ETDC: An Efficient Technique to Cleanse Data in the Data Warehouse

Saad B. Alotaibi
King Abdulaziz University
Saudi Arabia, Jeddah
salotaibi0395@stu.kau.edu.sa

ABSTRACT

Data cleansing can be considered to be an activity that is performed on the data sets of the data warehouse. The cleansing is done in order to enhance and collectively maintain data consistency and quality. The quality of data has a strong impact on a process such as: data mining, knowledge discovery and trend analysis that perform on existing data warehouse. In addition, the data quality effects on the enterprise decision making, planning, analysing and reporting, this process are very sensitive to quality of data. To get the useful result or information must be assure the data in data warehouse or database are consistent, accurate and unique. In this paper, we will discuss the quality of data problems that appear in the data, and provide an algorithm for cleansing the data in the data warehouse.

CCS Concepts

• Theory of computation~Design and analysis of algorithms
• Theory of computation~Incomplete, inconsistent, and uncertain databases

Keywords

Data warehouse, Data Cleaning, Data Cleansing.

1. INTRODUCTION

A data warehouse can be considered to be a complex organization that ideally stores large quantities of data. There are several procedures and processes that can be used in the process of building and maintaining a data warehouse. Any organization's data warehouse that is exponentially being used for the discovery of knowledge as well as trend analysis is maintained through deletions, insertions and regular updates. The users of the data warehouse exponentially characterizes the features and the quality of the data as being coherent, correct and accurate along the accessibility and the data newness. The quality of data however is noted to degrade with the customary updates that collectively have a huge impact on the processes such as the discovery of knowledge, data mining and the trend analysis that is performed on the data warehouse. The data warehouse of any enterprise consolidates the available data from several sources within the organization in an attempt to support organizational functions such as decision making, planning, analyzing and reporting. All these processes are very crucial as they require aspects of data quality as they entirely depend upon the consistency and the accuracy of the data. Degraded data stands to mislead the organizations as it causes the arriving of wrong conclusions as it leads to wastage of assets and resources. Today, data mining is increasingly being used by the companies that are related to finance, retail, marketing and communication. It helps these organizations to identify and detect all the possible variables that affect the processes on the operations. Such initiatives cause a great fortune in terms of human power, money and time as they are highly critical. The degraded form of data is crucial as it decreases the attempts of data reliability. When the organizational processes have been applied using the degraded and low quality data, they can eventually lead to lack of trust in the results. The whole purpose of undertaking to perform the reliability and accuracy of such data processes on the data warehouse is to arrive at a standstill which ultimately leads in resources' wastage. The solution to this is only by designing an automated program and cleansing too to ensure the enhancement and the maintenance of the quality of data. Automated data cleansing tool can be considered to be the most feasible, practical as well as cost-effective method of ensuring that data quality has been maintained at reasonable levels.

ICAIP 2017, August 25–27, 2017, Bangkok, Thailand

ACM ISBN 978-1-4503-5295-6/17/08…$15.00
DOI: https://doi.org/10.1145/3133264.3133296

2. DATA CLEANSING

Data cleansing ideally refers to all the activities that help in the process of determining and detecting all the corrupt, unwanted, inconsistent and faulty data and further correcting this data in an effort to increase its quality and effectiveness. When data has been stored in large quantities, this usually makes it prone to getting dirty. The storage systems such as data warehouse are often faced with a challenge as they hold great quantities of data and in this case, they are prone to getting degraded as a result of receiving faulty and low quality data from different sources. When there is dirty data in the warehouse, this leads to the rise of increased problems as this data warehouse is primary to the organizational functions of decision making, trend analysis, strategic planning, and discovery of knowledge. Higher operational costs and wrong decisions are more likely to occur with dirty data existing in the data warehouse. Once there is despoiled data in the data warehouse, this can lead to the management diverting their attention and not focusing on the core functions of the organization. Further, the despoiled data tarnishes the image in the organization as it has been identified to cause mistrust. To solve this, cleansing of data remains to be the paramount solution for the appropriate maintenance of the quality and the consistency level of the data. Despite that cleansing the data has been identified as the main initiative that enterprises can use to maintain the quality of data, these processes are usually prone to errors and they are highly time consuming. The manual

process of data cleansing is usually prone to increased human errors as well along with other problems which includes low speed and accuracy of the final results. Thus, to ensure that a corresponding and reliable method has been used in cleansing the data, an automation of the cleansing process is necessary. In this paper, the main focus upon the algorithm for the automated tool to be used in the identifying and detecting any corrupt data in the data-warehouse.

