Big Data Clustering with Privacy Preserving

Pakruddin .B[1] Shaik Arifa Banu[2] Suhana Tabassum[3] Swathi Kumari .M[4] Sadiya Firdose[5]
[1]Assistant Professor
[1,2,3,4,5]Department of Computer Science & Engineering
[1,2,3,4,5]HKBKCE, Bangalore, India

Abstract— Fuzzy clustering is an important technique used to cluster big data, pattern recognition and image processing. As big data is in explosive growth the clustering of big data is mandatory. However it is difficult to cluster large amount of heterogeneous data efficiently. To tackle this problem the proposed system clusters the data based on the keyword specified. This allows the data to be stored in the structured form which makes the data access easier. Further we preserve privacy on cloud by encrypting the data before uploading on the cloud with the help of AES encryption technique.

Key words: AES, PCM, FCM

I. INTRODUCTION

The social websites such as twitter, Facebook and Instagram are achieving hikes and has become popular, with the explosive growth big data, big data consist of different kinds of data which is heterogeneous in nature. Each object in big data set is multi model. The Big data sets includes various interrelated kind of object which is in the form of audio, image and texts resulting in high heterogeneity in terms of structured form which includes both unstructured data and structured data. The object with different type will carry different information where each object are interrelated with other object. For example, consider a small piece of sports video with meta-information where the video may include large number of ordered images to exhibit the procedure and Make use of meta-information, like notation and close text, to exhibit the extra information which cannot be exposed in the video, example the names of athletes. While the images with different information surrounded by texts may describe the same object from different perspectives. More over big data consist of large amount of data. Now a days the social websites like Facebook, twitter collects up to 600tera bytes of data per day.

In order to divide targets into various groups with respect to special metrics and formulating the objects with the same type of characteristics in the same group, clustering is developed. These clustering approach have been given the positive result in knowledge discovery and data manipulations. Nowadays there is a massive increment of big data, as it is the challenging task to cluster the big data and a lot of researchers and engineers are paying interest to cluster the big data. Let us take an example of existing system such as Timm. Suggested the two algorithms of possibilistic fuzzy clustering which will nullify the problem of cluster in PCM. GAO designed a graph-based co-clustering algorithm for big data by popularizing the old clustering method of image text. Zhang proposed a high-order clustering algorithm for big data by using the tensor vector space to model the correlations over the multiple modalities. By contrast, it is not an easy task to cluster the homogeneous data efficiently, because firstly it sequence the characteristics from the various modes and neglecting the complex correlation invisible in data sets and therefore the required result will not be produced. Secondly, it has a high time ramification, which make them suitable for small amount of data sets. Hence, it is difficult to cluster the big data efficiently.

These problems can be handled by the by the proposed system, Privacy Preserving High-Order PCM (PPHOPCM) which clusters the data. PCM plays an important role in fuzzy clustering by reflecting the characteristics of each object into separate clusters. It can also ignore the noise corruption during the process. Since PCM is been designed for small structured datasets, it cannot be used directly for big data clustering. PCM cannot conquer the complex correlation over multiple modalities of data object in heterogeneous. Therefore High Order PCM scheme is proposed which extends the above PCM by using the tensor space. Mathematically a multidimensional array is known as a tensor.

Heterogeneous data is been represented in big data analysis and data mining. Distributed HOPCM increases the efficacy of clustering big data by using the map reduce technique which employs the cloud servers to execute the HOPCM algorithm. Here the private is been disclosed during the execution of HOPCM, to protect the data on cloud. This is done by the proposed Privacy Preserving HOPCM algorithm which uses AES encryption. Which uses Taylor's theorem to update the membership matrix and clustering centers.

Here we demonstrate that HOPCM outperforms other algorithms in terms of clustering efficiency for big data. The main important experiments conducted on the heterogeneous data set are as follows:

1) NUS-WIDE
2) SNA-2

These two methods are used for assessing clustering accuracy and efficiency of algorithms by comparing with the three representative possibilistic c means algorithms. Namely, PCM, HOPCM, wPCM.

We can summarize the whole process as follows:

- A high order PCM algorithm is applied by optimizing the objective function in the tensor space for heterogeneous data clustering, which a conventional PCM fails to do.
- Cloud servers are employed to improve the clustering efficiency and also a distributed HOPCM algorithm is designed based on MAP-REDUCE technique.
- PPHOPCM algorithm is developed to protect the sensitive data while performing HOPCM on the cloud platform.
- This is done by using AES encryption method.

