Chapter 28
Enhancing Document Clustering Using Reweighting Terms Based on Semantic Features

Sun Park, Jin Gwan Park, Min A Jeong, Jong Geun Jeong, Yeonwoo Lee and Seong Ro Lee

Abstract This paper proposes a new document clustering method using the reweighted term based on semantic features for enhancing document clustering. The proposed method uses document samples of cluster by user to reduce the semantic gap between the user's requirement and clustering results by machine. The method can enhance the document clustering because it uses the reweighted term which can well represent an inherent structure of document set relevant to a user's requirement. The experimental results demonstrate that the proposed method achieves better performance than related document clustering methods.

Keywords Document clustering · Reweighting term · Sematic feature · Non-negative matrix factorization (NMF)

S. Park (✉) · J. G. Park · M. A. Jeong · Y. Lee · S. R. Lee
Mokpo National Univeristy, Mokpo, South Korea
e-mail: sunpark@mokpo.ac.kr

J. G. Park
e-mail: chrispj@nate.com

M. A. Jeong
e-mail: majung@mokpo.ac.kr

Y. Lee
e-mail: ylee@mokpo.ac.kr

S. R. Lee
e-mail: srlee@mokpo.ac.kr

J. G. Jeong
National Research Foundation of Korea, Seoul, South Korea
e-mail: jkjeong@nrf.re.kr

H.-K. Jung et al. (eds.), *Future Information Communication Technology and Applications*,
Lecture Notes in Electrical Engineering 235, DOI: 10.1007/978-94-007-6516-0_28,

28.1 Introduction

Traditional document clustering methods are based on bag of words (BOW) model, which represents documents with features such as weighted term frequencies (i.e., vector model). However, these methods ignore semantic relationship between the terms within a document set. The clustering performance of the BOW model is dependent on a distance measure of document pairs. But the distance measure cannot reflect the real distance between two documents because the documents are composed of the high dimension terms with relation to the complicated document topics. In addition, the results of clustering documents are influenced by the properties of documents or the desired cluster forms by user [1]. Recently, to overcome the problems of the vector model-based document clustering, knowledge based approaches are applied.

Knowledge based approaches can be either internal knowledge based or external knowledge based document clustering. Internal knowledge-based document clustering uses the inherent structure of the document set by means of a factorization technique [1–11]. These methods have been studied intensively and although they have many advantages, the successful construction of a semantic features from the original document set remains limited regarding the organization of very different documents or the composition of similar documents [1, 12]. This limitation becomes the cause of semantic gap between user's requirement and results of document clustering. External knowledge-based document clustering exploits the constructed term ontology from external knowledge database with regard to ontology as WordNet and Wikipedia [1–3].

In order to enhance the internal knowledge-based approaches, this paper proposes a document clustering method that uses the reweighted terms by semantic features of NMF and the selected sample document of cluster by user. The proposed method has the following advantages: First, the selected document samples by user can reduce the semantic gap between the user's goal and the clustered document by machine. Second, the reweighted terms based on the selected document and the semantic features of document set well represents the document cluster. Finally, the clustering method using the reweighted terms can enhance the performance of document clustering.

28.2 Related Works

28.2.1 Document Clustering

The factorization techniques for internal knowledge-based document clustering including non-negative matrix factorization (NMF) [4], concept factorization (CF) [5], adaptive subspace iteration (ASI) [6], and clustering with local and global regularization (CLGR) [7] have been proposed. Xu et al. proposed a document partitioning method based on the Non-negative Matrix Factorization (NMF) of the given document corpus [4]. Xu and Gong proposed a data clustering method which models each cluster as a linear combination of the data points and each data point as a linear combination of the cluster centers by concept factorization (CF) [5]. Li et al. proposed a document clustering algorithm Adaptive Subspace Iteration (ASI) via explicitly modeling the subspace structure associated with each cluster [6]. Wang and Zhang proposed the document clustering with Local and Global Regularization (CLGR). This method uses a local label predictors and a global label smoothness regularizer [7]. Park et al. proposed a document clustering methods which use non-negative matrix factorization and cluster refinement, weighted semantic features and cluster similarity, Latent Semantic Analysis (LSA) and fuzzy association, and NMF and fuzzy relationship without synonyms [8–11].

