ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 5, Issue 6, June 2018

PPHOPCM: Privacy-preserving High-order Possibilistic c-Means Algorithm for Big Data Clustering with Cloud Computing

[1] C Arun Kumar, [2] Dasari Nishanth Reddy, [3] Keerthana K N, [4] M Sai Chakradhar Reddy

***Abstract*-** **PCM is one of the methods used for C means clustering process in which there are two types of clustering, normal PCM clustering and important is high order PCM for big data clustering. The HOPCM method based on Map reduces for the large amount of the heterogeneous data is used. Finally a privacy preserving high-order possibilistic c-means algorithm to protect the private data on cloud by applying the background verification scheme to HOPCM a high-order PCM algorithm to tackle big data clustering by making the objective function minimal in the tensor space. Clustering is designed to separate objects into several different groups according to special metrics, making the objects with similar features in the same group. Clustering techniques have been successfully applied to knowledge discovery and data engineering. With the increasing popularity of big data, big data clustering is attracting much attention from data engineers and researchers.**

***Keywords*—** **Big data clustering, cloud computing, privacy preserving, possibilistic c-means, tensor space.**

I. INTRODUCTION

As personal computing technology and social websites, such as Facebook and Twitter, become increasingly popular, big data is in the explosive growth [1]. Big data are typically heterogeneous, i.e., each object in big data set is multi-modal [2]. Specially, big data sets include various interrelated kinds of objects, such as texts, images and audios, resulting in high heterogeneity in terms of structure form, involving structured data and unstructured data. Moreover, different types of objects carry different information while they are interrelated with each other [3]. For example, a piece of sport video with meta-information uses a large number of subsequent images to display the exercise process and uses some meta-information, such as annotation and surrounding texts, to show additional information which are not displayed in the video, for instance the names of athletes. Although the subsequent images pass on different information from the surrounding texts, they describe the same objects from different perspectives. Furthermore, big data are usually of huge amounts. For example, Facebook, the famous social websites, collects about 500 terabytes (TB) data every day [4]. These features of big data bring a challenging issue to clustering technologies.

Clustering is designed to separate objects into several different groups according to special metrics, making the objects with similar features in the same group [5, 6]. Clustering techniques have been successfully applied to knowledge discovery and data engineering [7]. With the increasing popularity of big data, big data clustering is attracting much attention from data engineers and researchers. For example, Gao et al. [8] designed a graph-based co-clustering algorithm for big data by generalizing their previous image-text clustering method. Chen et al. [9, 10] designed a Computer Science and engineering, Vemana institute of technology nonnegative matrix tri-factorization algorithm to cluster big data sets by capturing the correlation over the multiple modalities. Zhang et al. [11] proposed a high-order clustering algorithm for big data by using the tensor vector space to model the correlations over the multiple modalities. However, it is difficult for them to cluster big data effectively, especially heterogeneous data, due to the following two reasons. First, they concatenate the features from different modalities linearly and ignore the complex correlations hidden in the heterogeneous data sets, so they are not able to produce desired results. Second, they often have a high time complexity, making them only applicable to small data sets. Thus, they cannot cluster large amounts of heterogeneous data efficiently.

