Secure Clustering in a Distributed Network

S. Harippriya, T. Kalaikumaran, S. Karthik

Abstract— ***Data Mining plays a major role in storage of vast quantities of data. It extracts valuable knowledge, which helps organizations to obtain better results by pooling their data together. Distributed data mining is concerned about data that are shared among multiple organizations. A complementary approach to privacy-preserving data mining uses randomization techniques. Privacy-preserving data mining solutions have been presented both with respect to horizontally and vertically portioned databases, in which earlier data objects with the same attributes for the same data objects are owned by each party, respectively. The quality of a set of clusters can be measured using the value of an objective function which is taken to be the sum of the squares of the distances of each point from the centre of the cluster to which it is assigned.***

Index Terms— Arbitrarily partitioned Data, Data Mining.

I. INTRODUCTION

Distributed data mining is concerned about data that are shared among multiple organizations. Privacy-preserving distributed data mining deals with cooperative parties without revealing any of their individual data items. Large collections of data are mined for knowledge which improves the performance of the organization. Well known data mining tasks include clustering, prediction, association rule mining and outlier detection. Data mining is used in biomedical and DNA data analysis, financial data analysis, Identification of unusual patterns, and analysis of telecommunication data. A complementary approach to privacy-preserving data mining uses randomization techniques. Privacy-preserving data mining solutions have been presented both with respect to horizontally and vertically portioned databases, in which earlier data objects with the same attributes for the same data objects are owned by each party, respectively. Division of data into groups of similar objects is called Clustering. Certain fine details are lost by representing the data by fewer clusters but it achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, Statistics and numerical analysis. K-means clustering is a simple technique to group items into k clusters. There are many ways in which k clusters might potentially be formed. The quality of a set of clusters can be measured using the value of an objective function which is taken to be the sum of the squares of the distances of each point from the centre of the cluster to which it is assigned. It's required that value of this function to be as small as possible.

A more general direction for further work is to obtain privacy preserving data mining protocols for other algorithms over arbitrarily partitioned data. A common way for this to occur is through data aggregation. Data aggregation is when the data are accrued, possibly from various sources, and put together so that they can be analyzed.

II. LITERATURE SURVEY

Least Squares Quantization in PCM

The primary task in data mining is the development of models about aggregated data; they develop accurate models without access to precise information in individual data records? They consider the concrete case of building a decision-tree classifier from training data in which the values of individual records have been perturbed. The resulting data records look very different from the original records and the distribution of data values is also very different from the original distribution. It is not possible to accurately estimate original values in individual data records; they propose a-novel reconstruction procedure to accurately estimate the distribution of original data values. By using these reconstructed distributions, they are able to build classifiers whose accuracy is comparable to the accuracy of classifiers built with the original data.

Non-Cryptographic fault-tolerant computing in constant number of rounds of interaction.

It described mining and integrating data from multiple sources, there are many privacy and security issues. The security of the full privacy-preserving data mining protocol depends on the security of the underlying private scalar product protocol. The two private scalar product protocols, one of which was proposed in a leading data mining conference, are insecure and the other one is based on homomorphism encryption and improve its efficiency so that it can also be used on massive datasets. The goal is that one of the participants obtains the scalar product of the private vectors of all parties. It is often required that no information about the private vectors, except what can be deduced from the scalar product, will be revealed during the protocol.

Privacy Preserving Data Mining

The focus of the course is to understand what cryptographic problems can be solved, and under what assumptions. The main focus of the cryptography is the presentation of "feasibility results" It is typically not related to issues of efficiency. There are a number of significant differences between this course and its prerequisite.

Privacy Preserving Clustering by Data Transformation

The issue of privacy preserving data mining of two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information. They secure multi-party computation and as such, that can be solved by using known generic protocols. Data mining algorithms are typically

Manuscript received on April 11, 2013.

S. Harippriya, ME-Software Engineering, SNS College of Technology, Coimbatore, Tamil Nadu, India.

Dr. T. Kalaikumaran, HOD/CSE, SNS College of Technology, Coimbatore, Tamil Nadu. India.

