INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS…..

Implementation of Data Cleaning/Scrubbing in Data ware House for Efficient Data Quality

Priyanka Choudhary[1], **Dr. Sukhvir Singh**[2]

[1]N.C. College of Engineering, Kurukshetra University,
Israna, Panipat-132107, India
Choudhary.pri@gmail.com

[2]N.C. College of Engineering, Kurukshetra University,
Israna, Panipat-132107, India
boora_s@yahoo.com

Abstract: *The proposed work is about to optimized the data present in data warehouse. A Data warehouse can have data with several impurities such as duplicate data, incomplete data, unflustered data etc. In this proposed approach we have combined 3 approaches to resolve these impurities. These approaches include Duplicate Data Detection and Elimination, Apply Association Rule to collect related data and a fuzzy approach for the data classification. The reliability of data is because of its accuracy. A Data warehouse contains bulk of data. It includes the data taken from different data centers. Because of this data ware house can contain some data impurities. Data warehouse is responsible for all kind of management level decision related to data. Because of this there is the requirement to perform the cleaning on user data. To perform the knowledge acquisition and knowledge discovery we are here presenting an optimized approach to perform the data cleaning along with data association and the classification. The proposed approach is the rule based approach along with fuzzy decision. In this approach at first duplicate data and other impurities are cleaned from the database. After this the data integrity will be optimized by defining the parameters like strength and confidence. Finally a fuzzy decision is taken place to classify the data under some defined fuzzy rules.*

Keywords: *Data Detection and Elimination, Association Rule, fuzzy approach, knowledge discovery*

1. INTRODUCTION

Data cleansing, data cleaning, or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and then replacing, modifying, or deleting this dirty data.[1-2]

After cleansing, a data set will be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. [3]

Data quality- High-quality data needs to pass a set of quality criteria.

Accuracy: an aggregated value over the criteria of integrity, consistency, and density
Integrity: an aggregated value over the criteria of completeness and validity.
Completeness: achieved by correcting data containing anomalies.
Validity: approximated by the amount of data satisfying integrity constraints.
Consistency: concerns contradictions and syntactical anomalies.
Uniformity: directly related to irregularities and in compliance with the set 'unit of measure.
Density: the quotient of missing values in the data and the number of total values ought to be known [4-5]

A Data Warehouse is…

Data warehousing is a practical solution for providing strategic information. Database with the following distinctive characteristics: Separate from operational databases. Subject oriented: provides a simple, concise view on one or more

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS…..

selected areas, in support of the decision process. Organized by subject not by application. Constructed by integrating multiple, heterogeneous data sources.
Contains historical data: spans a much longer time horizon than operational databases. (Non volatile).
Read-Only access: periodic, infrequent updates. Stored collection of diverse data. A solution to data integration problem. Single repository of information.

2. LITERATURE REVIEW

In year 2009, Osman Abul, Harun G˙okc¸e, Ya˘gmur S¸engez, presents an evaluation framework which implements recent algorithms belonging to different approaches and a set of metrics to gauge the performance and problem difficulties.[6] The current work also presents an experimental study and its results where four algorithms and seven datasets are involved. Our results indicate that data distortion levels and runtime requirements are quite high, especially for difficult problem instances. Our conclusion is that there are new rooms for more sophisticated and tunable (w.r.t. effectiveness/efficiency tradeoff) algorithms.

In year 2010, Bi-Ru Dai and Li-Hsiang Chiang, propose an incremental mechanism and design a data structure in this paper to hide sensitive frequent patterns in the incremental environment. In this mechanism, the transaction data and sensitive patterns are stored in two types of trees. The proposed algorithm can efficiently find related transactions by links between these two types of trees. Experiment results show that the proposed method can efficiently hide.

In year 2011,[7] Anita A. Parmar,Udai Pratap Rao, Dhiren R. Patel, address such the problem of sensitive classification rule hiding. We propose a blocking based approach for sensitive classification rule hiding. First we find the supporting transactions of sensitive rules. Then we replace known values with unknown values ("?") in those transactions to hide a given sensitive classification rule. Finally the sanitized dataset is generated from which sensitive classification rules are no longer mined. We also discuss experimental results of our algorithm.[8]
Manya Sethi performed a work," *DATA WAREHOUSING AND OLAP TECHNOLOGY". DATA WAREHOUSING* and Online Analytical Processing (OLAP) are essential elements of decision support, which has increasingly become a focus of the database industry. Data Warehouse provides an effective way for the analysis and statistic to the mass data and helps to do the decision making. Many commercial products and services are now available and all of the principal database management system vendors now have offerings in these areas. The paper introduces the data warehouse and the online analysis process with an accent on their new requirements. I describe back end

3. NEW PROPOSED SCHEME

The theoretical backgrounds of the pruning scheme are based on the following two theorems which were presented.[9][10]

Theorem 1. A transaction can only be used to support the set of frequent (k+1)-itemsets if it consists of at least (k+1) candidate k-item sets.[11]

Theorem 2. An item in a transaction can be trimmed if it does not appear in at least k of the candidate k-item sets contained in the transaction

