

P-1924

**Secured Mining of Association Rule**

By

**Shanthi.P**

**Reg. No. 71204621048**

Of

**KUMARAGURU COLLEGE OF TECHNOLOGY  
COIMBATORE**

**A PROJECT REPORT**

Submitted to the

**FACULTY OF INFORMATION AND COMMUNICATION ENGINEERING**

*In partial fulfillment of the requirements*

*for the award of the degree*

*of*

**MASTER OF COMPUTER APPLICATIONS**

**June, 2007**

*2007*



*Certificate*

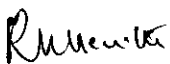
---

Kumaraguru College of Technology  
Coimbatore – 641006.

Department of Computer Applications

Bonafide Certificate

Certified that this project report titled **Secured Mining of Association Rule** is the bonafide work of **Ms. Shanthi P** who carried out the research under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.



Project Guide




Head of Department

Submitted for the University Examination held on 3.7.2007



Internal Examiner



External Examiner 3/7/07

# *Company Certificate*

---



**Date: 25<sup>th</sup> May 2007**

**To Whomsoever It May Concern**

Sub.: Student Project Completion reg.

This is to confirm the successful completion of project training of **Ms. Shanthi P – 04MCA48**, student at Master of Computer Applications department of **Kumaraguru College of Technology, Coimbatore** under the topic “**Secured Mining of Association Rules**”. The student’s behavior during the training period is good and the project duration is Jan 04, 2007 to May 25, 2007.

**For Youth Soft**

Team Leader (Programming)



*Abstract*

---

## Abstract

Protecting private data is an important concern. Data mining can extract important knowledge from large data collections. But these collections are split among the various parties. Privacy concerns may prevent one party from directly sharing the data and some type of information about the data with other parties. This proposed work addresses secure mining of association rules over horizontally partitioned data. This method incorporates cryptographic techniques to minimize information shared, while adding little overhead to the mining task.

This work begins with collecting data from various organizations and storing in a Global database. Then Global database is horizontally partitioned based on the value of a specific attribute in the database and distributed to various sites. The item sets satisfying the given association rule is encrypted and sent to other sites. Then find the union of locally supported itemsets without revealing the originator of the particular itemset.

Each site computes local and global support with the information received from other sites. If the support value of a site is larger than the threshold then it displays its item set. The global support value is computed in a secured manner without one site knowing the local support value of the other sites. Also the item sets are received in encrypted form without revealing the site information from which it originated. Thus the proposed work enables secured distributed mining of association rules on horizontally partitioned data.

# *Acknowledgement*

---



## Acknowledgement

I express my grateful thanks to our beloved principal, **Dr. Joseph V Thanikal** and our former principal **Dr. K.K.Padmanabhan**, Kumaraguru College of Technology, Coimbatore, for giving me an opportunity to take up this project.

I express my deep sense of gratitude to **Dr.S.Thangasamy**, Professor and Dean of Computer Science and Engineering department and **Dr.M.Gururajan**, Professor and Head of Department of Computer Applications for extending their help in providing all the facilities at college for the successful completion of the project.

I would like to express my gratitude to **Mrs. R. K. Kavitha**, Senior Lecturer, Department of Computer Applications, for her guidance, support, cooperation and valuable suggestions during the course of this project.

I also thank our external guide **Ms. S. Vaitheghi**, Youth Soft Technologies, Chennai, for providing me with adequate technical support and for his excellent guidance during the course of the project.

# *Contents*

---

## Table of Contents

<b>Topic</b>	<b>Page No.</b>
Abstract	iii
List of Tables	vii
List of Figures	viii
1. Introduction	1
1.1 System Overview	1
1.2 Company Profile	3
2. System Study and Analysis	5
2.1 Problem Statement	5
2.2 Literature Survey	5
2.3 Existing System	7
2.3.1 Drawbacks of the Existing System	8
2.4 Proposed System	8
2.4.1 Advantages of the Proposed System	8
2.5 Feasibility Analysis	9
2.5.1 Technical Feasibility	9
2.5.2 Operational Feasibility	9
2.5.3 Economic Feasibility	10
2.6 Application of the System	11
3. Development Environment	12
3.1 Hardware Requirements	12
3.2 Software Requirements	12

3.3 Programming Environment	13
3.3.1 JDBC	13
3.3.2 Oracle 9i	14
3.3.3 Java Swings	15
4. System Design and Development	17
4.1 Elements of Design	17
4.1.1 Modular Design	18
4.1.2 Output Design	22
4.1.3 Database Design	23
4.2 Table Structure	23
4.3 Data Flow Diagram	24
5. System Implementation and Testing	26
5.1 Implementation Overview	26
5.2 Software Testing	26
5.2.1 Unit Testing	27
5.2.2 Integration Testing	28
5.2.3 System Testing	29
6. Conclusion and Future Enhancement	30
6.1 Conclusion	30
6.2 Future Enhancement	31
Appendices	32
Reference	42

**List of Tables**

	<b>Table Description</b>	<b>Page No</b>
Table 4.2.1	Patient_details	23

## List of Figures

	<b>Figure Description</b>	<b>Page No</b>
Figure 4.1.1.1.1	Horizontal partition	19
Figure 4.1.1.2.1	Determining global candidate itemsets	20
Figure 4.1.1.3.1	Determining if itemset support exceeds threshold	21
Figure 4.3.1	Context Flow Diagram	24
Figure 4.3.2	Level 1 DFD	25

# *Introduction*

---

## CHAPTER 1

### INTRODUCTION

#### 1.1 SYSTEM OVERVIEW

The project titled "**Secured mining of association rule**" addresses secure mining of association rules over horizontally partitioned data.

Protecting private data is an important concern for society - several laws now require explicit consent prior to analysis of an individual's data. A simple loose notion of privacy is to protect only the actual data values within any transaction – as long as none of the data is known exactly, privacy is preserved. Data mining can extract important knowledge from large data collections. But sometimes these collections are split among various parties. Privacy concerns may prevent the parties from directly sharing the data and some types of information about the data. This project addresses secure mining of association rules over horizontally partitioned data. The methods incorporate cryptographic techniques to minimize the information shared, while adding little overhead to the mining task.

The goal of data mining is to extract or mine knowledge from large amounts data. However, data is often collected by several different sites. Privacy, legal and commercial concerns restrict centralized access to this data. Theoretical results from the area of secure multiparty computation in cryptography prove that assuming the existence of trapdoor permutations; one may provide secure protocols for any two party's computation as well as for any multiparty computation with honest majority. However, the general methods are far too inefficient and impractical for computing complex functions on inputs consisting of large sets of data. What remains open is come up with a set of



techniques to achieve this efficiency within a quantifiable security framework. The distributed data model considered is the homogeneous database scenario with the same set of data being collected is distributed in different sites. The dissertation presents several privacy preserving data mining algorithms operating over horizontally partitioned data. The set of underlying techniques solving independent sub-problems are also presented. Together, these enable the secure "mining" of knowledge.

In today's information age, data collection is ubiquitous, and every transaction is recorded some where. The resulting data sets can consist of terabytes or even beta bytes of data, so efficiency and scalability is the primary consideration of most data mining algorithms. Data mining technology has emerged as a means of identifying patterns and trends from large quantities of data. Most tools operate by gathering all data into a central site, then running an algorithm against that data. However, privacy concerns can prevent building a centralized warehouse and data may be distributed among several custodians, none of which are allowed to transfer their data to another site. The problem is that computing association rules. The goal is to produce association rules that hold globally while limiting the information shared about each site.

## 1.2 COMPANY PROFILE

Established in the year 2001 Youth Soft having its operation in Chennai - India, has experienced to serve global and local industrial outfits that belong to many functional Domains. Youth Soft acts as an independent application development division (Youth Soft - **Software Development Group**), which contributes a few list of application software to the Domestic and Foreign Business community.

Apart from domestic software development and support, Youth Soft has a principle tie up with M/s Valiant Ship Management Limited Hong Kong a multinational business entity involved in the Ship Management products, for IT support and specializing in Marine Asset Management, As Youth Soft - Marine Asset Management Information System Division).

Youth Soft is rigorously working towards improvising knowledge in the functional and technical areas. Youth Soft's Functional expertise extends to managing and developing Management Information Systems belonging to

- Materials Management
- Production and Maintenance Planning
- Educational Institution Management
- Portfolio Management and Safety Management procedures and protocols and middle level
- POS (Point of Sale) sector Management etc.

