Data Cleaning Rules Based on Conditional Functional Dependency

Manish Ranjan[1] and Akhilesh Bansiya [2]

[1]Computer Science Engineering Department, RKDF University, Bhopal, India
[21]Computer Science Engineering Department, RKDF University, Bhopal, India
[1] manishranjanmr709@gmail.com, [2] akhilesh2483@gmail.com

* Corresponding Author: Manish Ranjan
Manuscript Received: Manuscript Accepted:

Abstract: *Data discrepancies plague the majority of existing databases. To tackle this problem, initiatives to improve data quality are required. Two strategies for mining accurate conditional functional dependency rules from such databases to be used for data cleaning are proposed in this paper. The proposed strategies work by mining maximal closed frequent patterns first, then using lift measure to mine dependable conditional functional dependency rules. Furthermore, using the created rules, a data repairing procedure is proposed for resolving inconsistent tuples detected in the database. On both real-life and synthetic medical data sets, an extensive experimental research was done to confirm the effectiveness of the suggested strategies compared to existing methodologies.*
Keywords: *Data quality, Data repairing, Data mining, Data inconsistency, Data cleaning.*

I. Introduction

This Many of today's data management challenges stem from the increasing variety of scenarios where data is being exploited. This variety has a number of dimensions: breadth of data sources with varying representation and quality, breadth of use cases for a range of both technical and popular products, and breadth of people with varying skill sets who work with data sources to build data products. The human factors in this context present challenges and opportunities of increasing urgency for the technical community. Recent projections on labor markets predict dire shortfalls in the analytical talent available to realize the potential of "big data". Technologists can have significant impact here by developing techniques that dramatically simplify labor-intensive tasks in the data lifecycle.

Most of the materials in the first part of the tutorial come from our survey in Foundations and Trends in Databases. We describe a statistical perspective on qualitative data cleaning, where approaches either use techniques from Machine Learning to improve accuracy or efficiency or consider the effects of cleaning on subsequent numerical queries. We present these approaches within the same overall taxonomy of data cleaning and show that many qualitative techniques are amenable to such statistical analysis. By considering the qualitative models in a rigorous statistical framework, we can understand the trade-off between cleaning and the ultimate accuracy of inferences made from the data.

II. Related Work

Integrity constraint based data repairing is an iterative process consisting of two parts: detect and group errors that violate given integrity constraints (ICs); and modify values inside each group such that the modified database satisfies those ICs. However, most existing automatic solutions treat the process of detecting and grouping errors straightforwardly (e.g., violations of functional dependencies using string equality), while putting more attention on heuristics of modifying values within each group [1].

Some important data management and analytics tasks cannot be completely addressed by automated processes. These "computer-hard" tasks such as entity resolution, sentiment analysis, and image recognition, can be enhanced through the use of human cognitive ability. Human Computation is an effective way to address such tasks by harnessing the capabilities of crowd workers (i.e., the crowd). Thus, crowd sourced data management has become an area of increasing interest in research and industry. There are three important problems in crowd sourced data management. (1) Quality Control: Workers may return noisy results and effective techniques are required to achieve high quality; (2) Cost Control: The crowd is not free, and cost control aims to reduce the monetary cost; (3) Latency Control: The human workers can be slow, particularly in contrast to computing time scales, so latency-control techniques are required [2].

The human work involved in data transformation represents a major bottleneck for today's data-driven organizations. In response, we present Predictive Interaction, a framework for interactive systems that shifts the burden of technical specification from users to algorithms, while preserving human guidance and expressive power [3].

III. Importance of Data Cleaning

Any organization which uses its database for knowledge discovery and decision making will be required to keep its database updated and error free. The failure leads to loss of quality data and increase in operational costs. The data warehouse users use the features of data like coherency, correctness and accuracy of the data, which degrades with time and regular updates which in turn has an effect on the integrity of the data residing in a data warehouse.

These errors thus lead to poor decision making and errors in the trend analysis. Clean data is an essential requirement to any sales, marketing and distribution strategy. The avoidance of dirty data will help to decrease operational costs and time, thus leading to an improving brand image of an organization.