To tackle the above problems, this paper proposes a privacypreserving high-order PCM scheme (PPHOPCM) for big data clustering. PCM is one important scheme of fuzzy clustering [12, 13]. PCM can reflect the typicality of each object to different clusters effectively and it is able to avoid the corruption of noise in the clustering process [14]. However, PCM cannot be applied to big data clustering directly since it is initially designed for the small structured dataset. Specially, it cannot capture the complex correlation over multiple modalities of the heterogeneous data object. The paper proposes a high-order PCM algorithm by extending the conventional PCM algorithm in the tensor space. Tensor is called a multidimensional array in mathematics and it is widely used to represent heterogenous data in big data analysis and mining. In this paper, the proposed HOPCM algorithm represents each object by using a tensor to reveal the correlation over multiple modalities of the heterogeneous data object. To increase the efficiency for clustering big data, we design a distributed HOPCM algorithm based on MapReduce to employ cloud servers to perform the HOPCM algorithm. However, the private data tends to be in disclosure when performing HOPCM on cloud. Take the medical data which is a typical type of big data for example. A large amount of private information such as personal email address and diagnostic data is included in the medical records. The disclosure of the private information will threaten people's lives and property greatly. Therefore, to protect the private data on cloud, we propose a privacy preserving HOPCM scheme by using the BGV technique that is of high efficiency [15]. Unfortunately, BGV does not support the division operations and square root operations that are the necessary computation in the functions for updating the membership matrix and clustering centers in the HOPCM algorithm although it is a fully homomorphic encryption scheme. To tackle this issue, we use the Taylor's theorem to transform these functions to polynomial functions to remove these operations.

We conduct the experiments on the two representative big data sets, i.e., NUS-WIDE and SNAE2, to assess the clustering accuracy and efficiency of our algorithms by comparison with three representative possibilistic c-means clustering algorithms, namely HOPCM-15 [11], wPCM [14] and PCM [12]. Results demonstrate that HOPCM outperforms other algorithms in clustering accuracy for big data, especially for heterogeneous data. Furthermore, PPHOPCM can use cloud servers cluster big data efficiently without disclosure of the private data.

Therefore, our contributions are summarized as the following three aspects:

The conventional PCM algorithm cannot cluster heterogeneous data. Aiming at this problem, the paper proposes a high-order PCM algorithm by optimizing the objective function in the high-order tensor space for heterogeneous data clustering.

To employ cloud servers to improve the clustering efficiency, we design a distributed high-order possibilistic algorithm based on MapReduce.

To protect the sensitive data when performing HOPCM on the cloud platform, we develop a privacy-preserving high-order possibilistic c-means scheme by using the BGV encryption method.

The rest of the paper is organized as follows. Section II reviews the related work on the possibilistic c-means algorithm and big data clustering methods. HOPCM is illustrated in Section III and the distributed HOPCM scheme based on MapReduce is presented in Section IV. Section V illustrates the privacy-preserving HOPCM method and Section VI reports experimental results. The whole paper is concluded in Section VII.

II. RELATED WORK

This section reviews the related work on the possibilistic c-means algorithm and heterogeneous data clustering methods. As the preliminary, the PCM algorithm is described first, followed by the heterogeneous data clustering methods.

2.1 Possibilistic c-Means Algorithm

The possibilistic c-means algorithm is one of fuzzy clustering schemes. Different from the traditional clustering schemes which assign each object into only one group, fuzzy clustering schemes assign each object into multiple groups. Specially, the assignment of each object is typically a distribution over all the groups in the fuzzy clustering.

PCM is able to avoid the corruption of noise in the clustering big data sets. However, PCM is sensitive to initial parameters and usually produces a coincident clustering result [13]. Aiming at this problem, FPCM and PFCM were proposed by combining PCM and FCM [16]. Xie et al. [5] developed an enhanced PCM algorithm by grouping the data set into one main subset and one assistant subset to avoid the coincident result. In addition, PCM is not robust to the additional parameters. To tackle this problem, Yang et al. [17] proposed an unsupervised PCM scheme to improve the robustness of the conventional PCM algorithm. To cluster non-spherical data sets, some kernel-based possibilistic clustering algorithms have been proposed by mapping the objects of the data set into high order data space [18]. Other PCM variants include weighted PCM algorithm and sample-weighted PFCM algorithm [19, 20].

Although these algorithms can improve the performance of the conventional PCM clustering, they are all limited in the structured data clustering. Therefore, the paper proposes a high-order PCM algorithm to cluster heterogeneous data.

