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

J. Inf. Commun. Converg. Eng. 11(3): 179-184, Sep. 2013 **Regular paper**

Document Clustering Using Semantic Features and Fuzzy Relations

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

[1]Department of Computer Engineering, Honam University, Gwangju 506-714, Korea
[2]Institute of Information Science and Engineering Research, Mokpo National University, Muan-gun 524-729, Korea

Abstract

Traditional clustering methods are usually based on the bag-of-words (BOW) model. A disadvantage of the BOW model is that it ignores the semantic relationship among terms in the data set. To resolve this problem, ontology or matrix factorization approaches are usually used. However, a major problem of the ontology approach is that it is usually difficult to find a comprehensive ontology that can cover all the concepts mentioned in a collection. This paper proposes a new document clustering method using semantic features and fuzzy relations for solving the problems of ontology and matrix factorization approaches. The proposed method can improve the quality of document clustering because the clustered documents use fuzzy relation values between semantic features and terms to distinguish clearly among dissimilar documents in clusters. The selected cluster label terms can represent the inherent structure of a document set better by using semantic features based on non-negative matrix factorization, which is used in document clustering. The experimental results demonstrate that the proposed method achieves better performance than other document clustering methods.

Index Terms: Cluster label, Document clustering, Non-negative matrix factorization, Semantic features, WordNet

I. INTRODUCTION

The rapidly growing availability of a large quantity of textual data, such as online news, blogs, emails, and Internet bulletin boards, has created the need for effective document clustering methods. In addition, document clustering has been receiving increased attention as an important method for unsupervised document organization, automatic summarization, topic extraction, and information filtering or retrieval [1-7].

Recent studies on document clustering methods use machine learning [2, 5, 8-10] techniques, graph-based methods [7, 11], and matrix factorization-based methods [12-14]. Machine learning-based methods use a semi-supervised clustering model with respect to prior knowledge and documents' membership [2, 5, 8-10]. Graph-based methods model the given document set using an undirected graph in which each node represents a document [7, 11]. Matrix factorization-based methods use semantic features of the document sets for document clustering [12-15].

Traditional clustering methods are usually based on the bag-of-words (BOW) model. A disadvantage of the BOW model is that it ignores the semantic relationship among terms in the data set [5]. To resolve this problem, ontology or matrix factorization approaches are usually used. However, a major problem of the ontology approach is that it is usually difficult to find a comprehensive ontology that can

Received 04 March 2013, **Revised** 11 April 2013, **Accepted** 23 April 2013
***Corresponding Author** Sun Park (E-mail: sunpark@mokpo.ac.kr, Tel: +82-10-3919-0634)
Institute of Information Science and Engineering Research, Mokpo National University, 1666 Yeongsan-ro, Cheonggye-myeon, Muan-gun 524-729, Korea.

Open Access **http://dx.doi.org/10.6109/jicce.2013.11.3.179** print ISSN: 2234-8255 online ISSN: 2234-8883

cover all the concepts mentioned in a collection [5]. In addition, a matrix factorization approach might limit successful decomposition of semantic features from any data set as data objects viewed from extremely different viewpoints, or highly articulated objects [13, 16, 17].

In this paper, we propose a document clustering method using semantic features by non-negative matrix factorization (NMF) and fuzzy relations. The proposed method uses fuzzy relations between semantic features and terms in a document set to resolve the matrix factorization approach problem. The NMF can represent an individual object as the non-negative linear combination of partial information extracted from a large volume of objects [13, 16, 17]. NMF has great power to easily extract semantic features representing the inherent structure of data objects. The factorization result of NMF has a better semantic interpretation, and the clustering result can be easily derived from it [16]. Fuzzy relations [18] use the concept of fuzzy set theory [19] to model the vagueness in the information retrieval. The basic concept of fuzzy relations involves the construction of index terms from a set of documents [18].

The proposed method has the following advantages. First, it can extract important cluster label terms in a document set using semantic features by NMF. By this means, it can identify major topics and subtopics of clusters with respect to their semantic features. Second, it can remove the dissimilar documents in clusters using fuzzy relations between semantic features and document terms. Thus it can improve the quality of document clustering by assisting with the removal of dissimilarity information.

