ISSN: 0975-766X
CODEN: IJPTFI
Research Article

Available Online through
www.ijptonline.com

RECOVERING AND REBUILDING DEPARTED DATA APPROACH

Deepika.D*, Iswarya.R, Kaavya.D
Department of Information Technology, Sathyabama University- Chennai, Tamilnadu, India.
Email: sophiya2128@gmail.com

Received on 20-04-2016 *Accepted on 27-05-2016*

Abstract

One of the biggest drawback of storing data in database, if suppose the database is crashed then the entire data will be lossed. Then we cannot retrieve it from database. To provide a solution for this problem we introduce Data retrieval technique which we provide a backup while storing itself .Suppose if one database failed or crashed we can retrieve it from another storage. While storing the data will be stored in the encryption format. unique key will be generated before storing the file . This key will be attached to the data so that it can be retrieved more securely. While retrieving the data both password and private key has to be matched then only u can decrypt the data.

Introduction

In today world, storage is the main part where millions of people will store the data in database. What if the database has no backup at all. That is the point our project will be use Backup is essential for every data that is stored. Therefore we will create a back for every data that is stored by user. User can relax them selves after storing the data. Whatever happens You can retrieve your data with most efficient way. With all the unique key, password and the encryption technique you can retrieve your data in the most efficient format.

Related Works

The title is Dealing with Missing Values in Data presented by Jiri Kaiser in the year 2014. The paper Describes that Industrial data sets contain missing values due to various reasons. Equipment errors and incorrect Survey are done by every data entry procedures. By handling the analyzing of data and bias of Problems associated with missing values are loss of efficiency result in differences between missing and complete data. Selection of approach to missing values find missing data mechanism finds the various strategies for dealing with missing values. Analytical methods have their approach to handle missing values. and also an another option is Data Set Reduction. The problem of missing values can be handled by missing values imputation. It presents simple methods for missing values imputation using most common value, mean or median with closest fit approach and methods based on data mining algorithms like k-nearest neighbour, neural networks and association rules also shows their usability and presents issues with their applicability. The title is Dependencies of Conditional functionalities for Data Cleaning [6]. Presented by Philip Bohannon, Wenfei Fan, Floris Geerts. In the Year 2007. The paper Describes that Conditional functional dependencies (CFDs) proposed a class of constraints and study their applications in data cleaning. Traditional functional dependencies (FDs) developed mainly for schema design, These aim for capturing the consistency of data by incorporating bindings of semantically related values. Inference system analogous to Armstrong's axioms for FDs at consistency analysis. CFDs allow data bindings with large number of individual constraints may hold a table for complicating detection of constraint violations. Develop techniques for detecting CFD violations in SQL and novel techniques for checking multiple constraints in a single query. The performance of CFD-based methods is experimentally evaluated for inconsistency detection. It yields a constraint theory for CFDs and also a practical constraint-based method that move step toward for improving data quality.

The title is Getting Patterns and Relations from the World Wide Web [7]. Presented by Sergey Brin. In the Year 1999. The paper Describes that World Wide Web is a vast resource for information. It is extremely distributed. A particular type of data such as restaurant lists is scattered across thousands of independent information sources in many different formats. The problem of extracting a relation for such a data type from all sources automatically. Growing the target relation starting from a small sample is a technique which exploits the duality between set of patterns and relations. To test the technique they use to extract a relation of pairs from the World Wide Web. The World Wide Web provides a vast source of information of almost all types that ranging from DNA databases to resumes that lists of favorite restaurants. Many web servers and hosts that information is scattered using many different formats. Information were extracted from the World Wide Web and integrated into structured form, It would form an unprecedented source of information. The largest international directory of people include largest and most diverse databases of products with greatest bibliography of academic works, and many other useful resources.

The title is Analysis of Four Missing Data Treatment Methods for Supervised Learning [4]. Presented by Gustavo E. A. P. A. Batista and Maria Carolina Monard. In the Year 2003. The paper Describes that the relevant problem in data quality is presence of missing data. The frequent occurrence of missing data problem with many Machine Learning algorithms contain missing data with rather naive way. Missing data treatment should be carefully handled otherwise some portion error might be introduced into the knowledge induced. Analyze the use of nearest neighbor as an imputation method. Imputation Which denotes a procedure that replaces the missing values in a data set by some possible values. The approach of this advantage is that the missing data treatment is independent of the learning algorithm. Most suitable imputation method allows the user to select for each situation. Missing data imputation indicates the analysis based on the k-nearest neighbor algorithm can of internal methods with outperform used by C4.5 and CN2 treat missing data that can outperform the mean or mode imputation method of broadly used to treat missing values.

