ISSN: 2321-7782 (Online)

Volume 4, Issue 1, January 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Effective Integration and Analysis of Distributed Data using Secure Two Party Algorithm

R. Nithya[1]
Computer Science & Engineering
RVS Educational Trust's Group of Institutions
Dindigul, India,

K. Vimal[2]
Assistant Professor
Department of Computer Science, RVS Educational Trust's
Group of Institutions Dindigul, India

Abstract: To propose a method to securely integrate person specific sensitive data from two data providers whereby the integrated data still retains the essential information for supporting data mining tasks. The more real life scenarios are in need for simultaneous data sharing and privacy preservation of person specific sensitive data. It adopts differential privacy, a recently proposed privacy model that provides a provable privacy guarantee. A differentially private mechanism ensures that the probability of any output released data is equally likely from all nearly identical input data sets and thus guarantees that all outputs are insensitive to any individuals data. In particular, an algorithm is presented for differentially private data release for vertically partitioned data between two parties in the semi honest adversary model. To achieve this, first a two party protocol is presented for the exponential mechanism. This protocol can be used as a sub protocol by any other algorithm that requires the exponential mechanism in a distributed setting. Furthermore, two party algorithm is proposed that releases differentially private data in a secure way according to the definition of secure multiparty computation.

Key words: Essential information, two party algorithm, Vertically-partitioned, Differentially-private mechanism, Exponential mechanism.

I. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts cost or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Medical data mining

In 2011, the case of Sorrell V. IMS Health, decided by the Supreme Court of the United States ruled that pharmacies may share information with outside companies. This practice was authorized under the 1st Amendment of the Constitution protecting the freedom of speech.

Spatial data mining

Spatial data mining is the application of data mining methods to spatial data. The end objective of spatial data mining is to find patterns in data with respect to geography. So far, data mining and Geographic Information Systems (GIS) have existed as two separate technologies each with its own methods, traditions, and approaches to visualization and data analysis. Particularly, most contemporary GIS have only very basic spatial analysis functionality. The immense explosion in geographically referenced data occasioned by developments in IT, digital mapping, remote sensing, and the global diffusion emphasize the importance of developing data driven inductive approaches to geographical analysis and modelling.

Sensor data mining

Wireless sensor networks can be used for facilitating the collection of data for spatial data mining for a variety of applications such as air pollution monitoring. A characteristic of such networks is that nearby sensor nodes monitoring an environmental feature typically registers similar values. This kind of data redundancy due to the spatial correlation between sensor observations inspires the techniques for innetwork data aggregation and mining. By measuring the spatial correlation between data sampled by different sensors a wide class of specialized algorithms can be developed to develop more efficient spatial data mining algorithms.

Visual data mining

In the process of turning from analogical into digital, large data sets have been generated, collected, and stored discovering statistical patterns, trends and information which is hidden in data, in order to build predictive patterns. Studies suggest visual data mining is faster and much more intuitive than in traditional data mining.

Music data mining

Data mining techniques and in particular cooccurrence analysis has been used to discover relevant similarities among music corpora such as radio lists, CD databases for the purpose of classifying music into genres in a more objective manner.

Surveillance

Data mining has been used to fight terrorism by the U.S. government. Programs include the Total Information Awareness (TIA) program, Secure Flight known as Computer Assisted Passenger Prescreening System (CAPPS II), Analysis Dissemination Visualization Insight Semantic Enhancement (ADVISE) and the Multistate AntiTerrorism Information Exchange (MATRIX). These programs have been discontinued due to controversy over whether they violate the 4th Amendment to the United States Constitution although many programs that were formed under them continue to be funded by different organizations or under different names.

In the context of combating terrorism, two particularly plausible methods of data mining are pattern mining and subject based data mining.

Pattern mining

Pattern mining is a data mining method that involves finding existing patterns in data. In this context, patterns often mean association rules. The original motivation for searching association rules came from the desire to analyze supermarket transaction data that is to examine customer behaviour in terms of the purchased products. For example, an association rule bread ⇒ jam (80%) states that four out of five customers that bought bread also bought jam.

In the context of pattern mining as a tool to identify terrorist activity, the National Research Council provides the following definition. Pattern based data mining looks for patterns including anomalous data patterns that might be associated with terrorist activity these patterns might be regarded as small signals in a large ocean of noise. Pattern mining includes new areas such a Music Information Retrieval (MIR) where patterns seen both in the temporal and nontemporal domains are imported to classical knowledge discovery search methods.