The rest of the paper is organized as follows: Section II describes works related to document clustering methods. In Section III, we review NMF and fuzzy relations in detail. In Section IV, the proposed document clustering method is introduced. Section V shows the evaluation and experimental results. Finally, we conclude in Section VI.

II. RELATED WORKS

Traditional clustering methods can be classified into partitioning, hierarchical, density-based, and grid-based methods. Most of these methods use distance functions as object criteria and are not effective in high dimensional spaces [1, 3, 4, 6, 20].

Li et al. [20] proposed a document clustering algorithm called adaptive subspace iteration (ASI) using explicit modeling of the subspace structure associated with each cluster. Wang et al. [21] proposed a clustering approach for clustering multi-type interrelated data objects. It fully explores the relationship between data objects for clustering analysis. Park et al. [12] proposed a document clustering method using NMF and cluster refinement. Park et al. [15] proposed a document clustering method using latent semantic analysis (LSA) and fuzzy association. Xu et al. [14] proposed a document partitioning method based on the NMF of the given document corpus. Xu and Gong [13] proposed a data clustering method that models each cluster as a linear combination of the data points, and each data point as a linear combination of the cluster centers. Li and Ding [22] presented an overview and summary of various matrix factorization algorithms for clustering and theoretically analyzed the relationships among them.

Wang et al. [7] proposed document clustering with local and global regularization (CLGR). It uses local label predictors and a global label smoothness regularizer. Liu et al. [9] proposed a document clustering method using cluster refinement and model selection. It uses a Gaussian mixture model and expectation maximization algorithm to conduct initial document clustering. It also refines the initially obtained document clusters by voting on the cluster label of each document.

Ji and Xu [8] proposed a semi-supervised clustering model that incorporates prior knowledge about documents' membership for document clustering analysis. Zhang et al. [10] adopted a relaxation labeling-based cluster algorithm to evaluate the effectiveness of the aforementioned types of links for document clustering. It uses both content and linkage information in the dataset [11]. Hu et al. [5] proposed a document clustering method exploiting Wikipedia as external knowledge. They used Wikipedia for resolving the external ontology of document clustering problem. Fodeh et al. [2] proposed a document clustering method using an ensemble model combining statistics and semantics [7].

III. NMF AND FUZZY RELATION THEORY

A. NMF

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 i'th row vector and X_{ij} be the element of the i'th row and the j'th column.

NMF involves decomposing 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).

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

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.

We use the objective function that minimizes the Euclidean distance between each column of A and its approximation $\widetilde{A} = WH$, which was proposed in Lee and Seung [16, 17]. As an objective function, the Frobenius norm is used:

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

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

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu}\frac{(W^T A)_{\alpha\mu}}{(W^T WH)_{\alpha\mu}}, \quad W_{i\alpha} \leftarrow W_{i\alpha}\frac{(AH^T)_{i\alpha}}{(WHH^T)_{i\alpha}}. \quad (3)$$

Example 1. We illustrate the example using a NMF algorithm: 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, it decomposes the matrix A into the W and H matrices as in Fig. 1.

Fig. 2 shows an example of sentence representation using NMF. The column vector A_{*3} corresponding to the third sentence is represented as a linear combination of semantic feature vectors W_{*l} and semantic variable column vector H_{*3}.

The powers of the two non-negative matrices W and H are described as follows: all semantic variables (H_{lj}) are used to represent each sentence. W and H are represented sparsely. Intuitively, it make more sense for each sentence to be associated with some small subset of a large array of topics (W_{*l}), rather than just one topic or all the topics. In each semantic feature (W_{*l}), the NMF has grouped together semantically related terms [3].

$$\underset{A}{\begin{bmatrix}1&1&1&2\\0&2&1&0\\0&1&1&0\\2&1&1&1\end{bmatrix}} \approx \underset{W}{\begin{bmatrix}0&1.135&1.974\\2.239&0&0\\1.356&0.040&0.024\\1.175&2.271&0\end{bmatrix}} \times \underset{H}{\begin{bmatrix}0&0.853&0.523&0\\0.880&0&0.172&0.004\\0&0.506&0.409&0.759\end{bmatrix}}$$

Fig. 1. Result of the non-negative matrix factorization algorithm.

