Volume 4, Issue 3, March 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

An Efficient Incremental Clustering Method for Incremental Cloud Data

S. Nikkath Bushra
Research Scholar, Bharath University
Department Of Computer Science And Applicaton
St.Joseph's College Of Engineering, India

A. Chandra Sekar
Department Of Computer Science And Engineering
St.Josephs College Of Engineering
India

Abstract*— Cloud Computing is a technology that uses the internet and central remote servers to maintain data and applications. Cloud computing allows consumers and businesses to use applications without installation and access their personal files at any computer with internet access. This technology allows for much more efficient computing by centralizing data storage, processing and bandwidth. Typically, the data sets in these applications are anonymized using K-anonymity to assert the privacy of data owners but the privacy requirements can be violated when new data join over time. Traditional clustering algorithms are static in nature .In this paper new ways of Incremental clustering algorithms are discussed. This algorithm clusters data in dynamic form. The database is assumed to be clustered initially and every new element is added as without need of changing existing clustered database.

Keywords— *IGCA, Incremental Clustering, Anonymization, k-anonimity, Genetic clustering*

I. INTRODUCTION

Cloud computing comes into focus only when you think about what IT always needs: a way to increase capacity or add capabilities on the fly without investing in new infrastructure, training new personnel, or licensing new software. Cloud computing encompasses any subscription-based or pay-per-use service that, in real time over the Internet, extends IT's existing capabilities. The idea is not new, but this form of cloud computing is getting new life from Amazon.com, Sun, IBM, and others who now offer storage and virtual servers that IT can access on demand. Early enterprise adopters mainly use utility computing for supplemental, non-mission-critical needs, but one day, they may replace parts of the data center.

Due to the unreliability of the service and malicious attacks from hackers, there have concerns over the data security with cloud storage along with the widespread enthusiasm on cloud computing. Nowadays more and more events on cloud service outage or server corruption with major cloud infrastructure providers are reported . Data breaches from notable cloud services are also appear from time to time . For diverse motivations, the cloud service providers would also voluntarily analyze the customer's data. Therefore, the cloud is basically neither secure nor reliable from the cloud customers view point. It is difficult to anticipate the cloud customers to turn over control of their data to cloud servers solely based on economic savings and service flexibility without providing robust security, privacy and reliability guarantee.

In this paper various incremental clustering techniques for privacy preservation over incremental cloud data is presented. The k-anonymity model is utilized for ensuring the security in cloud computing environment since k-anonymity is an effective technique for privacy preservation over years. The major challenge of privacy preservation over the cloud data is handling the incremental data, because the cloud data may be updated continuously. Given a set of records as input choose some privacy sensitive attributes as quasi-identifiers for anonymization. After anonymization, the records are clustered based on different incremental clustering techniques and check the k-anonymity constraint and information loss for every cluster and modify the cluster based on the k-anonymity constraint.

Types of clustering

1. Non-Incremental Clustering- used for static datasets

Traditional clustering approaches usually analyze static datasets in which objects are kept unchanged after being processed, but many practical datasets are dynamically modified which means some previously learned patterns have to be updated accordingly. Re-clustering the whole dataset from scratch is not a good choice due to the frequent data modifications and the limited out-of-service time, so the development of incremental clustering approaches is highly desirable. Besides that, propositional clustering algorithms are not suitable for relational datasets because of their quadratic computational complexity.

2. Incremental Clustering – used for Dynamic Datasets

In recent years with data becoming larger and larger, the clustering algorithm not only needs to have high executing efficiency, but also is required to discover clusters with arbitrary shape and be insensitive to noise data they have linear time complexity, and can be used in clustering dynamic large databases . Clustering of very large document databases is useful for both searching and browsing. The periodic updating of clusters is required due to the dynamic nature of

databases. An algorithm for incremental clustering is introduced. The complexity and cost analysis of the algorithm together with an investigation of its expected behaviour are presented. Clustering techniques work with an item at a time and decide how to cluster it given the current set of points that have already been processed.

II. INCREMENTAL CLUSTERING

2.1 Incremental K-Means clustering

The incremental clustering is a technique that adds a new record to a corresponding cluster after clustering the initial set of records we given as input. We have to set the quasi-identifiers in the new record and the quasi-identifiers should be the same attributes we chosen before clustering the set of records we gave as input. After selecting the quasi-identifiers, we have to anonymize the data in the quasi-identifier attributes. The K-anonymity constraint after anonymizing the quasi-identifiers of the new record is check with each cluster and if it matches the k-anonymity constraint with any cluster, the new record will get merge with that cluster.

