PRIVACY PRESERVING MINING OF ASSOCIATION RULESIN HORIZONTALLY DISTRIBUTED DATABASES

JebaMoses.T,*M.Tech., MISTE*[*]

K.Elavarasi, *B.E.,MISTE*[**]

J.Jayachitra,*M.E.,MISTE*[***]

Abstract—

This paper, provide a secure mining in horizontally distributed databases by using association rules. We used Fast Distributed Mining (FDM) algorithm which is unsecured distributed version of the Apriori algorithm. Horizontally Distributed Databases divides the relation tuples, so that first subset of rows is present within the site 1 and another subset is present within the site 2. Original relation is obtained by taking union of all the sets. Association rules are if/then statements that help uncover relationships between unrelated data relational database and other information repository.

Index Terms—Privacy Preserving, Data Mining, Distributed Computation, Frequent Item sets, Association Rules.

[*] *Assistant Professor, Dept of Information Technology, IFET College of Engineering*

[**] *Senior Lecturer, Dept of Information Technology,IFET College of Engineering*

[***] *Associate Professor, Dept of Information Technology, IFET College of Engineering*

1. Introduction

Tostudyheretheproblemofsecureminingofassociationrulesinhorizontallypartitioneddata bases.Inthatsetting,thereareseveralsites(orplayers)thatholdhomogeneousdatabases,i.e.,databas esthatsharethesameschemabutholdinformationondifferententities.Thegoalistofindallassociatio nruleswithsupportatleast*s*andconfidenceatleast*c*,forsomegivenminimalsupportsize*s*andconfide ncelevel*c*,thatholdintheunifieddatabase,whileminimizingtheinformationdisclosedabouttheprivatedatabasesheldbythoseplayers.

Hereinweproposeanalternativeprotocolforthesecurecomputationoftheunionofprivatesu bsets.Theproposedprotocolimprovesuponthatin[1]intermsofsimplicityandefficiencyaswellaspr ivacy.Inparticular,ourprotocoldoesnotdependoncommutativeencryptionandobliviousransfer(whatsimplifiesitsignificantlyandcontributestowardsmuchreducedcommunicationandcomputati onalcosts).Whileoursolutionisstillnotperfectlysecure,itleaksexcessinformationonlytoasmallnu mber(three)ofpossiblecoalitions,unliketheprotocolof[1]thatdisclosesinformationalsotosomesin gle players.Inaddition,weclaimthattheexcessinformation thatourprotocolmayleakislesssensitivethantheexcess informationleakedbytheprotocolof[1].

2.TheFastDistributedMiningalgorithm

The FastDistributedMining(FDM)algorithmofCheungetal.[2],whichisanunsecureddistributedversi onoftheApriorialgorithm.Itsmainideaisthatany*s*-frequentitemsetmustbealsolocally*s*-frequentinatleastoneofthesites.Hence,inordertofindallglobally*s*-frequentitemsets,eachplayerrevealshislocally*s*-frequentitemsetsandthentheplayerscheckeachofthemtoseeiftheyare*s*-frequentalsoglobally.TheFDMalgorithmproceedsasfollows:

(1)Initialization

(2)CandidateSetsGeneration

(3)LocalPruning

(4)Unifyingthecandidateitemsets

(5)Computinglocalsupports

(6)BroadcastMiningResults

2.1. Overviewofthepaper

TheFDMalgorithmviolatesprivacyintwostages:InStep4,wheretheplayersbroadcasttheitemsetsthatarelocallyfrequentintheirprivatedatabases,andinStep6,wheretheybroadcastthesizesofthelocalsupportsofcandidateitemsets.KantarciogluandClifton[1]proposedsecureimplementationsofthosetwosteps.OurimprovementiswithregardtothesecureimplementationofStep4,whichisthemorecostlystageoftheprotocol,andtheoneinwhichtheprotocolof[1]leaksexcessinformation.KantarciogluandClifton'ssecureimplementationofStep4.Thendescribeouralternativeimplementationandproceedtoanalyzethetwoimplementationsintermsofprivacyandefficiencyandcomparethem.Ourprotocoloffersbetterprivacyandthatitissimplerandissignificantlymoreefficientintermsofcommunicationrounds,communicationcostandcomputationalcost.

