International Journal of Software Engineering and Its Applications
Vol.8, No.4 (2014), pp.1-10
http://dx.doi.org/10.14257/ijseia.2014.8.4.01

Big Text Data Clustering using Class Labels and Semantic Feature Based on Hadoop of Cloud Computing

Yong-Il Kim[1], Yoo-Kang Ji[2] and Sun Park[3*]

[1]Honam University, South Korea, [2]DongShin University, South Korea, [3]GIST, South Korea

[1]yikim@honam.ac.kr, [2]neobacje@gmail.com, [3]sunpark@nm.gist.ac.kr

Abstract

Clustering of class labels can be generated automatically, which is much lower quality than labels specified by human. If the class labels for clustering are provided, the clustering is more effective. In classic document clustering based on vector model, documents appear terms frequency without considering the semantic information of each document. The property of vector model may be incorrectly classified documents into different clusters when documents of same cluster lack the shared terms. To overcome this problem are applied by the knowledge based approaches. However, these approaches have an influence of inherent structure of documents on clustering and a cost problem of constructing ontology. In addition, the methods are limited to cluster suitable text document clustering from in exploding big text data on Cloud environment. In this paper, we propose a big text document clustering method using terms of class label and semantic feature based Hadoop. Class label term can well represent the inherent structure of document clusters by non-negative matrix factorization (NMF) based Hadoop. The proposed method can improve the quality of document clustering which uses the class label terms and the term weights based on term mutual information (TMI) with WordNet at a little cost. It also can cluster the big data size of document using the distributed parallel processing based on Hadoop. The experimental results demonstrate that the proposed method achieves better performance than other document clustering methods.

Keywords: *document clustering, NMF, semantic features, term weight, WordNet, term mutual information, big text data, Hadoop*

1. Introduction

Clustering of class labels can be generated automatically however there are different from labels specified by humans. The automatic class label is much lower quality than a maunal class label. If the specified class labels for clustering are provided with no human intervention, the clustering is more effective [1-3]. 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

* Corresponding author

ISSN: 1738-9984 IJSEIA

cluster forms by user [1]. Recently, to overcome the problems of the vector model-based document clustering, internal and external knowledge approaches are applied.

Internal knowledge-based document clustering uses the inherent structure of the document set by means of a factorization technique. The factorization techniques for document clustering including non-negative matrix factorization (NMF) [4-6], concept factorization (CF) [7], adaptive subspace iteration (ASI) [8], and clustering with local and global regularization (CLGR) [9] have been proposed, which can accurately identify the topics of document set from their semantic features. 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 [10]. External knowledge-based document clustering exploits the constructed term ontology from external knowledge database with regard to ontology. Recently, the term ontology techniques for document clustering are proposed such as term mutual information 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], *etc*. The 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 concepts mentioned in the documents collection, which is a cause of loss of information [1, 12]. Moreover, the ontology-based method takes higher cost to construct the ontology manually by knowledge engineers and domain experts.

In order to resolve the limitations of the knowledge-based approaches, this paper proposes a document clustering method that uses terms of class label by semantic features of NMF based on Hadoop and term weights by term mutual information (TMI) in connection with WordNet. The proposed method combines the advantages of the internal and external knowledge-based methods. In the proposed method, first, meaningful terms of class label for describing cluster topics of documents are extracted using NMF based on Hadoop. The extracted terms well represents the class label of document clusters by means of semantic features (*i.e.*, internal knowledge) having inherent structure of documents. Second, the term weights of documents are calculated using the TMI based on the synonyms of WordNet (*i.e.*, external knowledge) with respect to documents terms. The term weights can easily classify documents into an appropriate class label by extending the coverage of document with respect to class label. The method can cluster the big data size of document using the distributed parallel processing based on Hadoop.

The rest of the paper is organized as follows: Section 2, Hadoop framework are introduced for related works. In Section 3 describe the NMF algorithm in detail. In section 4 explains the proposed methods. In Section 5 shows the performance evaluation and experimental results of the proposed method. Finally, we conclude in Section 6.

2. Hadoop Framework

The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing. The Hadoop project includes the Apache Hadoop software library which is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures [15].

3. Non-negative Matrix Factorization

This section reviews NMF theory with algorithm and describes the advantage of semantic features by comparison between NMF and SVD (singular value decomposition) in Example 1. 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 Equation (1) [10].

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

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.