3. LITERATURE REVIEW

Cleansing the data in the data warehouse can be considered as a wide coverage area of study for the keen researchers. There exists a present schemer that helps in the detecting the multiple source problems. The schema proposes variant types of errors in the data sets. These errors include the noninclusion of a numerical value in the data type field. This could be a value that shows a different date of birth, a missing value for the date of birth. Not only does the automated schema identifies the potential errors, it also helps in the detection of quality data metric. This program also enables the process of finding the data quality problems when the organizations makes a switch from the old database to a new database. These proposals however have defined problems only. Our proposal based on smart technique that can be used in the finding of the most of errors in the data sets. We cover most of related works in the following:

Dr. Mortadha et al. [1] presents the enhancement of technique for cleansing the data in the data warehouse, through use the algorithm that can detects and corrects many types of errors. Moreover, expected the problems such as domain format errors, lexical errors, irregularities, constraint violation, integrity and duplicates. The proposed solution in this paper, was worked on the quantitative data and also in the data that has limited values.

A. Paul et-al. [2] provides a hybrid approaches that has name "HADCLEAN" to clean the data in the data warehouse. The proposed technique is combined of "PNRS and Transitive" closure algorithm. The idea of HADCLEAN algorithm is apply one algorithm in the data sets and then apply the other, to get the cleaned data. "The PNRS algorithm is applied on some attributes and gross errors like types, OCR errors etc. are removed". "Modified Transitive Closure algorithm is applied to remove duplicate records and fill missing records".

L. He et-al. [3] proposed the improved algorithm of the "Approximately Duplicate Records Cleansing" that depends on attributes selection after analysis the existing (SNM) that refer to "Sorted neighborhood method" and (MPN) that refer to "multipass sorted neighborhood method". The idea of the proposed algorithm is analysis the attributes and sorting a dataset many time to clustering the duplicate records. Moreover, the proposed algorithm gives the attributes weight to make comparing with high accurate.

T. Ohanekwu et al. [4] describes a technique that can eliminates the require to rely on the threshold matching through define a smart tokens that used in duplicates identification. The proposed algorithm used to match two inputs string. The previous matching depends on smart tokens that selected based on the very sensitive field of record.

Sh. Taneja et al. [5] proposed algorithm to handling the date type filed errors and furthermore any error occur in the date format. The proposed algorithm is a DFT that can transform or convert differ date format to unique consistent format, and that for avoiding any ambiguities.

Agusthiyar.R et al. [6] proposed framework for handling the errors in the heterogenous sources of the data at schema level. Moreover, the proposed framework can detect and remove the inconsistencies and errors in the simplified manner. In addition, the proposed framework deal to improve the data quality in multiple sources of data of the enterprise or organization that having different sources in different locations.

Xu Chu et-al. [7] proposed KATARA, it's the base of knowledge and gather the powered of data cleaning system that give the table, crowd and kb, explain table semantics into align it with a kb, identify incorrect and correct data, and then generates the top-k of possible repairs, into incorrect data. The KATARA has three modules including the pattern validation, pattern discovery and data annotation. For the pattern discovery module, it's responsible to discover the table patterns that located between a kb and the table. The pattern validation module, it's responsible for using the crowdsourcing in order to be selecting one "table pattern". Based on the selected "table pattern" the data annotation module, responsible to interacts with the crowd and the kb to annotate data. Moreover, it's responsible to generate the possible repairs in order to erroneous tuples.

P. Pahwa et-al. [8] addresses the dirty data problems, whether data is entered or detection in data warehouse. The paper provides algorithm that can detect the errors and the dirty data that existing in the data warehouse. The paper describes "The Alliance Rules" based on the mathematical concept that association rules for determining the faculty data, and also the dirty data in the data warehouse. The "Alliance Rules" used to identify and detect the errors in the data warehouse. The proposed algorithm deals with data duplicity errors of string type within the data warehouse in the different data marts. The "Alliance Rules" depend-on the data mining principle association rules extend or provide the solution for errors detections in the data sets. In addition, the errors are automatically detection. There is no manual intervention in the algorithm that proposed in the paper.

The existing work on the data warehouses has proposed the application of different methods of detection and identification of errors. These approaches that past researchers have presented are hitches that the sole researcher fails to address as these research focuses on the study of the error types and the error identification for the date data. They then focus on different formats of the date data that can be used to cleanse the data. This research however does not concentrate on some other data sort. Another principle hitch in this approach is that it couldn't distinguish the guile mistake. The algorithm provided required date of birth as the reference date as the premise of finish purifying procedure. The calculation proposed includes heaps of manual work amid its pre-processing stage. Since, pre-processing (i.e. introductory phase of algorithm) must be done physically. In the following section, we explain our proposed technique in detail.

