IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol.-7, Issue-3, March 2019 E-ISSN: 2347-2693

Effective Data Clustering and Efficient Security scheme in Cloud Computing

V. Prasathkumar[1*], K.Senthil[2], S.Vignesh[3], P.Ranjith Roshan [4], E.Prakash[5]

[1,2,3,4,5] Department of Information Technology, Sri Shakthi Institute of Engineering and Technology, Anna University, Coimbatore, India

[*]*Corresponding Author:* prasathkumar@siet.ac.in*Tel.: 9095330712*

DOI: https://doi.org/10.26438/ijcse/v7i3.955960 | Available online at: www.ijcseonline.org

Accepted: 09/Mar/2019, Published: 31/Mar/2019

***Abstract*—** As one important technique of fuzzy clustering in data mining and pattern recognition, the possibility c-means algorithm (PCM) has been widely used in image analysis and knowledge discovery. However, it is difficult for PCM to produce a good result for clustering big data, especially for heterogeneous data, since it is initially designed for only small structured data set. To tackle this problem, the paper proposes a high-order PCM algorithm for big data clustering by optimizing the objective function in the tensor space. Further, we design a distributed HOPCM method based on Map Reduce for very large amount of heterogeneous data. Experimental results indicate that PPHOPCM can effectively cluster numerous heterogeneous data using cloud computing without disclosure of private data.

***Keywords*—** Clustering Big data ,Cloud Computing, possibilistic -means algorithm, Privacy preserving , Tensor space

I. INTRODUCTION

As personal computing technology and social websites, such as Face book and Twitter, become increasingly popular, big data is in the explosive growth. Bigdata are typically heterogeneous each object in bigdata setismulti-modal. Specially, big data set include various interrelated kinds of objects, such as texts, images, and audios, resulting in high heterogeneity regarding structure form, involving structured data and unstructured data. Moreover, different types of objects carry different information while they are interrelated with each other. Furthermore, big data is usually of huge amounts. This feature of bigdata brings a challenging issue to clustering technologies. Clustering is designed to separate objects into several 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 attracted much attention from data engineers and researchers. For example, GAO designed a graph-based co-clustering algorithm for big data by generalizing their previous image-text clustering method. Chen et al. designed a non-negative matrix tri-factorization algorithm to cluster big data set by capturing the correlation over the multiple modalities. Zhang et al. proposed a high-order clustering algorithm for big data by using the tensor vectors 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. Here, a comprehensive survey has been made Section I contains the introduction of Effective Data Clustering and Efficient Security scheme in Cloud Computing Section II contain the literature survey, Section IIII represents the workflow of Effective Data and Security in Cloud computing Section IV addresses the methods. Finally, Section V concludes this article.

II. LITERATURE SURVEY

PAPER 1: An Incremental CFS Algorithm for Clustering Large Data in Industrial Internet of Things

In this paper, a latest clustering algorithm based on fast finding and searching of density peaks (CFS) is improved to cluster a large number of dynamic data in industrial Internet of Things; and CFS is proposed by Alex and Laio in Science 2014. It can find clusters of arbitrary shape and determine the number of clusters automatically. Some experiments have demonstrated its superiority in the efficiency and effectiveness over the previous algorithms for clustering large amounts of data. Specially, an incremental CFS algorithm based on k-mediums (ICFSKM) is designed for clustering dynamic data collected from industrial Internet of Things. The major contributions are threefold: 1) to integrate the clustering result of new arriving objects into the previous one, two clustering adjustment operations, i.e. cluster creating and cluster combining, are defined depending on the different scenarios. 2) k-mediods is employed to modify the clustering centers according to the new arriving objects. 3) An iterative procedure is designed to combine each two clusters with the difference less than a predefined threshold for the final clustering pattern. Finally, the performance of our scheme is evaluated on three popular UCI datasets and two real datasets collected from industrial Internet of Things by comparison with two representative incremental clustering algorithms, i.e., incremental fuzzy clustering with multiple medoids (IMMFC) and incremental affinity propagation algorithm based on message passing (I-APKN), in terms of clustering accuracy and time.