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 [10]. As an objective function, the Frobenius norm is used:

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

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:

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

Example 1) We illustrate an example of NMF and SVD decomposition [2, 10]. The non-negative matrix A is decomposed by *nnmf*() function of Matlab 7.8 into two non-negative matrices, W and H, as shown in Figure 1(a).

$$A \approx W \times H$$

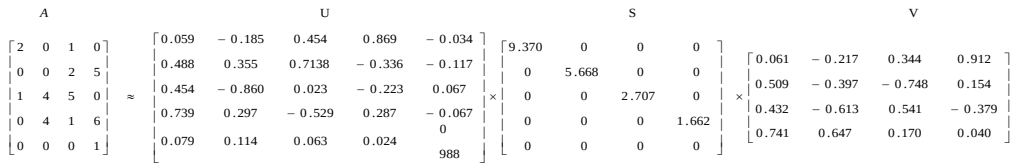
$$\begin{bmatrix} 2 & 0 & 1 & 0 \\ 0 & 0 & 2 & 5 \\ 1 & 4 & 5 & 0 \\ 0 & 4 & 1 & 6 \\ 0 & 0 & 0 & 1 \end{bmatrix} \approx \begin{bmatrix} 0 & 0 & 1.692 \\ 5.022 & 0.048 & 1.573 \\ 0 & 4.128 & 4.256 \\ 5.982 & 4.076 & 0 \\ 1.001 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 0 & 0 & 0.059 & 0.998 \\ 0 & 0.980 & 0.1981 & 0.007 \\ 0.371 & 0. & 0.929 & 0 \end{bmatrix}$$

(a) Result of NMF Decomposition

$$\begin{bmatrix} 0 \\ 0 \\ 4 \\ 4 \\ 0 \end{bmatrix} \approx 0 \times \begin{bmatrix} 0 \\ 5.022 \\ 0 \\ 5.982 \\ 1.001 \end{bmatrix} + 0.980 \times \begin{bmatrix} 0 \\ 0.048 \\ 4.128 \\ 4.076 \\ 0 \end{bmatrix} + 0 \times \begin{bmatrix} 1.692 \\ 1.573 \\ 4.256 \\ 0 \\ 0 \end{bmatrix}$$

$A_{*2} \quad H_{12} \; W_{*1} \qquad H_{22} \; W_{*2} \qquad H_{23} \; W_{*3}$

(b) Example of Column vector A_{*2} Representation using Semantic Features and Semantic Variables



(c) Result of SVD decomposition

Figure 1. Example of NMF and SVD

Figure 1(b) shows an example of the representation of a column vector corresponding to document by a linear combination of semantic feature and semantic variable. Figure 1(c) shows the result of SVD by *svd*() function of Matlab 7.8. There are no zero values and negative values in relation to the semantic feature matrices U and V in Figure 1(c). Unlike SVD, the semantic feature matrices *W* and *H* by NMF are sparse in Figure 1(a). Intuitively, the NMF can obtain semantic features that have a small semantic range rather than SVD. In other words, the sparse property of the semantic features of NMF can cover class labels by several terms of document to be associated with the semantic features. Thus, the semantic feature can easily identify class label terms to signify document cluster. Besides, it can help to distinguish the multiple meanings of the same term [10].

4. Proposed Document Clustering Method

This paper proposes a document clustering method using class label terms by NMF based on Hadoop and term weights based on TMI with WordNet. The proposed method consists of three phases: preprocessing, extracting class label terms, and clustering document, as shown in Figure 2. In the subsections below, each phase is explained in full.

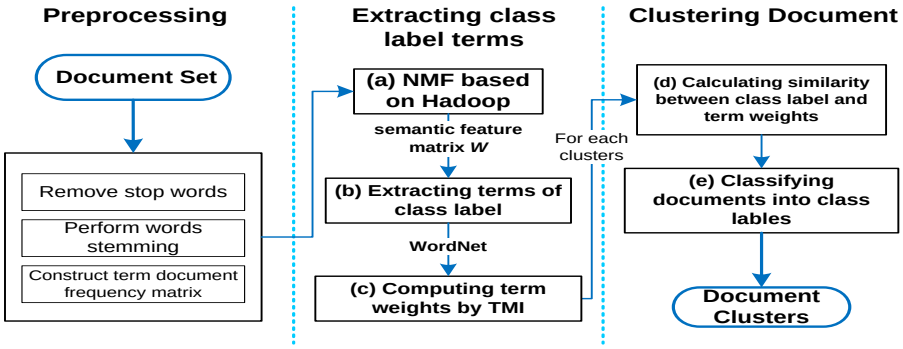


Figure 2. Document Clustering Method using Class Label Terms and Term Weights

4.1. Preprocessing

In the preprocessing phase, Rijsbergen's stop words list is used to remove all stop words, and word stemming is removed using Porter's stemming algorithm [16]. Then, the term document frequency matrix A is constructed from the document set.