Youth Soft also does frequent and periodical review of technical and functional expertise, and improvises to any latest trends and technology available in the market through perennial knowledge management programs. By which Youth Soft confers to a standard development promise that it can keep up to the

requirement of the client even if falling out of boundaries than the existing expertise.

### **1.2.1 Client List**

- United Ship Management Limited - HONGKONG
- Valiant Ship Management Limited - HONGKONG
- Bumi Armada Navigation Sdn Bhd – Kuala Lumpur – Malaysia
- Comgest international - Australia
- IT Maritime - Singapore
- GE Health - India
- IGP Group of Companies - Chennai
- RTC Private Limited - Chennai
- Chettinad Vidhyashram - Chennai
- VVM Higher Secondary School - Thiruchengode - Erode DT

•

# *System Study and Analysis*

---

## CHAPTER 2

### SYSTEM STUDY AND ANALYSIS

#### 2.1 PROBLEM STATEMENT

Data mining technology has emerged as a means of identifying patterns and trends from large quantities of data (Data warehouse). Most tools operate by gathering all data into a central site, then running an algorithm against that data. However, privacy concerns can prevent building a centralized warehouse – data may be distributed among several custodians, none of which are allowed to transfer their data to another site. This project addresses the problem of computing association rules within such a scenario.

The goal of the project is to identify whether the given association rule is supported globally, while limiting the information about local support count of each site and securely finding the confidence of the rule by identifying the global item set. It also aims at finding out the local item set satisfying the rule.

#### 2.2 LITERATURE SURVEY

S. Chawla, C. Dwork, and F. McSherry's, "Toward Privacy in Public Databases" states that the probability distribution approach replaces the data with another sample from the same (or estimated) distribution or by the distribution itself, and the value distortion approach disturbs the data elements or attributes directly by either additive noise, multiplicative noise, or some other randomization procedures.

A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke's, "Privacy Preserving Mining of Association Rules" states that an additive data distribution

technique for building decision tree classifiers. Each data element is randomized by adding some random noise chosen independently. The data miner reconstructs the distribution of the original data from its perturbed version and builds the classification models.

J.J. Kim and W.E. Winkler's, "Multiplicative Noise for Masking Continuous Data" states that the use of random additive noise is shown and pointed out that additive noise can be easily filtered out in many cases that may lead to compromising the privacy.

S. Chawla, C. Dwork, and F. McSherry's, "Toward Privacy in Public Databases" states that two basic forms of multiplicative noise have been well studied. One is to multiply each data element by a random number and the other one is to take a logarithmic transformation of the data first, add predefined multivariate Gaussian noise, and take the antilog of the noise-added data.

H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar's, "On the Privacy Preserving Properties of Random Data Perturbation Techniques", states that a randomized response method was developed for the purpose of data collection through interviews and Perturbation for categorical data was initially considered.

J. Vaidya and C. Clifton's "Privacy Preserving Association Rule Mining In Vertically Partitioned Data" states that each transaction is split across multiple sites, with each site having a different set of attributes for the entire set of transactions. In vertical partitioning, the relations at the individual sites must be joined to get the relation to be mined.

## 2.3 EXISTING SYSTEM

Previous work in privacy-preserving data mining has addressed two issues. In one, the aim is preserving customer privacy by distorting the data values [1]. The idea is that the distorted data does not reveal private information, and thus is “safe” to use for mining. The key result is that the distorted data, and information on the distribution of the random data used to distort the data, can be used to generate an approximation to the original data *distribution*, without revealing the original data *values*. The distribution is used to improve mining results over mining the distorted data directly, primarily through selection of split points to “bin” continuous data. Later refinement of this approach tightened the bounds on what private information is disclosed, by showing that the ability to reconstruct the distribution can be used to tighten estimates of original values based on the distorted data [1]. More recently, the data distortion approach has been applied to Boolean association rules [2]. Again, the idea is to modify data values such that reconstruction of the values for any individual transaction is difficult, but the rules learned on the distorted data are still valid. One interesting feature of this work is a flexible definition of privacy; e.g., the ability to correctly guess a value of ‘1’ from the distorted data can be considered a greater threat to privacy than correctly learning a ‘0’. The data distortion approach addresses a different problem from our work. The assumption with distortion is that the values must be kept private from whoever is doing the mining. We instead assume that *some* parties are allowed to see *some* of the data, just that no one is allowed to see *all* the data. In return, we are able to get exact, rather than approximate results.

The problem of privately computing association rules in *vertically* partitioned distributed data has also been addressed [3]. The vertically partitioned problem occurs when each *transaction* is split across multiple sites, with each site having a different set of attributes for the entire set of transactions. With horizontal partitioning each site has a set of complete transactions. In

relational terms, with horizontal partitioning the relation to be mined is the union of the relations at the sites. In vertical partitioning, the relations at the individual sites must be joined to get the relation to be mined. The change in the way the data is distributed makes this a much different problem from the one we address here, resulting in a very different solution.

### **2.3.1 Drawbacks of the Existing System**

*The drawbacks of the existing systems can be summarized as below:*

- Information must be kept private from all parties that are doing the mining.
- Only approximate data are obtained as a result of mining.
- In vertically partitioned approach, problem occurs when each transaction is split across multiple sites and the relations at the individual sites must be joined to get the relation to be mined.

## **2.4 PROPOSED SYSTEM**

The proposed system performs mining of association rule over horizontally partitioned data, in a secured manner, by limiting the information shared among the local sites. Association rule's support threshold and confidence are computed without revealing support count at individual sites. Item sets satisfying the association rule are transmitted in encrypted form.

### **2.4.1 Advantages of the Proposed System**

*The expected benefits of the Proposed System are as follows:*

- Privacy preserving
- Exchange of information without violating security policies.
- Secured multi-party computation forbids leakage of any information other than the final result.



- Secured computation of confidence of the association rule.
- Horizontally partitioning enables each site to possess the complete set of transactions.

## **2.5 FEASIBILITY ANALYSIS**

Feasibility analysis is the measure of how beneficial or practical the development of Information System will be to the Organization. Once the problem is explained information is gathered about the system to test whether the system is viable Technically, Financially and Operationally. Thus, feasibility study is carried out in three phases as follows:

### **2.5.1 Technical Feasibility**

Technical Feasibility is the measure of practicality of a specific technical solution and the availability of technical resources and expertise. It centers on the existing computer system (hardware, software, etc.) and to what extent it can support the new addition.

The proposed system is to be developed using Java, Oracle, which are some of the leading technologies in the market. These technologies work on all architectures i.e. on all available platforms. Hence if the system is to be executed on Linux platform later, the system can be ported across to it. This system works on any back-end (Oracle 9i, MS-Access). These features of the selected technologies are quite beneficial to the proper functioning of the system in different environments.

### **2.5.2 Operational Feasibility**

Operational Feasibility asks if the system will work when it is developed and installed. It checks for the support of the management, the current business methods, user's involvement and their attitude towards the proposed system, etc.

The proposed system can be applied in any knowledge discovering system requiring secured association rule mining. It supports in limiting information shared among various parties where the horizontally partitioned data are kept.

### **2.5.3 Economic Feasibility**

Economic Feasibility is the measure of the cost-effectiveness of the proposed system. The investment to be made in the proposed system must prove a good investment to the organization by returning benefits equal to or exceeding the costs incurred in developing the system.

The proposed benefits of the system will outweigh the costs to be incurred during system developed since the system does not require procurement of additional hardware facilities it is economically feasible. No extra cost is involved for securely transmitting data because of the deployment of simple encryption algorithm. Porting the system to work over any Back-end database and scaling the system to involve more number of sites, can be done without increase in cost.

## 2.6 APPLICATIONS OF THE SYSTEM

The few application areas where this concept of secured associative rule mining is applied are listed below.

- Knowledge discovery among intelligence services of different countries and collaboration among corporations without revealing trade secrets
- Within a single multi-national company, privacy laws in different jurisdictions may prevent sharing individual data.
- Mine health records to try to find ways to reduce the proliferation of antibiotic resistant bacteria. Insurance companies have data on patient diseases and prescriptions. Insurance companies will be concerned about sharing this data. Not only must the privacy of patient records be maintained, but insurers will be unwilling to release rules pertaining only to them.