Dr. S. Karthik, DEAN/CSE, SNS College of Technology, Coimbatore, Tamil Nadu. India.

complex and the input usually consists of massive data sets. The generic protocols are more efficient protocols are required. They focus on the problem of decision tree learning with the popular ID3 algorithm. Their protocol is considerably more efficient than generic solutions and demands both very few rounds of communication and reasonable bandwidth.

III. BACKGROUND WORK

Privacy Preserving Data Mining (PPDM)

Privacy preserving data mining has emerged as a very active research area. This field of research studies how knowledge or patterns can be extracted from large data stores while maintaining commercial or legislative privacy constraints. Quite often, these constraints pertain to individuals represented in the data stores. While data collectors strive to derive new insights that would allow them to improve customer service and increase their sales by better understanding customer needs, consumers are concerned about the vast quantities of information collected about them and how this information is put to use.

Privacy preserving data mining aims to settle these conflicting interests. The question how these two contrasting goals, mining new knowledge while protecting individuals' privacy, can be reconciled, is the focus of this research. It is to improve the trade of between privacy and utility when mining data. The objective is to create an innovative IT framework that leverages data (medical, imaging, treatment, biological, lab, genomic, psycho-social) that is already being collected in local databases of multiple institutions and makes this data available without any demand on the type and way in which the data is collected locally. The data extraction system needs to be flexible enough to cope with all the difference in local data availability, structure, type and meaning while at the same time present the data in a coherent model to facilitate analysis. The next step is to mirror the institution data into the local euro CAT database, which can connect to different data systems and will automatically collect data from standard databases and protocols (DICOM, HL7, etc.). The system will host the patient data as a mirror of the institution. Site-specific tools that expose the data at the euro CAT meta-levels have been developed for MAASTRO in preliminary work but will need to be customized for each participating center.

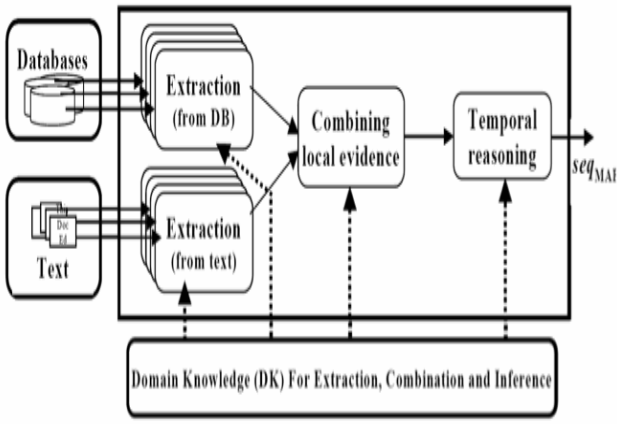


Fig 1: Privacy Preserving Data Mining

By developing and managing the data in this virtual hierarchy, it is also possible to develop privacy-preserving data mining algorithms which can be executed on various levels of the hierarchy, without any residual risk of data privacy non-conformance. To enhance security, and help overcome privacy regulations, the local euro CAT database will reside and remain within the institution. Furthermore, the euro CAT database simply holds a copy of data already present within the institution and does not interfere with the clinical process.

Clustering

Division of data into groups of similar objects is called Clustering. Certain fine details are lost by representing the data by fewer clusters but it achieves simplification. It models data by its clusters. Data modelling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. According to machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective clustering plays an important role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others. Clustering can be shown with a simple graphical.

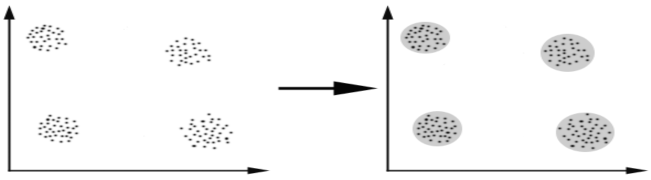


Fig 2: Clustering

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance*:* two or more objects belong to the same cluster if they are “close” according to a given distance (in this case geometrical distance).This is called distance-based clustering. Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

Arbitrarily Partitioned Data

Privacy-preserving data mining (PPDM) was introduced to enable conventional data mining techniques to preserve data privacy during the mining process. Much work has been done to explore privacy-preserving data mining on horizontally and/or vertically partitioned data involving multiple parties so that no single party holds the overall data of horizontally and vertically partitioned data.