3. KantarciogluandClifton

Protocol1istheprotocolthatwassuggestedbyKantarciogluandClifton[1]forcomputingtheunifiedlistofalllocallyfrequentitemsets,withoutdisclosingthesizesofthesubsetsnortheircontents. Theprotocolisappliedthesetofallitemsetsthatareglobally*s*-frequent. Refer to hereinafterasProtocolUNIFI-KC(<u>Uni</u>fyinglistsoflocally<u>F</u>requent<u>I</u>temsets—<u>K</u>antarciogluand<u>C</u>lifton).

ProtocolUNIFI-KCworksasfollows:First,eachplayeraddstohisprivatesubsetfakeitemsets,inordertohideitssize.Then,theplayersjointlycomputetheencryptionoftheirprivatesubsetsbyapplyingonthosesubsetsacommutativeencryption,whereeachplayeradds,inhisturn,hisownlayerofencryptionusinghisprivatesecretkey.Attheendofthatstage,everyitemsetineachsubsetisencryptedbyalloftheplayers;theusageofacommutativeencryptionschemeensuresthatallitemsetsare,eventually,encryptedinthesamemanner.Then,theycomputetheunionofthosesubsetsintheirencryptedform.Finally,theydecrypttheunionsetandremovefromititemsetswhichareidentifiedasfake.Wenowproceedtodescribetheprotocolindetail.

3.1.Securemultipartyprotocol

Aprotocolforcomputingthatfunctionwhichismuchsimplertounderstandandprogramandmuchmoreefficientthanthosegenericsolutions.ItisalsomuchsimplerthanProtocolUNIFI-KCandemployslesscryptographicprimitives.Ourprotocol(Protocol2)computesawiderrangeoffu nctions,whichwecallthresholdfunctions.

Twocommentsareinorder:

(1)Iftheindex$i$hadnotbeenpartoftheinputtothehashfunction(Steps2-3),thentwoequalcomponentsinP_1'sinputvector,say$s(i)=s(j)$,wouldhavebeenmappedtotwo equalsignatures,$s^{¢}(i)=s^{¢}(j)$.Hence,inthatcaseplayerP_2wouldhavelearntthatinP_1'sinputvectorthe$i$thand$j$thcomponentsareequal.Topreventsuchleakageofinformation,weincludetheindex$i$intheinputtothehashfunction.

(2)Aneventinwhich$s^{¢}(i)\hat{Î}Q^{¢}(i)$while$s(i)\hat{Ï}Q(i)$indicatesacollusion;Hashfunctionsaredesignedsothattheprob-abilityofsuchcollusionsisnegligible,whencetherisk ofacollusioncanbeignored.However,itispossibleforplayer$P_M$tocheckupfronttheselectedrandomkey.

ProtocolTHRESHOLDoperatescorrectlyifthein-equalityverificationsinStep7arecarriedoutcorrectly,since$(s(i)+s_M(i))\mathrm{mod}(M+1)$equalsthe $i$thcomponent$a(i)$inthesumvector.TheinequalityverificationiscorrectifProtocolSETINCiscorrect.

ThesusceptibilityofProtocolTHRESHOLD-Ctocoalitionsisnotverysignificantbecauseoftworeasons:

- Theentriesofthesumvector**a**donotrevealinformationaboutspecificinputvectors.Namely,knowingthat$a(i)=p$onlyindicatesthat$p$outofthe$M$bits$b_m(i)$,$1£m£M$,equal1,butitrevealsnoinformationregardingwhichofthe$M$bitsarethose.
- Thereareonlythreeplayersthatcancolludeinordertolearninformationbeyondtheintentionoftheprotocol.Suchasituationisfarlessseverethanasituationinwhichanyplayermayparticipateinacoalition,sinceifitisrevealedthatcollusiontookplace,thereisasmallsetofsuspects.

4. System architecture

Admin is used to view user details. Admin to view the item set based on the user processing details using association role with Apriori algorithm.

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository.

Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships.Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.