2.2 Big Data Clustering

Over the past few years, some algorithms have been proposed for big data clustering, especially for heterogeneous data sets. Early works focused on image-text co-clustering by information fusion [10]. Specially, many algorithms first extracted the image features and the text features separately, and then concatenated them into a single vector [21]. However, these methods are difficult to produce desired clustering results since they cannot capture the complex correlations over the bi-modalities of the objects by concatenating the features in linear way. To tackle this problem, Jiang and Tan [22] proposed two methods based on the vague information and the Fusion ART to learn the visual-textual correlations by measuring the image-text similarities.

Most of heterogeneous data clustering schemes are developed depending on graph theory. They usually transform the heterogeneous data clustering task into a graph partitioning problem. The most representative scheme of this type is the bipartite graph partition scheme proposed by Gao [8] for image-text clustering by interpreting the clustering task as a tripartite graph. Afterward, they extended this method for heterogeneous data clustering. The similar work is the isoperimetric co-clustering algorithm proposed by Rege et al. [23]. This algorithm clusters heterogeneous data by solving a set of linear equations. In addition, Cai et al. [24] developed a spectral clustering algorithm, which is a representative method based on graph theory, for heterogeneous data clustering by designing an iterative process to optimize a unified objective function. Another graph theory based method is spectral relational clustering (SRC) presented by Long et al. [25]. SRC first produces a collective clustering and then achieves the final result by deriving an iterative spectral clustering. The major weakness of the heterogeneous data clustering algorithms based on the graph theory is that they have a high time complexity. Therefore, they are often inefficient for large amounts of data.

Another kind of heterogeneous data clustering is based on the matrix factorization theory. For instance, Chen et al. [9] presented a clustering algorithm based on non-negative matrix factorization scheme for heterogeneous data clustering by minimizing the global reconstruction function of the relational matrix over multiple modalities [26]. Other heterogeneous data clustering techniques, such as the combinatorial Markov random fields (Comrafs), are depending on information theory [27]. Similar to the methods based on graph theory, these algorithms still have a high time complexity. For example, the computational complexity of Comrafs will increase significantly with the growing amount of heterogeneous data.

III. HIGH-ORDER POSSIBILISTIC C-MEANS ALGORITHM BASED ON TENSOR REPRESENTATION MODEL

In this part, we present the HOPCM scheme for heterogeneous data clustering based on the tensor data representation model. The tensor data model represents each object by using a tensor [28]. For example, a colorful image can be represented as a 3-order tensor RIw×Ih×Ic, where Iw,Ih, and Ic denote width, height and color space, representatively. Specially, an image with 560 × 480 in the RGB color space can be represented by R560×480×3. Furthermore, a piece of video with MPEG-4 format can be represented as a 4-order tensor RIw×Ih×Ic×If with If denoting the frames. The tensor model can represent any heterogeneous data object. More importantly, it can capture the complex correlations over the multiple modalities of each heterogeneous data object. The tensor-based representation models have been successfully used in big data analysis and mining in past few years [3, 11, 26]. Therefore, HOPCM extends the conventional possibilistic c-means algorithm using the tensor data representation model.

Similarly, we can get the equation for updating vi with the same format as Eq. (3).

Algorithm 1: The High-order Posssibilistic c-Means Algorithm.

```
Input: $X = \{X_1, X_2, \ldots, X_N\}$, $c$, $m$, $maxiter$
Output: $U = \{u_{ij}\}, V = \{v_i\}$
for $iteration = 1, 2, \ldots, maxiter$ do
    fo $i = 1, 2, \ldots, c$ d
        $v_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}}$;
        $\eta_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} \times d_{(T)ij}^{2}}{\sum_{j=1}^{n} u_{ij}^{m}}$;
    fo $i = 1, 2, \ldots, c$ d
        fo $j = 1, 2, \ldots, n$ d
            $u_{ij} = \frac{1}{(1 + (d_{(T)ij}^{2} / \eta_i)^{1/(m-1)})}$;
```

From Algorithm 1, the time complexity of HOPCM is controlled by calculating the tensor distance betweenxk and vi, which needs O(n × c) for each iteration. Therefore, HOPCM has a computational complexity of O(tn × c) with t iterations.

