International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 12 (2015) pp. 32681-32688
© Research India Publications
http://www.ripublication.com

Enhancing Privacy Preservation In Cloud Database Querying, Comparative Analysis By Cobweb & K-Means

Nazneen Mulani[a], Ambika Pawar[b], Preeti Mulay[c], Ajay Dani[d,]

[a]M.Tech Student, Symbiosis Institute of Technology, Pune, Gram-Lavale, Tal-Mulshi, 412115, India. Email-nazneen.mulani@sitpune.edu.in
Contact: +91 8007824322
[b]Researcher Scholar, Symbiosis Institute of Technology, Pune, Gram-Lavale, Tal-Mulshi, 412115, India Email-ambikap@sitpune.edu.in
[c]Research Guide, Symbiosis Institute of Technology, Pune, Gram-Lavale, Tal-Mulshi, 412115, India. Email-preeti.mulay@sitpune.edu.in
[d]Research Guide, Symbiosis Symbiosis Institute of Technology, Pune, Gram-Lavale, Tal-Mulshi, 412115, India. Email-ardani123@rediffmail.com

Abstract

Cloud computing is growing technology in enumerable ways. Hence preserving privacy is most challenging task in front of researchers. In this proposed system focus is given on finding vulnerable attributes to prevent brute force attack on cloud database. Incremental clustering: Cobweb and partitioning based clustering: k-means data clustering algorithms are applied on CRM and other databases to prove how privacy is preserved and how Cobweb performs better. Clustering of data helps to find the attributes which are vulnerable and they are considered as susceptible to brute force attack over Cloud environment to preserve the privacy. Privacy of database uploaded by clients and client's information db at backend with cloud service provider also needs protection.

Keywords: incremental clustering, partitioning based clustering privacy, Cobweb, k-means, db.

1 Introduction

Cloud Computing is a technology that is to provide resources and services over the internet to maintain data and applications. Services will be provided on subscription basis or pay per use basis. The Cloud computing [8] allows end users and businesses

to use applications without configuration and installation. This technology opens the path for much more efficient computing by centralizing data storage, processing and bandwidth. Cloud computing is divided into three segments: "application" "storage" and "connectivity." Each segment serves a different purpose and offers different products for businesses and individuals around the world.

In order to provide security over database querying at storage level and to prevent brute force attacks it is necessary to find the attributes of dataset on which attack can be done in future. Such attributes are treated as vulnerable attributes in this proposed system and clustering algorithms has been used to capture vulnerable attributes from datasets.

In real time, when any user fires a query to access data from Cloud database [2], it works as follows as shown in Figure 1. :

a. Database Owner encrypts database records and sends it to Server using Encrypt algorithm, when Server is offline.
b. While Database Owner runs Set Up algorithm to initialize some parameters and decryption key by running Extract algorithm with Database Owner.
c. When any user fires the query, it obtains search token and decryption key from Database Owner.
d. Then user sends token to Server who uses the token to search on each encrypted database records, for which Server runs Test Algorithm.

Certificate Authority handles the entire authority related database. To encrypt the database records Database Owner runs variant of k-means. In previous study, Variant of K-means [2] algorithm has been used to prevent the brute force attack. K, no. of clusters to form is required input from user side. While performing this research study, it is observed that k-means algorithm is non-optimized and there is requirement of another enhanced incremental-clustering approach [2]. The purpose of this study is to use Cobweb clustering with cloud to identify clusters in database efficiently, which indirectly reduces number of unnecessary comparisons of searching data on cloud and helps to find vulnerable attributes to preserve the privacy in real time cloud environment.

Next section 1.1, 1.2, 1.3 briefly describes comparative study of k-means and Cobweb algorithm and how Cobweb is more efficient than k-means. Section 2 presents related work, section 3 describes about the comparison of k-means and Cobweb algorithm with results. The proposed paper finally concludes how Cobweb algorithm is more optimized over k-means algorithm.