Advantages

• Simple, understandable
• Items automatically assigned to clusters

Drawbacks

• It is data dependent.
• It is based on greedy approach. So the final clusters may not represent a global optimization result but only the local one, and complete different final clusters can arise from difference in the initial randomly chosen cluster centers. • Need to specify k, the number of clusters, in advance.
• Unable to handle noisy data and outliers.
• Not suitable to discover clusters with non-convex shapes or clusters of very different size.

The efficiency of the algorithm: The complexity of the k-means algorithm is O (n, d, k) which involves the sample size, the number of dimensions and the number of partition The definition of "means" limits the application only to numerical variables

2.2 Incremental Genetic k-means clustering(IGCA)

Genetic algorithm is search heuristic usually applied in Optimization problems. The power of Genetic algorithm lies in its ability to perform parallel search in complex spaces. The behavior of genetic algorithms is highly influenced from natural evolution studied in biological sciences. In complex scenarios where the research problem under study involves a multi-dimensional search space and deterministic algorithms fail to meet time constraints, stochastic techniques like genetic algorithms are used.

The genetic operators that are used in GKA are

- Selection,
- The distance based mutation and
- K-means operator:

1) Coding: The natural way of coding such w into string is to consider a chromosome of length n and allow each allele in chromosome to take values from {1,2,…..,k}.
2) Initialization: Way of selecting initial population is random. Each allele in the population can be initialized to cluster number selected from uniform distribution over the set {1,2,…,k}.
3) Selection: Selection operator randomly select a chromosome from the previous population.
4) Mutation: The Mutation changes an allele value depending on the distances of the cluster centroids from the corresponding data point. It may be recalled that each allele corresponds to a data point and its value represents the cluster to which the data point belongs operator is defined such that the probability of changing an allele value to a cluster number is more if the corresponding cluster center is closer to the data point.

Drawback:

An algorithm with the above selection and mutation operators may take more time to converge, since the initial assignments are arbitrary and the
subsequent changes of the assignments are probabilistic. Moreover, the mutation probability is forced to assume a low value because high values of Pm
lead to oscillating behaviour of the algorithm

2.3 A Fast Incremental Clustering Algorithm

Fast incremental clustering algorithm can pre-bound the final clusters number, scan the original data set only once according to the memory capacity. When the number of the generated clusters is more than the constraint the two nearest clusters are merged and the radius threshold value changes dynamically high similarity within a cluster, low across clusters.

2.4 Incremental Hierarchical clustering Algorithm

In hierarchical clustering the goal is to produce a hierarchical series of nested clusters, ranging from clusters of individual points at the bottom to an all-inclusive cluster at the top. A diagram called a dendogram graphically represents this hierarchy and is an inverted tree that describes the order in which points are merged (bottom-up view) or clusters are split (top-down view). One of the attractions of hierarchical techniques is that they correspond to taxonomies that are very common in the biological sciences, e.g., kingdom, phylum, genus, species, … . (Some cluster analysis work occurs under the name of "mathematical taxonomy.") Another attractive feature is that hierarchical techniques do not assume any particular number of clusters. Instead any desired mber of clusters can be obtained by "cutting" the dendogram at the proper level. Finally, hierarchical techniques are thought to produce better quality clusters.

Another attractive feature is that hierarchical techniques do not assume any particular number of clusters. Instead any desired number of clusters can be obtained by "cutting" the dendogram at the proper level. Finally, hierarchical techniques are thought to produce better quality clusters.

1. Build a tree-based hierarchical taxonomy
2. Dendrogram
3.From a set of unlabeled examples.

Limitations and problems

1) No global objective function is being optimized.
2) Merging decisions are final.
3) Good local merging decisions may not result in good global results.
4) Agglomerative hierarchical clustering techniques have trouble with one of more of the following: noise and outliers, non-convex shapes, and a tendency to break large clusters.

III RELATED RESEARCH: A BRIEF REVIEW

This section shows some of the recent related researches about privacy issues in cloud computing and privacy preserving over incremental data sets

Xuyun Zhang *et al.* [21] have investigated the challenge about how to efficiently update huge volume incremental data sets to ensure privacy requirements of data owners and simultaneously achieve high data utility to data users. They have presented an efficient quasi-identifier index based approach for privacy preservation over incremental data sets on cloud. In their approach, QI-groups (QI: quasi-identifier) were indexed by the domain values in the current generalization level, which made it possible to access only a part of records in a data set in the presence of data updates rather than access all data records as required by existing approaches. To further improve the performance of quasi-identifier indexing, locality-sensitive hashing method was incorporated to place similar QI-groups on the same data storage nodes. Thus, the number of data nodes that a QI-group link across was reduced considerably with high probability. Based on the established indexes of an anonymized data set, they have designed an efficient quasi-identifier index based privacy preservation algorithm (QuIPP) for their approach.