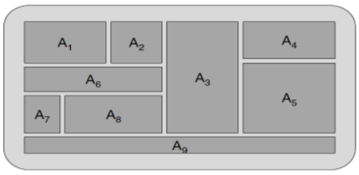


Fig 3: Arbitrarily Partitioned Data

For vertically partitioned data, two parties or more hold the different set of attributes for the same set of objects. In arbitrarily partitioned data, different disjoint portions are held by different parties.

For arbitrarily partitioned data,two parties hold data forming a (virtual) database consisting of their joint data. For horizontally partitioned data, two parties or more hold different objects for the same set of attributes. It means each object in the virtual database is completely owned by one party.

IV. SYSTEM ANALYSIS

Existing System

Privacy-preserving data mining (PPDM) was introduced, to enable conventional data mining techniques to preserve data privacy during the mining process. Much work has been done to explore privacy-preserving data mining on horizontally and/or vertically partitioned data involving multiple parties so that no single party holds the overall data. For horizontally partitioned data, two parties or more hold different objects for the same set of attributes. It means each object in the virtual database is completely owned by one party for vertically partitioned data, two parties or more hold the different set of attributes for the same set of objects. In arbitrarily partitioned data, different disjoint portions are held by different parties. While much data mining occurs on data within an organization, is it quite common to use data from multiple sources in order to yield more precise or useful knowledge. However, privacy and secrecy considerations can prohibit organizations from being willing or able to share their data with each other.

Proposed Work

It provide a privacy-preserving solution to an important data mining problem, that of clustering data. Privacy-preserving clustering has been previously addressed using randomization techniques. The k-means clustering algorithm is a well known iterative algorithm that successively refines potential clusters in an attempt to minimize the k-means objective function, which measures the goodness of a given clustering. A privacy preserving protocol for k-means clustering in the setting of arbitrarily partitioned data distributed between two parties.

The protocol is efficient and provides cryptographic privacy protection. In particular, the protocol provides the first privacy-preserving solution to k-means clustering for horizontally partitioned data. It also provide an analysis of the performance and privacy of the solution mapping the constraint satisfaction problem to an equivalent binary integer programming problem

V. METHODOLOGY

System Design

In this system architecture there is a system administrator who manages a list of accounts such as Gmail and yahoo. There is a database that is managed by the admin. There are Users who can create an account by using a username and password which is being stored in the database. In this we will be using k-means algorithm to encrypt the password that is being entered by the user. Thus the admin cannot track the cipher text. The user can use the private key to decrypt the password and access their account.

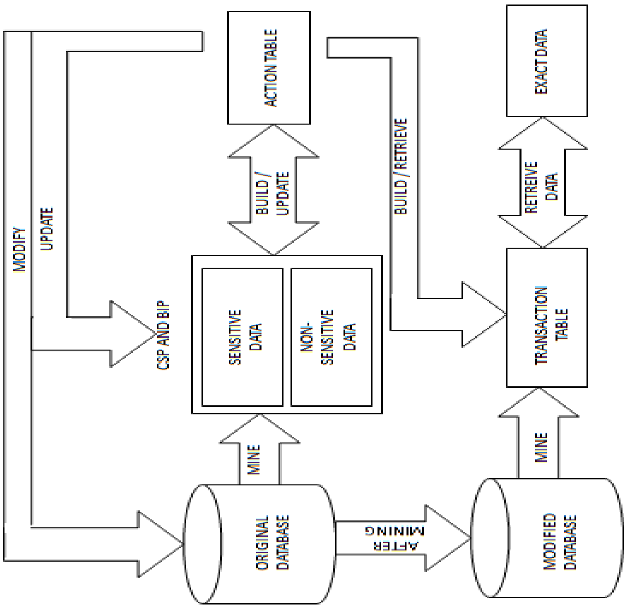


Fig4: Data Flow Diagram

Module Description

Account Creation

Admin can add Products for user using Product ID .When Authenticated user can process and create account. After creating an Admin will activate their Account. After Activation only the user can processed using that User id. After Activation Admin can view the details of purchased items with user ID and Product ID. After Purchasing, Admin will give away the delivery status like when the product will reach the user.