P-1924

*Development Environment*

---

## CHAPTER 3

### DEVELOPMENT ENVIRONMENT

#### 3.1 HARDWARE REQUIREMENTS

The hardware support required for deploying the application:-

##### 3.1.1 Server Configuration

Processor	: Pentium 3 Processor or above/Athlon Processor
RAM	: Minimum 512 MB
Hard Disk	: 20GB or more

##### 3.1.2 Client Configuration

Processor	: Pentium 3/4 Processor/Athlon Processor
RAM	: Minimum 128 MB

#### 3.2 SOFTWARE REQUIREMENTS

The software support required for deployment is:-

Architectural Support	: JDBC
Operating System	: Windows XP
Database	: Oracle 9i / MS-Access
Software for Development	: Java Swings

## 3.3 PROGRAMMING ENVIRONMENT

### 3.3.1 JDBC

The Java Database Connectivity Application Programming Interface (API) is an API currently being designed by Sun Microsystems that provides a Java language interface to the X/Open SQL Call Level Interface standard. This standard provides a DBMS-independent interface to relational databases that defines a generic SQL database access framework. The most visible implementation of the X/Open SQL CLI is Microsoft's ODBC (Open Database Connectivity). This API defines a common SQL syntax and function calls that can be used by developers to send SQL commands to and retrieve data from SQL databases. ODBC-enabled applications make use of database drivers (similar in concept to other device drivers) installed on the system that allow applications to talk to a vendor's database. Using this methodology, all of the DBMS-specific code is placed inside the ODBC driver and the application developer is shielded from implementation-specific problems in theory. Practically speaking, it is sometimes difficult to completely remove vendor-specific syntax from all ODBC operations, but in most cases, it is a relatively simple task to port ODBC to run on a new database server.

#### 3.3.1.1 JDBC overview

The JDBC API is expressed as a series of abstract Java interfaces within the `java.sql` package that will be provided as part of the JDK 1.1 release. Here are the most commonly used interfaces:

- **java.sql.DriverManager**:-Manages the loading and unloading of database drivers from the underlying system.
- **java.sql.Connection**:-Handles the connection to a specific database.

- **java.sql.Statement:**-Contains an SQL statement to be passed to the database; two subtypes in this interface are the PreparedStatement (for executing a precompiled SQL statement) and the CallableStatement (for executing a database stored procedure).
- **java.sql.ResultSet:**-Contains the record result set from the SQL statement passed to the database.

JDBC-enabled applets and applications make use of database drivers to connect to remote databases. What sets JDBC apart from ODBC is that these drivers can actually be applets themselves that get uploaded to the client system at runtime. Therefore, the overall Java model of a "thin client" querying a powerful database remains.

### 3.3.2 Oracle 9i

Oracle Corporation strives to comply with industry-accepted standards and participates actively in SQL standards committees. The strengths of SQL provide benefits for all types of users, including application programmers, database administrators, managers, and end users. Technically speaking, SQL is a data sublanguage. The purpose of SQL is to provide an interface to a relational database such as Oracle, and all SQL statements are instructions to the database.

#### 3.3.2.1 Features of Oracle 9i

ORACLE 9i provides statements for a variety of tasks, including:

- Querying data
- Inserting, updating, and deleting rows in a table
- Creating, replacing, altering, and dropping objects
- Controlling access to the database and its objects
- Guaranteeing database consistency and integrity
- Supports PL/SQL

### 3.3.3 Java Swings

Swing is a platform independent, Model-View-Controller GUI framework for Java. It follows a single-threaded programming model, and possesses the following traits:

- **Platform independence:** Swing is platform independent both in terms of its expression (Java) and its implementation (non-native universal rendering of widgets).
- **Extensibility:** Swing is a highly partitioned architecture which allows for the 'plugging' of various custom implementations of specified framework interfaces: Users can provide their own custom implementation(s) of these components to override the default implementations. In general, Swing users can extend the framework by: extending existing (framework) classes; providing alternative implementations of core components.
- **Component-Oriented:** Swing is a component-based framework. The distinction between objects and components is a fairly subtle point: concisely, a component is a well-behaved object with a known/specified characteristic pattern of behaviour. Swing objects asynchronously fire events, have 'bound' properties, and respond to a well known set of commands (specific to the component.)
- **Customisable:** The visual representation of a Swing component is a composition of a standard set of elements. Typically, users will programmatically customize a standard Swing component (such as a JTable) by assigning specific Borders, Colors, Backgrounds, etc., as the properties of the component. The core component will then use



these properties (settings) to determine the appropriate renderers to use in painting its various aspects.

- **Configurable:** Swing's heavy reliance on runtime mechanisms and indirect composition patterns allow it to respond at runtime to fundamental changes in its settings. For example, a Swing-based application can change its look and feel at runtime. Further, users can provide their own look and feel implementation, which allows for uniform changes in the look and feel of existing Swing applications without any programmatic change to the application code.
  
- **Lightweight UI:** The magic of Swing's configurability is also due to the fact that it does NOT use the native host OS's GUI controls for representation, but rather 'paints' its controls programmatically, through the use of Java 2D apis. Thus, a Swing component does NOT have a corresponding native OS GUI 'peer', and is free to render itself in any way that is possible with the graphics APIs.

# *System Design and Development*

---

## CHAPTER 4

### SYSTEM DESIGN AND DEVELOPMENT

#### 4.1 ELEMENTS OF DESIGN

System Design is the most creative and challenging phase in the development of a software system. Design implies to a description of the final system and the process by which it is developed. The first step is to determine what input data is needed for the system and then to design a database that will meet the requirements of the proposed system. The next step is to determine what outputs are needed from the system and the format of the output to be produced.

During the design of the proposed system some areas where attention is required are:

- What are the inputs required and the outputs produced?
- How should the data be organized?
- What will be the processes involved in the system?
- How should the screen look?

The steps carried out in the design phase are as follows:

- Modular Design
- Output Design
- Database Design

### **4.1.1 Modular Design**

It is always difficult for any System Development team to grasp a system without breaking it into several smaller systems. These smaller systems will be a part of the original system yet they will be independent in the sense that they will incorporate within them the major functionalities of the proposed system.

A software system is always divided into several subsystems which make it easier to develop and perform tests on the whole system. The subsystems are known as the modules and the process of dividing an entire system into subsystems is known as Decomposition.

The modules identified for the proposed system are as below:

- Horizontally partitioning & distribution of data among various sites
- Secured determining of global candidate itemsets
- Securely Finding Confidence of a Rule
- Testing support threshold
- Display the item sets

#### 4.1.1.1 Horizontal partitioning & Data distribution

- The data collected is partitioned based on any criteria as requested by the application and distributed among various sites. All sites have the same Schema (homogeneous databases), but each site has information on different entities.
- Each site has a set of complete transaction and the relations at the individual sites should be joined to get the relation to be mined.