28.2.2 Non-negative Matrix Factorization

This section reviews NMF theory with algorithm. In this paper, we define the matrix notation as follows: Let X_{*j} be j'th column vector of matrix X, X_{i*} be i'th row vector, and X_{ij} be the element of i'th row and j'th column. NMF is to decompose 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. (28.1) [12].

$$A \approx WH \tag{28.1}$$

where W is a $m \times r$ non-negative matrix and H is a $r \times n$ non-negative matrix. Usually r is chosen to be smaller than m or n, so that the total sizes of W and H are smaller than that of the original matrix A.

The objective function is used minimizing the Euclidean distance between each column of A and its' approximation $\tilde{A} = WH$, which was proposed by Lee and Seung [12]. As an objective function, the Frobenius norm is used:

$$\Theta_E(W,H) \equiv \|A - WH\|_F^2 \equiv \sum_{i=1}^{m} \sum_{j=1}^{n} \left(A_{ij} - \sum_{l=1}^{r} W_{il} H_{lj} \right)^2 \tag{28.2}$$

Updating W and H is kept until $\Theta_E(W,H)$ converges under the predefined threshold or exceeds the number of repetition. The update rules are as follows:

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{(W^T A)_{\alpha\mu}}{(W^T WH)_{\alpha\mu}}, \quad W_{i\alpha} \leftarrow W_{i\alpha} \frac{(AH^T)_{i\alpha}}{(WHH^T)_{i\alpha}} \tag{28.3}$$

28.3 The Proposed Method

This paper proposes a document clustering method using reweighting term based on semantic feature and estimation of term weighting. The proposed method consists of two phases: reweighting term and clustering document. In the subsection below, each phase is explained in full.

28.3.1 Reweighting Term

This section calculates term weighting with respect to the topics of the document clusters using semantic features of NMF and the selected sample documents relative to the cluster by user. The reweighted terms can well explain the topic of cluster that is derived from semantic feature of cluster structure by the estimation of term weighting.

The method of reweighting term is described as follows. First, let the number of cluster be set (it also can use to set the number of semantic feature r with connection to NMF), and then the sample documents regarding the clusters are selected by user. Second, preprocessing is performed (i.e., Rijsbergen's stop words list is used to remove all stop words, and word stemming is removed using Porter's stemming algorithm [13, 14]. Then, the term document frequency matrixes are constructed from the selected sample documents and document set.). Finally, the reweighting term g_a^{new} is calculated by using Eq. (28.4). However, we cannot directly calculate a new weight of a'th term. In order to solve this limitation, this paper proposes the Eq. (28.5), which it calculates the average weight of a'th row vector with regard to semantic features of document set by NMF a corresponding a'th term of the selected sample document .

$$g_a^{new} = g_a^{old} + \Delta g_a \tag{28.4}$$

where g_a^{new} is a new weight of a'th term, g_a^{old} is a weight of a'th term (i.e., initial value is 1.), Δg_a is variance in average weight of a'th row vector that is derived from Eq. (28.5).

$$\Delta g_a = E(\Delta g_a^i) = \frac{1}{n}\sum_{i=1}^{n} \Delta g_a^i = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{A_{ai}} \sum_{k \in I_i} \Delta H_{ki} W_{ak} \tag{28.5}$$

where $E()$ is variance, Δg_a^i is an average weight of a'th term and i'th document, n is the number of document in the document set, A_{ai} is a term frequency of a'th term and i'th document, I_i is term set k with respect to i'th variable column vector H_{*i} of document set corresponding $\Delta H_{ki} \neq 0$, ΔH_{ki} is variance in average of variable element of k'th term and i'th selected sample document.

Equation (28.6) is weight matrix G of reweighting term by using Eq. (28.4).

$$G = \begin{pmatrix} g_1 & 0 & \cdots & 0 \\ 0 & g_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & g_n \end{pmatrix} \tag{28.6}$$

where, if the a'th element of weight matrix G exists g_a^{new} a corresponding a'th term of A_{a*}, the element values is g_a^{new} otherwise 1.

28.3.2 Clustering Document

This section presents the clustering document using k means clustering method and reweighting terms of document set. The reweighting terms are calculated by using Eq. (28.7).

$$\tilde{A} = GA \tag{28.7}$$

where $\tilde{A}$ is reweighting term document frequency matrix, G is weight matrix, A is term document frequency matrix with relation to document set.