Apriori is designed to operate on database containing transactions. The purpose of the Apriori Algorithm is to find associations between different sets of data. It is sometimes referred to as "Market Basket Analysis". Each set of data has a number of items and is called a transaction. The output of Apriori is sets of rules that tell us how often items are contained in sets of data.

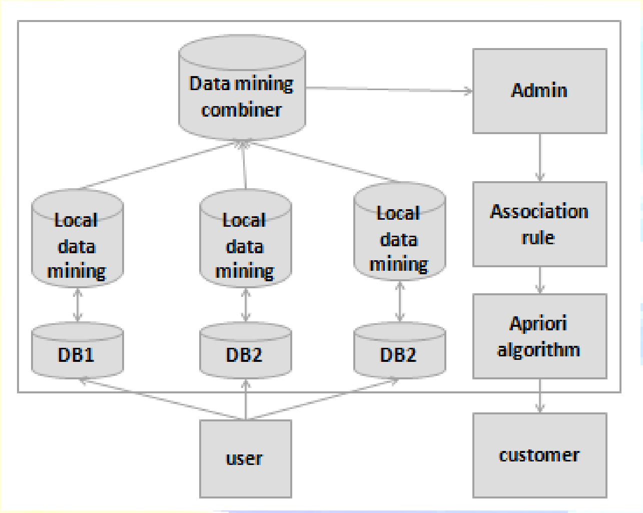


Fig.1 System architecture.

5. Privacy

WebeginbyanalyzingtheprivacyofferedbyProtocolUNIFI-KC.Thatprotocoldoesnotrespectperfectprivacysinceitrevealstotheplayersinformationthatisnotimpliedbytheirowninputandthefinaloutput.InStep11ofPhase1oftheprotocol,eachplayeraugmentsthesetX$_m$byfakeitemsets.Toavoidunnecessaryhashandencryptioncomputations,thosefakeitemsetsarerandomstringsinthecipher textdomainofthechosencommutativecipher.Hence,everyencrypted itemsetthatappearsintwodifferentlistsindicateswithhighprobabilityatrueitemsetthatislocally$s$-frequentinbothofthecorrespondingsites.Therefore,ProtocolUNIFI-KCrevealsthefollowingexcessinformation:

(1)P_1maydeduceforanysubsetoftheoddplayers,thenumberofitemsetsthatarelocallysupportedby allofthem.

(2)P_2maydeduceforanysubsetoftheevenplayers,thenumberofitemsetsthatarelocallysupportedbyallofthem.

(3)P_1maydeducethenumberofitemsetsthataresupportedbyatleastoneoddplayerandatleastone evenplayer.

(4)IfP_1andP_2collude,theyrevealforanysubsetoftheplayersthenumberofitemsetsthatarelocallysupportedbyallofthem.

5.1 Communication cost

By analyzethecommunicationandcomputationalcostsofProtocols UNIFI-KCandUNIFI.

Inevaluatingthecommunicationcost,weconsiderthreeparameters:Totalnumberofcommunicationrounds,totalnumberofmessagessent,andtheoverallsizeofthemessagessent.

The communication costs based on two protocols are,

(1) **CommunicationcostofProtocolUNIFI-KC:**AsProtocolUNIFI-KChashestheitemsetsand thenencryptsthem,*t*shouldbeatleasttherecommendedciphertextlengthincommutativeciphers.

(2) **CommunicationcostofProtocolUNIFI:**ProtocolUNIFIconsistsoffourcommunicationrounds(ineachoftheiterations):OneforStep2ofProtocolTHRESHOLDthatitinvokes;oneforStep4ofthatprotocol;oneforSteps4-5inProtocolSETINCwhichisusedfortheinequalityverificationsinProtocolTHRESHOLD;andoneforStep7inProtocolSETINC.

5.2 Computational cost

InProtocolUNIFI-KCeachoftheplayersneedstoperformhashevaluationsaswellasencryptionsanddecryptions.Asthecostofhashevaluationsissignificantlysmallerthanthecostofcommutativeencryption,wefocusonthecostofthelatteroperations.Sincecommutativeencryptionistypicallybasedonmodularexponentiation,theoverallcomputationalcostoftheprotocolisQ(Mt^3n)bitoperationsperplayer.