IV. DISTRIBUTED HIGH-ORDER POSSIBILISTIC CMEANS ALGORITHM BASED ON MAPREDUCE

In this part, to increase the efficiency of HOPCM for big data, we design a distributed high-order prossibilistic c-means algorithm (DHOPCM) based on MapReduce which is an efficient cloud computing programming model for massive data computation
[30].

Algorithm 1 shows that the most important steps of HOPCM is to calculate the membership matrix and the clustering centers. Therefore, we use the Map function to calculate the membership matrix and use the Reduce function to calculate the clustering centers. The whole scheme is outlined in Fig. 1.

From Eq. (8) for updating the membership matrix, only the object xi and clustering centers V = {v1,v2,...,vc} are required for calculating the membership values of the object xi towards each clustering center. Therefore, to reduce the storage of each computing node and the communication, we partition the membership matrix into p sub-matrices U = {U1,U2,...,Up} by columns. The dataset X is also partitioned into p subsets X = {X1,X2,...,Xp} accordingly. As shown in Fig. 1, in the Map phrase, we dispatch each sub-matrix, the corresponding subset and all the clustering centers to one computing node for updating the membership matrix.

Eq. (3) shows that it requires all the data objects to calculate the clustering center vi. Therefore, the communication will increase significantly if calculating the clustering centers directly in the Reduce function. So, two parameters, i.e.,
$\alpha i = \{\alpha i^{(1)},\alpha i^{(2)},...,\alpha i^{(p)}\}$ and $\beta i = \{\beta i^{(1)},\beta i^{(2)},...,\beta i^{(p)}\}$,

V. PRIVACY-PRESERVING HIGH-ORDER POSSIBILISTIC C-MEANS ALGORITHM BASED ON BGV

In the last part, we present a DHOPCM scheme based on MapReduce, which can significantly increase the efficiency for clustering big data by employing many cloud servers. However, private data usually suffers from disclosure when performing DHOPCM on cloud. Specially, there is huge scale of sensitive heterogeneous data, such as medical information and clinical charts, in medical area, which are vital to patients. Once they are leaked, patients lives and property will suffer a great threat [31].

To protect the private data, we devise a privacy-preserving HOPCM scheme (PPHOPCM) based on BGV in this section. The proposed scheme cannot only employ cloud servers to increase the clustering efficiency for large amounts of heterogeneous data, but also avoid the disclosure of the private data. BVG secure operations required for PPHOPCM are described first, followed by the details of the proposed scheme.

5.1 BGV Secure Operations

BGV is a leveled fully homomorphic encryption technique. It uses a Setup procedure to select a μ-bit modulus q and the following parameters: the dimension $n = n(\lambda,\mu)$, the degree $d = d(\lambda,\mu)$, the distribution $\chi = \chi(\lambda,\mu)$, and $N = \lceil(2n + 1)\log q\rceil$.

Furthermore, a key Switching procedure and a modulus Switching procedure are implemented in the BGV scheme. The former is used to reduce the dimension of the ciphertext while the latter is aims to reduce the noise.

The BGV technique has four major secure operations, i.e., encryption, decryption, secure addition and secure multiplication, required for implementing our proposed PPHOPCM scheme, listed as follows [15].

(1) Encryption: Encrypt a plaintext $m \in R2$ as a ciphertext $c \leftarrow m + ATr \in Rqn+1$.

(2) Decryption: Decrpt a ciphertext c to its plaintext $m \leftarrow ((< c,sj > \bmod q) \bmod 2)$ using the corresponding secret key sj.

(3) SecureAddition: Add two ciphertexts, i.e., c1 and c2, to their sum c4 on cloud by $c3 \leftarrow c1 + c2 \bmod qj$, and $c4 \leftarrow \mathrm{Refresh}(c3,\tau(sj' \rightarrow sj-1),qj,qj-1)$.