K.Kiran Kumar et al. [24] have developed a privacy-preserving public auditing system for data storage security. They designed the simulation by considering the single user. They utilized the public key based homomorphic authenticator and uniquely integrate it with random mask technique and automatic blocker. Extensive security and performance analysis showed the presented schemes were provably secure and highly efficient. Remya Rajan [25] has developed an efficient, secure and privacy preserving keyword search scheme which supported multiple users with low computation cost and flexible key management. It enabled the service provider to determine whether a document contains a specified keyword without getting any information about the document or keyword. It supported multi user requirements with user authentication and also avoided statistical attack on keywords. It also enabled the service provider to participate in partial decipherment thus reducing the users computational overhead. In the presented scheme, user authentication was provided before giving the secret key for decryption of document.

Anonimity Based method[3]before the micro data is published anonymity algorithm will process the data and send these anonymous data to service providers in the cloud .Then the service provider can integrate the auxiliary information to analyze the anonymous data in order to mine the knowledge they want. Quasi identifiers are anonymized .it is different from traditional cryptographic method which needs key to access the data. But using anonymity technique no need to Know the key so it is flexible and safe to protect individual's privacy in cloud computing services.

Bhushan Mahajan ,Swati Ganar[26] preserving confidentiality of data in cloud using dynamic anonymization to protect the sensitive data using two techniques M-Invariance proposed by X. Xiao and Y. Tao To overcome the problem in m-invariance, Feng Li and Shuigeng Zhou proposed a new generalization principle m-Distinct to effectively anonymize datasets with internal as well as external updates. M-Distinct uses m-unique which is used to maintain the sensitive values which are not different in separate publication . In order to maintain this indistinguishability of sensitive values, records must be partitioned carefully while releasing new publication. These techniques can be used as a better approach to secure a data in a cloud. K-anonymity can be used along with M-Distinct to provide dynamic anonymization. Then, use a key distribution approach which will authenticate the users and provide an anonymization table to valid users. Along with these methods, anonymization using other methods such as Hashing, Hiding, and Permutation can also be used to secure a cloud data.

IV COMPARITIVE STUDY

S.NO	Name of Technique	Advantages	Drawbacks
1.	QUIPP(Quasi index based privacy preservation) for incremental dataset	Access only a part of records in a data set in the presence of data updates rather than access all data records as required by existing approaches	Quasi index has to be identified
2.	M-Invariance using incremental clustering	Dynamic anonymization	Preserving current statics of attribute
3.	k-Anonimity Using Incremental Clustering	1)Indistinguishable against k-1 other data records 2)Loss of information is less	Does not guarantee against attackers with background knowledge
4.	Privacy preservation using incremental clustering	Time taken to update a record is less compared to traditional clustering	It may lead to relatively high data distortion.

V Conclusion

In this paper an incremental clustering technique for privacy preservation over incremental cloud data is proposed. With the help of the k-anonymity technique to assert the security in cloud computing environment and also used different incremental clustering techniques for initially clustering the set of records as input. The k-anonymity constraint and the information loss are by the k-means clustering technique. The new record given is then checked with each cluster and it is joined with a cluster based on the k-anonymity constraint.