The title is Missing Attribute Values With Three Approaches—A Rough Set Perspective. Presented by Jerzy W.Grzymala-Busse. In the Year 2004. The paper Describes that the missing attribute values of new approach based on the idea of an attribute-concept value. Two other approaches of missing attribute values, based on "do not care “conditions of lost values With rough set methodology of including attribute-value pair blocks, characteristic sets, and characteristic relations. Characteristic sets are used with elementary sets of characteristic relations with generalization of the indiscernibility relation. Lower and upper approximations with three definitions are used for induction of certain rules.

The title is The Relational Web of Data Integration[8].presented by Michael J. Cafarella, Alon Halevy, Nodira Khoussainova. In the Year 2009. The paper Describes that the missing attribute values of a new approach based on the idea of an attribute-concept value. Two other approaches of missing attribute values, based on "do not care “conditions. Mainly lost values are used using rough set methodology, in attribute-value pair blocks, characteristic sets, and characteristic relations. Elementary sets are characteristic sets with characteristic relations of generalization with the indiscernibility relation. Vast amount of the web contains a structured information such as HTML tables, HTML lists and deep-web databases; Re-purposing of data with combining gives a enormous potential in creative ways. Relational web raises several challenges of integrating data which are not addressed by current data integration systems.

The title is Distributional Similarity of Web Scale with Entity Set Expansion. The paper Describes that the Pair wise semantic similarity of computing between all words on the Web is a computationally challenging task. Highly scalable to make into process of parallelization and it should be plesent depends on distributional function implemented in the Map Reduce framework and deployed over a 200 billion word crawl of the Web. Pairwise between500 million terms is computed in 50 hours using 200 quad-core nodes. The task of automatic set expansion with learned similarity matrix that present a large empirical study to quantify the effect of expansion performance in corpus size, corpus quality, seed composition and seed size. Experimental test makes a public for set expansion analysis includes a collection of diverse entity sets extracted from Wikipedia. The title is Relations from Large Plain-Text Collections of extracting from snowfall[2]. Presented by Eugene Agichtein, Luis Gravano. In the Year 2010. The paper Describes that the Valuable structured data contain text documents which is hidden in regular English sentences. Relational table of data is exploited could used for answering precise queries or running data mining tasks. Extracting a technique for tables from data set that requires only a handful of training for users. It is used to generate extraction patterns which result in new small error being extracted from the document collection. Idea was build to present a Snowball system. Plain-text documents of snowball that introduces a novel strategies for generating pattern. At each iteration of the extraction process, Quality of these patterns evaluates the snowfall and small error without human intervention that keeps only the most reliable ones for the next iteration. Scalable ranking method and measurable way for our task present here thorough experimental evaluation of Snowball and comparable techniques with collection more than 300,000 newspaper document.

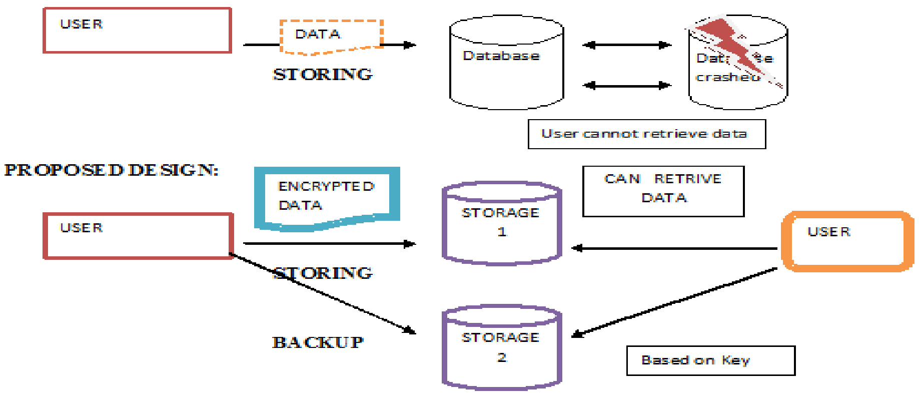
The title is The treatment of missing values with Effect in the classifier accuracy .Presented by Edgar Acuna, and Caroline Rodriguez. In the Year 2009. The paper Describes that the performance of a classifier constructed can be affected with the presence of missing values in a dataset. One used more frequently is deleting instances and also several methods have been proposed containing minimum one missing Value of feature. The effect on the wrong classification error rate will carry. The value of four methods of dealing with missing values will evaluate in twelve data sets with experiments: Mean, median, KNN imputation of the case deletion method. The KNN classifier are considered as classifiers. The Types are parametric classifier and a nonparametric classifier. The title is Effective Heuristic for Repairing Constraints by Value Modification using a cost based model Presented by Philip Bohannon, Philip Bohannon, Michael Flaster. In the Year 2005. The paper Describes that the various sources may contain inconsistencies of data gathered will damage integrity constraints. Due to minimum cost the changes will attempts to the constraint repair problem will cause and make satisfy the constraints. Repair cost is stated in terms of most previous work with tuple insertions and deletions, a database repair is considered for recent as a set of value modifications. Application of techniques which allows the novel cost framework from record linkage of search for good repairs. Minimal-cost repairs finds his model of NP-complete in the size to the database. Heuristic repair-construction will introduce an approach based on equivalence classes of attribute values. Two greedy algorithms can be defined using the approaches. Time cubic can be used in simple algorithm by using the size of the database. Duplicate-record detection which is used to develop optimizations that improve scalability. Synthetic and real data evaluate the framework of algorithms. Improved performance can be shown in proposed optimization at little with no cost in the repair of the data quality.