4.2. Extracting Class Label Terms

This section extracts class label terms in regard to the properties of the document clusters using NMF as in Figure 2(b). The terms of class label that can well explain the topic of

document cluster are derived by the semantic features of the NMF based on Hadoop. Semantic features of big text document for clustering are extracted by the modified Liu's distributed NMF (dNMF) method [17] based on distributed parallel processing on Hadoop MapReduce programming. The extracting method is described as follows. First, term document frequency matrix *A* is constructed by performing the preprocessing phase. Second, let the number of cluster (*i.e.*, the number of semantic feature r) be set, and then dNMF is performed on the matrix *A* to decompose the two sematic feature matrices *W* and *H*. Finally, matrix *W* and Equation (4) are used to extract class label terms. The column vector of matrix *W* corresponds to class label of cluster and the row vector of matrix *W* refers to terms of document, which the element of matrix *W* (*i.e.*, the semantic feature value) indicates how much the term reflects the cluster class labels. The equation of extracting terms of class labels is as follows.

$$CL^{p} \leftarrow A_{ij} \; if \;\; p = \arg\max_{1 \le j \le r} W_{ij} \; and \;\; W_{ij} \ge as^{j} \tag{4}$$

where CL^p is the term set of p'th class label of cluster, A_{ij} is the term corresponding to the semantic feature of i'th row and the j'th column in the matrix W. The average semantic feature value of j'th column vector, as^j, is as follows.

$$as^{j} = \frac{\sum_{i=1}^{m} W_{ij}}{nz} \tag{5}$$

where m is the number of i'th row (i.e., the number of terms), nz is the number of non-zero (*i.e.*, positive) elements of i'th row.

4.3. Computing Term Weights by TMI based on WordNet

In this section, the term weights are calculated by TMI (term mutual information) based on the synonyms of WordNet as in Figure 2(c). 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 [18]. Class label terms may be restricted from properties of document cluster and document composition. To resolve this problem, this paper uses term weight of documents by using the TMI on synonyms of WordNet. Term weights of the document are calculated by jing's TMI as in Equation (6) [11]. The Jing's TMI is as follows. In Equation (6), δ_{il} is to indicate semantic information between two terms. If term A_{lj} appears in the synonyms of A_{ij} by means of WordNet, δ_{il} will be treated in a same level for different A_{ij} and A_{lj}, otherwise, δ_{il} will be set zero.

$$\tilde{A}_{ij} = A_{ij} + \sum_{\substack{l=1 \\ i \ne l}}^{m} \delta_{il} A_{lj} \tag{6}$$

Example 2) Table 1 shows the six documents from a part of Figure 4.10 of [2]. Table 2 shows term document frequency matrix by performing the preprocessing phase from Table 1. Table 3 shows the sematic feature matrix W obtained through NMF from Table 2, and the result of average of non-zero element of semantic features vector as by using Equation (5). Table 4 shows the results of the extracted class label terms from Table 3, which matches the semantic feature values of more than the average semantic feature value as^j.

Table 1. Document Set of Composition of 6 Documents

documents	document contents
d1	A course on integral equations
d2	A tractors for semigroups and evolution equations
d3	Automatic differentiation of algorithms : theory, implementation, and application
d4	Geometrical aspects of partial differential equations

Table 2. Term Document Frequency Matrix from Table 1

document / term	d1	d2	d3	d4
course	1	0	0	0
integral	1	0	0	0
equations	1	1	0	1
tractors	0	1	0	0
semigroups	0	1	0	0
evolution	0	1	0	0
automatic	0	0	1	0
…	…	…	…	…
computational	0	0	0	0
algebra	0	0	0	0
commutative	0	0	0	0
oscillation	0	0	0	0
neutral	0	0	0	0
delay	0	0	0	0