While the support counting procedure is being executed, the whole database is streamed into the systolic array. However, not all the transactions are useful for generating frequent item sets. Therefore, we filter out items in the transactions according to Theorem 2 so that the database is reduced. In the architecture the trimming information records the frequency of each item in a transaction that appears in the candidate item sets.[12]The support counting and trimming information collecting operations are similar since they all need to compare candidate item sets with transactions. Therefore, in addition to transactions in the database, their corresponding trimming information is also fed into the systolic array in another pipe, while the support counting process is being executed. [13]

4. CONCLUSION AND FUTURE SCOPE

The reliability of data is because of its accuracy. A Data warehouse contains bulk of data. It includes the data taken from different data centers. Because of this data ware house can contain some data impurities. Data warehouse is responsible for all kind of management level decision related to data. Because of this there is the requirement to perform the cleaning on user data. To perform the knowledge acquisition and knowledge discovery we are here presenting an optimized approach to perform the data cleaning along with data association and the classification. The proposed approach is the rule based approach along with fuzzy decision. In this approach at first duplicate data and other impurities are cleaned from the database.[14]

In this present work we perform the work on data cleaning on a centralized dataset or the warehouse. The work can be improved by taking the concept of some other database systems such as Distributed Database or the Mobile Database System. More work can be done in direction of Efficiency. [15]

Databases can be larger in both depth and breadth:

More columns: Analysts must often limit the number of variables they examine when doing hands-on analysis due to time constraints. Yet variables that are discarded because they seem unimportant may carry information about unknown patterns. High performance data mining allows users to explore the full depth of a database, without preselecting a subset of variables.

More rows: Larger samples yield lower estimation errors and variance, and allow users to make inferences about small but important segments of a population.

4.1Suggestions for Future work:

In this present work we perform the work on data cleaning on a centralized dataset or the warehouse. The work can be improved by taking the concept of some other database systems such as Distributed Database or the Mobile Database System. More work can be done in direction of Efficiency.

REFERENCES

[1]. Anita A. Parmar, Udai Pratap Rao, Dhiren R. Pate, "Blocking based approach for classification Rule hiding to Preserve the Privacy in Databas", 2011 International Symposium on Computer Science and Society, 978-0-7695-4443-4/11 © 2011 IEEE

[2]. Aris Gkoulalas-Divanis Vassilios S. Verykios, "A Hybrid Approach to Frequent Itemset Hiding", 19th IEEE International Conference on Tools with Artificial Intelligence, 1082-3409/07 © 2007 IEEE

[3]. Aris Gkoulalas-Divanis Vassilios S. Verykios, "A Hybrid Approach to Frequent Itemset Hiding", 19th IEEE International Conference on Tools withArtificial Intelligence 1082-3409/07 @ 2007 IEEE

[4]. Aris Gkoulalas-Divanis, "Exact Knowledge Hiding through Database Extension", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 5, MAY 2009

[5]. B.Murugeshwari, Dr.K.Sarukesi, Dr.C.Jayakumar, "An Efficient method for knowledge hiding Through database extension", 2010 International Conference on Recent Trends in Information, Telecommunication and Computing, 978-0-7695-3975-1/10 © 2010 IEEE

[6]. Bi-Ru Dai, Li-Hsiang Chiang, "Hiding Frequent Patterns in the Updated Database", 978-1-4244-5943-8/10/ ©2010 IEEE.

[7]. Huanzhuo Ye, Di Wu, Shuai Chen An Open Data Cleaning Framework Based on Semantic Rules for Continuous Auditing 2010 IEEE

[8]. J. Jebamalar Tamilselvi and Dr. V. Saravanan, "A Unified Framework and Sequential Data Cleaning Approach for a Data Warehouse", ACM, IJCSNS Intenational Journal of Computer Science and Network Security, Vol.8 No.5, May 2008, Page(s): 117 – 121

[9]. J. Jebamalar Tamilselvi† and Dr. V. Saravanan IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.5, May 2008

[10]. Osman Abul Maurizio Atzori Francesco Bonchi Fosca Giannotti, "Hiding Sequences", 1-4244-0832-6/07/ ©2007 IEEE.

[11]. Osman Abul, Harun G¨okc¸e, Ya˘gmur S¸engez, "Frequent Itemsets Hiding: A Performance Evaluation Framework", 978-1-4244-5023-7/09/©2009 IEEE.

[12]. P. Eredics* and T.P. Dobrowiecki, Data Cleaning for an Intelligent Greenhouse 6th IEEE International Symposium on Applied

INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY
WINGS TO YOUR THOUGHTS…..

Computational Intelligence and Informatics • May 19–21, 2011.

[13]. R. Arora, P. Pahwa, S. Bansal, "Alliance Rules of Data Warehouse Cleansing", IEEE , International Conference on Signal Processing Systems, Singapore, May 2009, Page(s): 743 –747.

[14]. S. Chaudhuri, K. Ganjam, V. Ganti, "Data Cleaning in Microsoft SQL Server 2005", In Proceedings of the ACM SIGMOD Conference, Baltimore, MD, 2005.

[15]. S. Reddy, A. Lavanya, V. Khanna, L.S.S. Reddy, "Research Issues on Data Warehouse Maintenance", IEEE, ICACC '09. International Conference Advanced Computer Control, Singapore, Jan 2009, Page(s):623 – 627.