PAPER

Enhancing Document Clustering Using Condensing Cluster Terms and Fuzzy Association

Sun PARK[†a)], ***Member and*** **Seong Ro LEE**[††], ***Nonmember***

SUMMARY Most document clustering methods are a challenging issue for improving clustering performance. Document clustering based on semantic features is highly efficient. However, the method sometimes did not successfully cluster some documents, such as highly articulated documents. In order to improve the clustering success of complex documents using semantic features, this paper proposes a document clustering method that uses terms of the condensing document clusters and fuzzy association to efficiently cluster specific documents into meaningful topics based on the document set. The proposed method improves the quality of document clustering because it can extract documents from the perspective of the terms of the cluster topics using semantic features and synonyms, which can also better represent the inherent structure of the document in connection with the document cluster topics. The experimental results demonstrate that the proposed method can achieve better document clustering performance than other methods.

key words: *document clustering, non-negative matrix factorization (NMF), semantic features condensing cluster term, fuzzy association, synonyms, WordNet*

1. Introduction

There has been much research on information retrieval, information filtering, automatic summarization, and topic extraction in document clustering. As a effective technique, document clustering has received greater attention through the rapid growth of large sources of textual data, such as online news, blogs, Twitter, Facebook, emails, and messages boards [1], [6], [7], [17]–[19]. 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 the semantic relationship between the terms within a document set. The clustering performance of the BOW model is dependent on the distance measure of document pairs. However, the distance measure cannot reflect the real distance between two documents [6], [7], [18].

To resolve the shortcomings of the BOW model, ontological and semantic feature methods are usually employed. Ontological methods exploit external resources (or knowledge) that use the term ontology from WordNet and Wikipedia to improve the BOW term representation. However, it is often difficult to locate a comprehensive ontology that covers all concepts mentioned in the collection [6], [7], [18]. The semantic feature method uses the internal structure (or knowledge) of the document set, which can accurately identify the document set topics from their semantic features based on a factorization technique.

Recently, other techniques for document clustering including non-negative matrix factorization (NMF) [21], concept factorization [20], adaptive subspace iteration (ASI) [11], clustering with local and global regularization (CLGR) [19], and latent semantic analysis (LSA) [15] have been proposed, which are collectively referred to as factorization techniques. 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 [8], [10], [11], [13]–[16], [19]–[21].

To solve the internal structure restrictions in previous works [13]–[16], four document clustering methods have been proposed that use NMF with cluster refinement [13], weighted semantic features with cluster similarity [14], fuzzy association based on latent semantic analysis (LSA) [15], and fuzzy relationships depending on semantic features [16]. However, the results of these methods are influenced by the structure of the original document set [8], [20], [21].

In this paper, the focus is placed on internal knowledge methods. Internal knowledge methods use semantic features through representations of the inherent structure of document set to be derived from factorization methods. In order to resolve the limitations of the semantic feature methods, this paper proposes a document clustering method that uses the fuzzy association between the terms in the document set and the summarization of the document cluster.

In the proposed method, important terms for describing document cluster are extracted first using the semantic features of non-negative matrix factorization (NMF), which represents the inherent structure of the document set topic. The semantic feature can easily identify an appropriate document from the document set for the cluster topics. Second, to successfully cluster documents from the semantic features depending on the composition of the document set complexity, the extracted terms were expanded using synonyms from WordNet for suitable cluster topics. The expanded terms in the cluster topics can discriminate between documents that have a high similarity and ones that are irrelevant to the topic. It also resolves the cost problem surrounding the construction of ontology because the terms of

Manuscript received January 24, 2011.

[†]The author is corresponding author with Mokpo National University, R.411, B24, 61 Dorim-ri, Cheonggye-myeon, Jeonnam, Korea.

[††]The author is with Mokpo National University, Korea.

a) E-mail: sunpark@mokpo.ac.kr

DOI: 10.1587/transinf.E94.D.1227

the summarizing cluster are only expanded through the synonyms. Finally, the clustering results can be enhanced by exploiting the fuzzy association based on the terms of the condensing clusters and the terms of the documents. Fuzzy association assists in easily capturing similar documents in connection with the terms of representing cluster topics.

In the present study, the previous works [13]–[16] were modified because while they have advantages in clear identification of the cluster topic of compared with the internal knowledge methods [11], [19]–[21], they are restricted within the structure of original document set since the methods only use internal knowledge from the documents. Thus, the proposed method combines the advantages of the external and internal knowledge methods using the terms of the condensed document cluster using semantic features and synonyms from WordNet.