Table 3. *The as* and Semantic Feature Matrix W by NMF from Table 2

term	r1(cluster1)	r2(cluster2)	r3(cluster3)
course	0	0.321	0
integral	0	0.312	0
equations	0	1.576	0
tractors	0	0.434	0
semigroups	0	0.434	0
evolution	0	0.434	0
automatic	0.004	0	0.899
…	…	…	…
computational	0.999	0	0
algebra	1.999	0	0
commutative	0.999	0	0
oscillation	0	0.487	0.201
neutral	0	0.487	0.201
delay	0	0.487	0.201
as	0. 576	0.5767	0.689

Table 4. The Extracted Terms of Class Labels

r1	r2	r3
algorithms, geometric, ideals, varieties, introduction, computational	equations, different	automatic, different, algorithms, theory, implementation

4.4. Clustering Document using Similarity

This section presents the clustering documents using cosine similarity between class label terms and term weights of documents. The proposed method in Figure 2(d) and 2(e) is described as follows. First, the cosine similarity between class label terms and term weights is calculated. And then a document having a highest similarity value with respect to the class label is clustered into cluster label in connection with the document clusters [3, 16].

5. Experiments and Evaluation

This paper uses 20 Newsgroups data set for performance evaluation [19-22]. 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 [2-4, 7-9]. The cluster numbers for the evaluation method are set by ranging from 2 to 10, as shown in Table 5. 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 in Table 5. The KM is a document clustering using Kmeans method based on a traditional partitioning clustering technique [2]. The NMF, ASI, CLGR, RNMF, and FPCA methods are document clustering methods based on internal knowledge. The FAWDN and TMINMF methods are clustering methods based on combining the internal and external knowledge. TMINMF denotes the proposed method described within this paper. FAWDN denotes the previously proposed method using the WordNet and fuzzy theory [14]. FPCA is the previously proposed method using PCA (principal component analysis) and fuzzy relationship [6], and RNMF is the method proposed previously using NMF and cluster refinement [5]. NMF denotes Xu's method using non-negative matrix factorization [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 Table 5, the average normalized metric of TMINMF is 16.7% higher than that of KM, 13.5% higher than that of NMF, 12.9% higher than that of ASI, 7.68% higher than that of CLGR, 5.12% higher than that of RNMF, 3.27% higher than that of FPCA, and 2.44% higher than that of FAWDN.

Table 5. Evaluation Results Of Performance Comparison

K	2	3	4	5	6	7	8	9	10
KM	0.382	0.423	0.456	0.482	0.512	0.537	0.557	0.601	0.612
NMF	0.42	0.439	0.487	0.541	0.551	0.579	0.598	0.615	0.622
ASI	0.469	0.477	0.483	0.511	0.524	0.568	0.604	0.621	0.641
CLGR	0.478	0.523	0.574	0.589	0.604	0.635	0.647	0.658	0.664
RNMF	0.498	0.525	0.607	0.622	0.642	0.654	0.669	0.681	0.704
FPCA	0.51	0.543	0.624	0.666	0.657	0.67	0.689	0.71	0.7
FAWDN	0.523	0.549	0.634	0.667	0.661	0.668	0.694	0.735	0.712
TMINMF	0.532	0.554	0.654	0.694	0.71	0.721	0.712	0.745	0.741
average	0.477	0.504	0.565	0.597	0.608	0.629	0.646	0.671	0.675

6. Conclusion

This paper proposes the enhancing document clustering method using class label terms and term weights. The proposed method uses the semantic features by internal knowledge of NMF to extract the class label terms, which are well represented within the important class labels of the documents cluster. To resolve the limitation of the semantic features with respect to be influenced by internal structure of documents, the method uses TMI (term mutual information) to calculate term weights of documents based on external knowledge of WordNet. In addition, it uses a similarity between the class label terms and term weights to improve the quality of the document clustering. It also can cluster the big data size of document using the distributed parallel processing based on Hadoop. It was demonstrated that the normalized mutual information is higher than other document clustering methods for 20 Newsgroups test collections using the proposed method.