(4) SecureProduct: Multiply two ciphertexts, i.e., c1 and c2, to their product c4 on cloud by $c3 \leftarrow c1 \otimes c2 \bmod qj$, and $c4 \leftarrow \mathrm{Refresh}(c3,\tau(sj' \rightarrow sj-1),qj,qj-1)$.

Compared with other encryption schemes, there are two major advantages of the BGV technique. First, it is a fully homomorphic encryption scheme which supports an arbitrary number of addition operations and multiplication operations simultaneously. Second, BGV is more efficient than most of other encryption schemes. So, BGV is used to encrypt the private data in this paper.

5.2 Complexity Analysis

Now, we estimate the computation complexity and the communication complexity of the PPHOPCM scheme. We use ADD and MUL to represent the time cost of one addition operation and one multiplication, representatively.

Computation Cost. Assume that the dataset X = {x1,x2,...,xk}, each represented by a T-order tensor Algorithm 4: The Privacy-preserving High-order Posssibilistic c-Means Algorithm.

VI. EXPERIMENTS

To estimate the clustering accuracy and efficiency of our schemes, we perform the proposed algorithms on the cloud platform including 20 nodes, each with 3.2 GHz Core i7 CPU and 8GB memory. We first compare our HOPCM algorithm with HOPCM-15, wPCM and PCM in clustering accuracy on two representative big data sets, i.e., NUS-WIDE and SNAE2. And then, we evaluate the clustering efficiency of the PPHOPCM algorithm by comparison with HOPCM and DHOPCM. At last, we estimate the scalability of PPHOPCM and DHOPCM based on speedup.

6.1 Data Sets and Evaluation Criteria

Two representative big data sets, i.e., NUS-WIDE and SNAE2, are used to estimate the clustering accuracy and efficiency of our schemes. NUS-WIDE is downloaded from Flikr.com [32]. It consists more than 260,000 images which are grouped by 81 classes. All the images are annotated by some texts, constituting a heterogeneous dataset. To evaluate the robustness of our proposed schemes, we sampled 80, 000 representative images which can be averaged to 8 subsets, each grouped by 14 categories, from NUSWIDE. SNAE2, downloaded from Youtube, includes 1800 pieces of videos, grouped by four classes, i.e., sport, news, advertisement and entertainment. Each video consists 100 frames, represented by a 4-order tensor in our schemes.

6.2 Performance Evaluation of HOPCM

The task of this experiment is to evaluate the clustering accuracy of HOPCM in terms of E∗ and ARI by comparison with three representative possibilistic clustering algorithms, i.e., HOPCM-15, wPCM, and PCM. HOPCM-15 is proposed by Zhang et al. [11] for heterogeneous data clustering while wPCM is a weighted PCM scheme. Different from our HOPCM scheme, HOPCM-15 learns features from heterogeneous data using improved auto-encoder model before clustering. For wPCM and PCM, we concatenate the attributes of each modality to form a single vector for clustering heterogeneous data.

Table 2 shows the clustering result on the NUS-WIDE based on E∗ while Table 3 presents the result based on ARI. From the results, we can observe that HOPCM produces the lowest values of E∗ and the highest values of ARI in most cases. Specially, HOPCM yields the E∗ value of 2.72 and the ARI value of 0.91, representatively, on the whole dataset. Such results imply that HOPCM performs best in clustering accuracy. PCM and wPCM perform worst in terms of E∗ and ARI since they cannot capture the complex correlations over the multiple modalities of heterogeneous objects only by concatenating their attributes. HOPCM-15 usually performs similarly with HOPCM in clustering the NUS-WIDE dataset, sometimes even better than HOPCM. This is because HOPCM-15 calculates the similarity between each object and the clustering centers by adopting tensor distance that can reveal the distributions for some heterogeneous objects. However, HOPCM-15 performs significantly less efficiently than HOPCM, which will be shown subsequently.

We perform the four algorithms on the different proportions of the NUS-WIDE dataset to evaluate the clustering efficiency. The result is shown in Fig. 2.