PAPER 2: Non-Negative Matrix Factorization for Semi supervised Heterogeneous Data Co clustering

Coclustering heterogeneous data has attracted extensive attention recently due to its high impact on various important applications, such us text mining, image retrieval, and bioinformatics. However, data co clustering without any prior knowledge or background information is still a challenging problem. In this paper, we propose a Semi supervised Non-negative Matrix Factorization (SS-NMF) framework for data co clustering. Specifically, our method computes new relational matrices by incorporating user provided constraints through simultaneous distance metric learning and modality selection. Using an iterative algorithm, we then perform tri factorizations of the new matrices to infer the clusters of different data types and their correspondence. Theoretically, we prove the convergence and correctness of SS-NMF coclustering and show the relationship between SS-NMF with other well-known coclustering models. Through extensive experiments conducted on publicly available text, gene expression, and image data sets, we demonstrate the superior performance of SS-NMF for heterogeneous data coclustering.

PAPER 3: Semi-Supervised Heterogeneous Fusion for Multimedia Data Co-Clustering

Co-clustering is a commonly used technique for tapping the rich meta-information of multimedia web documents, including category, annotation, and description, for associative discovery. However, most co-clustering methods proposed for heterogeneous data do not consider the representation problem of short and noisy text and their performance is limited by the empirical weighting of the multi-modal features. In this paper, we propose a generalized form of Heterogeneous Fusion Adaptive Resonance Theory, called GHF-ART, for co-clustering of large-scale web multimedia documents. By extending the two-channel Heterogeneous Fusion ART (HF-ART) to multiple channels, GHF-ART is designed to handle multimedia data with an arbitrarily rich level of meta-information. For handling short and noisy text, GHF-ART does not learn directly from the textual features. Instead, it identifies key tags by learning the probabilistic distribution of tag occurrences. More importantly, GHF-ART incorporates an adaptive method for effective fusion of multimodal features, which weights the features of multiple data sources by incrementally measuring the importance of feature modalities through the intra-cluster scatters. Extensive experiments on two web image data sets and one text document set have shown that GHF-ART achieves significant better clustering performance and is much faster than many existing state-of-the-art algorithms.

Existing system

Proposed a high-request bunching calculation for huge information by utilizing the tensor vectors space to show the relationships over the various modalities. In any case, it is troublesome for them to bunch huge information adequately, particularly heterogeneous information, because of the accompanying two reasons. To begin with, they link the highlights from various modalities straight and overlook the mind boggles relationships covered up in the heterogeneous informational indexes, so they are not ready to deliver wanted outcomes. Second, they regularly have a high time multifaceted nature, making them just pertinent to little informational collections. Accordingly, they can't bunch a lot of heterogeneous information effectively.

Proposed System

Proposes a protection safeguarding high-request PCM plot (PPHOPCM) for huge information grouping. PCM is one vital plan of fluffy grouping PCM can mirror the commonality of each item to various bunches viably and it can keep away from the defilement of commotion in the bunching procedure. Be that as it may, PCM can't be connected to huge information grouping straightforwardly since it is at first intended for the little organized data set. Exceptionally, it can't catch the minding boggling relationship over numerous modalities of the heterogeneous information object. The paper proposes a high-request PCM calculation by broadening the ordinary PCM calculation in the tensor spaces. Tensor is known as a multidimensional exhibiting in arithmetic and it is generally used to speak to heterogeneous information in enormous information examination and mining.