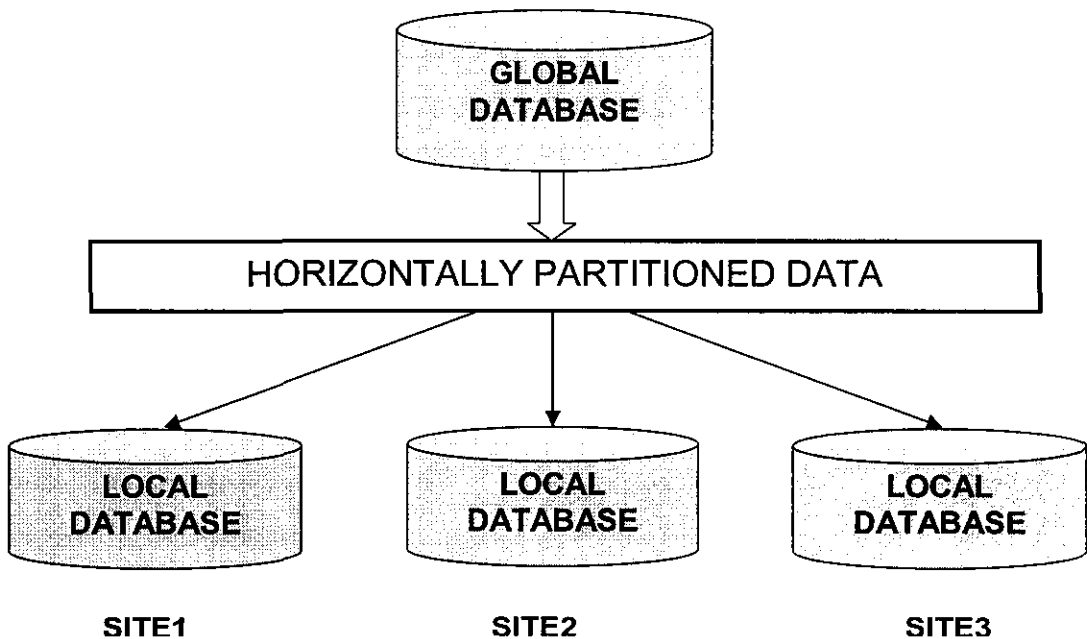
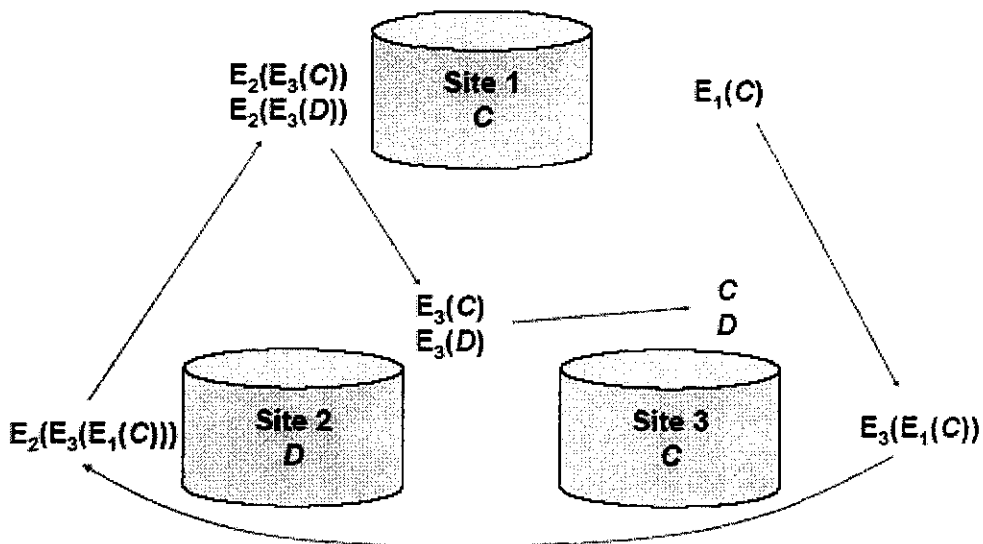


Figure 4.1.1.1.1 Horizontal partition

#### 4.1.1.2 Secured determining of global candidate itemsets

An association rule is an implication of the form  $AB \Rightarrow C$ . For example: (sex, age, hospital)  $\Rightarrow$  (disease). Let us assume that a transaction database  $DB$  is horizontally partitioned among  $n$  sites (namely  $S_1, S_2, \dots, S_n$ ) where  $DB = DB_1 \cup DB_2 \cup \dots \cup DB_n$  and  $DB_i$  resides at site  $S_i$  ( $1 = i = n$ ). The itemset  $X$  has *local* support count of  $ABC.sup_i$  at site  $S_i$  if  $ABC.sup_i$  of the transactions contains  $ABC$ . The *global* support count of  $ABC$  is given as  $ABC.sup = \sum_{i=1}^n ABC.sup_i$ .



**Figure 4.1.1.2.1 Determining global candidate itemsets**

Each party encrypts its own frequent itemsets (e.g., Site 1 encrypts itemset  $C$ ). The encrypted itemsets are then passed to other parties, until all parties have encrypted all itemsets. These are passed to a common party to eliminate duplicates, and to begin decryption. (In the figure, the full sets of itemsets are shown to the left of Site 1, after Site 1 decrypts.) This set is then passed to each party, and each party decrypts each itemset. The final result is the global candidate itemsets ( $C$  and  $D$  in the figure).

#### 4.1.1.2 Securely Finding Confidence of a Rule

The global support and confidence of an association rule  $AB \Rightarrow C$  knowing only the local supports of  $AB$  and  $ABC$ , and the size of each database are computed as below

$$\begin{aligned}
 support_{AB \Rightarrow C} &= \frac{\sum_{i=1}^{sites} support\_count_{ABC}(i)}{\sum_{i=1}^{sites} database\_size(i)} \\
 support_{AB} &= \frac{\sum_{i=1}^{sites} support\_count_{AB}(i)}{\sum_{i=1}^{sites} database\_size(i)} \\
 confidence_{AB \Rightarrow C} &= \frac{support_{AB \Rightarrow C}}{support_{AB}}
 \end{aligned}$$

Note that this doesn't require sharing any individual transactions. Since each site knows  $support\_count_{AB}(i)$  and  $support\_count_{ABC}(i)$ .

#### 4.1.1.3 Testing support threshold

Each of the locally supported itemsets is tested to see if it is supported globally.

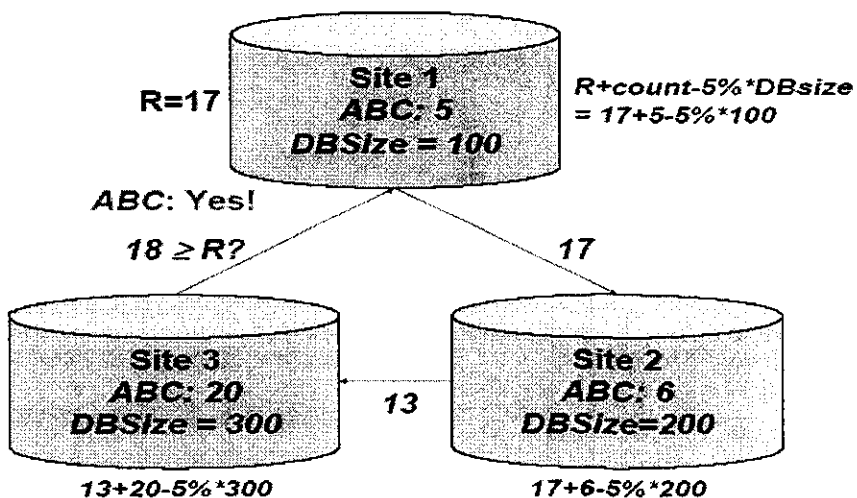


Figure 4.1.1.3.1 Determining if itemset support exceeds 5% threshold

In the figure, the itemset ABC is known to be supported at one or more sites, and each computes their local support. The first site chooses a random value R, and adds to R the amount by which its support for ABC exceeds the minimum support threshold. This value is passed to site 2, which adds the amount by which its support exceeds the threshold (note that this may be negative, as shown in the figure.) This is passed to site three, which again adds its excess support. The resulting value (18) is tested using a secure comparison to see if it exceeds the Random value (17). If so, itemset ABC is supported globally.

#### 4.1.1.4 Displaying the itemsets

Server sends minimum support threshold for the rule to all the local sites where the partitioned data is distributed. The local sites having support count exceeding the minimum threshold value returns the candidate itemsets to the server. The server displays the candidate itemsets.

#### 4.1.3 Output Design

Reports are generated as output for the users to view and take print-outs. Different reports are generated for different criteria. The reports present in the system are:

- Supported candidate itemsets list report

*Supported candidate itemsets* list report produces a list of itemset which has support count exceeding the minimum threshold value.



#### 4.1.4 Database Design

A database is a collection of inter-related data stored with minimum redundancy to serve many users quickly and efficiently. The general objective of database design is to make the data access easy, inexpensive and flexible to the user. An elegantly designed database can play a strong foundation for the whole system.

The details about the relevant data for the system are first identified. According to their relationship, tables are designed through the following method.

- The data type for each data item in the table is decided.
- The tables are then normalized.

The tables are normalized so that they can provide better response time, have data integrity, avoid redundancy and be secure.

#### 4.2 Table Structure

<b>Table No. 4.2.1      Table Name: Patient_details</b>			
This table gives the schema of the global database			
<b>No.</b>	<b>Field Name</b>	<b>Type</b>	<b>Remarks</b>
1	ID	NUMBER(5)	PK
2	NAME	VARCHAR2(18)	NOT NULL
3	SEX	CHAR	Check(F/M)
4	AGE	NUMBER(3)	Check(0<age<=100)
5	HOSPITAL	VARCHAR2(18)	Hospital's name
6	DISEASE	VARCHAR2(18)	Patient's disease
7	CITY	VARCHAR2(18)	Patient's hometown

### 4.3 Data Flow Diagrams

Data flow diagrams are graphical representation depicting information regarding the flow of control and the transformation of data from input to output. The DFD may be used to represent the system or software at any level of abstraction. In fact, DFD can be partitioned into levels. A Level 0 DFD called Context Level Diagram represents the entire software system as a single bubble with its interactions.