The remainder of the paper is organized as follows. Section 2 describes the related works regarding document clustering methods, and Sect. 3 reviews the NMF and fuzzy association methods. Section 4 presents the proposed document clustering method, while Sect. 5 describes the proposed document clustering algorithm. Then, Sect. 6 shows the evaluation and experimental results, and Sect. 7 concludes the paper.

2. Related Works

Generally, clustering methods fall into three types: partitioning, hierarchical, and density-based clustering. Partitioning clustering directly clusters the document set into k disjointed cluster labels where the documents in one cluster label are more closely related than documents in another cluster label. Hierarchical clustering successively groups documents that are close to one another, until all groups are merged into one through building cluster trees. Density-based clustering gathers the neighborhood documents of one document set in a cluster label using density conditions. However, most of these methods use distance functions as object criteria based on the BOW model and are not effective in high dimensional spaces in relation to document clustering [1]–[3], [6], [7], [11], [17].

Recently, a knowledge-based document clustering method, which is used to increase the efficiency of document clustering, has been proposed; this method can be divided into external knowledge and internal knowledge. The external knowledge uses external resources that constructs the knowledge ontology from external sources such as WordNet [5], Mesh [23], Wikipedia [6], [7], etc. However, these methods have a high cost due to the organization of the ontology with the appropriate information connected with the topics from the ontology coverage limitation and due to information loss in the document set [7]. Internal knowledge methods use semantic features by representing the inherent structure of document set that will be derived using factorization methods. However, the results of these methods are influenced by the structure of the original document set [8], [20], [21].

In the internal knowledge methods approach, Li et al. [11] proposed 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. [21] proposed the document partitioning method based on non-negative matrix factorization (NMF) in given document corpuses. The results from the addressed methods have a stronger semantic interpretation than 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 [20]. To resolve the limitation of the NMF method, Xu and Gong [20] modeled each concept as a linear combination of data points and cluster centres called concept factorization. Li and Ding [10] presented an overview and summary of various matrix factorization algorithms for clustering, and analyzed their relationships theoretically. To overcome the problems of the partitioning methods [19], Wang and Zhang used clustering with local and global regularization (CLGR), which uses local label predictors and global label smoothness regularizers. They achieved satisfactory results because the CLGR algorithm uses fixed neighborhood sizes. However, the different neighborhood sizes deteriorate the final clustering results [19].

3. NMF and Fuzzy Association

3.1 Non-Negative Matrix Factorization

This paper defines the matrix notation as follows. Let the j'th column vector and the i'th row of matrix X be X_{*j} and X_{i*}, respectively. Thus, assume that X_{ij} shows the element of the i'th row and the j'th column in the same matrix X.

The NMF can represent an individual object as the non-negative linear combination of the section of information extracted from a large volume of objects [8], [9], [20], [21]; furthermore, it can easily extract the semantic features representing the inherent structure of data objects [8]. NMF algorithm is summarized as follows. Let the NMF decompose a $m \times n$ matrix X into a non-negative matrix W and a non-negative matrix H. The W and H matrices, having semantic features related to the inherent structure of the original matrix X, can be expressed as follows:

$$X \approx WH \tag{1}$$

where W and H are $m \times r$ and $r \times n$ non-negative matrices, respectively, and r is the number of semnatic features. Usually, r is chosen to be smaller than m or n, so that the size of W and H are smaller than that of the original matrix X. To distinguish the semantic feature matrices, matrices W and H are have been called the semantic feature matirx W and the semantic variable matrix H by Lee and Seung [8], [9].

To factorize the original matrix, the NMF uses an objective function that minimizes the Euclidean distance between two non-negative matrices, and then updates the rules. As an objective function, the Frobenius norm is used as follows [8], [9]:

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

Then, W and the H are updated until $\Theta_E(W,H)$ converges under the predefined tolerance. The update rules are as follows:

$$H_{\alpha\beta} \leftarrow H_{\alpha\beta}\frac{(W^TX)_{\alpha\beta}}{(W^TWH)_{\alpha\beta}},\, W_{i\alpha} \leftarrow W_{i\alpha}\frac{(XH^T)_{i\alpha}}{(WHH^T)_{i\alpha}} \quad (3)$$

An example of matrices W and H is illustrated and demonstrates how they are composed of semantic features from the original matrix X in Example 1. Example 1 uses Eq. (2) and (3) to exemplify the NMF algorithm result as follows.

Example 1. Let r be 3, the number of repetitions be 50, and the tolerance be 0.001. When the initial elements of the W and H matrices are 0.5, matrix X is decomposed into the W and H matrices, as shown in Fig. 1.