III. WORK FLOW DIAGRAM

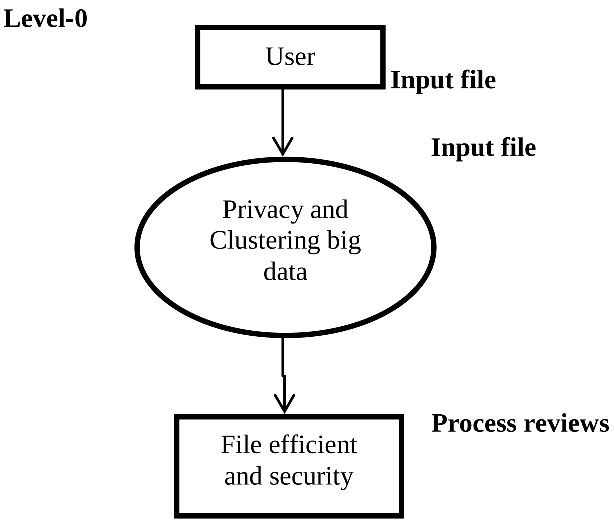


Figure 1.Level-0 Dataflow Diagram

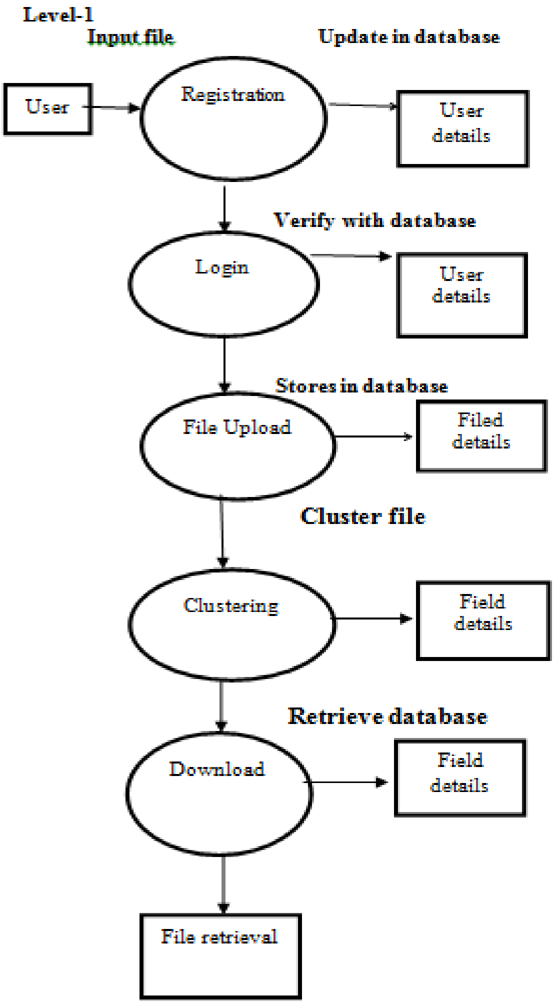


Figure 2. .Level-1Dataflow Diagram

IV. **Modules**

- User Interface Design
- Uploading data
- Encryption of data
- Data Search
- Clustering of data
- Top k result

1. User Interface Design

The fundamental thought of this module is to structure the UI for clients in the undertaking. The login page is to plan for information proprietor and information client. After the information proprietor logins into the framework, the page showed which enables the information proprietor to accomplish the encoded document transfer to the framework. At the point when the client logins to the framework, the framework enables the client to include the unscrambling key and characteristics for recovery of indicated record. Before getting to the record from framework, the client must enlist into the framework.

2. Uploading data

The information proprietor has an accumulation of n records to re-appropriate onto the cloud server in encoded structure and anticipates that the cloud server should give watchword recovery administration to information proprietor himself or other approved clients. To accomplish this, the information proprietor needs to manufacture an accessible list from a gathering of catchphrases removed out of documents, and after that redistributes both the encoded file and scrambled records onto the cloud server.

3. Encryption of data

We propose another accessible encryption conspire, in which novel advances in cryptography network are utilized, including information encryption and the vector space display. In the proposed plan, the information proprietor scrambles the accessible record with information encryption. At the point when the cloud server gets an inquiry comprising of multi key-words, it figures the scores from the scrambled record put away on cloud and after that profits the encoded scores of documents to the information client. Next, the information client decodes the scores and selects the best k most elevated scoring records' identifiers to demand to the cloud server. The recovery takes a two-round correspondence between the cloud server and the information client.

4 .Data Search

The information client is approved to process multi keyword recovery over the redistributed information. The registering power on the client side is constrained, which implies that activities on the client side ought to be streamlined. The approved information client at first produces an inquiry. For security thought, which catchphrases the information client has sought must be covered. In this manner, the information client scrambles the question and sends it to the cloud server that profits the pertinent documents to the information client. Thereafter, the information client can decode and make utilization of the documents.