Subject based data mining

Subject based data mining is a data mining method involving the search for associations between individuals in data. In the context of combating terrorism, the Council provides the following definition. Subject based data mining uses an initiating individual or other datum that is considered, based on other information, to be of high interest and the goal is to determine what other persons or financial transactions or movements are related to that initiating datum.

Knowledge grid

Knowledge discovery on the grid generally refers to conducting knowledge discovery in an open environment using grid computing concepts allowing users to integrate data from various online data sources as well make use of remote resources for executing their data mining tasks. The earliest example was the Discovery Net developed at Imperial College London which won the Most Innovative Data Intensive Application Award at the ACM SC02 conference and exhibition based on a demonstration of a fully interactive distributed knowledge discovery application for a bioinformatics application. Other examples include work conducted by researchers at the University of Calabria who developed Knowledge Grid architecture for distributed knowledge discovery based on grid computing.

Components of data mining algorithms

Each data mining method can be characterized in terms of four aspects.

- The models or patterns that are used to describe what is searched for in the data set. Typical models are dependency rules, clusters and decision trees.

- The scoring functions that are used to determine how well a given dataset fits the model. This is comparable to the similarity functions used in information retrieval.

- This method is applied in order to find data in the dataset that scores well with respect to the scoring function. Normally this requires efficient search algorithms that allow identifying those models that fit the data well according to the scoring functions.

- Finally the scalable implementation of the method for large datasets. Indexing techniques and efficient secondary storage management are applied.

II. Objective

The main contribution is to present a two party protocol for the exponential mechanism.Using this protocol as a sub protocol of the main algorithm and it can also be used by any other algorithm that uses the exponential mechanism in a distributed setting. The first two party data publishing algorithm for vertically partitioned data that generates an integrated data table satisfying differential privacy. The algorithm also satisfies the security definition in the secure multi party computation (SMC) literature. It also shows that the differentially private integrated data table preserves information for a data mining task. In particular, the proposed two party algorithm provides similar data utility for classification analysis when compared to the single party algorithm and it performs better than the recently proposed two party algorithm.

III. Existing System

This method investigates incremental detection of errors in distributed data. Given a distributed database D, a set ∑ of conditional functional dependencies, the set V of violations of the CFDs in D, and updates ΔD to D, it is to find with minimum data shipment, changes ΔV to V in response to ΔD.

The need for the study is evident since real life data is often dirty, distributed and frequently updated. It is often prohibitively expensive to recompute the entire set of violations when D is updated. It shows that the incremental detection problem is NPcomplete for database D that is partitioned either vertically or horizontally, even when ∑ and D are fixed.

It shows that it is bounded, there exist algorithms to detect errors such that their computational cost and data shipment are both linear in the size of ΔD and ΔV, independent of the size of the database D. It provides such incremental algorithms for vertically partitioned data and horizontally partitioned data and show that the algorithms are optimal. Further propose optimization techniques for the incremental algorithm over vertical partitions to reduce data shipment.

A. SYSTEM ARCHITECTURE

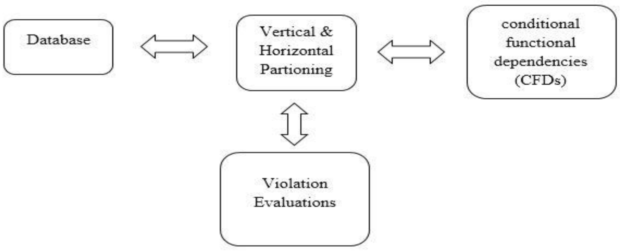

Database
Vertical & Horizontal Partioning
conditional functional dependencies (CFDs)
Violation Evaluations

Fig. 1 System Architecture For Existing System

B. DISADVANTAGES

- It is expensive to recomputed the entire set of violations when database is updated.
- Loss of information due to the poor anonymization.
- It does not satisfies the security definition of the adversary model.

IV. Proposed System

Privacy preserving data publishing addresses the problem of disclosing sensitive data when mining for useful information. Among the existing privacy models, €-differential privacy provides one of the strongest privacy guarantees. This approach addresses the problem of private data publishing, where different attributes for the same set of individuals are held by two parties.

In this model, parties follow the algorithm but may try to deduce additional information from the received messages. Therefore at any time during the execution of the algorithm no party should learn more information about the other parties data than what is found in the final integrated table which is differentially private.