A column vector corresponding to the j'th document X_{*j}, can be represented as a linear combination of the semantic feature vectors W_{*l} and the semantic variable H_{lj}, as follows:

$$X_{*j} = \sum_{l=1}^{r} H_{lj}W_{*l} \quad (4)$$

Figure 2 shows an example of the document representation in relation to the semantic features as shown in Eq. (4). The column vector X_{*3} in Fig. 1, corresponding to the third document, is represented as a linear combination of the semantic feature vectors W_{*l} and an element of the semantic variable vector H_{*3}.

The advantages of the NMF are that all semantic variables (H_{lj}) are used to represent each document. W and H are sparsely represented. Intuitively, the sparse property of semantic features indicates that it is sensible for each document to be associated with a small subset of a large array of topics (W_{*l}), rather than just one topic or all topics. This means that some semantic features in documents cover the cluster. Thus, semantic features can easily identify the topic of a particular document cluster. For each semantic feature (W_{*l}), NMF groups the semantically related terms together [8], [20], [21].

$$\underset{X}{\begin{bmatrix}0&3&2&0\\0&0&1&0\\4&0&0&2\\0&0&1&0\end{bmatrix}} \approx \underset{W}{\begin{bmatrix}0&1.752&0.089\\0&0.726&1.423\\1.730&0&0.002\\0.865&0&0.001\end{bmatrix}} \times \underset{H}{\begin{bmatrix}0&0&2.312&0\\1.618&0.004&0&0.004\\0.549&0.698&0.003&0.698\end{bmatrix}}$$

Fig. 1 Result of the NMF algorithm.

$$\underset{X_{*3}}{\begin{bmatrix}2\\1\\0\\1\end{bmatrix}} \approx \underset{H_{13}}{2.312}\times\underset{W_1}{\begin{bmatrix}0\\0\\1.730\\0.865\end{bmatrix}} + \underset{H_{23}}{0}\times\underset{W_2}{\begin{bmatrix}1.752\\0.726\\0\\0\end{bmatrix}} + \underset{H_{33}}{0.003}\times\underset{W_3}{\begin{bmatrix}0.089\\1.423\\0.002\\0.001\end{bmatrix}}$$

Fig. 2 Example of document representation using semantic feature vectors and semantic variable vectors.

3.2 Fuzzy Association

Fuzzy association applied in information retrieval is used in the proposed method for the clustering document. Fuzzy association [4], which constructs the index terms from a document set, uses the fuzzy set theory [18], [22] to model the vagueness within the information retrieval. Fuzzy association in document clustering is formalized within the fuzzy set theory and based on the definition of fuzzy association. It uses the association between the terms to improve the clustering results in the document set. The construction of fuzzy association between the terms is defined as follows [4], [22].

Definition 1. A fuzzy association between two finite sets $X = \{x_1, \dots, x_u\}$ and $Y = \{y_1, \dots, y_v\}$ is formally defined as a binary fuzzy association $f : X \times Y \rightarrow [0, 1]$, where u and v represent the numbers of elements in X and Y, respectively.

Definition 2. Given a set of index terms, $T = \{t_1, \dots, t_k\}$, and a set of documents, $D = \{d_1, \dots, d_n\}$, each t_i is described by a fuzzy set $h(t_i)$ of documents; $h(t_i) = \{F(t_i, d_j) \mid \forall d_j \in D\}$, where $F(t_i, d_j)$ is the significance, or membership, degree of t_i, in d_j.

Definition 3. The fuzzy related terms (RT) association is based on the evaluation of the co-occurrences of t_i and t_j in the set D and can be defined as follows:

$$RT(t_i, t_j) = \frac{\sum_k \min(F(t_i, d_k), F(t_j, d_k))}{\sum_k \max(F(t_i, d_k), F(t_j, d_k))} \quad (5)$$

4. Proposed Document Clustering Method

This study proposes a document clustering method using fuzzy association dependent on the terms of a summarizing document cluster using NMF and WordNet. The proposed method consists of three phases: preprocessing, extracting terms, and clustering document, as shown in Fig. 3. In the subsections below, each phase is explained in full.

4.1 Preprocessing

In the preprocessing phase, van Rijsbergen's stop words list is used to remove all stop words, and word stemming is removed using Porter's stemming algorithm [2], [17]. Then, the term document matrix A is constructed from the document set [1], [2], [17]. Let A be $m\times n$ terms by the documents matrix, where m is the number of terms and n is the number of documents in the document set.

