ABSTRACT

Data imputation aims at filling in missing attribute values in databases. Most existing imputation methods to string attribute values are inferring-based approaches, which usually fail to reach a high imputation recall by just inferring missing values from the complete part of the data set. Recently, some retrieving-based methods are proposed to retrieve missing values from external resources such as the World Wide Web, which tend to reach a much higher imputation recall, but inevitably bring a large overhead by issuing a large number of search queries. In this paper, we investigate the interaction between the inferring-based methods and the retrieving based methods. We show that retrieving a small number of selected missing values can greatly improve the imputation recall of the inferring-based methods.

Keywords—WWW, Data imputation

I. INTRODUCTION

Data mining is used for extracting information from huge set of data.

The information or knowledge extracted so can be used for any of the applications. Such as, Market Analysis, Fraud Detection, Customer Retention, Production Control, Science Exploration, etc. Data incompleteness is a pervasive data quality problem in all kinds of databases. The process of filling in missing attribute values is well-known as Data Imputation. So far, plenty of imputation techniques have been developed for missing quantitative data, which is either continuous data such as temperature, salary, age, etc. Data with a relatively small (finite) scope of values such as weather, gender, country, etc. Only limited attention has been paid to non quantitative data which is pure string data with a large scope of values, such as email, phone, company, address, etc. However, pure string data takes up a large part of the missing data in many databases.
6. Deploying and Updating Models

The following diagram describes the relationships between each step in the process, and the technologies in Microsoft SQL Server that you can use to complete each step.

Defining the Problem

The first step in the data mining process, as highlighted in the following diagram, is to clearly define the problem, and consider ways that data can be utilized to provide an answer to the problem.

Preparing Data

The second step in the data mining process, as highlighted in the following diagram, is to consolidate and clean the data that was identified in the Defining the Problem step.

Exploring Data

The third step in the data mining process, as highlighted in the following diagram, is to explore the prepared data.

Building Models

The fourth step in the data mining process, as highlighted in the following diagram, is to build the mining model or models. You will use the knowledge that you gained in the Exploring Data step to help define and create the models.

Exploring and Validating Models

The fifth step in the data mining process, as highlighted in the following diagram, is to explore the mining models that you have built and test their effectiveness.

Deploying and Updating Models

The last step in the data mining process, as highlighted in the following diagram, is to deploy the models that performed the best to a production environment.

II. METHODOLOGIES

MODULES

- User Interface Design
- Filter Generation
- Threshold Queries
- Public DB Management
- Semantic Security

Modules Description:

User Interface Design

In this module we design the windows for the project. These windows are used to send a message from one peer to another. We use the Swing package available in Java to design the User Interface. Swing is a widget toolkit for Java. It is part of Sun Microsystems’ Java Foundation Classes an API for providing a graphical user interface for Java programs. In this module mainly we are focusing the login design page with the Partial knowledge information. Application Users need to view the application they need to login through the User Interface GUI is the media to connect User and Media Database and login screen where user can input his/her user name, password and password will check in database, if that will be a valid username and password then he/she can access the database.

Filter Generation

You can always see the result of the mix of all these components in the one file Continuous aggregation queries over dynamic data are used for real time decision making and timely business intelligence. In this paper we consider queries where a client wants to be notified over distributed data crosses a specified file. Get an overview of the process of creating portlets, learn about the concepts of the APIs used to develop portlets, and view the samples to get you started. Also, learn about integrating features such as single sign-on, cooperative sharing of information using the property broker.

Threshold Queries

The performance comparison of our threshold query protocols can be summarized in Complexity client and Complexity server, where enc. And dec. stand for encryption and decryption of bit, add. And multi. denote the homomorphic addition and multiplication of bits, and ADD. represents the homomorphic. The performance of our generic construction depends on the performance of the underlying basic constructions.
Public DB Management

Databases systems are central to most organizations’ information systems strategies. At any organizational level, users can expect to have frequent contact with database systems. Therefore, skill in using such systems – understanding their capabilities and limitations, knowing how to access data directly or through technical specialists, knowing how to effectively use the information such systems can provide, and skills in designing new systems and related applications – is a distinct advantage and necessity today. In public DB management always our data stored in encrypt format because attackers don’t understand this data.

Semantic Security

Semantic security provides measures for preventing, detaining or minimizing effects of semantic attacks. Traditional approaches to information system security focused on protecting systems and the information stored, processed and distributed on them. The goal of this project is to develop techniques to detect inconsistencies or irregularities (Behavior that breaches the rule, custom or morality) in online information, identify false information and evaluate the reliability of information sources and track those sources. A semantic attack is one in which the attacker modifies electronic information in such a way that the result is incorrect, but looks correct to the casual or perhaps even the attentive viewer. IRIA developed a categorization of semantic attacks, as well as implementing a set of techniques for detecting semantic attacks.

III. SYSTEM ALOGRITHM

Retrieving-inferring based approaches:

As observed from our experiments, the cost of an inferring operation is negligible compared to that of a web-based retrieving operation. Based on this observation, we propose an interactive retrieving and inferring data imputation approach, TRIP, which can benefit from the high recall of retrieving-based approach and the efficiency of inferring-based approach. While an inferring step in TRIP fills in all currently inferable missing values to the greatest extent, the succeeding retrieving step retrieves a set of selected missing values that make some unfilled missing values become inferable for the next inferring step. Inferring and retrieving are alternately performed until no more missing values can be imputed.

IV. CONCLUSION

We propose an innovative architecture that guarantees confidentiality of data stored in public databases. On the basis of the state of the art fully homomorphic encryption techniques, we have presented constructions Performance Comparison for disjunctive, conjunctive, and complement threshold queries based on keyword frequency and then a construction for a generic threshold query based on keyword frequency. Our protocols are semantically secure as long as the underlying fully homomorphic encryption scheme is semantically secure. Our construction for disjunctive threshold query is able to search for documents.

REFERENCES