II. LITERATURE SURVEY

In the year 1997, [1] the system proposed fuzzy possibilistic c-means (FPCM) framework and this algorithm which produces the membership matrix and clustering data centers. FPCM limit the typical values so as the overall data points to form a cluster. The row sum limit generates the artificial

values for large data sets. This system also proposed the model called possibilistic fuzzy c means (PFCM) structure. PFCM generates both possibilities and membership at the same time, also with the point prototypes or clustering centers for all clusters. PFCM is a crossing of possibilistic c-means (PCM) and fuzzy c-means (FCM), which averts the several issues of PCM, FCM and FPCM. PFCM figure out the noise sensitivity defect of FCM, overtakes the clustering issues of PCM and terminates the row sum constrains of PFCM. By deriving the first order condition for extreme point of PFCM function and can be used as basis for optimization approach to determine the minima of PFCM function. FCM and PCM can be compared with PFCM in terms of different numerical examples. These examples will show that PFCM can be compared with both the previous models. Hence the PFCM prototypes are comparative of little sensitive to outliers, but PFCM will not be able to cluster the large amount of data.

The system [2] proposed the weighted possibilistic c-means algorithm (WPCM). The weight will identify the possibility of a given vector which is acceptable to all the clusters. By giving less weight values to outliers, it will decrease the effects on noisy data of clustering procedures. Here, it is viewed as the possibilistic c-means algorithm is a important method of the weighted possibilistic c-means algorithm, if vector weight is assigned to one method. Different methods for finding the weight values are shown. The accuracy of the algorithm is determined with the help of a Gaussian random number generator with outliers and unreal data set that contains outliers. However, finding the weighted values will still give the greatest uncertainty which has been the case for different work prior to the evolution of WPCM approach.

Big data contains huge amount of data, variety of data and requires processing with a high velocity. Deep learning works in vector space and is applied for image classification, feature learning in speech recognition and also in processing of language. Here deep learning is used feature learning in big heterogeneous data. It uses tensor distance as the average sum of squares encourages intermediate representation. It also trains the parameters in auto tensor encoder, from vector space to the tensor space of high order. Tensor auto encoder model contains of an input layer, an output layer and a hidden layer. Each layer is represented by a tensor. High order back-propagation algorithm is used for efficient computation of the partial derivatives. However several experiments demonstrate that the tensor auto encoder model is efficient for feature learning for heterogeneous data [3].

Therefore clustering of data should be done for efficient processing of data. [MTC] multitask clustering is good at clustering each individual task and also collects the relationships among the tasks from the data automatically. It is used in solving [DMTFC] Discriminative multitask feature clustering, and convex relationships of discriminative multitask. It aims to learn about the shared featured representation and task relationships. It selects the solutions by optimizing via cutting plane scheme, optimizing via extended level method, optimizing via alternative method. The discriminative multitask clustering specifies four assumptions namely DMT feature clustering works under multivariate models across tasks, while the DMT Relationship clustering models the relationship of the task. MIP problems are briefed to convex optimization problems. Therefore multi-domain sentiments datasets show the effectiveness of this system [4].

Big data concerns to huge amount of data, different kinds of data. The speed of unprocessed data learning have many important issues. To overcome these issues, a deep computation model is introduced for learning characteristics on big data effectively. Owning to the large amount of data in the intelligent and high computational ramification, the heavy computation model finds it uneasy to execute in real-time with limited computing power and memory storage. Therefore, to enhance the efficiency of big data future learning, a privacy preserving deep computation model is introduced by offloading the valuable operations to the cloud. Privacy reference becomes obvious because there are a huge number of confidential data by several applications in the smart city, such as private data of governments or proprietary information of organization. To preserve the private data, the system uses the AES encryption scheme to encrypt the private data and apply cloud servers to execute the high-order algorithm on the encrypted data effectively for heavy computation model preparation. In this system only the encryption operations and the decryption operations are performed by the user while all the calculation jobs are executed on the cloud. Observational output show that the system is enhanced by around 2.5 times in the grooming high octane equate to the conventional heavy computation model without exposing the private data using the cloud computing involves ten nodes. Significantly, this strategy is extremely scalable by connecting more cloud servers that is especially eligible for big data. Developing the smart city is important to improve the efficiency, dependability, and security of a smart city. Smart city comprises of intelligent transfers, smart grid, intelligent protection etc. With the evolution of these subjects, past decades have proved the extraordinary increments of smart cities. With the large implementation of different mobile gadgets like sensors and RFID, information are being gathered at new rate in the smart city. Hence, is sarcastic for smart city development, supervising and dominating to design big data modeling and analytic technologies. As a fundamental technique of big data, feature learning can discover the underlying structure of big data to provide intelligent decision for developing smart city systems [4].