5. Clustering of data

Bunching is intended to isolate objects into a few distinct gatherings as per extraordinary measurements, making the articles with comparative highlights in a similar gathering. Bunching strategies have been effectively connected to information disclosure and information designing. With the expanding prominence of huge information, enormous information grouping is drawing in much consideration from information designers and scientists.

6. Top k result

The substantial number of information clients and reports in the cloud, it is important to permit multikey-word in the pursuit inquiry and return archives in the request of their pertinence with the questioned catchphrases. Scoring is a characteristic method to weight the pertinence. In light of the pertinence score, records would then be able to be positioned in either ascending or descending. A few models have been proposed to score and rank records in IR people group. Among these plans, we embrace the most broadly utilized one tied weighting, which includes two properties term recurrence and converse archive recurrence.

Algorithm

Elliptic curve cryptography (ECC)

Elliptic curve cryptography (ECC) is an approach to public-key cryptography based on the algebraic structure of elliptic curves over finite fields. Elliptic curves are also used in several integer factorization algorithms that have applications in cryptography, such as Lenstra elliptic curve factorization.

Steps:

ECC domain parameters over GF(q), are a sextuple:

T = (q, a, b, G, n, h)

- $q = p$ or $q = 2^m$
- a and $b \in GF(q)$

$$y^2 \equiv x^3 + ax + b \pmod{p} \quad \text{for } q = p > 3$$
$$y^2 + xy = x^3 + ax^2 + b \quad \text{for } q = 2^m \geq 1$$

- a base point $G = (x_G, y_G)$ on $E_{(a,b)(}GF(q))$,
- a prime n which is the order of G

(The order of a point P on an elliptic curve is the smallest positive integer r such that rP = O.)

h = #E/n. where #E represents number of points on elliptic curve and is called the curve order.

ECC Key Generation

A public key $Q = (x_Q, y_Q)$ associated with a domain parameter (q, a, b, G, n, h) is generated for an entitiy A using the following procedure :

- Select a random or pseudo-random integer d in the interval [1,n-1].
- Compute Q = dG.
- A's public key is Q; A's private key is d.

ECC Key Validation

A public key $Q = (x_Q, y_Q)$ associated with a domain parameter (q, a, b, G, n, h) is validated for an entitiy A using the following procedure :

- Check that $Q \neq O$
- Check that x_Q and y_Q are properly represented elements of GF(q).
- Check that Q lies on the elliptic curve defined by a and b.
- Check that nQ = O.

Elliptic Curve Digital Signature Algorithm (ECDSA)

- Proposed by Abdalla, Bellare and Rogaway in 1999.
- Entity A has domain parameters D = (q, a, b, G, n, h) and

public key Q_A and private key d_A. And entity B has authentic copies of D and Q_A.

To sign a message m, A does the following:

- Select a random integer k from [1,n-1].
- Compute $kG = (x_1, y_1)$ and $r = x_1 \bmod n$. If r = 0 then go to step 1.
- Compute $k^{-1} \bmod n$. Compute e = SHA-1(m).
- Compute $s = k^{-1}\{e + d_A \cdot r\} \bmod n$.

THE SIMPLE EXPLANATION FOR ELLIPTIC CURVE CRYPTOGRAPHIC ALGORITHM (ECC)

Elliptic Curve Cryptography (ECC) was discovered in 1985 by Victor Miller (IBM) and Neil Koblitz (University of Washington) as an alternative mechanism for implementing public-key cryptography.

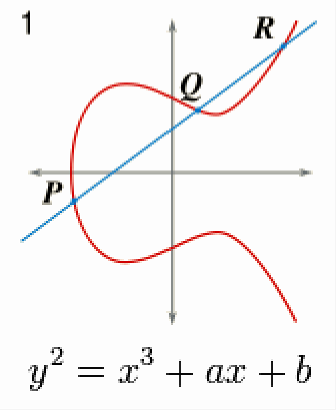
This assume that those who are going through this article will have a basic understanding of cryptography (terms like encryption and decryption) .