Data Privacy

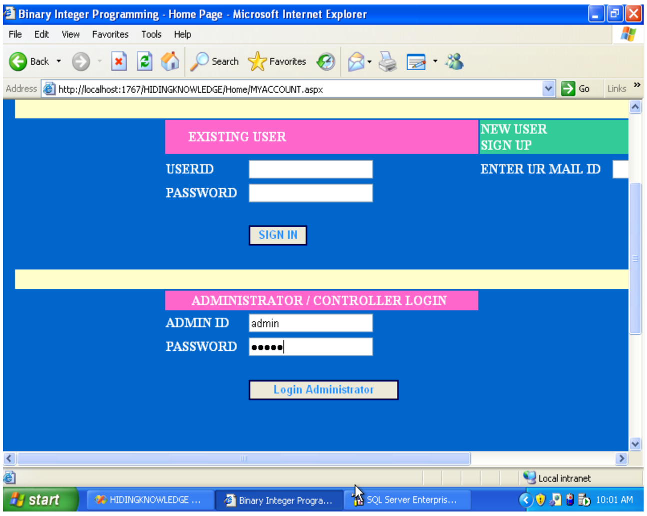
Information privacy or data privacy is the relationship between collection and dissemination of data, technology, the public expectation of privacy, and the legal and political issues surrounding them. Privacy concerns exist wherever personally identifiable information is collected and stored – in digital form or otherwise. Improper or non-existent disclosure control can be the root cause for privacy issues.

Binary Integer Programming

The need to hide sensitive information before sharing databases has long been recognized. In the context of data mining, sensitive information often takes the form of item sets that need to be suppressed before the data is released. The problem is to minimizing the number of nonsensitive item sets lost while concealing sensitive ones. It is shown to be an intractably large version of an NP-hard problem. Consequently, a two-phased procedure that involves the solution of two smaller NP-hard problems is proposed as a practical and effective alternative.

VI. PERFORMANCE EVALUATION

The Account Creation Module has been created with the products that are available in a company. Administrator has all the control of the website. He is responsible for User Registration Module which is done Fig 5.


Binary Integer Programming - Home Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost:1767/HIDINGKNOWLEDGE/Home/MYACCOUNT.aspx
Go Links
Local intranet
start HIDINGKNOWLEDGE ... Binary Integer Progra... SQL Server Enterpris... 10:01 AM

EXISTING USER

USERID

PASSWORD

SIGN IN

NEW USER SIGN UP

ENTER UR MAIL ID

ADMINISTRATOR / CONTROLLER LOGIN

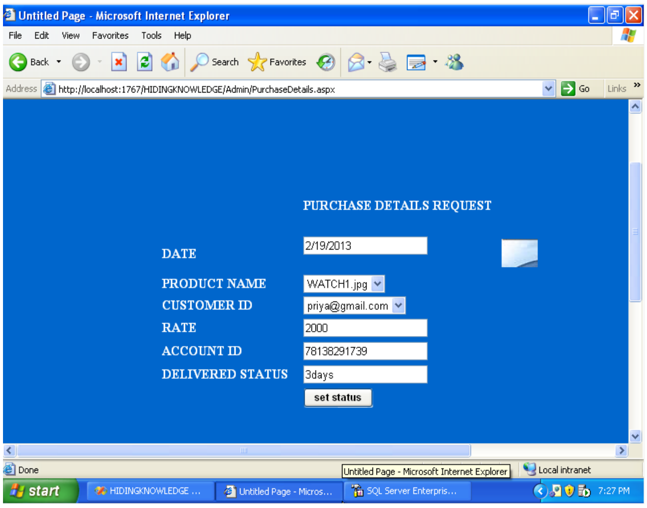
ADMIN ID admin

PASSWORD •••••

Login Administrator

Fig: 5

In Fig 6, the Purchase Details Request has been made by the user after the successful registration in the website with the help of Admin.