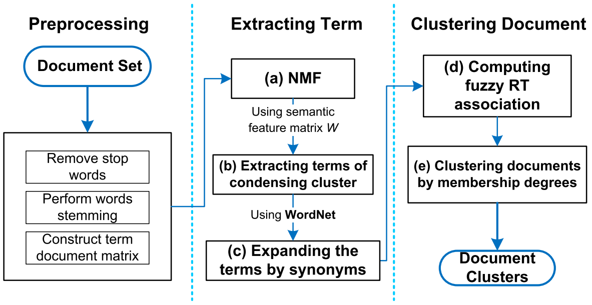
Preprocessing
Extracting Term
Clustering Document
Document Set
Remove stop words
Perform words stemming
Construct term document matrix
(a) NMF
Using semantic feature matrix W
(b) Extracting terms of condensing cluster
Using WordNet
(c) Expanding the terms by synonyms
(d) Computing fuzzy RT association
(e) Clustering documents by membership degrees
Document Clusters

Fig. 3 Document clustering method using terms of condensing cluster and fuzzy association.

4.2 Extracting Terms

This section extracts the terms that can well represent the document cluster properties using the NMF and synonyms from WordNet. The extracting term phase consists of the extracting the terms of the condensing cluster and expanding the terms using synonyms. The extracted terms are used in the clustering document phase for the clustering document with respect to a topic in the document cluster.

4.2.1 Extracting Terms of the Condensing Cluster

In this section, the terms of the condensing cluster that can well summarize the document cluster topic are extracted using the semantic features of the NMF. The extracting terms of the condensing cluster method in Fig. 3 (b) is described as follows. Let r be the number of clusters, and then the preprocessing phase is performed. Next, NMF is performed on the document set A to obtain the two non-negative matrices W and H. Matrix W is used to extract the terms of condensing cluster. The term A_{ic} is assigned to the condensing cluster term CT^p if $p = \arg\max_{1\le c\le r}\{W_{ic}\}$ and $W_{ic} \geq av_{sfv}$. Here, CT is the term sets of the condensing cluster, $CT = \{CT^1, CT^2, \ldots, CT^r\}$, from each cluster r with respect to the extracting and expanding terms. The average semantic feature value, av_{sfv}, is as follows:

$$av_{sfv} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{r} W_{ij}}{n \times r} \tag{6}$$

The semantic feature matrix W represents the inherent structure of the document set in relation to its terms. The semantic feature values also indicate how much the term reflects the cluster topics. In this paper, the average semantic feature value is used to extract the cluster terms because the terms that correspond to a very small value of the semantic feature are meaninglessness in relation to the cluster topics.

An example of the extracting terms of the condensing cluster are illustrated in Example 2 for exemplification of the proposed method. Example 2 using the NMF and Eq. (6) exemplifies the results of the extracting terms.

Table 1 Term document matrix.

Term \ Document	*d1*	*d2*	*d3*	*d4*	*d5*	*d6*	*d7*
t1	3	4	0	0	0	0	0
t2	1	2	0	0	0	4	3
t3	3	1	0	0	2	5	4
t4	0	0	4	2	1	2	2
t5	0	0	1	1	1	1	1
t6	1	0	3	2	2	0	0
t7	0	0	1	1	1	0	0

Table 2 Semantic features matrix W by NMF from Table 1.

	r1	*r2*	*r3*
t1	0	0	5.0010
t2	4.8976	0	1.7653
t3	6.7613	0.3297	2.0926
t4	2.5637	4.4180	0
t5	1.4070	1.5653	0
t6	0.0441	4.0667	0.5706
t7	0.0142	1.6287	0

Table 3 Result of extracting terms of condensing cluster from Table 2.

	Terms of condensing cluster
$CT^1(r1)$	*t2, t3*
$CT^2(r2)$	*t4, t6*
$CT^3(r3)$	*t1*

Example 2. Table 1 shows the term document matrix with respect to seven documents and seven terms. Table 2 shows the semantic features matrix W obtained through the NMF from Table 1. In Table 2, r is the number of clusters corresponding to the number of semantic feature column vectors, and t is the term that corresponds the semantic feature values in the row vector. The terms of the condensing cluster with the top rank semantic feature values are extracted in each row in Table 2. The terms of the condensing cluster are selected over the average semantic feature value (i.e. $av_{sfv} = 1.7679$). Table 3 shows the results of the extracted terms of the condensing cluster from Table 2 using the proposed method. In Table 3, the terms with the top rank semantic value are able to well cover the cluster topic, therefore the term of cluster t demonstrates the relatability of the cluster topic r.