InProtocolTHRESHOLD,whichProtocolUNIFIcalls,eachplayerneedstogenerate($M-$

1)n(pseudo)random($\log_2 M$)-

bitnumbers(Step1).Then,eachplayerperforms$(M-1)n$additionsofsuchnumbersinStep1aswellasinStep3.PlayerP_1hastoperformalso$(M-2)n$additionsinStep5.Therefore,thecomputationalcostforeachplayerisQ$(Mn\log_2 M)$bitoperations.Inaddition,Players1and$M$needtoperform$n$hashevaluations.ComparedtoacomputationalcostofQ(Mt^3n)bitoperationsperplayer,weseethatProtocolUNIFIoffersasignificantlyimprovementwithrespecttoProtocolUNIFI-KCalsointermsofcomputationalcost.

6. Experimentalsetup

TocomparedtheperformanceoftwosecureimplementationsoftheFDMalgorithm. Inthefirstimplementation(denotedFDM-KC),weexecutedtheunificationstep(Step4inFDM)usingProtocolUNIFI-KC,wherethecommutativecipherwas1024-bitRSA[8];inthesecondimplementation(denotedFDM)weusedourProtocolUNIFI,wherethekeyed-hashfunctionwasHMAC[4].Inbothimplementations,weimplementedStep5oftheFDMalgorithminthesecuremannerthatwasdescribedinSection3.Wetestedthetwoimplementationswithrespecttothreemeasures:

1) Totalcomputationtimeofthecompleteprotocols(FDM-KCandFDM)overallplayers.ThatmeasureincludestheAprioricomputationtime,andthetimetoidentifythegloballys-frequentitemsets,asdescribedin Section3.(ThelattertwoproceduresareimplementedinthesamewayinbothProtocolsFDM-KCandFDM.)
2) 2)Totalcomputationtimeoftheunificationprotocolsonly(UNIFI-KCandUNIFI)overallplayers.
3) Totalmessagesize.

We ran three experiment set, where each set tested the dependenceoftheabovemeasuresonadifferentparameter:

-N-thenumberoftransactionsintheunifieddatabase,

-M- the number of players, and

- S- the threshold support size.

6.1 **Related work**

Previousworkinprivacypreservingdatamininghasconsideredtworelatedsettings.One,inwhichthedataownerandthedatamineraretwodifferententities,andanother,inwhichthedataisdistributedamongseveralpartieswhoaimtojointly performdataminingontheunifiedcorpusofdatathatthey hold.

Inthefirstsetting,thegoalistoprotectthedatarecordsfromthedataminer.Hence,thedataowneraimsatanonymizingthedatapriortoitsrelease.Themainapproachinthiscontextistoapplydataperturbation[3],[4].Theideaisthat theperturbeddatacanbeusedtoinfergeneraltrendsinthedata,withoutrevealingoriginalrecordinformation.

Inthesecondsetting,thegoalistoperformdataminingwhileprotectingthedatarecordsofeachofthedataownersfromtheotherdataowners.Thisisaproblemofsecuremulti-partycomputation.Theusualapproachhereiscryptographicratherthanprobabilistic.LindellandPinkas[7]showedhow tosecurelybuildanID3decisiontreewhenthetrainingsetisdistributedhorizontally.Linetal.[6]discussedsecureclusteringusingtheEMalgorithmoverhorizontallydistributeddata.Theproblemofdistributedassociationruleminingwas studiedin[5],[9],[10]intheverticalsetting,whereeachpartyholdsadifferentsetofattributes,andin[1]inthehorizontalsetting.Alsotheworkof[8]consideredthisprobleminthehorizontalsetting,buttheyconsideredlarge-scalesystemsinwhich,ontopofthepartiesthatholdthedatarecords(resources)therearealsomanagerswhicharecomputersthatassisttheresourcestodecryptmessages.