Today, cloud computing has to play a vital role in big data modeling and analytic. It has been successfully applied in industrial products and commercial fields that take advantage of big data. For example, with cloud computing, Google offers a wide variety of real-time services such as real-time searching real-time translation and voice recognition. Cloud computing provides us with strong computing power and massive storage space. Therefore, it is an effective method to improve the efficiency of training deep computation model for big data feature learning by offloading the expensive operations to the cloud. However, privacy concerns bring forward in the cloud computing since there exist a large number of private data collected from the smart city, such as population and economic information. These data may contain sensitive data of governments or proprietary information of the enterprises once they are disclosed,

people's lives and property will be seriously threatened. Especially in the smart city, disclosure of sensitive data is not only a privacy issue but of legal concerns according to privacy protection laws such as the Health Insurance Portability and Accountability Act (HIPAA). Therefore, we focus on the privacy preserving deep computation model with the cloud computing. The privacy preserving deep computation model poses a number of issues and challenges, especially for big data feature learning by incorporating the computing of the cloud. We discuss the key challenges in three aspects as follows:

1) To protect the private data and intermediate results it requires secure computation of various operations needed by the deep computation model.
2) To improve the efficiency of deep computation model training and big data feature learning, it requires to choose the efficient full homomorphic encryption scheme according to the major operations of the algorithms in the privacy preserving deep computation model.
3) To produce the correct result on the cipher texts using the full homomorphic encryption scheme, the sigmoid function is required to approximate as a new function involving only addition operations and multiplication operations [6].

To enhance the perfection of learning results, it performs more than one user to collaborate via Joint Back Propagation neural network learning with respective of their Data sets. BG back propagation is an efficient method for learning neural networks which is been used in most of the application. BGN has two stages that is (a) feed forward and (B) error back propagation. The Feed forward neural network ensures that the connection between the nodes do not form a cycle. Error back propagation is used for the calculation of the weight to be used on the network. The collaborative learning improves learning efficiency when compared to learning data set locally. The system what it does is the party uploads their data on to the cloud, but the party doesn't want to disclose his/her private data to others. The participating parties can carry not only their data set they can also carry others data set too. The existing system provides the collaboration of parties in the way of considering two parties or by data partition. The system provides the solution to problem by utilizing the power of cloud computing that is where each party uploads the data in the form of cipher text locally to cloud which is been encrypted. Then the cloud performs most of the operations on the cipher text without knowing the original private data, the communication and computation cost on each party is minimal and independent to the number of participants. The system has to face three challenges i.e.,

1) The private data of each participant is to be protected and the required result is generated during the BPN learning process, and it requires various operation to have a secure computation.
2) It ensures that the computation and communication cost for each is affordable.
3) The training data set and for collaborative training which can be owned by different parties and partitioned not only in single way it can be done in arbitrary way [7].

III. PROBLEM STATEMENT

Clustering is necessary to group the similar items into a separate matrix in order to make the processing of data easier. Today's growth in big data leads to great difficulty in clustering the data, especially the heterogeneous form of data. To cluster the datasets, the process of capturing of correlation among different modalities and to identify the hidden complexities is required for an efficient clustering. By concatenating the characteristics of different modalities linearly, the complex correlation that are been hidden are been ignored in the datasets. This makes the results to be inaccurate. With high time complexity it is been used only for the small datasets, hence heterogeneous data in large amount cannot be clustered effectively. There is a possibility of data getting accessed by an unauthorized user, hence there is no privacy given to data access.

IV. OBJECTIVES

The main objectives of Big Data Clustering with Privacy Preserving is to retrieve the user required information from the large amount of data which is stored in cloud as well as the database. It also includes the following targets-
To cluster the heterogeneous data, which involves text, image, video, audio, etc.

1) It also improves the clustering efficiency.
2) To protect the data on cloud platform which is encrypted with the help of AES encryption algorithm.
3) To increase the clustering accuracy for big data, especially for heterogeneous data.
4) To retrieve the data effectively from the application.