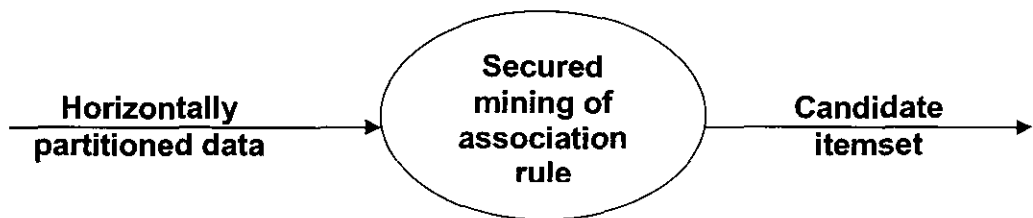


Figure 4.3.1: Context Flow Diagram

The Level 1 DFD will explain the major modules in the whole system, i.e., how the data flow between each of these modules. The interaction of each process is shown.

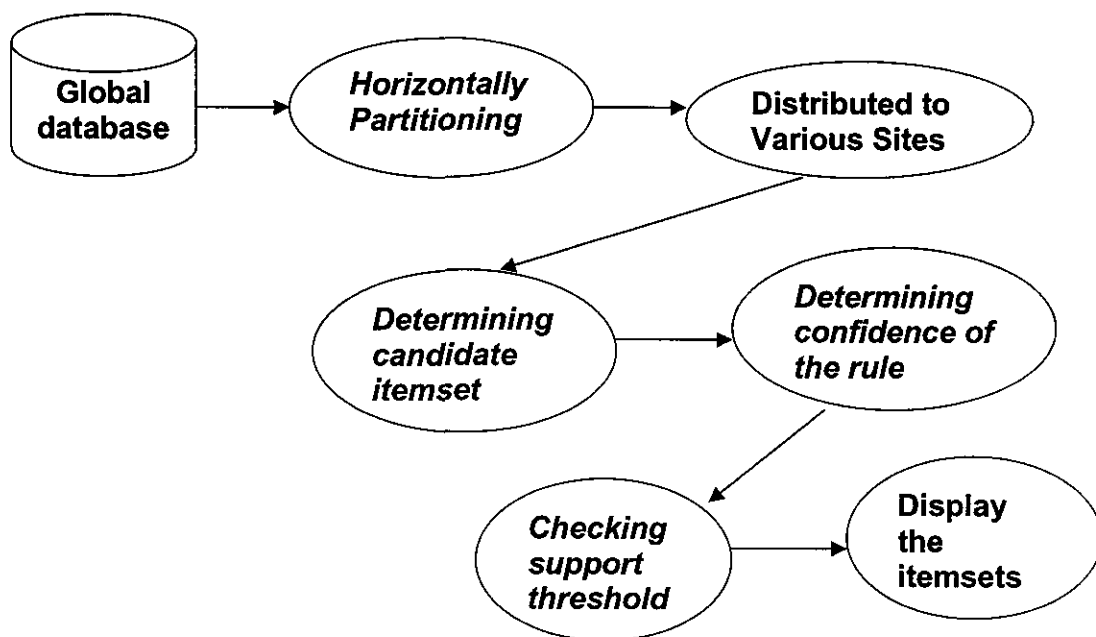


Figure 4.3.2: Level 1 DFD

*Implementation*

---

## CHAPTER 5

### SYSTEM IMPLEMENTATION AND TESTING

#### 5.1 IMPLEMENTATION OVERVIEW

Implementation includes all those activities that take place to convert from the old system to the new. The new system may be totally new, replacing an existing system.

*In this project as explained in the System design stage the secured mining of association rule is implemented in JAVA. The itemsets are transmitted in an encrypted form thus preserves privacy of individual site's data. The global support of the rule is determined.*

Without revealing the local support value of each site to other sites, using the steps followed in the project, any association rule can be mined securely without violating the security policies. Test cases are performed and its results are matching with expected results.

#### 5.2 SOFTWARE TESTING

Testing is a critical element of software quality and assurance and represents the ultimate review of specification design and coding. It is a vital activity that has to be enforced in the development of any system. This could be *done in parallel during all the phases of system development. The feedback received from these tests can be used for further enhancement of the system under consideration. The testing phase conducts test using the Software*

Requirement Specification as a reference and with the goal to see whether the system satisfies the specified requirements.

Test cases are generated for each screen. These test cases will cover every possibility which could result in both positive and negative results. These test plans are maintained for any further testing done on the system. The test plan stores information such as, the test script/input, expected output, actual output, comments and the name of the tester. This plan will be followed for all types of testing done in the system.

The main types of tests carried out on this project are:

- Unit Test
- Integration Test
- System Test

### **5.2.1 Unit Testing**

Module or Unit Testing is the process of testing all the program units that make up a system. Unit testing focuses on an individual module thus allowing one to uncover all the errors made logically and while coding in the module.

In this project, each process is tested separately as a unit. Initially the flow of control and data passing through each process is checked. When considering a module as a unit, the flow of data and control through the whole module is tested. The result is stored in the test plan. In a process, each control is further tested in unit testing. Once the errors are rectified, the testing procedure is repeated with same test cases to ensure this hasn't produced new errors. Hence this is a continuous process.

Test cases were generated to test the control flow of each unit or module. Almost all cases needed for testing control flows have been generated.

Test Cases for the Login Screen:-

Sr.No	Test Case	Expected Result	Observed Result	Status
1	User id : Admin Password: XYZ	Invalid Login	Invalid Login	Pass
2	User id : Admin Password: server	Login successful	Login Successful	Pass
3	User id : Xyz, Password: abc	Invalid Login	Invalid Login	Pass

### 5.2.2 Integration Testing

Integration testing tests the process of integrating the various modules to form the completed system. Integration starts with a set of units each individually tested in isolation and ends when the entire application has been built. Integration testing verifies that the combined units function together correctly. It facilitates in finding problem that occur at interface or communication between the individual parts.

This system follows top-down integration testing. Modules were linked to the main menu in a sequence as required in the real time operating mode of the system. Menu items were created as and when required for the integration. For eg. The distributed data at each site is verified, if it has sufficient data records satisfying the rule, in order to generate correct results.

### **5.3.3 System Testing**

System testing is actually a series of different tests, whose primary purpose is to fully exercise the computer-based system. This helps in verifying that all the system elements have been properly integrated and perform the allocated functions. It verifies the entire product after having integrated all software and hardware components, and validates it according to the original project requirement. The system testing takes into consideration the hardware, and the software. The proposed system is able to be run on any back-end database. The project is tested against recovery from errors.



*Conclusion*

---

## CHAPTER 6

### CONCLUSION AND FUTURE ENHANCEMENT

#### 6.1 CONCLUSION

Cryptographic tools can enable data mining that would otherwise be prevented due to security concerns. This project gives procedures to mine distributed association rules on horizontally partitioned data. It is shown that distributed association rule mining can be done efficiently under reasonable security assumptions.

The system performs mining of association rule over horizontally partitioned data, in a secured manner, by limiting the information shared among the local sites. It protects the individual data privacy, but it does require that each site disclose what rules it supports and how much it supports each potential global rule.

There are few application areas where this concept of secured associative rule mining is applied. Knowledge discovery among intelligence services of different countries and collaboration among corporations without revealing trade secrets. Within a single multi-national company, privacy laws in different jurisdictions may prevent sharing individual data. Mine health records to try to find ways to reduce the proliferation of antibiotic resistant bacteria.