Survey of Existing System:

In Existing system the data that is stored in database will not be backed up. And the data will not be secured while storing data. So the main disadvantage is data cannot be retrieved if the database is crashed or the data deleted due to some issues. So the user will suffer from data retrieval options. Another major drawback is security, hacker can easily hack the data if the two level securities is not provided. Finally the data can be only inserted but not retrieved.

Normally if a data is stored in online, it will get stored in one place. Our concern is all about backing up data in two different place so that it can be retrieved securely and as well as efficiently. For example if u take online cloud Storage of such as Amazon cloud, Microsoft cloud. Once you stored in these cloud you will be worried about backup details. Even if cloud owner try to access your data they can be easily hack it by password hackers.

One of the main disadvantages of existing system is the level of security that it has had. Suppose if we store out data in online for example email, cloud we can easily retrieve the file that has be sent to us .Maximum there are only one level of security known as password .If it hacked, our data can be easily retrieved. In existing system our data will be stored in normal mode. For example if we store "hi how are u", it will be stored in the database as same way. This also one of the biggest drawback in security. In proposed system we create a backup of data while stored by the user. If one database crashed we can retrieve it from another database.



This data will be stored in cipher text format so that hacker cannot hack the data. And data will be retrieved only if the Unique key and password matched, otherwise we cannot retrieve the data.

Let’s take some of the example in proposed system. In proposed mode all our data will be stored in 16 bit encryption node. Which can be done using cipher package from java. Which we have used two mode encrypt mode and decrypt mode. We are comes to the security levels where we have provide two levels of security. First one is unique key .This generated randomly from java code, for this we have used random number java program to generate. It is 16 bit key , which will be unique for every data. Finally your password will be attached to the data and get stored. There should be a difference between our password, it can be lengthy password so that it will be more secured. Our database name is network secure and we will be accessing through mysql query browser. So in a real time application like email, if a data is deleted or crashed we cannot able to retrieve it. But in our proposed system we can able to retrieve the data.

1. User Interface Design

In user interface design module we design a registration window for the project. This window is used to send a message from one peer to another. We use a swing package available in Java to design user interface design. Swing is a widget toolkit for Java. It is a part ‘Sun Microsystems Java Foundation Classes an API for providing a GUI for Java programs. In this module mainly we are focusing the login design page with the Partial knowledge information. So all the user will be registering his/her information in database. Information consists of name, email id, age ,username, password, city, country. These information will stored in each column wise in database table. Authorized Users can only be able to view the application they need to login through the User Interface GUI is the media to connect User and Media database through login screen where user can input his/her user name, password and the password will be checked in database, if that will be a valid username and password then he/she can access the database. Only the registered user can able to login. If the user is not register it will show error. All the user must register and then login.

1.1Authorised person

In our application not all user can able to login directly .we have authentication process which only registered user can able to login the application. All these details will be in our mysql query browser

1.2 Input Expected Output For User Interface Design:

Input : User Login name and Password

Output : If Valid user Open the window otherwise error page.

2. Storage Management:

In Storage management user will upload data in database. You can select any specific data which you fell so important for your lifelong you can upload it for secure reason so you can see it on anytime from anywhere you want. While uploading user data will create backup in two database. So if one database crashed user can retrieve data from another database. From this storage user can store data more efficiently.