$$\underset{A_{*3}}{\begin{bmatrix}1\\1\\1\\1\end{bmatrix}} \approx 0.532\times \underset{W_{*1}}{\begin{bmatrix}0\\2.239\\1.356\\1.175\end{bmatrix}} + 0.172\times \underset{W_{*2}}{\begin{bmatrix}1.135\\0\\0.040\\2.217\end{bmatrix}} + 0.409\times \underset{W_{*3}}{\begin{bmatrix}1.974\\0\\0.024\\0\end{bmatrix}}$$

A_{*3} H_{13} W_{*1} H_{23} W_{*2} H_{33} W_{*3}

Fig. 2. Example of sentence representation using semantic features and semantic variables.

B. Fuzzy Relations Theory

In this section, we give a brief review of fuzzy relations theory [1, 18, 19], which is used in document clustering. The fuzzy set is defined as follows:

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

Definition 2. *Given a set of index terms, $T = \{t_1, \ldots, t_2\}$, and a set of documents, $D=\{d_1, \ldots, d_v\}$, each t_i is represented by a fuzzy set $h(t_i)$ of documents, $h(t_i) = \{F(t_i,d_j) \mid \forall d_i \in D\}$, 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) relation 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.*

$$RT(t_i,t_j) = \frac{\sum_k \min(F(t_i,d_k),F(t_j,d_k))}{\sum_k \max(F(t_i,d_k),F(t_j,d_k))}. \quad (4)$$

A simplification of the fuzzy RT relation based on the co-occurrence of terms is given as follows:

$$r_{i,j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}} \quad (5)$$

where $r_{i,j}$ represents the fuzzy RT relation between 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.

Definition 4. *The membership degrees between each document to each of the cluster sets can be defined as follows*:

$$\mu_{i,j} = \sum_{\forall t_a \in d_i}[1 - \prod_{\forall t_b \in CT^j}(1 - r_{a,b})] \quad (6)$$

where $\mu_{i,j}$ is the membership degree of d_i belonging to CT^j, $r_{a,b}$ is the fuzzy relation between term $t_j \in d_j$ and term $t_b \in CT^j$. CT is the term set with respect to representing a cluster topic.

IV. PROPOSED DOCUMENT CLUSTERING METHOD

In this section, we propose a method that clusters documents by semantic features by NMF and fuzzy relations.

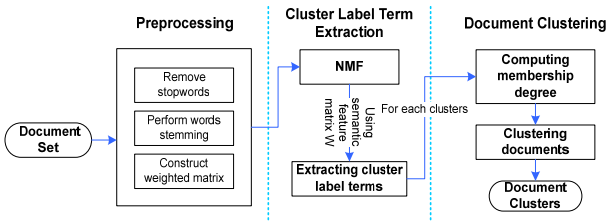

Fig. 3. Document clustering method using semantic features and fuzzy relations. NMF: non-negative matrix factorization.

The proposed method consists of the preprocessing phase, cluster label extraction phase, and the document cluster phase. We next give a full explanation of the three phases shown in Fig. 3.

A. Preprocessing

In the preprocessing phase, we remove all stop-words by using Rijsbergen's stop-words list and perform word stemming by Porter's stemming algorithm [3, 6]. Then we construct the term-frequency vector for each document in the document set [1, 3, 6].

Let $T_i = [t_{1i}, t_{2i} \ldots t_{ni}]^T$ be the term-frequency vector of document i, where elements t_{ji} denote the frequency in which term j occurs in document i. Let A be an $m \times n$ terms by documents matrix, where m is the number of terms and n is the number of documents in a document set.

B. Cluster Label Term Extraction by NMF

In the cluster label terms extraction phase, we use semantic features by NMF [15-17] to extract cluster label terms. The proposed cluster label term extraction method is described as follows. First, the preprocessing phase is performed, and then the term-document frequency matrix is constructed. Table 1 shows the term-document frequency matrix with respect to 7 documents and 6 terms. Table 2 shows the semantic features matrix W by NMF from Table 1. The cluster label terms having the top semantic values are selected in each column for the cluster label terms in Table 2. Table 3 shows the extracted cluster label terms from Table 2.