V. SYSTEM ARCHITECTURES

The system architecture is a conceptual model that specifies structure, behavior and other aspects of a system. The general architecture of Big Data Clustering with Privacy Preserving is shown in the figure below. Admin will upload any files such as text, audio, video, image etc. to the cloud from the application, the data will be updated in the database as well, the uploaded data will be in the form encrypted format, the admin will provide the keyword for each data while uploading.

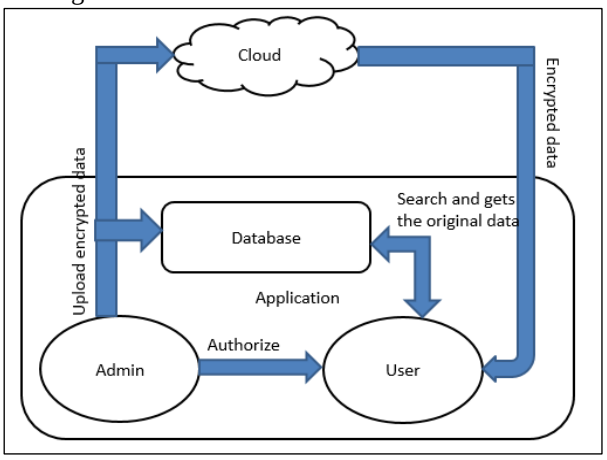

Fig. 5.1: system architecture of Big Data Clustering with Privacy Preserving

The user has to register in the application and should get the authorization from the admin only then the user will

be able to download the data from the application. For each user the unique secret key is generated. The user should know the secret key, password and keyword to download the data. If any unauthorized user tries to download the data directly from the cloud then he/she will be able to download the encrypted data. Hence, the privacy is preserved by accessing the application.

VI. MODULE DESCRIPTION

A. Uploading of data

The admin logs in to application with a login id and password. Once the admin logs into the application a list of options appear, which includes file uploading, downloading, viewing the user requests, files and the user search history. The admin uploads the file with the file name and the keyword. The file can be of any type which includes audio, video, image or text file. The file is encrypted and then uploaded successfully on the cloud and stored in the database as well. The admin can view the user requests and authorize the users whom he wants to provide access to the application.

B. Clustering of data

The clustering of data can be done based on the keyword specified by the admin. All the related data to that keyword is grouped into a single cluster. A number of clusters are been created based on the keywords provided. This type of clustering method makes the processing of uploaded heterogeneous data. With this method data is been stored in a structured form so that it makes data access simpler. The new user enrolls into the application by providing all his details. Once the user is permitted successfully he /she can enter the application by using the user login id. The user fetches the keyword of the particular cluster, if he wants download the data. The user can download the data based on the random secret key generated by the algorithm.

C. Privacy preserving

Initially we perform AES encryption technique to encrypt the data before it is uploaded. The encrypted data is then updated in the database as well as in the cloud. If the user wants to access the data he must be authorized to application. An unauthorized user cannot enter into the application. If he wants to download from the cloud he gets an encrypted data instead of original data. The privacy is preserved on the cloud. The user can download the data successfully using the secret key generated by the algorithm. And each key is unique to the different user. So that the data cannot be disclosed and it can be more confidential.

VII. CONCLUSION AND FUTURE SCOPE

In the Proposed system we are clustering the different types of data i.e.; huge amount of heterogeneous data also known as Big data. Big data includes audio, video, images etc….The clustering of data can be done based on keywords specified by the admin. Based on these keywords provided all the data related to that keyword is grouped into a single cluster .Similarly the clustering is done for all the other data set based on the specified keywords .Based on the keyword provided by the admin all the data related to the keyword is grouped into a single cluster. Similarly the clustering is done for different data sets based on the keywords specified we also employee cloud servers for uploading the clustered data. Initially we perform AES encryption technique to encrypt the data to preserve its privacy Before the data is being uploaded on the cloud .once the data is uploaded successfully on the cloud.it is updated in the MySQL database as well .Since the data uploading is done by the admin the admin himself has the authority to authorize the users who can enter the application and access and download the data. If the user is not authorized by the admin he/she will not be able to view the original data content. Instead they can only access the encrypted data. In the future work, this proposed application will be further validated on huge data sets.