In particular, an algorithm is presented for differentially private data release for vertically partitioned data between two parties in the semi honest adversary model. To achieve this, first present a two party protocol for the exponential mechanism. This protocol can be used as a subprotocol by any other algorithm that requires the exponential mechanism in a distributed setting. Furthermore, a two party algorithm is proposed that releases differentially private data in a secure way according to the definition of secure multiparty computation. Experimental results on real life data suggest that the proposed algorithm can effectively preserve information for a data mining task.

Secure Multiparty Computation

Two probability distributions are computationally indistinguishable if no efficient algorithm can tell them apart. The output distribution of every efficient algorithm is oblivious whether the input is taken from the first distribution or from the second distribution. Many of the protocols as in the case of the proposed algorithm involve the composition of secure subprotocols in which all intermediate outputs from one subprotocol are inputs to the next subprotocol. These intermediate outputs are either simulated given the final output and the local input for each party or computed as random shares. Random shares are meaningless information by themselves. Shares can be combined to reconstruct the result.

Protocol for Exponential mechanism

It presents a two party protocol for the exponential mechanism together with a detailed analysis. The exponential mechanism chooses a candidate that is close to optimum with respect to a utility function while preserving differential privacy. In the distributed setting, the candidates are owned by two parties and therefore a secure mechanism is required to compute the same output while ensuring that no extra information is leaked to any party. The protocol outputs a winner candidate depending on its score using the exponential mechanism. The scores of the candidates can be calculated using different utility functions. Given the scores of all the candidates exponential mechanism selects the candidate with the highest probability where value is the sensitivity of the chosen utility function.

Differentially Private Mechanism

The database is distributed between two parties which would like to perform data analysis on their joint data. In this setting, it would like to guarantee two sided differential privacy protecting the data of both parties. That is each party view of the protocol should be a differentially private function of the other parties input.

Generalized Problem Example

Generalize this problem as follows: A bank A and a loan company B have different sets of attributes about the same set of individuals identified by the common identifier attribute (ID), such that bank A owns DA (ID, Job, Balance), while loan company B owns DB (ID, Sex, Salary). These parties want to integrate their data to support better decision making such as loan or credit limit approvals. In addition to parties A and B, their partnered credit card company C also has access to the integrated data so all three parties A, B, and C are data recipients of the final integrated data. Parties A and B have two concerns. First, simply joining DA and DB would reveal sensitive information to the other party. Second, even if DA and DB individually do not contain person specific or sensitive information the integrated data can increase the possibility of identifying the record of an individual.

A. SYSTEM ARCHITECTURE

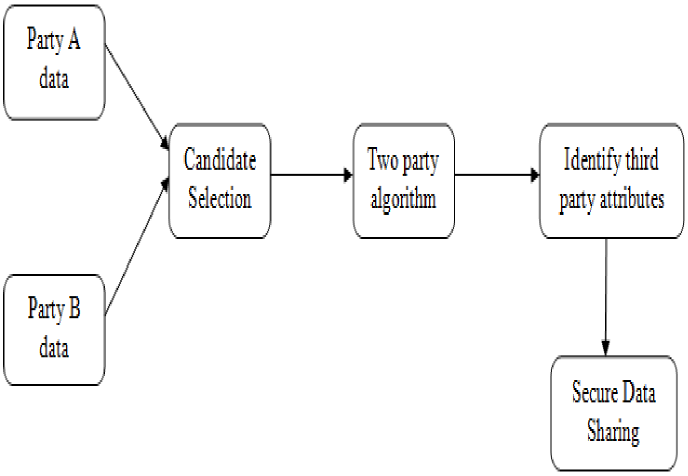


Fig. 2 System Architecture For Proposed System

B. MODULES

1. Two party data collection
2. Candidate selection
3. Two party algorithm implementation
4. Third party attributes identification
5. Sharing of data over third party

C. MODULES DESCRIPTION

1): TWO PARTY DATA COLLECTION

This module is for collecting the data records with splitting of data for two parties. Both parties have a common primary key for identification and further integration. The data consists of sensitive information and nonsensitive information. The sensitive information should be kept privacy during maintaining and sharing of data to others.