4.2.2 Expanding the Terms Using Synonyms

Document sets belong to a particular topic and can be related to several topics. The topics overlap among the related topics and are not completely independent of each other due to restrictions within the overlap properties in the cluster topics. The condensing cluster terms may restrict the cluster documents using the topic properties and document composition. To overcome the limitations of condensing the cluster terms, they are expanded using synonyms from WordNet. The concept behind this approach attempts to expand terms by exploring the document set for more relevant the

Table 4 Result of synonyms ordered using the estimated noun frequency from "term" by WordNet v2.1.

Order	Sense of term (expanded terms of "term")	Explanation of sense
1	Word	a unit of language that native speakers can identify
2	Time period, period	an amount of time
3	Statement	a message that is stated or declared
4	Quantity	something that has a magnitude and can be represented in mathematical expressions by a constant or a variable
5	Constituent, grammatical constituent	a word or phrase or clause forming part of a larger grammatical construction
6	Point, point in time	an instant of time; "at that point I had to leave
7	State	a sculpture representing a human or animal

properties of the cluster topics. WordNet, which is created and maintained by Princeton University, is a lexical database based on psycholinguistic principles for the English language. English words in WordNet are encoded into concepts in terms of sets of synonyms, called synsets, which provide various semantic relationships between the synonym sets [12].

The expanding terms method in Fig. 3 (c) is described as follows. The extracted term is expanded through the synonyms from WordNet as a basic function with respect to the rank of the estimated noun frequency. Then, the synonyms of the terms from the condensing cluster matrix CT are constructed using the expanding terms for each cluster. In this paper, only synonyms of nouns are used for the expanding terms since many terms are expanded through the verb synonyms over a range of cluster topics.

Example 3. An example of the expanding terms is illustrated using WordNet release 2.1. Table 4 shows the synonyms ordered by the estimated noun frequency of the "term".

4.3 Clustering Document by Fuzzy Association

This section presents the clustering of documents using fuzzy association and the condensing cluster terms. The document clustering results can be improved because fuzzy association assists clustering to identify a highly similar documents with respect to the terms of the summarizing cluster. The proposed method is described in Figs. 3 (d) and 3 (e).

In Fig. 3 (d), the term correlation matrix M is constructed using the fuzzy RT association presented in Eq. (7). It uses the relationship between the expanded terms of the condensing cluster and the terms of the document set. Table 5 shows an example of the term correlation matrix using Eq. (7) from Table 1. A simplification of the fuzzy RT association [4] of Eq. (5) based on the co-occurrence of terms is given as follows:

$$fa_{i,j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}} \tag{7}$$

where $fa_{i,j}$ represents the fuzzy RT association between the terms i and j, $n_{i,j}$ is the number of documents containing both i'th and j'th terms, n_i is the number of documents including the i'th term, and n_j is the number of documents including the j'th term.

Table 5 Term correlation matrix from Table 1 by the fuzzy RT association.

Term	*t1*	*t2*	*t3*	*t4*	*t5*	*t6*	*t7*
t1	1	0.5	0.4	0	0	0.2	0
t2	0.5	1	0.8	0.2857	0.2857	0.1428	0
t3	0.4	0.8	1	0.4286	0.4286	0.2857	0.1428
t4	0	0.2857	0.4286	1	1	0.5	0.6
t5	0	0.2857	0.4286	1	1	0.5	0.6
t6	0.2	0.1428	0.2857	0.5	0.5	1	0.75
t7	0	0	0.1429	0.6	0.6	0.75	1

Table 6 Results of document clustering from Table 5.

	Documents
C^1	*d1, d2, d6, d7*
C^2	*d3, d4, d5*

In Table 5, the term correlation matrix is a 7 × 7 symmetric matrix, whose element, m_{ij}, has a value on the interval $[0, 1]$, with 0 indicating no relationship and 1 indicating a full relationship between terms t_i and t_j. Therefore, m_{ij} is equal to 1 for all $i = j$ since a term has the strongest relationship with itself [4].

In Fig. 3 (e), the membership degree $\mu_{i,j}$ is calculated using Eq. (8). Then, a document d_i is clustered into cluster C^j where the membership degree $\mu_{i,j}$ is the maximum. Here, fuzzy association is used to capture the relationships between different terms within documents; each pair of terms has an associated value in order to distinguish itself from a document cluster. This, ambiguity can be avoided in term usage for effective document clustering. Table 6 shows the results of document clustering using Eq. (8) from Tables 3 and 5.