VII. CONCLUSION

In this paper, we proposed a high-order PCM scheme for heterogeneous data clustering. Furthermore, cloud servers are employed to improve the efficiency for big data clustering by designing a distributed HOPCM scheme depending on MapReduce. One property of the paper is to use the BGV technique to develop a privacy-preserving HOPCM algorithm for preserving privacy on cloud. Experimental results show PPHOPCM can cluster big data by using the cloud computing technology without disclosing privacy. In fact, for the large scale of heterogeneous data that does not require to be protected, the DHOPCM is more suitable since it is more efficient than PPHOPCM. The efficiency of PPHOPCM and DHOPCM can be further improved when using more cloud servers, making them more suitable for big data clustering, since they are of high scalability demonstrated by the experimental results.

In this work, the proposed schemes are preliminarily evaluated on two representative heterogeneous datasets. In the future work, the proposed algorithms will be further validated on larger actual datasets.

REFERENCES

[1] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, ”Data Mining with Big Data,” IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, Jan. 2014.

[2] B. Ermis, E. Acar, and A. T. Cemgil, ”Link Prediction in Heterogeneous Data via Generalized Coupled Tensor Factorization,” Data Mining and Knowledge Discovery, vol. 29, no. 1, pp. 203-236, 2015.

[3] Q. Zhang, L. T. Yang, and Z. Chen, ”Deep Computation Model for Unsupervised Feature Learning on Big Data,” IEEE Transactions on Services Computing, vol. 9, no. 1, pp. 161-171, Jan. 2016.

[4] N. Soni and A. Ganatra, ”MOiD (Multiple Objects Incremental DBSCAN) - A Paradigm Shift in Incremental DBSCAN,” International Journal of Computer Science and Information Security, vol. 14, no. 4, pp. 316-346, 2016.

[5] Z. Xie, S. Wang, and F. L. Chung, ”An Enhanced Possibilistic c-Means Clustering Algorithm EPCM,” Soft Computing, vol. 12, no. 6, pp. 593-611, 2008.

[6] Q. Zhang, C. Zhu, L. T. Yang, Z. Chen, L. Zhao, and P. Li, ”An Incremental CFS Algorithm for Clustering Large Data in Industrial Internet of Things,” IEEE Transactions on Industrial Informatics, 2015. DOI: 10.1109/TII.2017.2684807.

[7] X. Zhang, ”Convex Discriminative Multitask Clustering,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 1, pp. 28-40, Jan. 2015.

[8] B. Gao, T. Liu, T. Qin, X. Zheng, Q. Cheng, and W. Ma, ”Web Image Clustering by Consistent Utilization of Visual Features and Surrounding Texts,” in Proceedings of the 13th Annual ACM International Conference on Multimedia, 2005, 112-121.

[9] Y. Chen, L. Wang, and M. Dong, ”Non-Negative Matrix Factorization for Semisupervised Heterogeneous Data Coclustering,” IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1459-1474, Oct. 2010.

[10] L. Meng, A. Tan, and D. Xu, ”Semi-Supervised Heterogeneous Fusion for Multimedia Data Co-Clustering,” IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 9, pp. 2293-2306, Aug. 2014.

[11] Q. Zhang, L. T. Yang, Z. Chen, and Feng Xia, ”A High-Order Possibilistic-Means Algorithm for Clustering Incomplete Multimedia Data,” IEEE Systems Journal, 2015, DOI: 10.1109/JSYST.2015.2423499.

[12] R. Krishnapuram and J. M. Keller, ”A Possibilistic Approach to Clustering,” IEEE Transactions on Fuzzy Systems, vol. 1, no. 2, pp. 98-110, May 1993.

[13] R. Krishnapuram and J. M. Keller, ”The Possibilistic c-Means Algorithm: Insights and Recommendations,” IEEE Transactions on Fuzzy Systems, vol. 4, no. 3, pp. 385-393, Aug. 1996.

[14] Q. Zhang and Z. Chen, ”A Weighted Kernel Possibilistic c-Means Algorithm Based on Cloud Computing For Clustering Big Data,” International Journal of Communication Systems, vol. 27, no. 9, pp. 1378-1391, 2014.