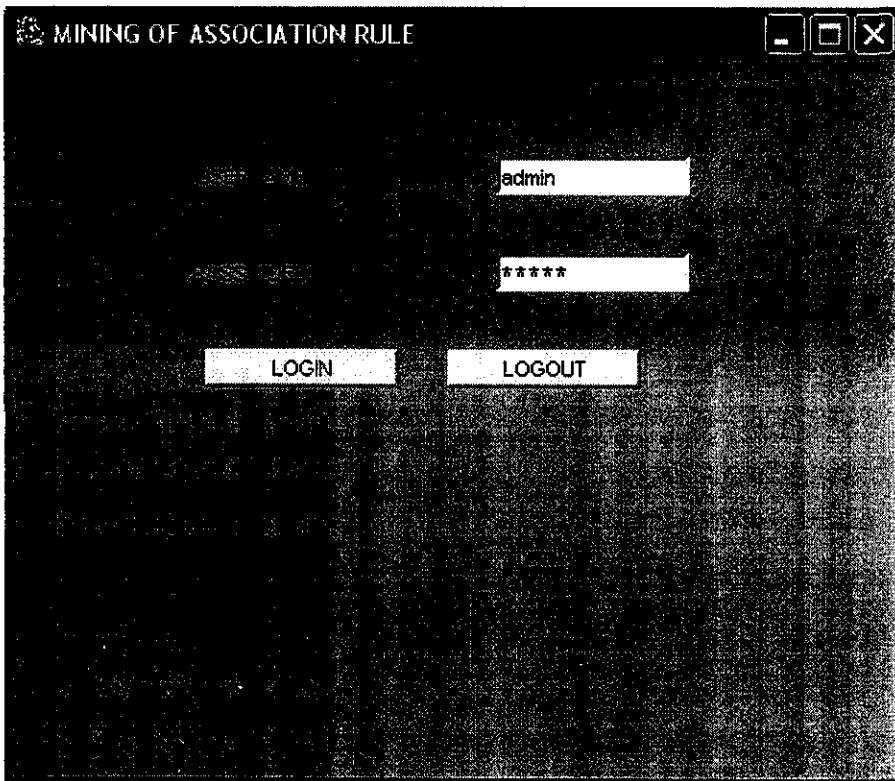
## 6.2 FUTURE ENHANCEMENTS

In future the need for mining of data where access is restricted by privacy concerns will increase. Secure algorithms for classification, clustering, etc. are required. Another possibility is secure *approximate* data mining algorithms. In summary, it is possible to mine globally valid results from distributed data without revealing information that compromises the privacy of the individual sources. Such privacy preserving data mining can be done with a reasonable increase in cost over methods that do not maintain privacy. Continued research will expand the scope of privacy-preserving data mining, enabling most or all data mining methods to be applied in situations where privacy concerns would appear to restrict.

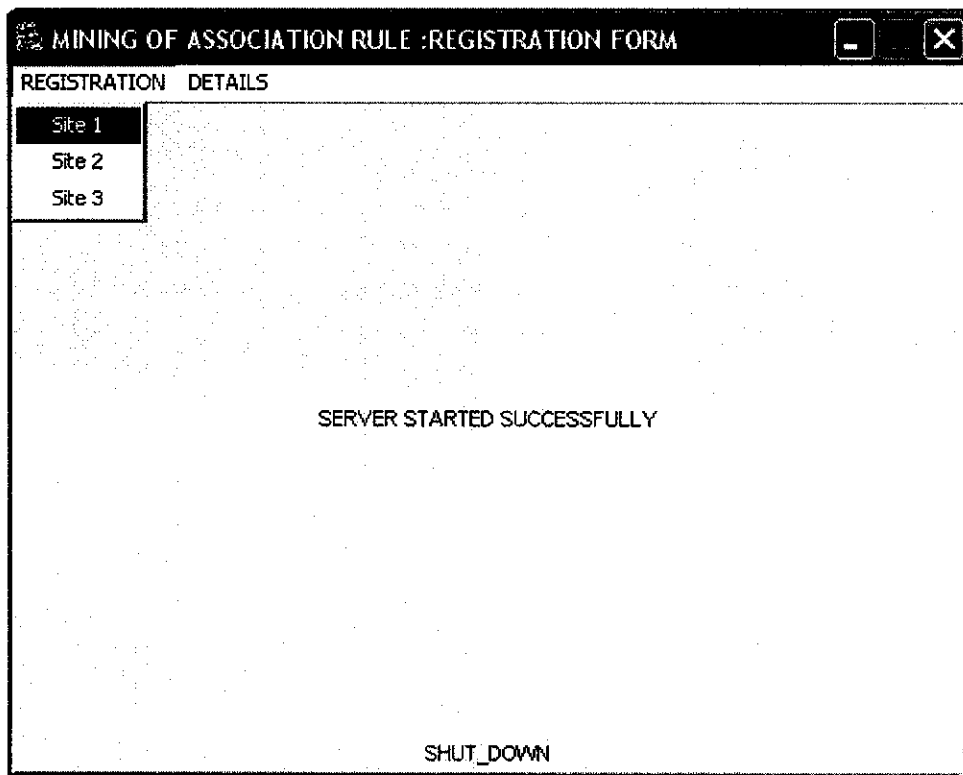
# *Appendices*

---

## APPENDICE

Login Screen

## Main Menu



## REGISTRATION FORM

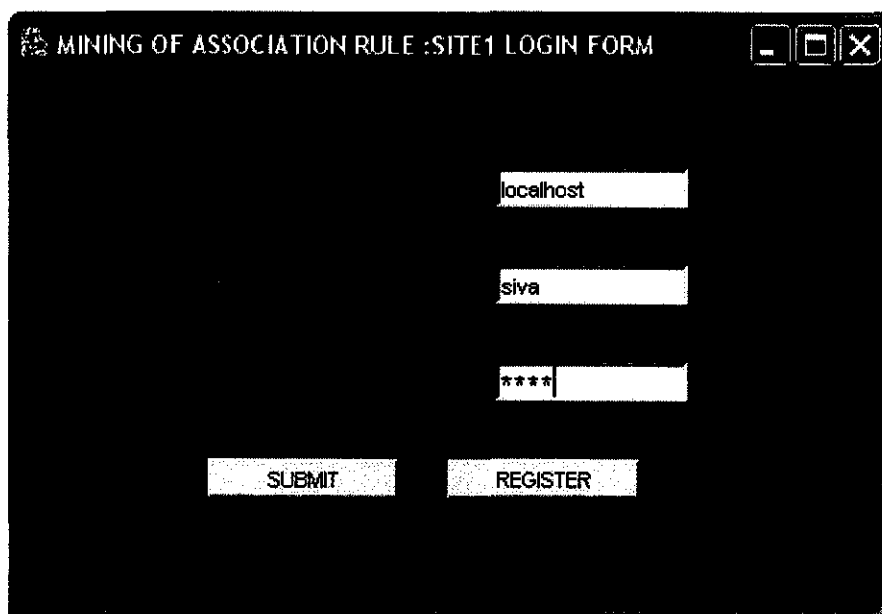
The image shows a screenshot of a software application window titled "MINING OF ASSOCIATION RULE :REGISTRATION FORM". The window contains a "REGISTRATION DETAILS" form with the following fields:

Name	Fathima.B
Age	46
Sex	F
	lifeline
	coimbatore
	ifever

A "Send" button is located below the "Age" field. A "Message" dialog box is open in the foreground, displaying the text "Successfully Added" and an "OK" button.

SHUT\_DOWN

## CLIENT1 FORMS



MINING OF ASSOCIATION RULE :SITE1 LOGIN FORM

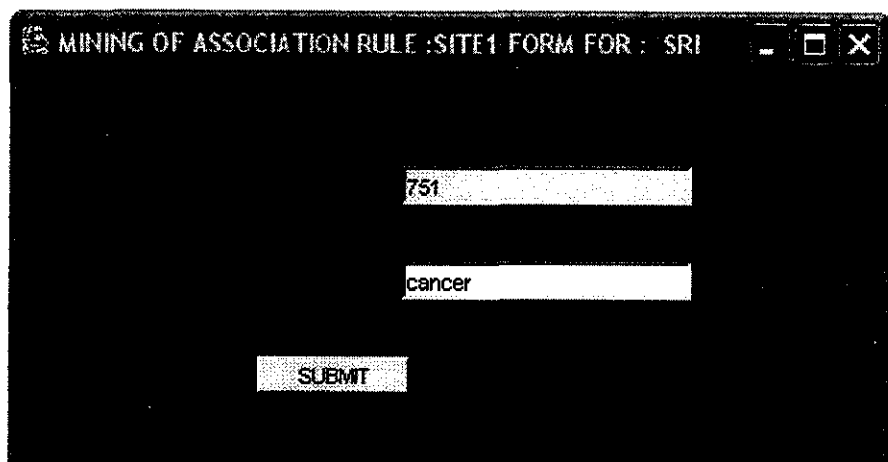
localhost

siva

\*\*\*\*

SUBMIT REGISTER

This is a screenshot of a web browser window displaying a login form. The title bar reads "MINING OF ASSOCIATION RULE :SITE1 LOGIN FORM". The form contains three input fields: the first contains "localhost", the second contains "siva", and the third contains "\*\*\*\*" (masked password). Below the input fields are two buttons: "SUBMIT" and "REGISTER".



MINING OF ASSOCIATION RULE :SITE1 FORM FOR : SRI

751

cancer

SUBMIT

This is a screenshot of a web browser window displaying a form. The title bar reads "MINING OF ASSOCIATION RULE :SITE1 FORM FOR : SRI". The form contains two input fields: the first contains "751" and the second contains "cancer". Below the input fields is a single button labeled "SUBMIT".