1.1 Cobweb Clustering over k-means

Cobweb algorithm builds clusters by incrementally adding instances to a tree, merging them with an existing cluster if this leads to a higher “Category Utility” (CU) value than when the instance would get its own cluster. [5] If the need arises, an existing cluster may also be split up into two new clusters, if this is beneficial to the CU value. The resulting set of clusters is called a “dendogram”. The replacement of Cobweb over k-means is effective process in cloud preservation and there is no need

of k-selection algorithm along with Cobweb like Variant of k-means. This clustering algorithm hierarchically generates the clusters for given data sets by engaging with Database Owner. To summarize [2,4,5]:

- Cobweb is Conceptual incremental clustering,
- Cobweb does not require inputs from user including value of k, distance measure, assumed centroids etc. also hopping of cluster members consistently in k-means affects quality of clusters and may not get the same cluster after re-execution, may affect complexity.
- Cobweb can be applied on Categorical data rather than only numerical data.
- Unlike k-means Cobweb converges with every dataset.
- Cobweb is based on Category utility function which forms the entire dendogram for clustering the input data. This restricted amount of evidence implies probabilistic reasoning. Only those instances are selected which has certain probability of belonging to a particular cluster
- Cobweb minimizes the CPU burst as well as I/O burst time.
- Also reduces the memory operations while constructing the dendogram, which increases the faster accessibility in distributed environment.

For input Datasets, Cobweb constructs dendogram i.e. “classification tree” based on CU formula. In incremental approach for many times CU results into unnecessary skew tree hierarchy for given datasets. Kim. P and Choi. J [4] demonstrated the modified category utility suppresses the formation of spurious nodes and increases classification quality without a loss of learning accuracy. This can be added as future work to make Cobweb more efficient.

1.2 Case study: Sugar CRM

EMIS [9] is the UK’s market-leading primary care software provider with more than 39 million patient records entrusted to its systems. EMIS clinical systems are already used by over 5,000 healthcare organizations across the UK, from GP practices and out-of-hours services, to community care and sexual health services. By using the same system, everyone can access the same information about their patients-no matter where they are treated-making the prospect of integrated cares a reality. Such huge amount of data is handled by SugarCRM [9] on Cloud Server to provide better availability, flexibility and productivity in Healthcare Organization. In such case, Security is another major factor in storing Patients Records on Cloud Server. EMIS stores large collection of bulky data from individual partner and provide open access to all partners of EMIS. Each individual partner treat thousands of patients, and each individual patient is having own personal as well as health checkup records like date of visit, height, weight, sugar level, blood pressure and so on. As there is no guarantee that all this data is securely stored and get accessible to all partners of EMIS through cloud server, to provide privacy and security from unwanted activity in EMIS, privacy preservation technique can be applied efficiently. Others algorithm like Encryption and Decryption, Cobweb runs incrementally over the given datasets, it is having ability to automatically adjust the number of classes in a partition. This will be

applicable for the all new patients also which are newly joined to EMIS System. For future records Cobweb will flexibly construct tree hierarchy.

EMIS deals with the millions of records, in such case to randomly choose factor 'k' i.e. no of cluster is difficult task. Cobweb also helps at Network level to reduce the computing costs in many operations. This enriches Centralized customer management, providing easy and fast access for clients and staff, Improved account-based reporting to support further development of products and services in EMIS.

2 Related Works

Privacy Preservation in Cloud is considered as a major issue in cloud security. Till date lots of solutions to preserve the privacy in Cloud has been introduced by tracing the different parameters. Technique proposed in [6] presents an anonymous privilege control scheme to address not only the data privacy problem in cloud storage, but also the user identity privacy issues in existing access control schemes.

Data sharing in Cloud Computing paper [1] focuses on providing a dependable and secure cloud data sharing service that allows users dynamic access to their data on cloud. In order to achieve this, we propose an effective, scalable and flexible privacy preserving data policy with semantic security; by utilizing cipher text policy attribute based encryption (CP-ABE) combined with identity-based encryption (IBE) techniques.

Paper on Security and privacy for storage and computation in cloud computing [7] propose a privacy cheating discouragement and secure computation auditing protocol, or SecCloud, which is a first protocol bridging secure storage and secure computation auditing in cloud and achieving privacy cheating discouragement by designated verifier signature, batch verification and probabilistic sampling techniques.

3 Results using data mining tool Weka

WEKA is open source software implemented under the GNU General Public License. It is a collection of machine learning algorithms for data mining tasks, where algorithms are applied directly to a dataset. Weka 3.7 has been used to aid the required investigations. Dataset on E-commerce, CRM and secure transactions given by Eurostat [10] is used for the analysis with Weka. Total 31 instances year wise represents up to date Percentage of enterprises having received orders via computer mediated networks at particular Geo Location. Figure 2 shows 31 instances in graphical representation with state of Mexican and data percentage.

Weka classifies the training instances into clusters according to the cluster representation and computes the percentage of instances falling in each cluster as shown in Figure 3 and Figure 4.