The membership degrees [4] between each document in each cluster set are defined as follows:

$$\mu_{i,j} = \sum_{\forall t_a \in d_i} \left[1 - \prod_{\forall t_b \in CT^j} (1 - fa_{a,b}) \right], \tag{8}$$

where $\mu_{i,j}$ is the membership degree of d_i belonging to C^j, and $fa_{a,b}$ is the fuzzy association between term $t_a \in d_i$ and term $t_b \in CT^j$. CT is a set of condensing cluster terms.

In Table 6, the terms of condensing cluster from Table 3 are used without the expanded terms for a simple explanation in relation to the clustering document process.

5. Proposed Document Clustering Algorithm

In Sect. 4, the proposed method that can extract the condensing cluster terms and the documents clustered for document clustering were explained. Therefore, the following document clustering algorithm is proposed in connection with Sect. 4.

```
Algorithm: ClusterDocument(A, r, m, n)
Input: The term document matrix A, the number of clusters r,
       the number of terms m, the number of documents n.
Output: The average semantic feature value $av_{sfv}$,
        the set of cluster terms CT, the set of clustering document C.
1: Perform the preprocessing of matrix A;
2: Perform the NMF and calculate $av_{sfv}$;
3: for i←1 to m do
4:     if $p = \arg\max_{1 \le c \le r} \{W_{ic}\}$ and $W_{ic} \ge av_{sfv}$ then $CT^p$←i'th term;
5:     end if
6: end
7: Expand the cluster label terms by synonyms into CT;
8: Calculate the fuzzy RT association and the membership degree;
9: for i←1 to n do
10:    if $q = \arg\max_{1 \le e \le r} \{\mu_{i,e}\}$ then $C^q$←$d_i$;
11:    end if
12: end
```

In lines 2 to 6, the extracting terms of the condensing cluster phase use the semantic features according to the NMF. In line 7, the expanding terms phase uses the synonyms from WordNet. In lines 8 to 12, the clustering document phase uses fuzzy association between the expanded terms and the terms of document set.

6. Performance Evaluation

To evaluate the proposed method, the Reuters[†] document corpora composed of 21,578 documents, which are grouped into 135 clusters, was used. Documents in Reuters maintain multiple cluster labels with documents in each cluster having a broader variety of content [19]. Mixed documents were randomly chosen from multiple clusters of the Reuters documents and selected k documents among the remainder. The k documents were applied to the proposed clustering process. The result is evaluated by comparing the similarity between the clusters using clustering methods and those of Reuters.

A normalized mutual information metric $\overline{MI}$ was used to measure the document clustering performance [11], [15], [19]–[21]. 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 $MI(C, C')$ 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 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. (10) [11], [15], [19]–[21]:

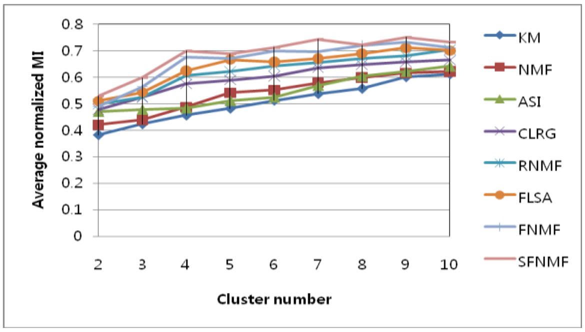

Fig. 4 Evaluation results of performance comparison.

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

The evaluation was conducted by comparing it with seven document clustering methods using the same data corpora, and then implemented eight different document clustering methods: SFNMF, FNMF, FLSA, RNMF, KM, NMF, ASI, and CLRG. The KM is a general clustering method using k-means based on a traditional partitioning clustering technique for document clustering [1], [3], [17]. The SFNMF, FNMF, FLSA, RNMF, NMF, ASI, and CLRG methods are internal knowledge methods based on factorization techniques for document clustering. SFNMF denotes the proposed method described within this paper; FNMF denotes the previously proposed method using the NMF and fuzzy relationship [16]. FLSA is the previously proposed method using LSA and fuzzy association [15], and RNMF is the method proposed previously using NMF and cluster refinement [13]. NMF denotes Xu's method using non-negative matrix factorization [21]. ASI is Li's method using adaptive subspace iteration [11]. Lastly, CLRG denotes Wang's method using local and global regularization [19].