Table 1. Term-document frequency matrix

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

Table 2. Semantic features matrix W by non-negative matrix factorization from Table 1

	r1	*r2*	*r3*
t1	0	1.8455	0.4791
t2	0	0	2.4913
t3	0.2884	2.6364	0
t4	2.8135	0	0.0048
t5	2.1483	0.0834	0.0341
t6	1.1197	0	0

Table 3. Result of cluster label terms extraction from Table 2

	Cluster label terms
C1	*t4, t5*
C2	*t3, t1*
C3	*t2, t1*

C. Clustering Document by Fuzzy Relations

The document clustering phase is described as follows. First, we construct the term correlation matrix M with respect to the relationship between cluster label terms and terms of the document set using the fuzzy RT relation by Eq. (5). The term correlation matrix is an n by n symmetric matrix whose element, m_{ij}, has the value on the interval [0, 1] in which 0 indicates no relationship and 1 indicates a full relationship between the terms t_i and t_j. Therefore, m_{ij} is equal to 1 for all $i = j$. Since a term has the strongest relationship to itself [18]. Table 4 shows the term correlation matrix using Table 1 and Eq. (5).

Second, a document d_i is clustered into the cluster C^j, where the membership degree $\mu_{i,j}$ is the maximum by Eq. (6). The term t_a in i is associated with cluster C^j if the terms k_b's in CT^j (for cluster C^j) are related to the term t_a [18]. Table 5 shows the result of document clustering using Table 4 and Eq. (6).

Table 4. Term correlation matrix from Table 2 by the fuzzy related terms relation

	t1	*t2*	*t3*	*t4*	*t5*	*t6*
t1	1	0.5	0.5	0	0	0
t2	0.5	1	0.4	0.17	0.14	0
t3	0.5	0.4	1	0.29	0.29	0
t4	0	0.17	0.29	1	1	0.4
t5	0	0.14	0.29	1	1	0.4
t6	0	0	0	0.4	0.4	1

Table 5. Result of document clustering from Table 4

	Document
C1	*d4, d5, d6, d7*
C2	*d1*
C3	*d2*

V. EXPERIMENTS AND EVALUATION

We use the *Reuters* (http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html) document corpora to evaluate the proposed method. The *Reuters* corpus has 21,578 documents, which are grouped into 135 clusters [7]. We use the normalized mutual information metric MI for measuring the document clustering performance [7, 13, 14].

We have conducted a performance evaluation by testing the proposed method and comparing it with 6 other representative data clustering methods using the same data corpus. We implemented 7 document clustering methods: FNMF, FLSA, RNMF, KM, NMF, ASI, and CLRG. In Fig. 4, Fisher NMF (FNMF) denotes our proposed method. Feature LSA (FLSA) denotes our previous proposed method using LSA and fuzzy relations [15]. Robust NMF (RNMF) denotes our previous method by using NMF and cluster refinement [14]. KM denotes the partitioning method using traditional *k*-means [1, 3, 4, 6]. NMF demotes Xu's method using non-negative matrix factorization [14]. ASI denotes Li's method using adaptive subspace iteration [20]. CLRG denotes Wang's method using local and global regularization [7].

The evaluation results are shown in Fig. 4. The evaluations were conducted for the cluster numbers ranging from 2 to 10. For each given cluster number *k*, 50 experiment runs were conducted on different randomly chosen clusters, and the final performance values were obtained by averaging the values from the 50 experiments.

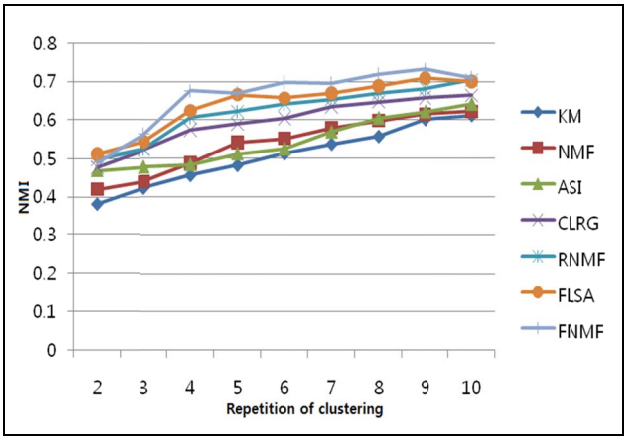

Fig. 4. Evaluation results of performance comparison. NMI: normalized mutual information, KM: k-means, NMF: non-negative matrix factorization, ASI: adaptive subspace iteration, CLRG: clustering with local and global regularization, RNMF: robust NMF, FLSA: feature latent semantic analysis, FMNF: Fisher NMF.