For the given data set the k-means forms the clusters based on input taken from user before applying clustering algorithm. While after applying the Cobweb on same data set in incremental manner, It generates hierarchical clustering, where clusters are described probabilistically. While performing clustering of the GeoPlaces data, some class attributes is ignored (using the ignore attributes panel) in order to

allow later classes to clusters evaluation as shown in Fig.2 and Fig 3. After applying k-means and Cobweb on Eurostat dataset by considering Data% as major attribute, the clusters as shown in figures states on x-axis and GeoTime on y-axis. This indicates that clusters formed using Cobweb is better than the k-means Clustering due to tightly coupled members and more inter-cluster distance. Also, by considering number of clusters and time taken by both algorithm following results are obtained as shown in Table 1. For every iteration time taken by k-means clustering is greater than time taken by Cobweb clustering algorithm.Where as in case of number of clusters, Cobweb forms better clusters as compared to k-means.

Thus, it concludes that Cobweb is more efficient than k-means in terms of finding vulnerability and preserving privacy. Vulnerable attribute collected using Weka is validated with Data Owner in privacy preservation solution.

4 Conclusions

Cobweb overcomes most of disadvantages inhibited by k-means clustering. It reduces efforts of specifying the input parameter for further formation of total clusters and works efficiently to reduce the time factor in existing privacy preservation solution system, prevents the brute force attack on cloud database. Important in terms of CPU time, access time between the time at which user fires the query and the time at which it user actually receives particular records and best utilization of available network bandwidth. Cobweb proved to be most optimized solution for preserving privacy due to dynamic cloud setup.

Future work of this paper includes focus on cloud service providers db and BaaS.

Table. 1. Eurostat Data Set Analysis in WEKA

Iteration	Total instances	k-means/time in sec	COBWEB/time in sec
1	31	11/0.01	14/0.01
2	31+63	12/0.02	15/0.01
3	31+63+95	11/0.03	17/0.02