The equation of an elliptic curve is given as,

$$y^2 = x^3 + ax + b$$

Few terms that will be used,

E -> Elliptic Curve
P -> Point on the curve
n -> Maximum limit (This should be a prime number)



The fig show are simple elliptic curve.

Key Generation

Key generation is an important part where we have to generate both public key and private key. The sender will be encrypting the message with receiver's public key and the receiver will decrypt its private key.

Now, we have to select a number **'d'** within the range of **'n'**. Using the following equation we can generate the public key

Q = d * P

d = The random number that we have selected within the range of (1 to n-1). P is the point on the curve.

'Q' is the public key and 'd' is the private key.

Encryption

Let 'm' be the message that we are sending. We have to represent this message on the curve. This have in-depth implementation details. All the advance research on ECC is done by a company called certicom.

Consider 'm' has the point 'M' on the curve 'E'. Randomly select 'k' from [1 - (n-1)].

Two cipher texts will be generated let it be C1 and C2.

C1 = k*P

C2 = M + k*Q

C1 and C2 will be send.

Decryption

We have to get back the message 'm' that was send to us,

M = C2 – d * C1

M is the original message that we have send.

Proof

How do we get back the message,

M = C2 – d * C1

'M' can be represented as 'C2 – d * C1′

C2 – d * C1 = (M + k * Q) – d * (k * P) (C2 = M + k * Q and C1 = k * P)

= M + k * d * P – d * k *P

(cancel out k * d * P)= M

(Original Message)

V. CONCLUSION AND FUTURE SCOPE

In this paper, we proposed a high-request PCM plot for heterogeneous information bunching. Besides, cloud servers are utilized to improve the efficiency for dispersed HOPCM conspire contingent upon MapReduce. huge information grouping by planning .One properties of the paper is to utilize the BGV system to built up a protection safeguarding HOPCM calculation for saving security on cloud. PPHOPCM can effectively cluster numerous heterogeneous data using cloud computing without disclosure of private data. Test results show PPHOPCM can group huge information by utilizing the distributed computing innovation without revealing protection. Truth is told, for the substantial size of heterogeneous information that does not require to be ensured, the DHOPCM is more appropriate since it is more efficient than PPHOPCM. Proposes a protection safeguarding high-request PCM plot (PPHOPCM) for huge information grouping. The paper proposes a high-request PCM calculation by broadening the ordinary PCM calculation in the tensor spaces. Tensor is known as a multidimensional exhibiting in arithmetic and it is generally used to speak to heterogeneous information in enormous information examination and mining. The efficiency of PPHOPCM and DHOPCM can be additionally improved when utilizing more cloud servers, making them increasingly reasonable for enormous information grouping, since they are of high adaptability showing by the exploratory outcomes. In this work, the proposed plans are to begin with assessed on two agent heterogeneous datasets. Later on work, the proposed calculations will be additionally approved on bigger genuine 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.SoniandA.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 Semi supervised Heterogeneous Data Co clustering," 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

Authors Profile

Mr. V.Prasathkumar pursed Master of Science from Bannari Amman Institute of Technology, Anna University, Chennai in 2014 He is currently working as an Assistant Professor in Department of Information Technology, Anna University,Chennai since 2014. He is a member of The Institutions of Engineers(India) Since 2016. He has pul 5 papers in reputed international journal. He has 5 years of teaching experience.

Mail Id: prasathkumar@siet.ac.in

K.SENTHIL,
B .Tech, Department of Information Technology ,
Sri Shakthi Institute of Engineering and Technology,
Coimbatore Tamilnadu.
santhossenthil24997@gmail.com

S.VIGNESH,
B.Tech, Department of Information Technology ,
Sri Shakthi Institute of Engineering and Technology,
Coimbatore Tamilnadu.
vignesh1398@gmail.com

P.RANJITH ROSHAN,
B.Tech, Department of Information Technology ,
Sri Shakthi Institute of Engineering and Technology,
Coimbatore Tamilnadu.
rangithroshan33@gmail.com

E.PRAKASH,
B.Tech, Department of Information Technology ,
Sri Shakthi Institute of Engineering and Technology,
Coimbatore Tamilnadu.
prakashmopi@gmail.com