MINING OF ASSOCIATION RULE :SITE1 AUTHENTICATION


751

SUBMIT

MINING OF ASSOCIATION RULE :SITE1 AUTHENTICATION

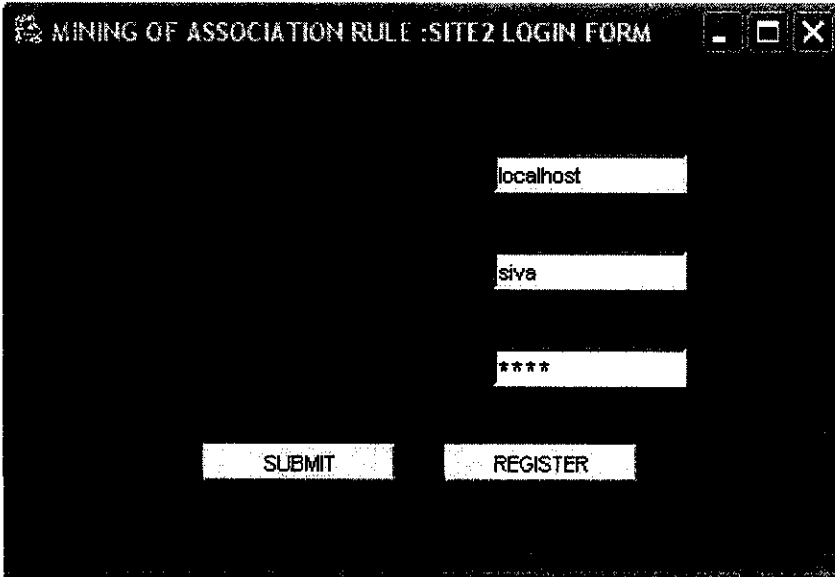
751

ALERT MESSAGE

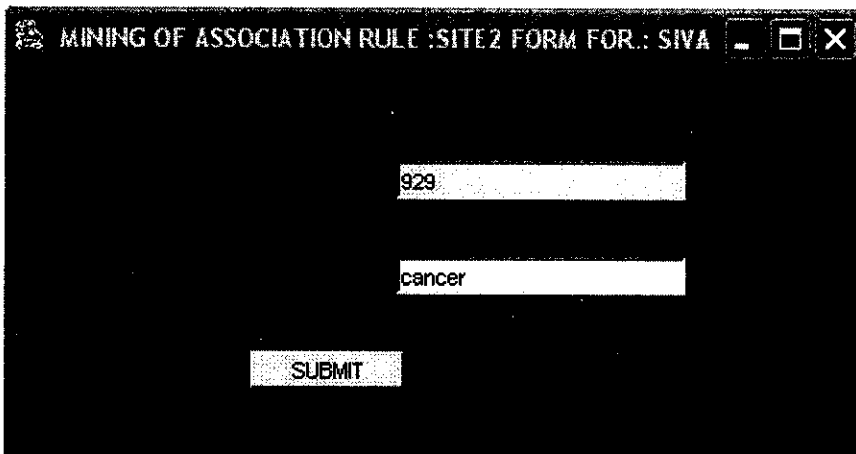
 DISEASE 'cancer' : HAVING GLOBAL SUPPORT  
YOUR SUPPORTED VALUE IS : 45

OK

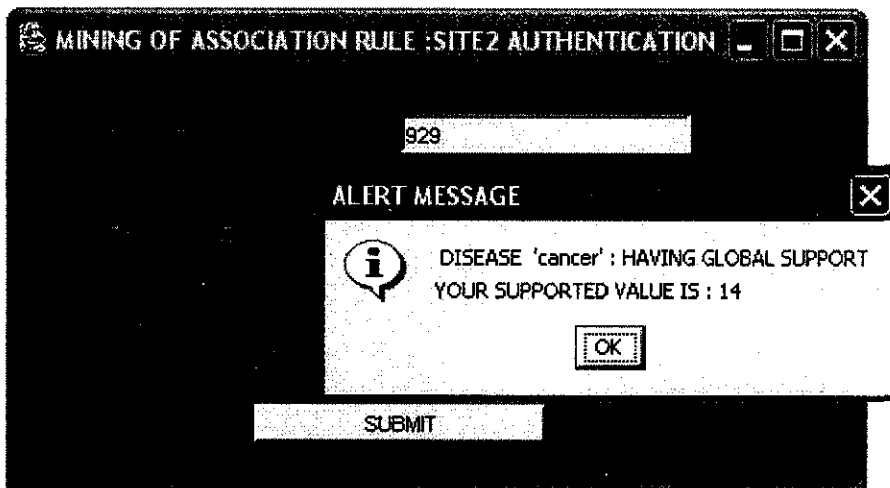
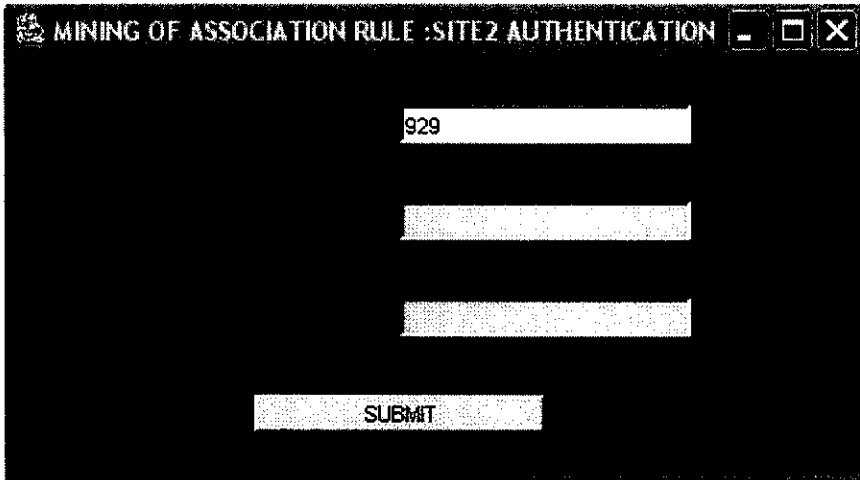
SUBMIT

**CLIENT2 FORMS**

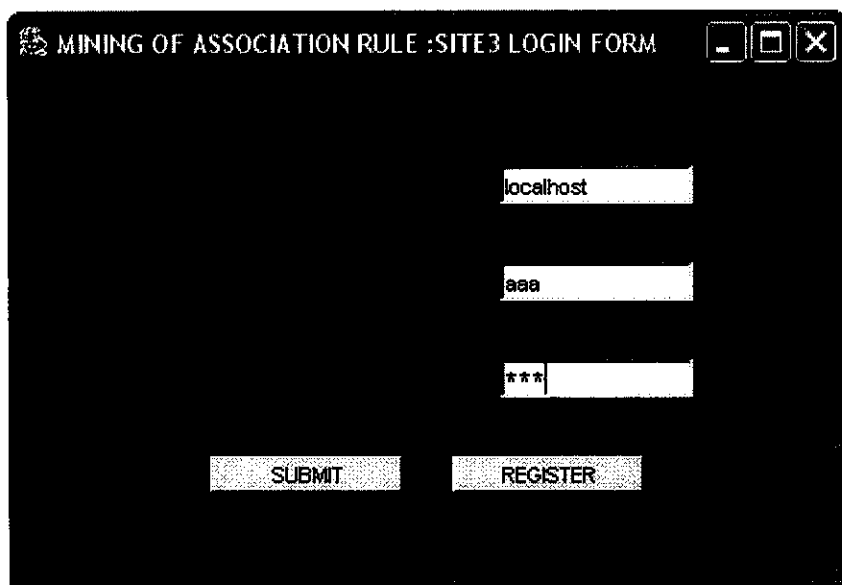
A screenshot of a web browser window titled "MINING OF ASSOCIATION RULE :SITE2 LOGIN FORM". The window contains a login form with three input fields: the first contains "localhost", the second contains "siva", and the third contains four asterisks "\*\*\*\*". Below the input fields are two buttons: "SUBMIT" and "REGISTER".



A screenshot of a web browser window titled "MINING OF ASSOCIATION RULE :SITE2 FORM FOR.: SIVA". The window contains a form with two input fields: the first contains "929" and the second contains "cancer". Below the input fields is a single button labeled "SUBMIT".