The evaluation study was conducted for the cluster numbers ranging from 2 to 10, as shown in Fig. 4. For each given cluster number k, 50 experiments were performed on different randomly chosen clusters, and the final performance values were averaged the values obtained from running experiments. As seen in Fig. 4, the average normalized metric $\overline{MI}$ of SFNMF is 26.08% higher than that of KM, 21.39% higher than that of NMF, 20.64% higher than that of ASI, 12.96% higher than that of CLRG, 9.24% higher than that of RNMF, 6.53% higher than that of FLSA, and 3.52% higher than that of FNMF.

To better understand the reason why the proposed method is more effective than the general clustering method [1], [3], [17] and internal knowledge methods [16], [21], the influence of external knowledge with synonyms is analyzed from the clustering methods in Fig. 5, which shows

[†] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

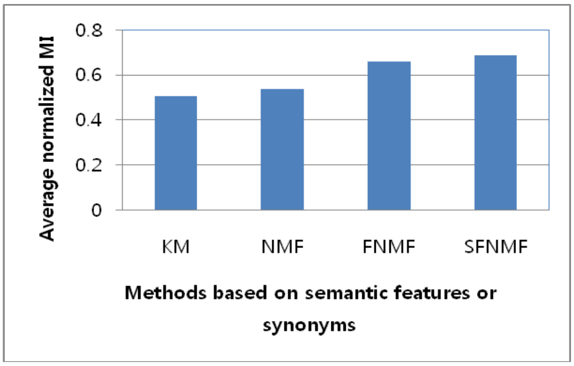

Fig. 5 Evaluation results from the comparison of semantic features or synonyms.

the comparison results of the average normalized metric $\overline{MI}$ with respect to the four clustering methods. Here, the KM uses only the distance measure between documents for clustering documents based on the BOW model. The NMF uses only the semantic features that reflect the relationship between the documents and cluster topics; however, it does not reflect the concern between the document features and cluster topics. The FNMF uses the fuzzy relationship between the documents and semantic features representing the inherent structure of the cluster topics for clustering documents. The SFNMF uses fuzzy association between the document set terms and condensing cluster topic terms for the clustering document.

In Fig. 5, the comparison results of the FNMF and SFNMF are stronger than those of the KM and NMF. The FNMF sufficiently represents the hidden topics of the clusters using fuzzy relationships based on semantic features; however, the vector distance of KM and semantic features of NMF are not sufficient to reflect the latent topics of the documents to the cluster. The SFNMF showed the best performance, because it uses the terms of the summarizing cluster topics that rely on the semantic features of the internal knowledge and the synonyms of the external knowledge, which can reflect the fitness of the documents to the cluster topics. Also, it efficiently clusters documents using fuzzy association depending on the terms of the condensing cluster. Thus, the proposed method is able to more successfully identify similar documents in each cluster, when compared with other clustering methods.

7. Conclusion

This paper presents a document clustering method using fuzzy association and the terms of the condensing cluster topics based on semantic features and synonyms. The proposed method uses the semantic features depending on the NMF and synonyms from WordNet to extract terms for the summarizing cluster, which are well covered within the major topics of the document set. It also uses fuzzy association between the terms of the document set and the terms of the condensing cluster to improve the quality of the document clustering. It was demonstrated that the normalized mutual information is higher than the internal knowledge and general clustering methods for Reuters test collections using the proposed method.

Acknowledgement

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 (2010-0028295).

References

[1] S. Chakrabarti, Mining the Web: Discovering Knowledge from Hypertext Data, Morgan Kaufmann, 2003.

[2] S.J. Fodeh, W.F. Punch, and P.N. Tan, "Combining statistics and semantics vis ensemble model for document clustering," Proc. 24th Annual ACM Symposium on Applied Computing (SAC'09), pp.1446–1450, Honolulu, Hawaii, March 2009.

[3] W.B. Frankes and B.Y. Ricardo, Information Retrieval, Data Structure & Algorithms, Prentice-Hall, 1992.

[4] C. Haruechaiyasak, M.L. Shyu, and S.C. Chen, "Web document classification based on fuzzy association," Proc. 25th Annual International Computer Software and Applications Conference (COMPSAC'02), pp.487–492, Oxford, England, Aug. 2002.

[5] J. Han and M. Kamber, Data Mining Concepts and Techniques, Second ed., Morgan Kaufmann, 2006.

[6] T. Hu, H. Xiong, W. Zhou, S.Y. Sung, and H. Luo, "Hypergraph partitioning for document clustering: A unified clique perspective," Proc. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08), pp.871–872, Singapore, July 2008.