Untitled Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost:1767/HIDINGKNOWLEDGE/Admin/PurchaseDetails.aspx
Go Links
Done
Untitled Page - Microsoft Internet Explorer
Local intranet
start HIDINGKNOWLEDGE ... Untitled Page - Micros... SQL Server Enterpris... 7:27 PM

PURCHASE DETAILS REQUEST

DATE 2/19/2013

PRODUCT NAME WATCH1.jpg

CUSTOMER ID priya@gmail.com

RATE 2000

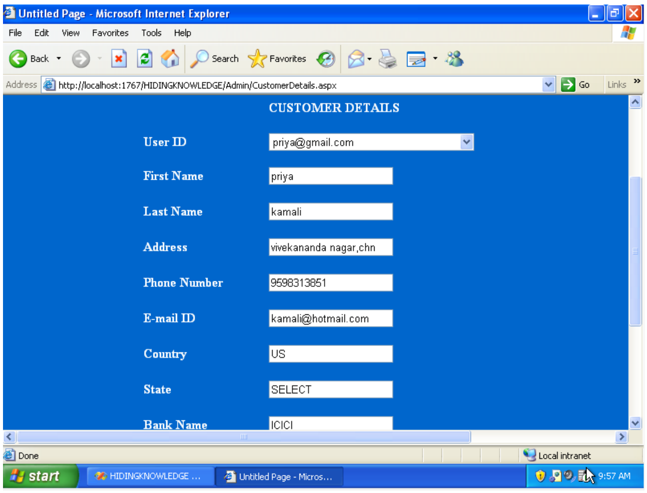
ACCOUNT ID 78138291739

DELIVERED STATUS 3days

set status

Fig: 6

In Fig 7, the customers are entitled to enter their personal details along with the contact address for communication.


Untitled Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost:1767/HIDINGKNOWLEDGE/Admin/CustomerDetails.aspx
Go Links
Done
Local intranet
start HIDINGKNOWLEDGE ... Untitled Page - Micros... 9:57 AM

CUSTOMER DETAILS

User ID priya@gmail.com

First Name priya

Last Name kamali

Address vivekananda nagar,chn

Phone Number 9598313851

E-mail ID kamali@hotmail.com

Country US

State SELECT

Bank Name ICICI

Fig: 7

We present a privacy preserving protocol for k-means clustering in the setting of arbitrarily partitioned data distributed between two parties. Our protocol is efficient and provides cryptographic privacy protection. In particular, our protocol provides the first privacy-preserving solution to k-means clustering for horizontally partitioned data.

VII. CONCLUSION

This paper has two main contributions. First, we introduce the idea of arbitrarily partitioned data, which is a generalization of both horizontally and vertically partitioned data. Protocols in this model can be applied to both horizontally and vertically partitioned data, as well as to data anywhere in between. Second, we provide a privacy preserving k-means clustering algorithm over arbitrarily partitioned data.

As previously discussed, our algorithm potentially leaks some information through the intermediate cluster assignments, even though the intermediate cluster centers themselves are not revealed. It is not clear except for some rare scenarios such as the one described that these values can yield anything useful, but obviously it would be desirable to find an efficient algorithm that it not leak the intermediate cluster assignments. A more general direction for Future work is to obtain privacy preserving data mining protocols for other data mining algorithms over arbitrarily partitioned data.

ACKNOWLEDGMENT

The authors would like to thank the editor-in-chief, the associate editor and anonymous referees for their comments.