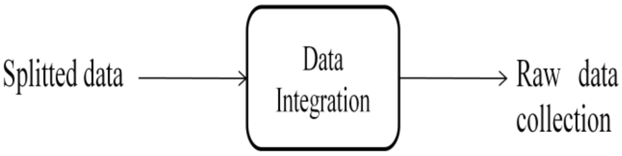


Fig. 3 Two Party Data Collection

These distributed data can be integrated to enable better data analysis for making better decisions and providing high quality services. For example, data can be integrated to improve medical research, customer service or homeland security. Data integration between autonomous entities should be conducted in such a way that no more information than necessary is revealed between the participating entities. At the same time new knowledge that results from the integration process should not be misused by adversaries to reveal sensitive information that was not available before the data integration.

2): CANDIDATE SELECTION

This module takes the output of last module. This module is to find out the candidates from the two party splitted data. The sensitive information are taken into account and used for further splitting it is called candidate selection. Both parties will have the candidate selection and the original data are grouped with the candidates.

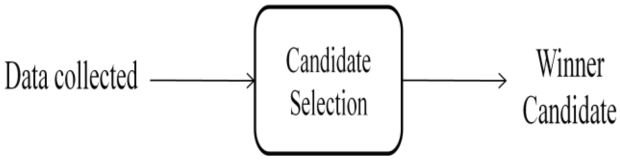


Fig. 4 Candidate Selection

The steps are

- The raw data collected is then undergoes candidate selection.
- It involves distributed exponential mechanism algorithm.
- It identifies winner candidate using exponential function and Random Value Protocol (RVP) .
- The sensitive information are taken into account for candidate selection.
- Winner candidates are grouped with original data.

3): TWO PARTY ALGORITHM IMPLEMENTATION

This module describes the two party algorithm for differentially private is data release for vertically partitioned data. This module present a Distributed Differentially private anonymization algorithm based on Generalization (DistDiffGen) for two parties. The algorithm first generalizes the raw data and then adds noise to achieve differential privacy.

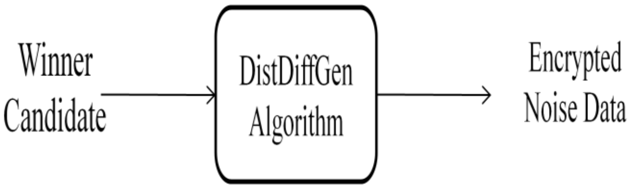


Fig. 5 Two Party Algorithm Implementation

The Distributed Differentially private anonymization algorithm based on Generalization has three functions

- Generalizing raw data: Involves UCut to split candidates based on specialization and uses DistExp Agorithm to find candidates.
- Computing the count: Involves computing the true count given by both parties.
- Computing the noisy count: Laplace noise generated by Guassian random variables added to the true count with winner candidate.

4): THIRD PARTY ATTRIBUTES IDENTIFICATION

The results from two party algorithm will deliver to the third party who is requesting the data. This module identifies the third party requirements and set attributes for each third party. It provides token for each third party as per their attributes requirements. Through this token the third party can get the data as per the attributes chosen.

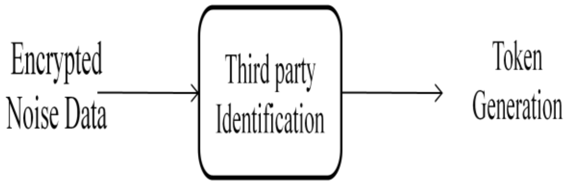


Fig. 6 Third Party Attributes Identification

The steps are

- Third party is first registered with its information.
- Each Third party attribute request is then identified.
- The Encrypted noisy data is then provided as token to each third party as per their attribute requirements.

5): SHARING OF DATA OVER THIRD PARTY

This module describes about the sharing of data over third party. For secure sharing of the data, the attribute based sharing technique is used. Each third party should have a token that describes about the attributes of the third party the token is act as a key to view the original data. This method limits the data which going outside to other third parties.