VIII. SCREENSHOTS

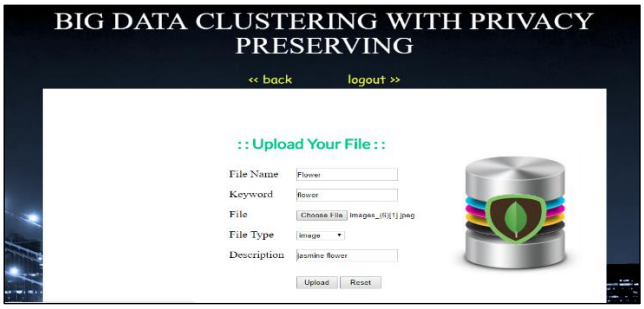

Fig 8.1: Uploading data from application

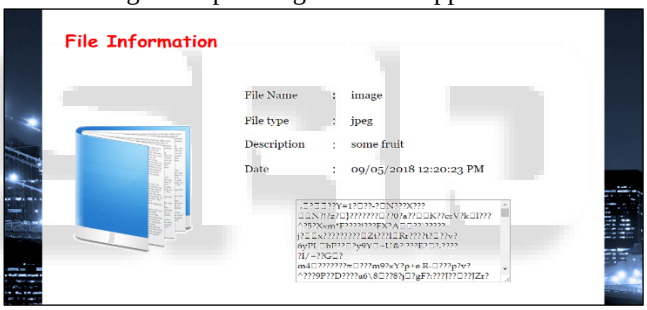

Fig 8.2: Encrypted data in application

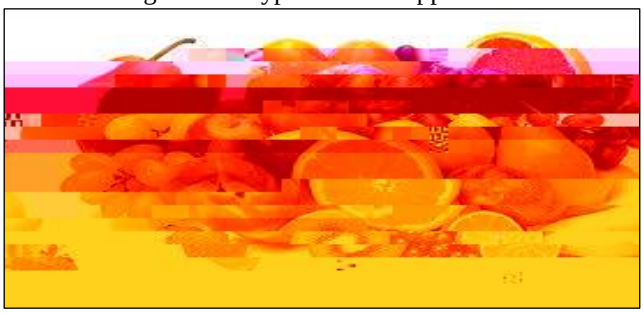

Fig 8.3: Encrypted image on cloud

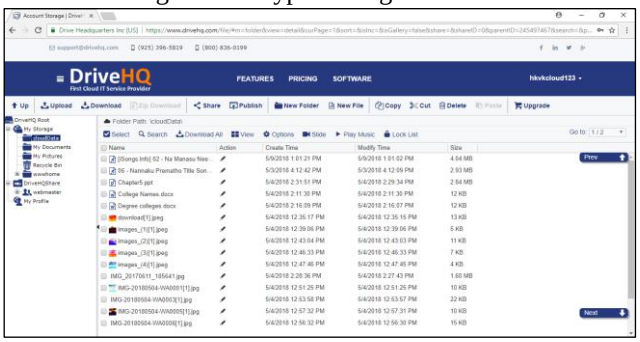

Fig. 8.4: Uploaded data on cloud

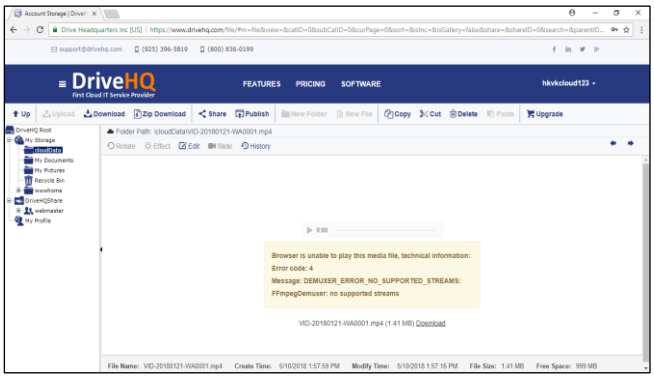

Fig. 8.5: Encrypted video on cloud

REFERENCES

[1] N.R.Pal, K.Pal, J. M. Keller, and J. C. Bezdek, "A PossibilisticFuzzyMeansClusteringAlgorithm,"IEEETransactionsonFuzzySystems,vol. 13, no. 4, pp. 517-530, Aug. 2005.

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

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

[5] 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.

[6] 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.

[7] Palak Sachar Computer Science and Engineering CT Group of Institution, Jalandhar Punjab, India Palakchp4@gmail.com" Social Media Generated Big Data Clustering Using Genetic Algorithm" IEEE Transactions on Knowledge and Data Engineering, Vikas Khullar Computer Science and Engineering CT Group of Institution, Jalandhar Punjab, India Vikas.khullar@gmail.com jan. 2017.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

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

[24] 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.