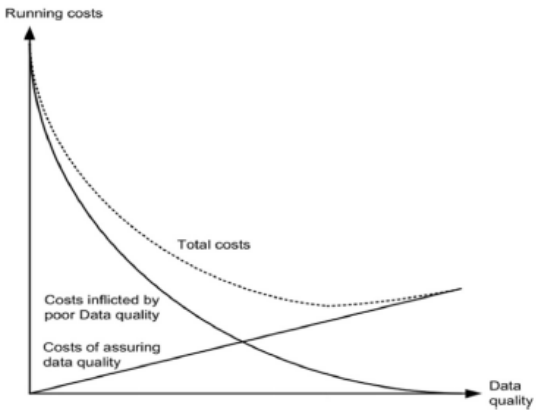

Figure 1: Total costs acquired by quality of data in the organization.

Dirty data costs American businesses billions of dollars each year, resulting in poor decision-making [4]. Figure 1 shows a link between the expenses of poor data quality and the costs of maintaining good data quality. There is a balance between the expenses of poor data quality and the price of data quality assurance. Increasing poor data quality has a detrimental impact on the resultant high-quality data, prompting people to consider ways to ensure and attain data quality.

As a result, as part of the data cleaning process, it's critical to concentrate on finding and correcting data inconsistency issues.Guaranteeing high-quality, dependable data, in particular, is a competitive advantage for all industries, necessitating correct data cleansing solutions.

Data cleaning (or data scrubbing) is a crucial step in the data preprocessing process. It's great for keeping track of inconsistencies before mining and analysing data [5]. Data cleaning is used in a variety of applications, including data warehousing, data quality management, and database knowledge discovery.

IV. Proposed techniques

Conventional approaches, such as traditional FDs, which were intended primarily for schema design, have various flaws. Other methods provide redundant, unreliable rules that are not scalable in large datasets with many properties. Furthermore, alternative solutions necessitate user participation in the cleaning process, which is costly and ineffective. The presented strategies are intended to overcome the limitations of existing data quality techniques in a range of application fields.

IV.A Steps Involved in Data Cleaning

Data cleansing is usually a two-step process including detection and then correction of errors in a data set. The steps involved in Data Cleansing are:

a) Identification of errors-records could have incomplete or corrupted data.
b) Perform error verification-whether it is truly an error or not. This situation occurs in organizations where there exists a usage of organizational jargons [6].
c) Extract the data to be cleaned-the data is extracted and stored in a temporary table, operations are performed and the data is repaired and verified, then it is replaced in the target table.
d) Perform data cleaning-which can be done automatically or manually.

Manual process is however avoided as it is highly time consuming and tedious in nature. It is limited by human capabilities like speed, accuracy in error detection and correction. Thus leading to more error prone performances and degrading the quality of data, which in turn leads to increase in operational costs and hence poor decision making. It is extremely important to categorize the data according to the rate of its criticality.

a) Critical errors-needs to be immediately addressed i.e.; error reporting ,verification and cleansing
b) Non-critical errors can be temporarily ignored.

IV.B Alliance Rules Algorithm

The need of this algorithm raised in the data mart system of an organization which involved a customer bill generation. The storage of large amount of information about the customers suffers from the problem of dirty data. A data warehouse is formed by merging data from different sources which can have different field formats. A data mart of a telecommunication system may involve sections like bill generation, account section, personal information section etc. [7] which on merging can lead to several inconsistencies and hence dirty data.

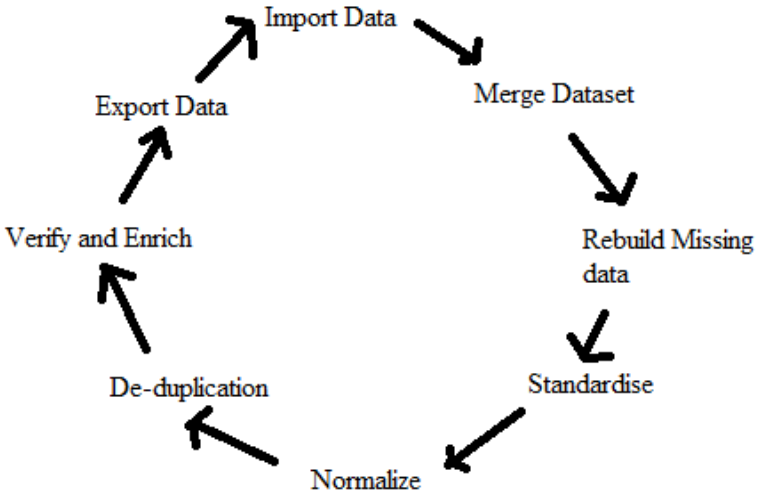

Figure 2: Data Cleaning Steps

This algorithm addresses the errors and issues occurred in the 'name' field of the data warehouse. The name field being an important aspect in a customer based organization forming an integral part of the bill generation and their strategies, any duplicity or field mismatch can lead to organization mistrust and wastage of time.