[7] X. Hu, X. Zhang, C. Lu, E.K. Park, and X. Zhou, "Exploiting wikipedia as external knowledge for document clustering," Proc. 15th ACM SIGKDD Conference On Knowledge Discovery and Data Mining (KDD'09), pp.389–396, Paris, France, June 2009.

[8] X. Ji, W. Xu, and S. Zhu, "Document clustering with prior knowledge," Proc. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06), pp.405–412, Seattle, USA, Aug. 2006.

[9] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol.401, pp.788–791, Oct. 1999.

[10] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," Advances in Neural Information Processing Systems, vol.13, pp.556–562, April 2001.

[11] T. Li and C. Ding, "The relationships among various nonnegative matrix factorization method for clustering," Proc. 6th IEEE International Conference on Data Mining (ICDM'06), pp.362–371, Hong Kong, China, Dec. 2006.

[12] X. Liu, Y. Gong, W. Xu, and S. Zhu, "Document clustering with cluster refinement and model selection capabilities," Proc. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02), pp.191–198, Tampere, Finland, 2002.

[13] T. Li, S. Ma, and M. Ogihara, "Document clustering via adaptive subspace iteration," Proc. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'04), pp.218–225, UK, July 2004.

[14] G. Miller, "WordNet: A lexical databassed for English," CACM, vol.38, no.11, pp.39–41, 1995.

[15] S. Park, D.U. An, B.R. Cha, and C.W. Kim, "Document clustering with cluster refinement and non-negative matrix factorization," Proc. 16th International Conference on Neural Information Processing (ICONIP'09), pp.281–288, Bangkok, Thailand, Dec. 2009.

[16] S. Park, D.U. An, and I.C. Choi, "Document clustering using weighted semantic features and cluster similarity," Proc. 3rd IEEE

International Conference on Digital Game and Intelligent Toy Enhanced Learning (DIGITEL'10), pp.185–187, Kaohsiung, Taiwan, April 2010.

[17] S. Park, D.U. An, B.R. Cha, and C.W. Kim, "Document clustering with semantic feature and fuzzy association," Proc. International Conference on Information Systems, Technology and Management (ICISTM'10), pp.167–175, Bangkok, Thailand, March 2010.

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

[19] B.Y. Ricardo and R.N. Berthier, Moden Information Retrieval, ACM Press, 1999.

[20] J. Wang, H. Zeng, Z. Chen, H. Lu, L. Tao, and W.Y. Ma, "ReCoM: Reinforcement clustering of multi-type interrelated data objects," Proc. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03), pp.274–281, Toronto, Canada, Aug. 2003.

[21] F. Wang and C. Zhang, "Regularized clustering for documents," Proc. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07), pp.95–102, Amsterdam, July 2007.

[22] W. Xu and Y. Gong, "Document clustering by concept factorization," Proc. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'04), pp.202–209, UK, July 2004.

[23] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," Proc. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'03), pp.267–273, Toronto, Canada, Aug. 2003.

[24] L.A. Zadeh, D. Dubois, H. Prade, and R.R. Yager, Fuzzy Sets, in editions, Readings in Fuzzy Sets for Intelligent Systems, Morgan Kaufmann, 1993.

[25] D. Zhang and Y. Dong, "Semantic, hierarchical, online clustering of web search results," Proc. Asia Pacific Web Conference (APWEB), pp.67–78, Hangzhou, China, April 2004.

[26] X. Zhang, X. Hu, and X. Zhou, "A comparative evaluation of different link types on enhancing document clustering," Proc. ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08), pp.555–562, Singapore, July 2008.

Sun Park is a research professor at Research Faculty Institute of Information Science and Engineering Research, Mokpo National University, Korea. He received the Ph.D. degree in Computer & Information Engineering from Inha University in 2007, the M.S. degree in Information & Communication Engineering from Hannam University in 2001, and the B.S. degree in Computer Engineering from Jeonju University in 1996. Prior to becoming a researcher at Mokpo National University, he has worked as a postdoctoral at Chonbuk National University, and professor in Dept. of Computer Engineering, Honam University, Korea. His research interests include Data Mining, Information Retrieval, and Information Summarization.

Seong Ro Lee received th B.S. degree in electronics engineering from Korea University, Seoul, Korea, in 1987, respectively, and the M.S., and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, Korea, 1990 and 1996, respectively. In September 1997, he joined the Division of Electronics Engineering, Mokpo National University, Jeonnam, Korea. His research interests include digital communication system, mobile and satellite communications system, applications of telematics, USN and embedded system. He serves as chairman of detection and estimation committee for the Korea Information and Communications Society.