The k mean algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high but inter-cluster similarity is low [13, 14]. In this paper, we use cosine similarity for cluster distance measure with association to k means as Eq. (28.8)

$$dist(\tilde{A}_{*_a}, \tilde{A}_{*_b}) = 1 - csim(\tilde{A}_{*_a}, \tilde{A}_{*_b}) \tag{28.8}$$

$$csim(\tilde{A}_{*_a}, \tilde{A}_{*_b}) = \frac{\sum_{i=1}^{m} \tilde{A}_{ia} \times \tilde{A}_{ib}}{\sqrt{\sum_{i=1}^{m} \tilde{A}_{ia}^2} \times \sqrt{\sum_{i=1}^{m} \tilde{A}_{ib}^2}} \tag{28.9}$$

where $\tilde{A}_{*_a}$ and $\tilde{A}_{*_b}$ are a'th and b'th column vectors of reweighting term document frequency matrix $\tilde{A}$, respectively. These vectors have non-negative values so that are $0 \leq csim() \leq$ and $0 \leq dist() \leq 0$.

28.4 Experiments and Evaluation

This paper uses 20 Newsgroups data set for performance evaluation [15]. To evaluate the proposed method, mixed documents were randomly chosen from the 20 Newsgroups documents. Normalized mutual information metric used to measure the document clustering performance [1–11].

Normalized mutual information metric $\overline{MI}$ as used to measure the document clustering performance [1–11]. To measure the similarity between the two sets of document clusters $C = \{c_1, c_2, \ldots, c_K\}$ and $C' = \{c'_1, c'_2, \ldots, c'_K\}$, the following mutual information metric $MI(\mathrm{C},\mathrm{C}')$ was used:

$$MI(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)} \tag{28.10}$$

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 ci as well as 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. The normalized metric, $\overline{MI}$, which takes values between zero and one, was used as shown in Eq. (28.11) [1–11]:

$$\overline{MI}(C, C') = \frac{MI(C, C')}{\max(H(C), H(C'))}. \tag{28.11}$$

The cluster numbers for the evaluation method are set by ranging from 2 to 10. For each given cluster number K, 50 experiments were performed on different randomly chosen clusters, and the final performance values averaged the values obtained from running experiments.

In this paper, the eight different document clustering methods are implemented as Fig. 28.1. The RT, KM, NMF, CF, ASI, CLGR, FPCA, and RNMF methods are document clustering methods based on internal knowledge. The TR denotes the proposed method described within this paper. The KM is a document clustering using k means method based on a traditional partitioning clustering technique [13, 14]. NMF denotes Xu's method using non-negative matrix factorization [4].

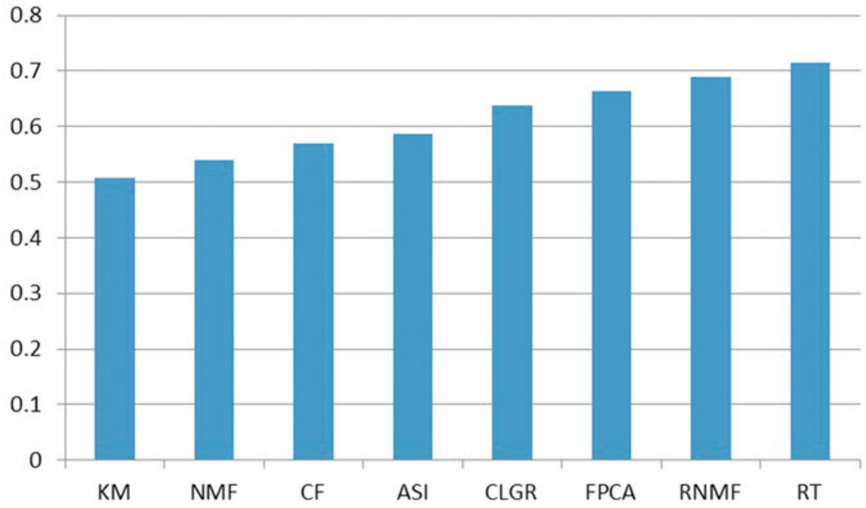

Fig. 28.1 The evaluation results with respect to the average normalized mutual information of clustering methods

CF denotes Xu and Gong's method using concept factorization which models each cluster as a linear combination of the data points and each data point as a linear combination of the cluster centers [5]. ASI is Li's method using adaptive subspace iteration [6]. Lastly, CLGR denotes Wang's method using local and global regularization [7]. FPCA is the previously proposed method using principal component analysis (PCA) and fuzzy relationship [10], and RNMF is the method proposed previously using NMF and cluster refinement [11].

