C')$ are the entropies of C and C', respectively. The metric does not need to locate the corresponding counterpart in C', and the value is maintained for all permutations. A normalized metric MI, which takes values between zero and one, was used as shown in Eq.

(10) [2-4, 7-9]:

$$\overline{MI}(C,C') = \frac{MI(C,C')}{\max(H(C),H(C'))} \quad (10)$$

In this study, seven different document clustering methods were implemented, as shown in Fig. 1. The NMF, ASI, CLGR, RNMF, and FPCA methods are document clustering methods based on internal knowledge, and the FAWDN and TMINMF methods are clustering methods based on a combination of the internal and the external knowledge. WNMF denotes the proposed method described in this paper. FAWDN denotes the previously proposed method that is based on WordNet and fuzzy theory [14]. FPCA is the previously proposed method based on principal component analysis (PCA) and fuzzy relationships [6], and RNMF is the method proposed previously using NMF and cluster refinement [5]. NMF denotes Xu's method using NMF [4]. ASI is Li's method using adaptive subspace iteration [8]. Lastly, CLGR denotes Wang's method using local and global regularization [9].

As seen in Fig. 1, the average normalized metric of WNMF is 14.99% higher than that of NMF, 14.48% higher than that of ASI, 9.21% higher than that of CLGR, 6.66% higher than that of RNMF, 4.80% higher than that of FPCA, and 3.98% higher than that of FAWDN.

VI. CONCLUSION

In this paper, we proposed an enhanced text document clustering method using NMF and WordNet. The proposed method uses the semantic features of the document on the basis of the internal knowledge of NMF for extracting the cluster terms, which are well represented within the important inherent structure of the document cluster. In order to solve the limitation of the internal knowledge-based clustering methods with respect to the influence of the internal structure of documents, the proposed method uses TMI of WordNet to calculate the term weights of documents. Further, this method uses a similarity between the cluster terms and the term weights to improve the quality of the text document clustering. It was demonstrated that the value of the normalized mutual information metric is higher in the case of the proposed method than in the case of the other text document clustering methods for a dataset of 20 newsgroups.

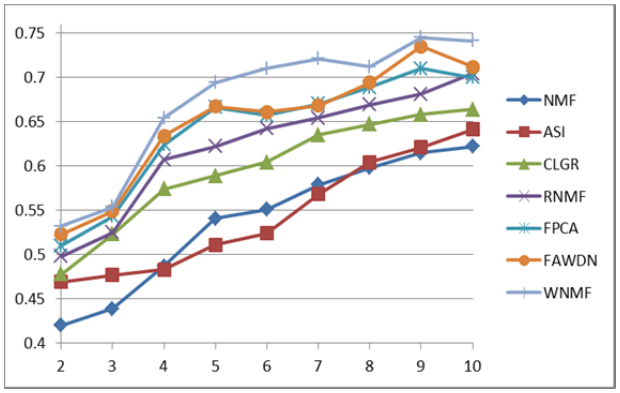

Fig. 1. Evaluation results of performance comparison. NMF: nonnegative matrix factorization, ASI: adaptive subspace iteration, CLGR: clustering with local and global regularization, RNMF: cluster refinement NMF, FPCA: fuzzy and principal component analysis, FAWDN: NMF based fuzzy and WordNet, WNMF: cluster based WordNet and NMF.

REFERENCES

[1] J. Hu, L. Fang, Y. Cao, H. J. Zeng, H. Li, Q. Yang, and Z. Chen, "Enhancing text clustering by leveraging Wikipedia semantics," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Singapore, pp. 179-186, 2008.

[2] S. Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data*. Boston, MA: Morgan Kaufmann Publishers, 2003.

[3] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology behind Search*, 2nd ed. New York, NY: Addison-Wesley, 2011.

[4] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in information Retrieval*, Toronto, Canada, pp. 267-273, 2003.

[5] S. Park, D. U. An, B. Cha, and C. W. Kim, "Document clustering with cluster refinement and non-negative matrix factorization," in *Proceedings of the 16th International Conference on Neural Information Processing*, Bangkok, Thailand, pp. 281-288, 2009.

[6] S. Park and K. J. Kim, "Document clustering using non-negative matrix factorization and fuzzy relationship," *Journal of Korea Navigation Institute*, vol. 14, no. 2, pp. 239-246, 2010.

[7] W. Xu and Y. Gong, "Document clustering by concept factorization," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, UK, pp. 202-209, 2004.

[8] T. Li, S. Ma, and M. Ogihara, "Document clustering via adap-tive subspace iteration," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, UK, pp. 218-225, 2004.

[9] F. Wang, C. Zhang, and T. Li, "Regularized clustering for documents," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, pp. 95-102, 2007.

[10] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-

negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788-791, 1999.

[11] L. Jing, L. Zhou, M. K. Ng, and J. Z. Huang, "Ontology-based distance measure for text clustering," in *Proceedings of 2006 SIAM International Conference on Data Mining*, Bethesda, MD, 2006.

[12] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou, "Exploiting Wikipedia as external knowledge for document clustering," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Paris, France, pp. 389-396, 2009.

[13] H. H. Tar and T. T. S. Nyaunt, "Ontology-based concept weighting for text documents," *World Academy of Science, Engineering and Technology*, vol. 57, pp. 249-253, 2011.

[14] S. Park and S. R. Lee, "Enhancing document clustering using condensing cluster terms and fuzzy association," *Journal of IEICE Transactions on Information and Systems*, vol. 94D, no. 6, pp. 1227-1234, 2011.

[15] W. B. Frakes and R. Baeza-Yates, *Information Retrieval: Data Structures & Algorithms*. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[16] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, 1995.

[17] The 20 newsgroups data set [Internet], Available: http://qwone.com/~jason/20Newsgroups/.

Chul-Won Kim

is a professor at Honam University, South Korea from 1998. He received his Ph.D. degree in Computer Engineering from Kwangwoon University, South Korea in 1997. His research interests include image processing, multimedia information retrieval, and multimedia processing.

Sun Park

is a research professor at Networked Computing System Lab., GIST, South Korea. He received his Ph.D. degree in Computer and Information Engineering from Inha University in 2007. Prior to becoming a researcher at GIST, he worked as a research professor at the Institute Research of Information Science and Engineering at Mokpo National University, a postdoctoral researcher at Chonbuk National University, and a professor in the Department of Computer Engineering at Honam University, Korea. His research interests include data mining, information retrieval, future internet, cloud systems, and IoT.