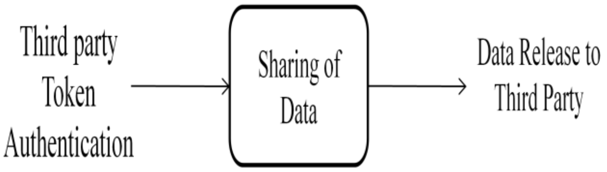
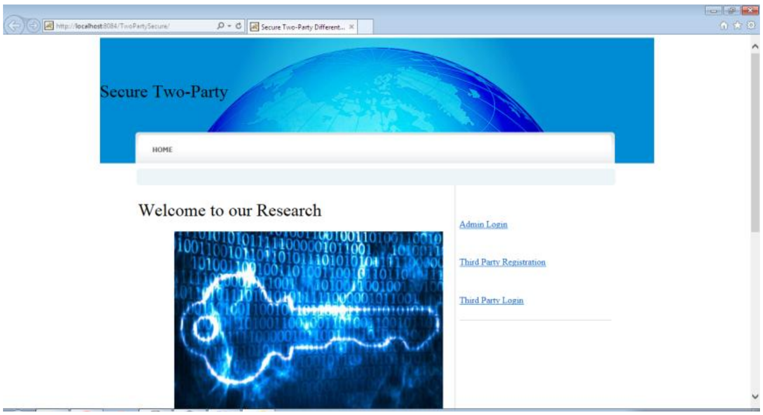


Fig. 7 Sharing Of Data Over Third Party

The steps are

- Token generated obtained by third party is then authenticated for security of the data.
- Authentication is done with the token generated by the two parties.
- Data is released or shared over the third party.

V. SCREENSHOT


http://localhost:8084/TwoPartySecure/
Secure Two-Party Different...
HOME

Secure Two-Party

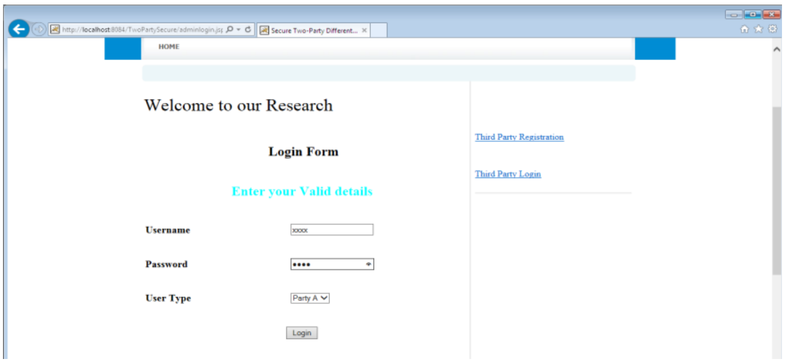
Welcome to our Research

Admin Login

Third Party Registration

Third Party Login

Fig. 8 Home Page of Secure Two Party


http://localhost:8084/TwoPartySecure/adminlogin.jsp
Secure Two-Party Different...
HOME
Login

Welcome to our Research

Login Form

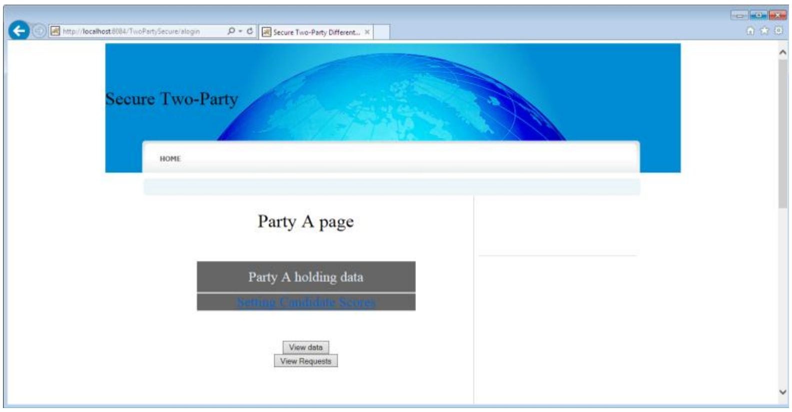
Enter your Valid details

Username	xxxx
Password	••••
User Type	Party A

Third Party Registration

Third Party Login

Fig. 9 Login Page


http://localhost:8084/TwoPartySecure/alogin
Secure Two-Party Different...
HOME
View data
View Requests