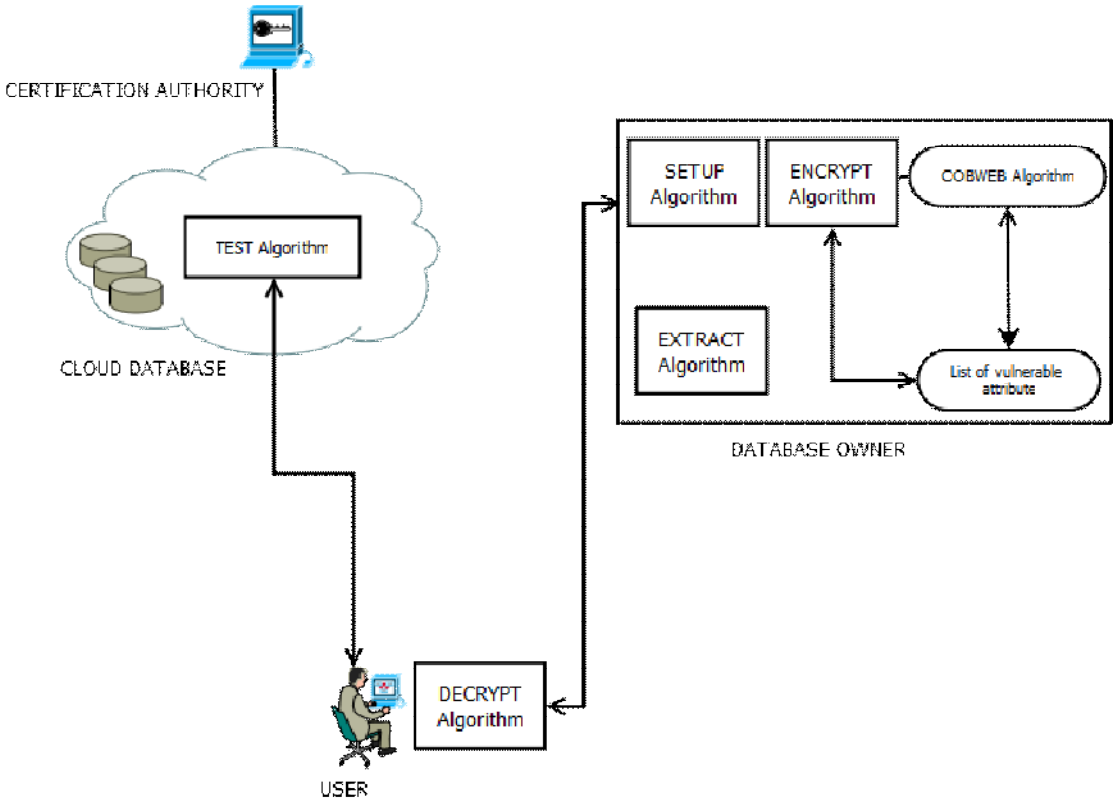

Fig. 1. Privacy Preservation Cloud Model

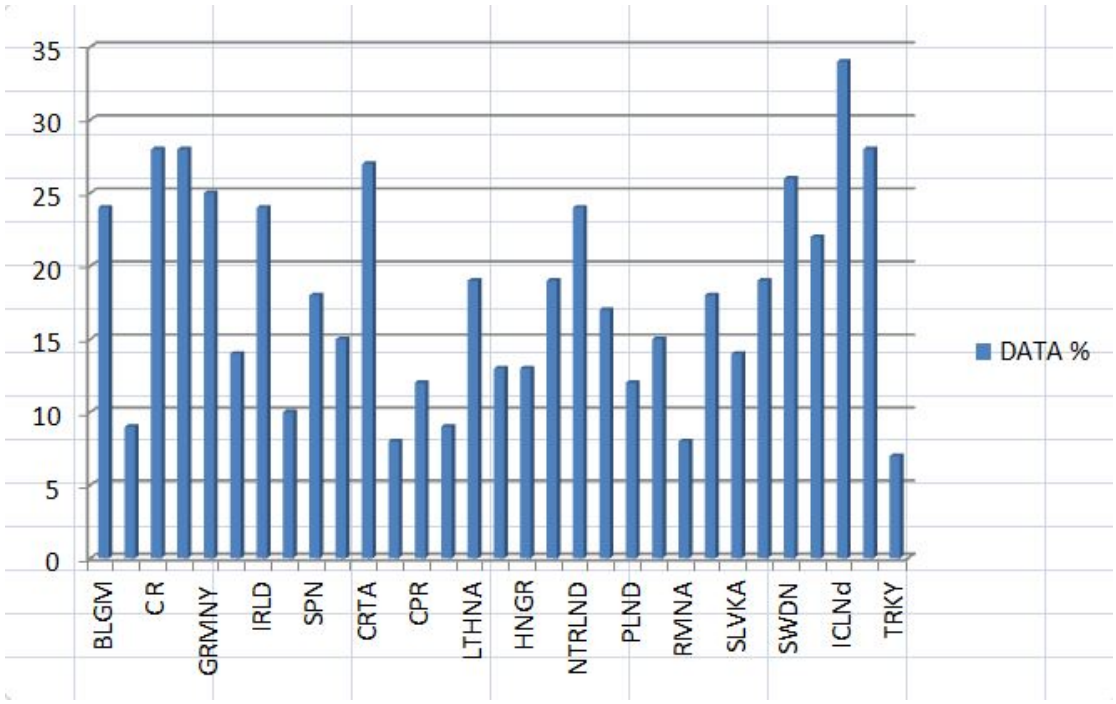

Fig. 2. Data Percentage across different location in 2014

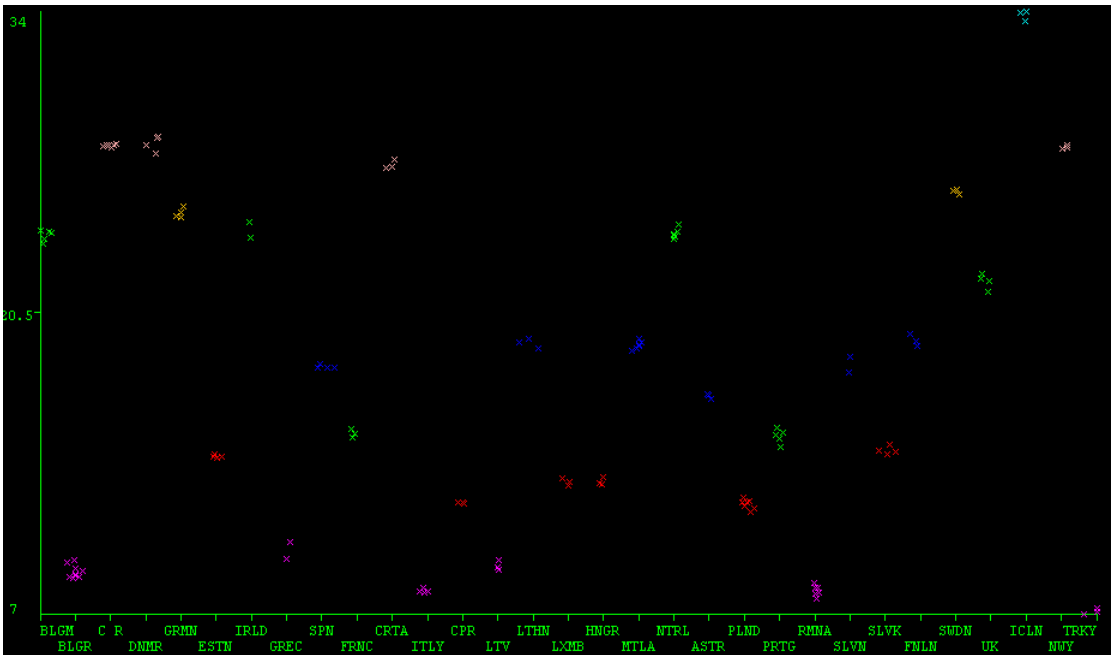

Fig. 3 Clusters for Eurosat using k-means

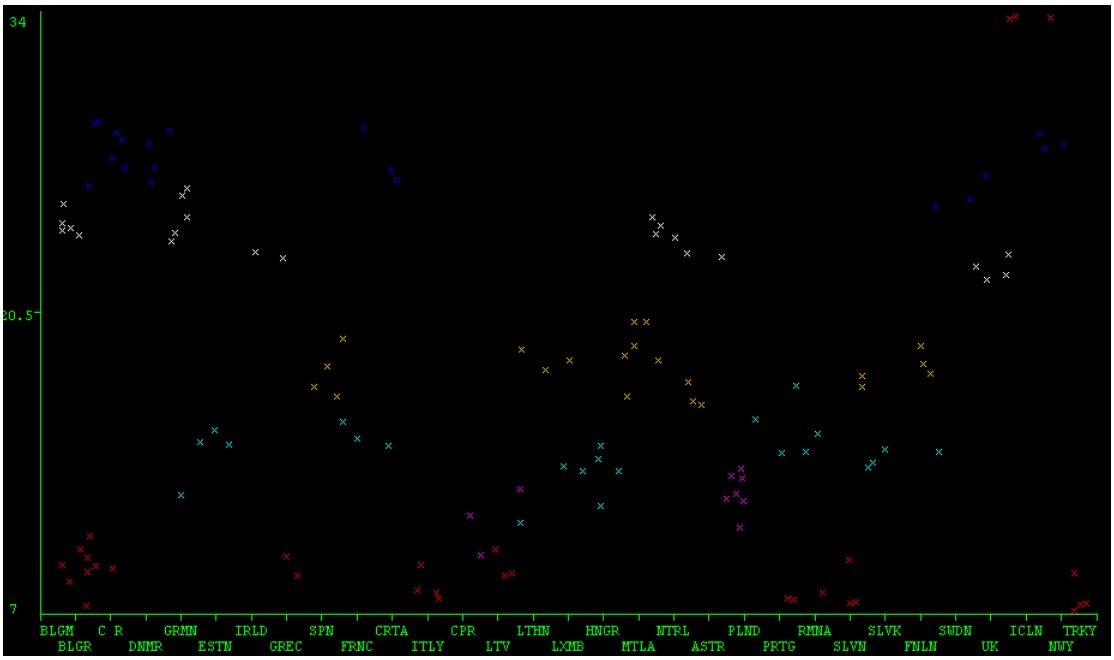

Fig. 4 Clusters for Eurostat using COBWEB

References

[1]. Suriyapriya1, M., Joicy, A., "Attribute Based Encryption with Privacy Preserving In Clouds", IJRITCC, pp.231-236.

[2]. Pawar, A., Dani, A., 2014, "Enhancing Privacy-Preserving Cloud Database Querying by Preventing Brute Force Attacks", International Journal of Computer, Information Science and Engineering, Vol:8 No:1, pp. 51-57.

[3]. Eurostat Database [online] (http://ec.europa.eu/eurostat/data/database)

[4]. Kim, P., Choi, J., "Incremental Conceptual Clustering Using a Modified Category Utility".

[5]. Fisher, Douglas H., 1998, "Knowledge Acquisition via incremental Conceptual Clustering", pp. 139-172.

[6]. Wan, M., Jungo, T., "Privacy Preserving Cloud Data Access With Multi-Authorities" pp. 1-9.

[7]. Lifei w., Haojin Z.(et al). "Security and privacy for Storage and Computation in Cloud" Computing-information Sciences, pp. 371-386.

[8]. Wikiinvest Definition of Cloud Computing [online]. (http://www.wikinvest.com/concept/Cloud_Computing) U. S. (10 May 2015)

[9]. Case.Study: SugarCRM[Online](http://www.sugarcrm.com/case-studies), U.K.. (21 April 2015)