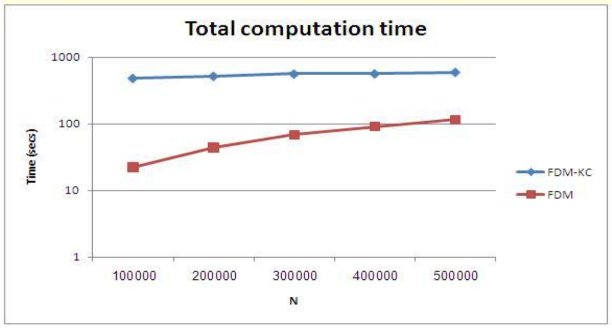


Fig.2 Computation cost

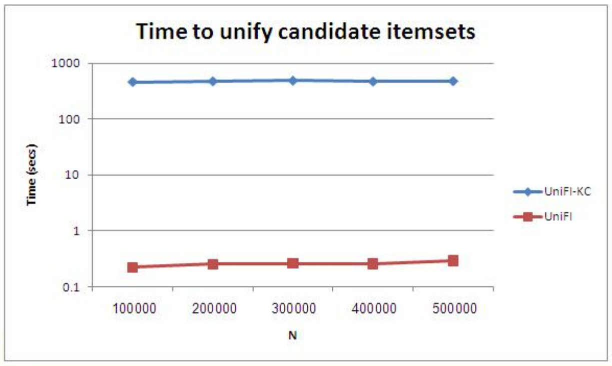


Fig.3 Communication cost

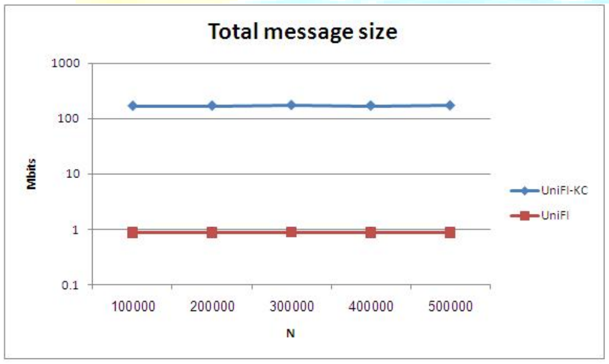


Fig.4 Number of transaction N

8. Result and Discussion

The overall result of Privacy Preserving Mining of Association Rule in Horizontally Distributed Database is discussed with help of screen shots.

The home page contains three different parties. They are client, data owner and server. Each party contains own login page to view their content.

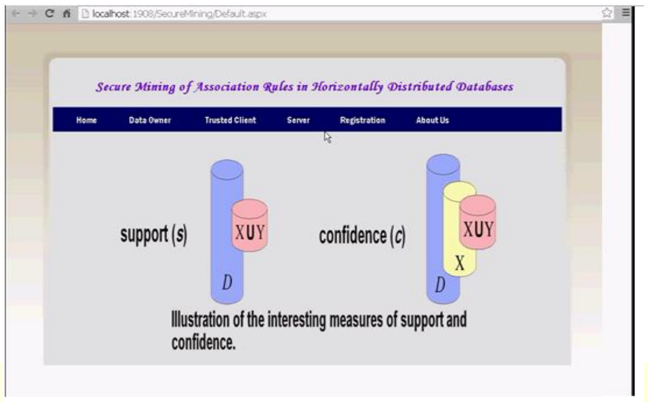
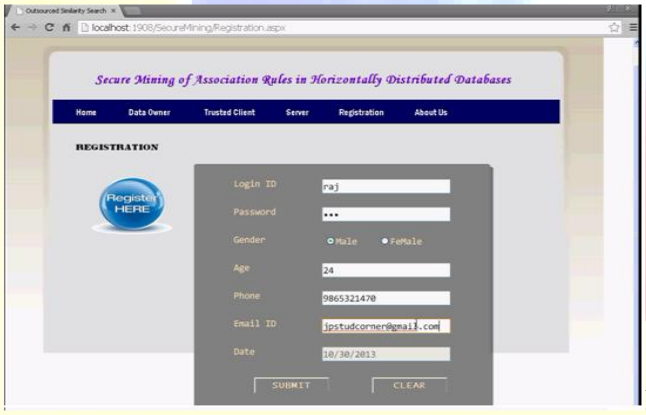

localhost:1908/SecureMining/Default.aspx
Secure Mining of Association Rules in Horizontally Distributed Databases
Home
Data Owner
Trusted Client
Server
Registration
About Us
support (s)
XUY
D
confidence (c)
XUY
X
D
Illustration of the interesting measures of support and confidence.