2.1 Cipher mode

In our application data will be stored in cipher mode. which has two sublevels encrypt and decrypt mode. so while storing the data it will converted into 16 bit encryption and stored. So we have providing much more secured data then it real time application.

2.2 Input Expected Output For Filter Generation

Input : In this process store our data in one Folder

Output : Each and every system connecting with DB their value resources.

3. Data Generic Queries

In Data Generic Querie module user will search the data based on the queries. These queries will be given by the user during upload data process. So that the queries will be attached to that particular data. It is very useful to find whether that data is available in database or not.

3.1 Performance:

The performance comparison of our Data Genric 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.

3.2 Input Expected Output for Data Genric Queries

Input : The server will take care of the connected resources

Output: The server got the number of requests from the clients separately.

4. Public DB Management:

In public database management data will be stored in two database backup .And Private key will be attached to data while storing in database. This private key will be given to the user while uploading the data. So to retrieve data both private key and password has to be matched. If the Key Matching is done successfully then the Data can be retrived.

4.1 Technical specialist

Databases systems are central to most organizations like information systems strategies. At any organizational level the users can expect to have frequent contact with uploaded files.

Therefore technical specialist use some skill for using such systems to understanding their capabilities and limitations, knowing how to access data directly and also should know how to effectively use the information in such systems that are provided, and skills for designing new systems is a distinct advantage and necessity today.

In public DB management always our data stored in encrypt format because attackers should not understand this data.

4.2 Input Expected Output for Public DB Management

Input : In Public server any user data will be stored in encrypted format.

Output : Third party doesn't understand the encrypted format data.

5. Semantic Security

In Semantic security all the data will be converted into encrypted format so that user cannot hack the data. Semantic security provided for preventing, detaining or minimizing effects of semantic attacks. It is one of the most used approaches to information system security focused on protecting systems and the information stored, processed and distributed among them.

It 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.

5.1 Semantic attacks

A semantic attack is one in which the attacker modifies technical information in such a way that the result is incorrect, but looks like correct and casual or perhaps even to the attentive user. IRIA developed a categorization of semantic attacks and also implementing a set of techniques for detecting semantic attacks. To decrypt data private key has to be given correctly. To retrieve a data key matching is necessary.

5.2 Key attaching process

While data storing in encryption mode , unique key will be attached to the data. For retrieving the data the key matching has to be done successfully. Based on this only the data will be retrieved .

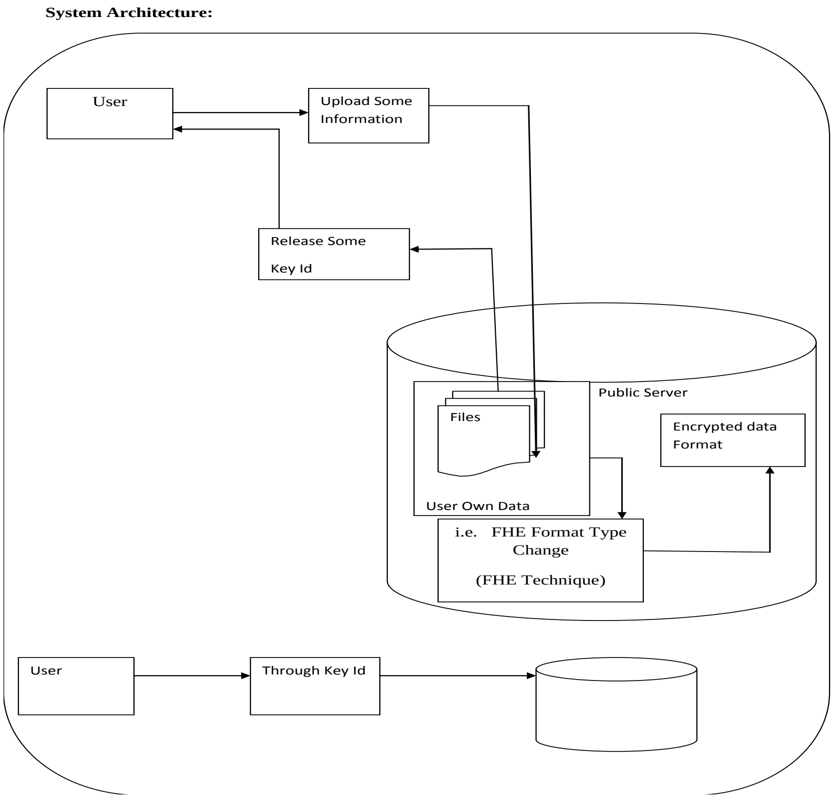
5.3 Input Expected Output For Semantic Security

Input : Data resource allocation in self organizing users

Output: To retrieve a data from a public server what the things need to be done by the requester. Public server respond only genuine key only.