## CLIENT3 FORMS



MINING OF ASSOCIATION RULE :SITE3 LOGIN FORM

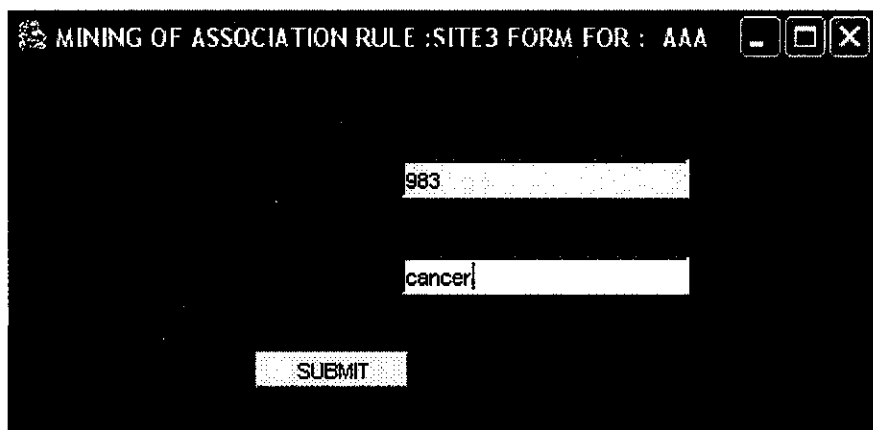
localhost

aaa

\*\*\*

SUBMIT REGISTER

This is a screenshot of a web browser window. The title bar reads "MINING OF ASSOCIATION RULE :SITE3 LOGIN FORM". The page content includes three text input fields stacked vertically. The first field contains the text "localhost", the second contains "aaa", and the third contains three asterisks "\*\*\*". Below these fields are two buttons: "SUBMIT" on the left and "REGISTER" on the right.



MINING OF ASSOCIATION RULE :SITE3 FORM FOR : AAA

983

cancer

SUBMIT

This is a screenshot of a web browser window. The title bar reads "MINING OF ASSOCIATION RULE :SITE3 FORM FOR : AAA". The page content includes two text input fields stacked vertically. The first field contains the text "983" and the second contains "cancer". Below these fields is a single button labeled "SUBMIT".

MINING OF ASSOCIATION RULE :SITE3 AUTHENTICATION


983

SUBMIT

MINING OF ASSOCIATION RULE :SITE3 AUTHENTICATION

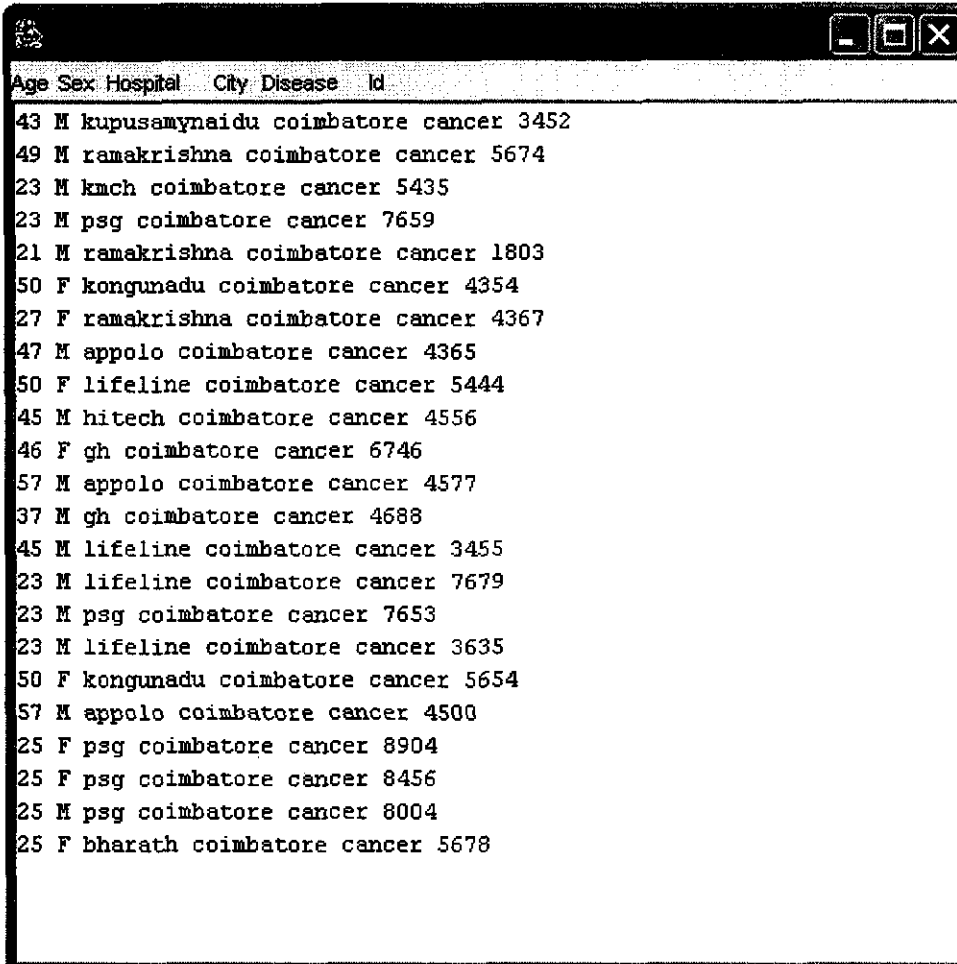
983

ALERT MESSAGE

 DISEASE 'cancer' : HAVING GLOBAL SUPPORT  
YOUR SUPPORTED VALUE IS : 17

OK

SUBMIT

**Supported Candidate Itemset Report**

Age	Sex	Hospital	City	Disease	Id
43	M	kupusamynaidu	coimbatore	cancer	3452
49	M	ramakrishna	coimbatore	cancer	5674
23	M	kmch	coimbatore	cancer	5435
23	M	psg	coimbatore	cancer	7659
21	M	ramakrishna	coimbatore	cancer	1803
50	F	kongunadu	coimbatore	cancer	4354
27	F	ramakrishna	coimbatore	cancer	4367
47	M	appolo	coimbatore	cancer	4365
50	F	lifeline	coimbatore	cancer	5444
45	M	hitech	coimbatore	cancer	4556
46	F	gh	coimbatore	cancer	6746
57	M	appolo	coimbatore	cancer	4577
37	M	gh	coimbatore	cancer	4688
45	M	lifeline	coimbatore	cancer	3455
23	M	lifeline	coimbatore	cancer	7679
23	M	psg	coimbatore	cancer	7653
23	M	lifeline	coimbatore	cancer	3635
50	F	kongunadu	coimbatore	cancer	5654
57	M	appolo	coimbatore	cancer	4500
25	F	psg	coimbatore	cancer	8904
25	F	psg	coimbatore	cancer	8456
25	M	psg	coimbatore	cancer	8004
25	F	bharath	coimbatore	cancer	5678

*Reference*

---

## REFERENCE

1. R. Agrawal and R. Srikant, "**Privacy-Preserving Data Mining**," in Proceedings of the 2000 ACM SIGMOD Conference on Management of Data , pp. 439-450.
2. A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke (2002), "**Privacy Preserving Mining Of Association Rules**," in The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 217-228.
3. J. Vaidya and C. Clifton (2002), "**Privacy Preserving Association Rule Mining In Vertically Partitioned Data**," in The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 639-644.
4. Herbert Schildt, "**The Complete Reference, Java2**", Tata McGraw-Hill Publishing Company, 2002.
5. Murat Kantarcoglu and Chris Clifton, "**Privacy Preserving Association Rule Mining In Horizontally Partitioned Data**".
6. S. Chawla, C. Dwork, and F. McSherry, "**Toward Privacy in Public Databases**", Proc. Second Theory of Cryptography Conf. (TCC'05), Feb. 2005.3
7. J.J. Kim and W.E. Winkler, "**Multiplicative Noise for Masking Continuous Data**", Technical Report Statistics #2003-01, Statistical Research Division, US Bureau of the Census, Washington D.C., Apr. 2003.
8. H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "**On the Privacy Preserving Properties of Random Data Perturbation Techniques**", Proc. IEEE Int'l Conf. Data Mining, Nov. 2003.
9. [http://en.wikipedia.org/wiki/Java\\_Database\\_Connectivity](http://en.wikipedia.org/wiki/Java_Database_Connectivity).
10. [http://en.wikipedia.org/wiki/Swing\\_\(Java\)](http://en.wikipedia.org/wiki/Swing_(Java))