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", Proceeding of SIGIR'08, Singapore, **(2008)** July, pp. 179-186.
[2] S. Chakrabarti, "Mining the web: Discovering Knowledge from Hypertext Data", Morgan Kaufmann Publishers, **(2003)**.
[3] B. Y. Ricardo and R. N. Berthier, "Moden Information Retrieval: the concepts and technology behind search", Second edition, ACM Press, **(2011)**.
[4] W. Xu, X. Liu and Y. Gon, "Document Clustering Based On Non-negative Matrix Factorization", Proceeding of SIGIR'03, Toronto Canada, **(2003)** August, pp. 267-274.
[5] S. Park, D. U. An, B. R. Cha and C. W. Kim, "Document Clustering with Cluster Refinement and Non-negative Matrix Factorization", Proceeding of the 16th ICONIP'09, 281-288, Bangkok, Thailand, **(2009)** December.
[6] S. Park and K. J. Kim, "Document Clustering using Non-negative Matrix Factorization and Fuzzy Relationship",The Journal of Korea Navigation Institute, vol. 14, no. 2, **(2010)** April, pp. 239-246.
[7] W. Xu and Y. Gong, "Document Clustering by Concept Factorization", Proceeding of SIGIR'04, UK, **(2004)** July, pp. 202-209.
[8] T. Li, S. Ma, M. Ogihara, "Document Clustering via Adaptive Subspace Iteration", Proceeding of SIGIR'04, UK, **(2004)** July, pp. 218-225.
[9] F. Wang and C. Zhang, "Regularized Clustering for Documents", Proceeding of SIGIR'07, Amsterdam, **(2007)** July, pp. 95-102.
[10] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization", Nature, vol. 401, **(1999)** October, pp. 788-791.
[11] L. Jing, L. Zhou, M. K. Ng and J. Z. Huang, "Ontology-based Distance Measure for Text Clustering", Proceeding of SIAM International conference on Text 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 proceeding of KDD'09, Paris, France, **(2009)** June, pp. 389-396
[13] H. H. Tar and T. T. S. Nyaunt, "Ontology-based Concept Weighting for Text Documents", World Academy of Science, Engineering and Technology, vol. 81, **(2011)**, pp. 249-253.
[14] S. Park and S. R. Lee, "Enhancing Document Clustering Using Condensing Cluster Terms and Fuzzy Association", Journal of IEICE TRANS, Information and System, vol. E94-D, no. 6, **(2011)** June, pp. 1227-1234.
[15] The Apache Hadoop project, "http://hadoop.apache.org/", **(2013)**.
[16] W. B. Frankes and B. Y. Ricardo, "Information Retrieval: Data Structure & Algorithms", Prentice-Hall, **(1992)**.
[17] C. Liu, H. C. Yang, J. Fan, L. W. He and Y. M. Wang, "Distributed Nonnegative Matrix Factorization for Web-Scale Dyadic Data Analysis on MapReduce", Proceeding of the International World Wide Web Conferene Comittee, USA, **(2010)**, pp. 1-10.
[18] G. Miller, "WordNet: A lexical databassed for English", CACM, vol. 38, no. 11, **(1995)**, pp. 39-41.
[19] The 20 newsgroups data set. http://people.csail.mit.edu/jrennie/20Newsgroups/, **(2012)**.
[20] A. Ngoumou and M. F. Ndjodo, "A Rewriting System Based Operational Semantics for the feature Oriented Resue Method", International Journal of Software Engineering and Its Applications, vol. 7, no. 6, **(2013)**, pp. 41-60.

[21] S. Mal and K. Rajnish, "New Quality Inheritance Metrics for Object-Oriented Design", International Journal of Software Engineering and Its Applications, vol. 7, no. 6, **(2013)**, pp. 185-200.
[22] S. H. Lee, J. G. Lee and K. I. Moon, "A preprocessing of Rough Sets Based on Attribute Variation Minimization", International Journal of Software Engineering and Its Applications, vol. 7, no. 6, **(2013)**, pp. 411-424.

Authors

Yong-Il Kim, received a B.S. in computer science from Chonnam University in 1984, and M.S. degree in computer science from Korea Advanced Institute Science and Technology (KAIST), Korea in 1986. From March 2002, he has joined as an Associate Professor at the Honam University, Gwangju, Korea. His research interests in data mining, big data, and intelligent agents. He is a member of IEEE.

Yoo-Kang Ji, he is a Visiting Professor at Dept. of Information & communication Eng., Dongshin Univ., South KOREA. He received the B.S., M.S., and Ph.D. degree in the Dept. of Information & Communication Eng. from DongShin Univ., KOREA in 2000, 2002 and 2006 respectively. He has worked professor in Dept. of Information & Communication Eng. DongShuin Univ. Mar. 2006 to Aug. 2009 His research interests in Mobile S/W, Networked Video and Embedded System.

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