REFERENCES

1. Agrawal.R and Srikant.R(2000)'Privacy preserving data mining',In Proc.ACM SIGMID Conf. on Management of Data,pages 439-450.ACM Press
2. Ali Inan, Yucel Saygin, Erkay Savas, Ayca Azgın Hintoglu, Albert Levi,2006. 'Privacy Preserving Clustering on Horizontally Partitioned Data', Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)
3. Chris Clifton,(2001)'Privacy Preserving Distributed Data Mining'
4. Geetha Jagannathan, Krishnan Pillaipakkamnatt and Rebecca N. Wright,'A New Privacy-Preserving Distributed k-Clustering Algorithm'
5. Goethals.B,Laur.S,Lipmaa.H, and Mielikainen.T,(2004)'On Secure scalar product computation for privacy-preserving data mining'.In The 7th Annual International Conf. in Information Security and Cryptology
6. Golreich.O(2004)'Foundations of Cryptography',Vol II Cambridge University Press
7. Krishna Prasad.P and Pandu Rangan.C(2007) 'Privacy Preserving BIRCH Algorithm for Clustering over Arbitrarily Partitioned Databases' R. Alhajj et al. (Eds.): ADMA 2007, LNAI 4632, pp. 146–157, 2007. © Springer-Verlag Berlin Heidelberg
8. Lindell.Y and Pinkas.B(2000)'Privacy preserving data mining',Lecture Notes in Computer Science,1880
9. Lloyd.S.P(1982)'Least squares quantization in PCM',IEEE Transactions on Information Theory,28:129-137
10. MacQueen.J,(1967)'Some methods for classification and analysis of multivariate observations'.In Proc.Fifth Berkeley Symposium on Mathematical Statistics and probability,volume 1,pages 2 81-296
11. Maneesh Upmanyu, Anoop M. Namboodiri, Kannan Srinathan, and C.V. Jawahar(2010)'Efficient Privacy Preserving K-Means Clustering', H. Chen et al. (Eds.): PAISI 2010, LNCS 6122, pp. 154–166. © Springer-Verlag Berlin Heidelberg
12. Oliveria.S and Zaiane.O.R(2003)'Privacy preserving clustering by data transformation' In Proc.18th Brazilian Symposium on Databases, pages 304-318
13. Prakash.V.S,Shanmugam.A,Murugesan.P (2012) 'Efficient Cluster Based Privacy Preservation Data Perturbation Technique in Multi-Partitioned Datasets',European Journal of Scientific Research, ISSN 1450-216X Vol. 86 No 2 September, 2012, pp.254-263
14. Shuguo HAN, and Wee Keong NG(2007)'Multi-Party Privacy-Preserving Decision Trees for Arbitrarily Partitioned Data' International Journal Of Intelligent Control And Systems Vol. 12, No. 4, 351-358

15. Vaidya.J and Clifton.C(2003)'Privacy preserving k-means clustering over vertically partitioned data' In Proc. 9th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining.ACM Press.

AUTHORS PROFILE

S. Harippriya, received B.E degree on Computer Science and Engineering from Anna University of Technology, Coimbatore, Tamilnadu , INDIA in 2011 and pursuing M.E Software Engineering in SNS College of Technology affiliated to Anna University, Chennai.Her Research includes in Data Mining in privacy preserving. Her Published one international Journal.

Dr. T. Kalaikumaran, is presently Professor & Head in the Department of Computer Science & Engineering, SNS College of Technology affiliated to Anna University- Coimbatore, Tamilnadu, India. He received M.E. degree in Computer Science and Engineering from the Anna University, Chennai and Ph.D. degree from Anna University, Chennai.. He is interested in the research areas of data mining, spatial data mining, machine learning, uncertain data classification and clustering, pattern recognition, database management system and informational retrieval system. He has published 3 international journals. He is a member of CSI and IEEE.

Dr. S. Karthik, is presently Professor & Dean in the Department of Computer Science & Engineering, SNS College of Technology, affiliated to Anna University- Coimbatore, Tamilnadu, India. He received the M.E degree from the Anna University Chennai and Ph.D degree from Ann University of Technology, Coimbatore. His research interests include network security, web services and wireless systems. In particular, he is currently working in a research group developing new Internet security architectures and active defense systems against DDoS attacks. Dr.S.Karthik published more than 35 papers in refereed international journals and 25 papers in conferences and has been involved many international conferences as Technical Chair and tutorial presenter. He is an active member of IEEE, ISTE, IAENG, IACSIT and Indian Computer Society.