As seen in Fig. 28.1, the average normalized metric of RT is 20.8 % higher than that of KM, 17.58 % higher than that of NMF, 14.48 % higher than that of CF, 12.88 % higher than that of ASI, 7.74 % higher than that of CLGR, 5.06 % higher than that of FPCA, and 2.44 % higher than that of RNMF.

28.5 Conclusion

This paper presents a document clustering method using the reweighted term based on semantic features for enhancing document clustering. The proposed method uses document samples of cluster by user to reduce the semantic gap between the user's requirement and clustering results by machine. The method can enhance the document clustering because it uses the reweighted term which can well represent an inherent structure of document set relevant to a user's requirement. It was demonstrated that the normalized mutual information is higher than the internal knowledge based clustering methods for 20 Newsgroups data set using the proposed method.

Acknowledgments This work was supported by Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2009-0093828). "This research was supported by the The Ministry of Knowledge Economy (MKE), Korea, under the Information Technology Research Center (ITRC) support program supervised by the National IT Industry Promotion Agency (NIPA)" (NIPA-2012-H0301-12-2005).

References

1. Hu X, Zhang X, Lu C, Park EK, Zhou X (2009) Exploiting wikipedia as external knowledge for document clustering. In: Proceeding of the 15th ACM SIGKDD conference on knowledge discovery and data mining (KDD'09). Paris, France, pp 389–396
2. Hu T, Xiong H, Zhou WS, Sung Y, Luo H (2008) Hypergraph partitioning for document clustering: a unified clique perspective. In: Proceeding of the ACM SIGIR conference on research and development in information retrieval (SIGIR'08). Singapore, pp 871–872
3. Park S, Kim KJ (2010) Document clustering using non-negative matrix factorization and fuzzy relationship. J Korea Navig Inst 14(2):239–246
4. Xu W, Liu X, Gong Y (2003) Document clustering based on non-negative matrix factorization. In: Proceeding of the ACM SIGIR conference on research and development in information retrieval (SIGIR'03). Toronto, Canada

5. Xu W, Gong Y (2004) Document clustering by concept factorization. In: Proceeding of the ACM SIGIR conference on research and development in information retrieval (SIGIR'04). UK, pp 202–209
6. Li T, Ma S, Ogihara M (2004) Document clustering via adaptive subspace iteration. In: Proceeding of the ACM SIGIR conference on research and development in information retrieval (SIGIR'04). UK, pp 218–225
7. Wang F, Zhang C (2007) Regularized clustering for documents. In: Proceeding of the ACM SIGIR conference on research and development in information retrieval (SIGIR'07). Amsterdam, pp 95–102
8. Park S, An DU, Cha BR, Kim CW (2009) Document clustering with cluster refinement and non-negative matrix factorization. In: Proceeding of the 16th international conference on neural information processing (ICONIP'09). Bangkok, Thailand
9. Park S, An DU, Choi IC (2010) Document clustering using weighted semantic features and cluster similarity. In: Proceeding of the 3rd IEEE international conference on digital game and intelligent toy enhanced learning (DIGITEL'10). Kaohsiung, Taiwan
10. Park S, An DU, Cha BR, Kim CW (2010) Document clustering with semantic feature and fuzzy association. In: Proceeding of the international conference on information systems, technology and management (ICISTM'10). Bangkok, Thailand
11. Park S, Kim KJ (2010) Document Clustering using non-negative matrix factorization and fuzzy relationship. J Korea Navig Inst 14(2):239–246
12. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401:788–791
13. Frankes WB, Ricardo BY (1992) Information retrieval, data structure & algorithms. Prentice-Hall, Englewood Cliffs
14. Ricardo BY, Berthier RN (1999) Moden information retrieval. ACM Press, New York
15. The 20 newsgroups data set (2012). http://people.csail.mit.edu/jrennie/20Newsgroups/