VI. CONCLUSION

In this paper, we have presented a document clustering method using semantic features by NMF and fuzzy relations. The proposed method in this paper has the following advantages. First, it can identify dissimilar documents between clusters by fuzzy relations with respect to cluster label terms and documents, thereby improving the quality of document clustering. Second, it can easily extract cluster label terms that cover the major topics of a document well using semantic features by NMF. Experimental results show that the proposed method outperforms 6 other summarization methods.

REFERENCES

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

[2] S. J. Fodeh, W. F. Punch, and P. N. Tan, "Combining statistics and semantics via ensemble model for document clustering," in *Proceeding of the 24th Annual ACM Symposium on Applied Computing*, Honolulu, HI, pp. 1446-1450, 2009.

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

[4] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, 2nd ed. Boston, MA: Morgan-Kaufmann, 2006.

[5] 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.

[6] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. New York, NY: ACM Press, 1999.

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

[8] X. Ji and W. Xu, "Document clustering with prior knowledge," in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, WA, pp. 405-412, 2006.

[9] X. Liu, Y. Gong, W. Xu, and S. Zhu, "Document clustering with cluster refinement and model selection capabilities," in *Proceeding of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland, pp. 191-198, 2002.

[10] X. Zhang, X. Hu, and X. Zhou, "A comparative evaluation of different link types on enhancing document clustering," in *Proceeding of the 31th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Singapore, pp. 555-562, 2008.

[11] T. Hu, H. Xiong, W. Zhou, S. Y. Sung, and H. Luo, "Hypergraph partitioning for document clustering: a unified clique perspective," in *Proceeding of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Singapore, pp. 871-872, 2008.

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

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

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

[15] S. Park, D. U. An, B. R. Cha, and C. W. Kim, "Document clustering with semantic features and fuzzy association," in *Proceeding of the 4th International Conference on Information Systems, Technology and Management*, Bangkok, Thailand, pp. 167-175, 2010.

[16] 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.

[17] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*. Cambridge, MA: MIT Press, pp. 556-562, 2001.

[18] C. Haruechaiyasak, M. L. Shyu, S. C. Chen, and X. Li, "Web document classification based on fuzzy association," in *Proceedings of the 26th International Computer Software and Applications Conference on Prolonging Software Life: Development and Redevelopment*, Oxford, UK, pp. 487-492, 2002.

[19] L. A. Zadeh, "Fuzzy sets," in *Readings in Fuzzy Sets for Intelligent Systems*. San Francisco, CA: Morgan-Kaufmann, 1993.

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

[21] J. Wang, H. Zeng, Z. Chen, H. Lu, L. Tao, and W. Y. Ma, "ReCoM: reinforcement clustering of multi-type interrelated data objects," in *Proceeding of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, pp. 274-281, 2003.

[22] T. Li and C. Ding, "The relationships among various nonnegative matrix factorization method for clustering," in *Proceeding of the 6th International Conference on Data Mining*, Hong Kong, China, pp. 362-371, 2006.

Chul-Won Kim
received a Ph.D. degree in Computer Engineering from Kwangwoon University in 1997. He is a Professor at Honam University. His research interests include XML retrieval, multimedia information retrieval, and multimedia processing.

Sun Park
is a research professor at the Institute of Research on Information Science and Engineering, Mokpo National University, Korea. He received a Ph.D. degree in Computer & Information Engineering from Inha University in 2007, an M.S. degree in Information & Communication Engineering from Hannam University in 2001, and a B.S. degree in Computer Engineering from Jeonju University in 1996. Prior to becoming a researcher at Mokpo National University, he worked as a postdoctoral researcher at Chonbuk National University and professor in the Department of Computer Engineering, Honam University, Korea. His research interests include big data mining, information retrieval, IT fusion, and information summarization.