Here the records are matched using the attribute keys. Using key the records are grouped and matched as a set of related records and then errors are detected and corrected after detailed analysis. This helps to fill in the blank cells (fields) and also remove the duplicity errors and redundancies.

The modified transitive closure makes use of more than one key to match and group the related records. Primary secondary and tertiary key concept is used to group the related records, and when the records are matched blanks are filled and redundancies are removed.

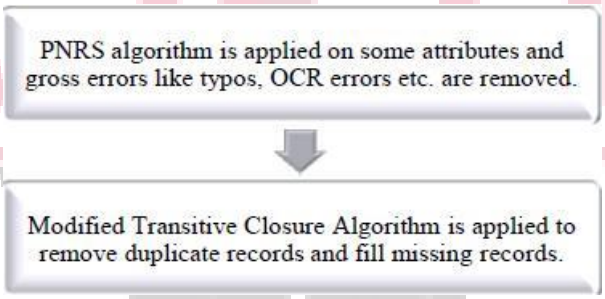

Figure 3: Flowchart of HADCLEAN

The algorithm of proposed methodology is Improved Heuristic Algorithm with Multiple FDs is as follows:

Input: database D and set $\sum$ of FDs

Output: FT-consistent database D' (After cleaning/repairing dataset)

Step 1: while $\exists t^{\varphi} \in D \setminus \hat{I}_{\Sigma}$ s. t. t^{φ} is FT-consistent with $\hat{I}_{\varphi}$ **do**

// Search every single or multiple tuple in given dataset for error detection

Step 2: $C \leftarrow \{ t^{\varphi} \in D \setminus \hat{I}_{\Sigma}$ s. t. t^{φ} is FT-consistent with $\hat{I}_{\varphi} \}$;

// extract each tuple with their cost

Step 3: $t^{\varphi} \leftarrow$ the one in C with the smallest tuple cost;

// extract tuple whose has smallest cost

Step 4: $\varphi \leftarrow$ the FD s. t. t^{φ} is FT-consistent with $\hat{I}_{\varphi}$;

//remove dependencies on multiple level

Step 5: $I_{\varphi} \leftarrow \hat{I}_{\varphi} \cup \{t^{\varphi}\}$;

// create new tuple group in FD

Step 6: Update $t^{'\varphi} \in N\left(t^{\varphi}\right)$ to $t_b^{'\varphi}$;

// update tuple group in dataset

Step 7: Join I_{Σ}^{*} to get the targets;

// update dependencies choose as target set.

Step 8: For each tuple t ∈ D do

//Now determine new generated tuple in dataset

Step 9: Modify t to its closest target;

//Determine closet property of FD and modify FD

Step 10: return $D^{'}$;

// obtain repaired or cleaned dataset

References

[1] Shuang Hao, Nan Tang, Guoliang Li, Jian He, Na Ta and Jianhua Feng, "A Novel Cost-Based Model for Data Repairing", IEEE Transactions on Knowledge and Data Engineering, 2017.

[2] Guoliang Li, Jiannan Wang, Yudian Zheng and Michael J. Franklin "Crowdsourced Data Management: A Survey", IEEE Transactions on Knowledge and Data Engineering, 2016.

[3] Jeffrey Heer, Joseph M. Hellerstein and Sean Kandel, "Predictive Interaction for Data Transformation", 7th Biennial Conference on Innovative Data Systems Research (CIDR '15) January 4-7, 2015, Asilomar, California, USA.K. Elissa, "Title of paper if known," unpublished.

[4] Fan W, Ma S, Tang N, Yu W. Interaction between record matching and data repairing. J Data Inf Qual 2014 May 1;4(4):16.

[5] Mezzanzanica Mario, Boselli Roberto, Cesarini Mirko, Mercorio F. Automatic synthesis of data cleansing activities. In: Proceedings of the 2nd international conference on data technologies and applications; 2013. p. 138e49.

[6] Guoliang Li, Jiannan Wang, Yudian Zheng and Michael J. Franklin "Crowdsourced Data Management: A Survey", IEEE Transactions on Knowledge and Data Engineering, 2016.

[7] Michele Dallachiesa, Amr Ebaid, Ahmed Eldawy, Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani and Nan Tang, "NADEEF: A Commodity Data Cleaning System", SIGMOD'13, June 22–27, 2013, New York, New York, USA.