Fig.5 Home Page

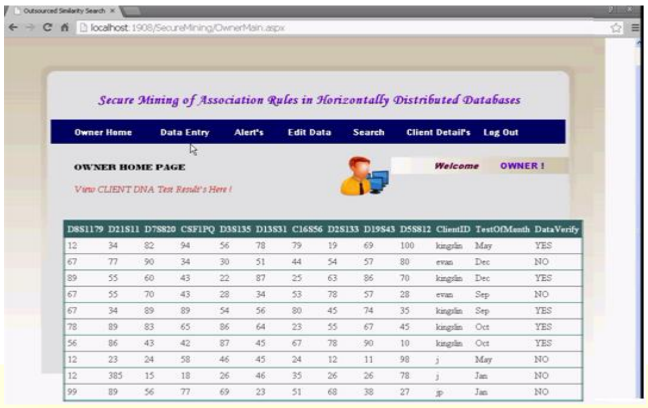
The new client wants to register to get login id and password. The data owner enters the data for particular then only record id is created. The data owner can see a client details and their content. They have right to edit the content but other cannot have rights to change the content (client, server). If server is try to change the content while verification the alert message will send to both the data owner and the client. Through the alert message they understand server trying to change the content.


Outsourced Similarity Search
localhost:1908/SecureMining/Registration.aspx
Secure Mining of Association Rules in Horizontally Distributed Databases
Home
Data Owner
Trusted Client
Server
Registration
About Us
Register HERE
SUBMIT
CLEAR

REGISTRATION

Login ID	raj
Password	•••
Gender	Male ● FeMale
Age	24
Phone	9865321470
Email ID	jpstudcorner@gmail.com
Date	10/30/2013

Fig.6 Registration page



Secure Mining of Association Rules in Horizontally Distributed Databases

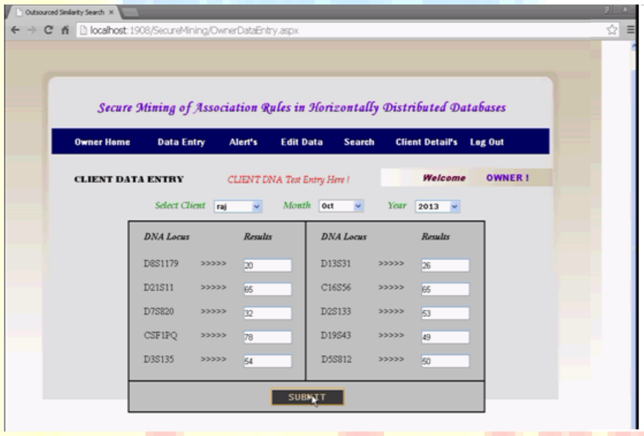
OWNER HOME PAGE

View CLIENT DNA Test Result's Here !

D8S1179	D21S11	D7S820	CSF1PQ	D3S135	D13S31	C16S56	D2S133	D19S43	D5S812	ClientID	TestOfMonth	DataVerify
12	34	82	94	56	78	79	19	69	100	kingslin	May	YES
67	77	90	34	30	51	44	54	57	80	evan	Dec	NO
89	55	60	43	22	87	25	63	86	70	kingslin	Dec	YES
67	55	70	43	28	34	53	78	57	28	evan	Sep	NO
67	34	89	89	54	56	80	45	74	35	kingslin	Sep	YES
78	89	83	65	86	64	23	55	67	45	kingslin	Oct	YES
56	86	43	42	87	45	67	78	90	10	kingslin	Oct	YES
12	23	24	58	46	45	24	12	11	98	j	May	NO
12	385	15	18	26	46	35	26	26	78	j	Jan	NO
99	89	56	77	69	23	51	68	38	27	jp	Jan	NO

Fig.7 Data owner view the client details.

The data owner has only the rights to create client id, record id and key to decrypt the data which is entered by data owner.When the server tries toedit the content but the content will not be change only the alert message will send.