References

[1] Siani, P., S. Yun and M. Miranda, "A privacy manager for cloud computing," Cloud Computing, vol.5931, pp. 90-106, 2009.
[2] Jian, W., Z. Yan, J. Shuo and L. Jiajin,"Providing privacy preserving in cloud computing," International Conference on Test and Measurement, 5-6 Dec., Coll. of Inf. Sci. and Technol., Donghua Univ., Shanghai, China, vol.2,pp. 213-216,2009.
[3] Jian, W., L. Yongcheng, J. Shuo and L. Jiajin,"A survey on anonymity-based privacy preserving," International Conference on E-Business and Information System Security, 23-24 May, Coll. of Inf. Sci. Technol., Donghua Univ., Shanghai, pp: 1-4,2009.
[4] M. Srivatsa and L. Liu. Secure event dissemination in publish-subscribe networks. Proc.of IEEE ICDCS, 2007.
[5] K. Minami, A. J. Lee, M. Winslett, and N. Borisov,"Secure aggregation in a publish-subscribe system," Proc. of the 7th ACM WPES, pp. 95–104, 2008.
[6] M Nabeel, N Shang, E Bertino,"Privacy-Preserving Filtering and Covering in Content Based Publish Subscribe System," https://www.cerias.purdue.edu/assets/pdf/bibtex_archive/2009-15-report.pdf
[7] Blog service hosted by google crashes review.http://hostwisely.com/blog/blog-service-hosted-by-google-crashes/.
[8] "Microsoft cloud data breach heralds things to come," http://www.techworld.com.au/article/372111/.
[9] "Summary of the amazon ec2 and amazon rds service disruption in the us east region," http://aws.amazon.com/message/65648/.
[10] Zack Whittaker, "Amazon web services suffers partial outage," http://www.zdnet.com/blog/btl/amazon-web-services-suffers-partial-outage/79981.
[11] Mathew J. Schwartz. 6 worst data breaches of 2011, 2011. http://www.informationweek.com/news/security/attacks/232301079.
[12] Aaron Souppouris. Linkedin investigating reports that 6.46 million hashed passwords have leaked online, 2012.
[13] Darlene Storm. Epsilon breach: hack of the century? 2011.
[14] A. Weiss, "Computing in the Clouds," Networker, vol.1, no.4, pp.16-25, Dec. 2007.
[15] Michael Armbrust, Armando Fox, and et al.,"Above the clouds: A Berkeley view of cloud computing," Technical Report UCB-EECS-2009-28, University of California, Berkeley.
[16] M. Armbrust et al., Above the Clouds: A Berkeley View of Cloud Computing, Berkeley, California, USA, 2010, http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html, accessed on January 21, 2011.
[17] S. Pearson, Y. Shen, M. Mowbray,"A privacy manager for cloud computing," in: Proceedings of The 1st International Conference on Cloud Computing, Beijing, China, December 1–4, pp. 90–106,2009.
[18] S. Pearson, "Taking account of privacy when designing cloud computing services," in: Proceedings of the 2009 ICSE (International Conference on Software Engineering) Workshop on Software Engineering Challenges of Cloud Computing, Vancouver, Canada, May 23, pp. 44–52,2009.
[19] Meena Dilip Singh,P. Radha Krishna,Ashutosh Saxena,"A Cryptography Based Privacy Preserving Solution to Mine Cloud Data," Proceedings of the Third Annual ACM Bangalore Conference, 2010.

[20] Rajeev Bedi, Mohit Marwaha, Tajinder Singh, Harwinder Singh and Amritpal Singh, "Analysis of different privacy preserving Cloud storage frameworks," International Journal of Computer Science & Information Technology (IJCSIT), Vol.3, No.6, Dec 2011.
[21] Xuyun Zhang, Chang Liu, Surya Nepal, Jinjun Chen,"An efficient quasi-identifier index based approach for privacy preservation over incremental data sets on cloud, "Journal of Computer and System Sciences, 2012.
[22] Gaofeng Zhang, Yun Yang, Xiao Liu,Jinjun Chen,"A Time-series Pattern based Noise Generation Strategy for Privacy Protection in Cloud Computing,"2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 2012.
[23] Gaofeng Zhang,YunYang, Jinjun Chen,"A historical probability based noise generation strategy for privacy protection in cloud computing," Journal of Computer and System Sciences,vol.78,pp.1374–1381, 2012.
[24] K.Kiran Kumar, K.Padmaja,P.Radha Krishna, "Automatic Protocol Blocker for Privacy-Preserving Public Auditing in Cloud Computing," Cloud Computing,Vol.PP,No.99, 2012.
[25] Remya Rajan,"Efficient and privacy preserving multi User keyword search for cloud Storage services," International Journal of Advanced Technology & Engineering Research (IJATER), Vol.2, No.4, July 2012.
[26] Bhushan Mahajan ,Swati Ganar,"Review paper on preserving confidentiality of data in cloud using data anonimyzation ,"International Journal of Computer Science and Network (IJCSN) Volume 1, Issue 6, October 2012.

Author Profile

S. Nikkath Bushra received her Bachelor's degree from the University of Madras, Chennai, Tamil Nadu ,India . M.C.A from the University of Madras, Chennai, Tamil Nadu, India, M.E degree (Computer Science and Engineering) from the Sathyabama University ,Chennai, Tamilnadu India and M.Phil from Mother Teresa University, Kodaikanal, Tamil Nadu, India from 2008 to 2009, she worked as a Database Administrator in IBM. She is currently an Associate Professor in the Department of Computer Science and application at St. Joseph's College Of Engineering, Chennai and She is a research scholar of Bharat University, Chennai, Tamil Nadu, India. Her research interests are in cloud computing, privacy preservation in cloud computing and Data mining.

Chandra Sekar A, received his B.E degree from Angala Amman College of Engineering and Technology affiliated to Bharathidasan University, M.E degree from A.K. College of Engineering affiliated to Madurai Kamaraj University, and Ph.D in Information and Communication Faculty (Computer science & Engineering) from Anna University, Chennai, India. He is currently working as a Professor in the Department of Computer Science & Engineering in St. Joseph's College of Engineering, Chennai. His area of interest includes Network Security and Analysis of Algorithms. Life Member in ISTE and fellow membership from International Science congress Association.