System Architecture

System Architecture Diagram



System Architecture Explanation

It bring a large overhead by issuing a large number of search queries. The interaction between the methods are investigated. By retrieving a small number of selected missing values can greatly improve the imputation recall of the inferring-based methods. The interactive retrieving-inferring data imputation approach performs retrieving and inferring alternately in filling the missing attribute values in a dataset. To ensure the high recall at the minimum cost, it faces a challenge of selecting the least number of missing values for retrieving to maximize the number of inferable values. To identify an optimal retrieving-inferring scheduling scheme in Deterministic Data Imputation (DDI). Retrieving-based methods are used to retrieve missing values from external resources such as the World Wide Web, which tend to reach a much higher imputation .And the optimality of the generated scheme is theoretic retrieving-based methods are extremely used to retrieve missing values from external resources such as the optimal scheme is not feasible to be achievably analyzed with proofs. Also analyze with an example that the optimal scheme is not feasible to be achieved.

Future Enhancement:

However, pure inferring-based imputation method often fails to fill in some missing values as not every missing values are found based on the inference rules corresponding to those constraints. We say a missing value is inferable if there is at least one way to infer its value from the other existing or inferable values.

Conclusion:

We propose a hybrid retrieving-inferring data imputation approach to alternately perform retrieving and inferring in imputing missing values in a database. Finally we can store the data without worrying whether it can be crashed, deleted or backup storage. Our project provides backup data while user storing in database. And user can retrieve it in a more efficient manner with unique key and password matching process.

Acknowledgement:

We would like to express our sincere gratitude to our honourable Chancellor,

Col. Dr. JEPPIAAR, M.A., B.L., Ph.D., for his guidance and advice.

We sincerely thank our beloved Directors, **Dr. MARIE JOHNSON, B.E., M.B.A., M.Phil., Ph.D.,** and **Dr. MARIAZEENA JOHNSON, B.E., M.B.A., M.Phil., Ph.D.,** for all the help and facilities provided to complete this project successfully. We acknowledge our Vice Chancellor, **Dr.B.SHEELA RANI, M.S.(By Research), Ph.D.,** for the encouragement. We also thank our Registrar, **Dr.S.S.RAU, Ph.D.,**and Controller of Examinations, **Dr.K.V.NARAYANAN, Ph.D.,**for their timely support. We would like to thank **Dr. Head of Faculty of Computing** and **Head of Department Information Technology** for the help and support.

We thank our internal guide **Ms.Kaavya.D** (Internal) for her constant encouragement and valuable suggestions in completing this project and also for the timely help in the improvement of the project at all stages. We extend our heartfelt thanks to all the department staffs, our families, friends and well-wishers for constructive opinions, encouragement and valuable cooperation for the success of this project.

Reference:

1. SAbiteboul, R.Hull,and V. Vianu. Foundations of databases. Addison-Wesley Reading, 1995.
2. E.Agichtein and L.Gravano. Snowball: Extracting relations from large plaintext collections. In ACM DL, pages 85–94, 2000.
3. J.Barnard and D.Rubin. Small-sample degrees of freedom with multiple imputation. Biometrika, 86(4):948–955, 1999.
4. G.Batista and M.Monard. An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence, 17(5-6):519–533, 2003.
5. P.Bohannon, W.Fan, M.Flaster, and R. Rastogi.A cost-based model and effective heuristic for repairing constraints by value modification. In SIGMOD, pages 143–154, 2005.
6. P.Bohannon, W. Fan, F.Geerts, X.Jia, and A. Kementsietsidis. Conditional functional dependencies for data cleaning. In ICDE, pages 746–755, 2007.
7. S.Brin. Extracting patterns and relations from the world wide web. The World Wide Web and Databases, pages 172–183, 1999.
8. M.Cafarella, A.Halevy, and N. Khoussainova.Data integration for the relational web. PVLDB, 2(1):1090–1101, 2009.
9. M.Cafarella, A. Halevy, D. Wang, E. Wu, and Y.Zhang. Webtables: exploring the power of tables on the web. PVLDB, 1(1):538–549, 2008.
10. H.Elmeleegy, J.Madhavan, and A. Halevy. Harvesting relational tables from lists on the web. PVLDB, 2(1):1078–1089, 2009.

Corresponding Author:
Deepika.D*,
Email: *sophiya2128@gmail.com*