Secure Mining of Association Rules in Horizontally Distributed Databases

CLIENT DATA ENTRY

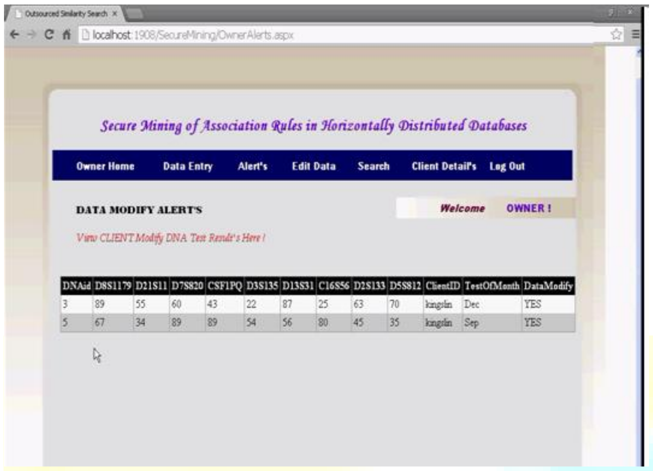
CLIENT DNA Test Entry Here !

Select Client: raj Month: Oct Year: 2013

DNA Locus		Results	DNA Locus		Results
D8S1179	>>>>>	20	D13S31	>>>>>	26
D21S11	>>>>>	65	C16S56	>>>>>	65
D7S820	>>>>>	32	D2S133	>>>>>	53
CSF1PQ	>>>>>	78	D19S43	>>>>>	49
D3S135	>>>>>	54	D5S812	>>>>>	50

Fig.8 Client data entry.

The data owner has only the rights to create client id, record id and key to decrypt the data which is entered by data owner. When the server tries toedit the content but the content will not be change only the alert message will send.


Outsourced Similarity Search
localhost:1908/SecureMining/OwnerAlerts.aspx
Secure Mining of Association Rules in Horizontally Distributed Databases
Owner Home
Data Entry
Alert's
Edit Data
Search
Client Detail's
Log Out
Welcome OWNER !

DATA MODIFY ALERT'S

View CLIENT Modify DNA Test Result's Here !

DNAid	D8S1179	D21S11	D7S820	CSF1PQ	D3S135	D13S31	C16S56	D2S133	D5S812	ClientID	TestOfMonth	DataModify
3	89	55	60	43	22	87	25	63	70	kingslin	Dec	YES
5	67	34	89	89	54	56	80	45	35	kingslin	Sep	YES

Fig.9 Alert message.

The data owner can view the data in three different ways. They are encrypted hierarchical index search, metric preserving transformation and flexible distance based hashing.

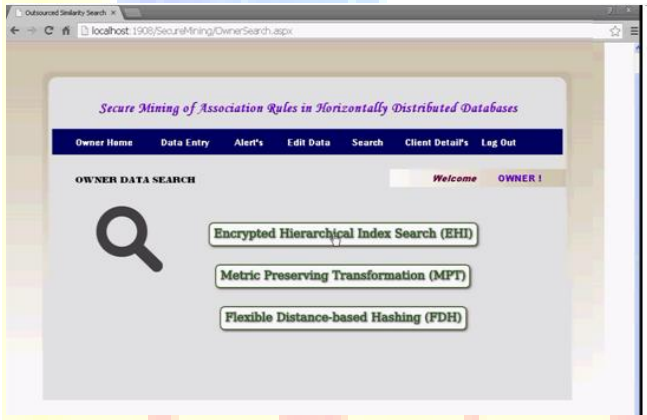
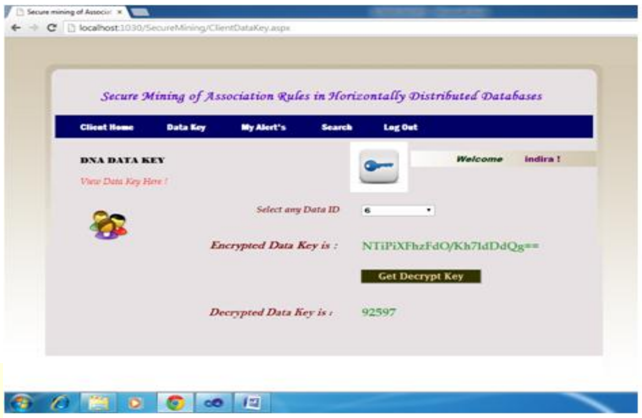