[15] Q. Zhang, L. T. Yang, and Z. Chen, ”Privacy Preserving Deep Computation Model on Cloud for Big Data Feature Learning,” IEEE Transactions on Computers, vol. 65, no. 5, pp. 1351-1362, May 2016.

[16] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, ”A Possibilistic Fuzzy c-Means Clustering Algorithm,” IEEE Transactions on Fuzzy Systems, vol. 13, no. 4, pp. 517-530, Aug. 2005.

[17] M. Yang and C. Lai, ”A Robust Automatic Merging Possibilistic Clustering Method,” IEEE Transactions on Fuzzy Systems, vol. 19, no. 1, pp. 26-41, Feb. 2011.

[18] M. Filippone, F. Masulli, and S. Rovette, ”Applying the Possibilistic cMeans Algorithm in Kernel-Induced Spaces,” IEEE Transactions on Fuzzy Systems, vol. 18, no. 3, pp. 572-584, Jun. 2010.

[19] A. Schneider, ”Weighted Possibilistic c-Means Clustering Algorithms,” in Proceedings of the 9th IEEE International Conference on Fuzzy Systems, 2000, pp. 176-180.

[20] B. Liu, S. Xia, Y. Zhou, and X. Han, ”A Sample-Weighted Possibilistic Fuzzy Clustering Algorithm,” Acta Electronica Sinica, vol. 30, no. 2, pp. 371-375, 2012.

[21] R. Zhao and W. Grosky, ”Narrowing the Semantic Gap Improved TextBased Web Document Retrieval Using Visual Features,” IEEE Transactions on Multimedia, vol. 4, no. 2, pp. 189-200, Jun. 2002.

[22] T. Jiang and A.-H. Tan, ”Learning Image-Text Associations,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 2, pp. 161-177, Feb. 2009.

[23] M. Rege, M. Dong, and J. Hua, ”Graph Theoretical Framework for Simultaneously Integrating Visual and Textual Features for Efficient Web Image Clustering,” in Proceedings of the 17th international conference on World Wide Web, 2008, pp. 317-326.

[24] X. Cai, F. Nie, H. Huang, and F. Kamangar, ”Heterogeneous Image Feature Integration via Multi-Modal Spectral Clustering,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1977-1984.

[25] B. Long, X. Wu, Z. Zhang, and P. Yu, ”Spectral Clustering for Multi-Type Relational Data,” in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 585-592.

[26] Q. Gu and J. Zhou, ”Co-Clustering on Manifolds,” in Proceedings ofthe 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , 2009, pp. 359-367.

[27] R. Bekkerman, M. Sahami, and E. Learned-Miller, ”Combinatorial Markov Random Fields,” in Proceedings of the 17th European Conference on Machine Learning, 2006, pp. 30-41.

[28] L. Kuang, F. Hao, L. T. Yang, M. Lin, C. Luo, and G. Min, ”A
Tensor-based Approach for Big Data Representation and Dimensionality Reduction,” IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 3, pp. 280-291, Sept. 2014.

[29] Y. Liu, Y. Liu, and K. Chan, ”Tensor Distance based Multilinear Localitypreserved Maximum Information Embedding,” IEEE Transactions on Neural Network, vol. 21, no. 11, pp. 1848-1854, Nov. 2010.

[30] J. Dean and S. Ghemawat, ”MapReduce: Simplified Data Processing on Large Clusters,” Communications of the ACM, vol. 51, no. 1, pp. 107-113, 2008.

[31] J. Yuan and S. Yu, ”Privacy Preserving Back-propagation Neural Network Learning Made Practical with Cloud Computing,” IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 1, pp. 212-221, Jan. 2014.

[32] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, ”NUSWIDE: a Real-World Web Image Database from National University of Singapore,” in Proc. of ACM International Conference on Image and Video Retrieval, 2009, pp. 1-9.