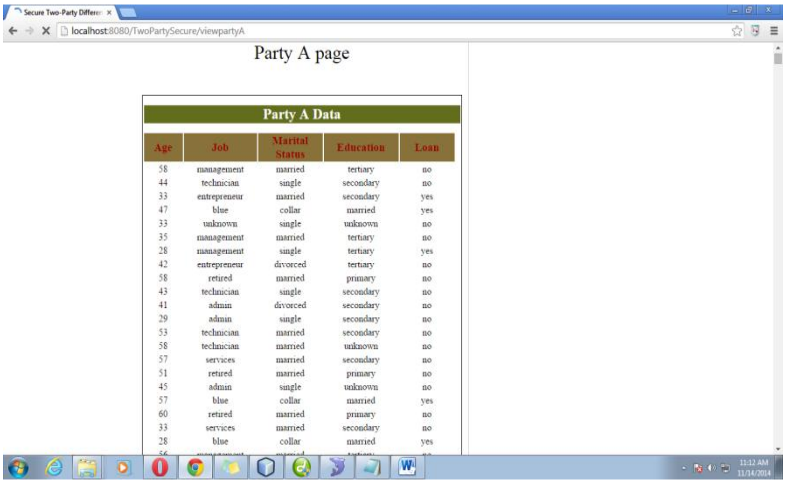
Secure Two-Party

Party A page

Party A holding data

Setting Candidate Scores

Fig.10 Data Sets


Secure Two-Party Differen
localhost:8080/TwoPartySecure/viewpartyA
11:12 AM
11/14/2014

Party A page

Party A Data

Age	Job	Marital Status	Education	Loan
58	management	married	tertiary	no
44	technician	single	secondary	no
33	entrepreneur	married	secondary	yes
47	blue	collar	married	yes
33	unknown	single	unknown	no
35	management	married	tertiary	no
28	management	single	tertiary	yes
42	entrepreneur	divorced	tertiary	no
58	retired	married	primary	no
43	technician	single	secondary	no
41	admin	divorced	secondary	no
29	admin	single	secondary	no
53	technician	married	secondary	no
58	technician	married	unknown	no
57	services	married	secondary	no
51	retired	married	primary	no
45	admin	single	unknown	no
57	blue	collar	married	yes
60	retired	married	primary	no
33	services	married	secondary	no
28	blue	collar	married	yes

Fig. 11 Party information

VI. CONCLUSIONS

Differentially private data proposed an algorithm to securely integrate person specific sensitive data from two data providers whereby the integrated data still retains the essential information for supporting data mining tasks. However, data integration between autonomous entities should be conducted in such a way that no more information than necessary is revealed between the participating entities. At the same time, new knowledge that results from the integration process should not be misused by adversaries to reveal sensitive information that was not available before the data integration.

References

1. P. A. Bernstein and Dah Ming. W. Chiu, Using semi-joins to solve relational queries, J. ACM, Jan. 1981.
2. J. A. Blakeley, P. A. Larson, and F. W. Tompa, Efficiently updating materialized views, in Proc. ACM SIGMOD, New York, NY,USA, 1986.
3. A. Gupta and J. Widom, Local verification of global integrity constraints in distributed databases, in Proc. ACM SIGMOD,Washington, DC, USA, 1993.
4. Nam Huyn, “Maintaining global integrity constraints in distributed databases,” Constraints, 1997.
5. M. Arenas, L. E. Bertossi, and J. Chomicki, “Consistent query answers in inconsistent databases,” in Proc. PODS, Philadelphia,PA, USA, 1999.
6. S. Agrawal, S. Deb, K. V. M. Naidu, and R. Rastogi, “Efficient detection of distributed constraint violations,” in Proc. ICDE, Istanbul, Turkey, 2007.
7. W. Fan, F. Geerts, X. Jia, and A. Kementsietsidis, “Conditional functional dependencies for capturing data inconsistencies,” ACM Trans. Database System Jun 2008.
8. A. Kementsietsidis, F. Neven, D. Craen, and S. Vansummeren, Scalable multi-query optimization for exploratory queries over federated scientific databases,in Proc. VLDB, Auckland, New Zealand, 2008.
9. J. Li, A. Deshpande, and S. Khuller, Minimizing communication cost in distributed multi-query processing, in Proc. ICDE, Shanghai, China, 2009.
10. W. Fan, F. Geerts, S. Ma, and H. Müller, Detecting inconsistencies in distributed data, in Proc. ICDE, Long Beach, CA, USA, 2010.

AUTHOR(S) PROFILE

R. Nithya, received his B.E degree in PSNA College of Engineering and Technology ,Dindigul, India and currently pursuing M.E degree in, India RVS School of Engineering, Dindigul. Her research interests include Data Mining, Advanced Database and Operating System .

K. Vimal, presently he is working as Assistant Professor in Computer Science and Technology at RVS Educational Trust’s Group of Institutions Dindigul, TamilNadu, India. Already he has worked as Assistant Professor in National Engineering College, KovilPatti past 7 years.