Outsourced Similarity Search
localhost:1908/SecureMining/OwnerSearch.aspx
Secure Mining of Association Rules in Horizontally Distributed Databases
Owner Home
Data Entry
Alert's
Edit Data
Search
Client Detail's
Log Out
OWNER DATA SEARCH
Welcome OWNER !
Encrypted Hierarchical Index Search (EHI)
Metric Preserving Transformation (MPT)
Flexible Distance-based Hashing (FDH)

Fig.10 Owner data search.

To get key for decrypt the data (to view in original format). The key is uniquely generated for each record. The key can view by three of them called data owner, client and server. The key also changed only by data owner. The client and the server are view and use the key but not edit.



Secure Mining of Association Rules in Horizontally Distributed Databases

DNA DATA KEY

View Data Key Here !

Welcome indira !

Select any Data ID: 6

Encrypted Data Key is : NTiPiXFhzFdO/Kh7ldDdQg==

Decrypted Data Key is : 92597

Fig.11 To get key.

8. Conclusion

I have proposedaprotocolforsecureminingofassociationrulesinhorizontallydistributeddatabasesthatimprovessignificantlyuponthecurrentleadingprotocolintermsofprivacyandefficiency.Oneofthemainingredientsinthisproposedprotocolisanovelsecuremulti-partyprotocolforcomputingtheunion(orintersection)ofprivatesubsetsthateachoftheinteractingplayersholds.Anotheringredientisaprotocolthatteststheinclusionofanelementheldbyoneplayerinasubsetheldbyanother.Thoseprotocolsexploitthefactthattheunderlyingproblemisofinterestonlywhenthenumberofplayersisgreaterthantwo. Anefficientprotocolforinequalityverificationsthatusestheexistenceofasemi honestthirdparty.Suchaprotocolmightenabletofurtherimproveuponthecommunicationandcomputationalcostsofthesecondandthirdstagesoftheprotocol. While the solutionisstillnotperfectlysecure,itleaksexcessinformationonlytoasmallnumber(three)ofpossiblecoalitions,unliketheprotocolof Fastalgorithmsforminingassociationrulesinlargedatabasesthatdisclosesinformationalsotosomesingle players.

References:

[1]M.KantarciogluandC.Clifton.Privacypreservingdistributedminingofassociationrulesonhorizontallypartitioneddata.*IEEETransactionsonKnowledgeandDataEngineering*,16:1026–1037,2004.

[2]D.W.L.Cheung,J.Han,V.T.Y.Ng,A.W.C.Fu,andY.Fu.Afastdistributedalgorithmforminingassociationrules.In*PDIS*,pages31–42,1996.

[3]R.AgrawalandR.Srikant.Privacy-preservingdatamining.In*SIGMODConference*,pages439–450,2000.

[4]A.V.Evfimievski,R.Srikant,R.Agrawal,andJ.Gehrke.Privacypreservingminingofassociationrules.In*KDD*,pages217–228,2002.

[5]M.Kantarcioglu,R.Nix,andJ.Vaidya.Anefficientapproximateprotocolforprivacy-preservingassociationrulemining.In*PAKDD*,pages515–524,2009.

[6]X.Lin,C.Clifton,andM.Y.Zhu.Privacy-preservingclusteringwithdistributedEMmixturemodeling.*Knowl.Inf.Syst.*,8:68–81,2005.

[7]Y.LindellandB.Pinkas.Privacypreservingdatamining.In*Crypto*,pages36–54,2000.

[8]A.Schuster,R.Wolff,andB.Gilburd.Privacy-preservingassociationruleminininginlarge-scaledistributedsystems.In*CCGRID*,pages411–418,2004.

[9]J.VaidyaandC.Clifton.Privacypreservingassociationrulemininginverticallypartitioneddata.In*KDD*,pages639–644,2002.

[10]J.Zhan,S.Matwin,andL.Chang.Privacypreservingcollaborativeassociationrulemining.In*DataandApplicationsSecurity*,pages153–165,2005.