4. THE PROPOSED TECHNIQUE (ETDC)

We proposed in this paper new idea in data cleansing area. Actually, the idea is a smart technique that can identify of most data errors including: redundant data, data formats errors, data duplicity, numeric values errors, Arabic language mistakes, symbol errors. The technique works within the data gathering from multi sources and before the loading into the target (Big Storage Systems). We focus on the Arabic language because no more accurate techniques working with it. Moreover, our technique deals with the numeric and symbol errors. Our technique working as follows:

1. Take the table form the sources X, Y, …, N.

2. Extract the data from the field.
3. Identify the error types within the following types:
    a. Redundant data
        i. Convert the words into the root for identifying the meaning of word to avoid the different representation for same meaning.
        ii. The converting techniques basedon Stanford API (“Tagger, Segmenter and Parser”).
        iii. Check the filed meaning for all tables to detect the redundant data. iv. Remove the redundant data, and store it in the temporary table.
    b. Data formats errors
        i. Check the type of data.
        ii. Unifying the types by converting the data types into standard type.
        iii. Store it in the temporary table.
    c. Numeric values errors
        i. Check the data in the filed if it’s string such as (one, two, ...etc.) convert it into (1, 2, …etc.).
        ii. Store it in the temporary table.
    d. Arabic language mistakes
        i. Check the Arabic meaning basedon Stanford api and unifying the text into standard word such as “كبيير” change it into “كبير”
        ii. Store it in the temporary table.
    e. Symbol errors
        i. Check the filed type and identify the symbol such as (1-1-2017) change it into (1/1/2017), also (1.000) change it into (1,000) and so on.
        ii. Store it in the temporary table.
    f. Misspelling (Google API)
        i. Check the misspelling based on Google Translate Api and correct the words.
        ii. Store it in the temporary table.

The following figure 1. explain how the technique works.

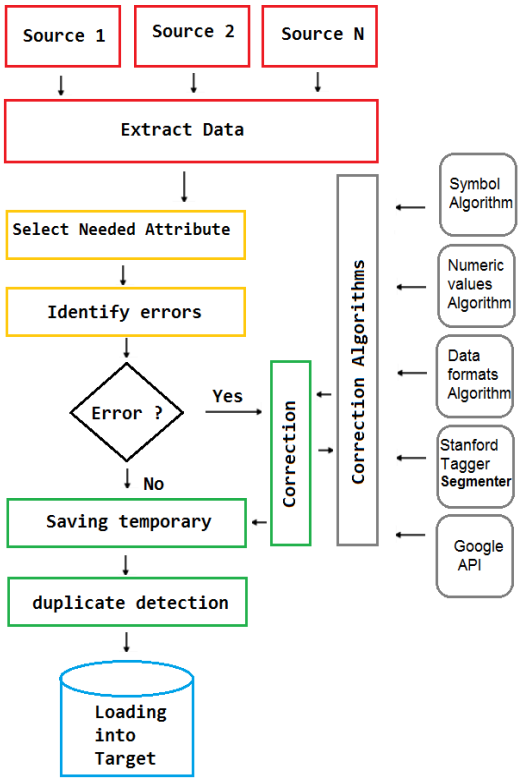

Figure 1. ETDC technique flowchart

5. RESULT

We assume the sources are from multi branches of stores or any enterprise that provide the customer registrations, and we need to build one storage to combine all users of all sources.

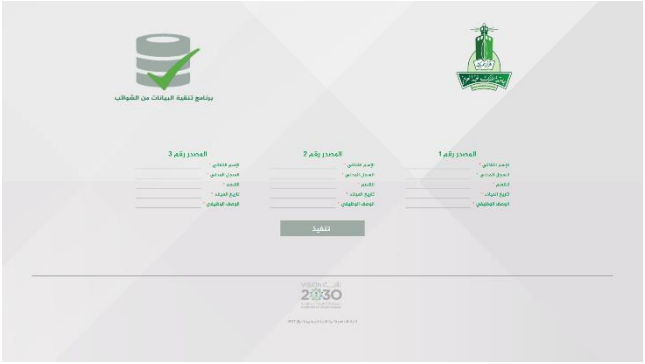

Figure 2. Main page before the processing

The figure 2. gathering the data of three sources and check the users’ tables to combinate it into one storage system. The following figure 3. show the data before the cleansing.

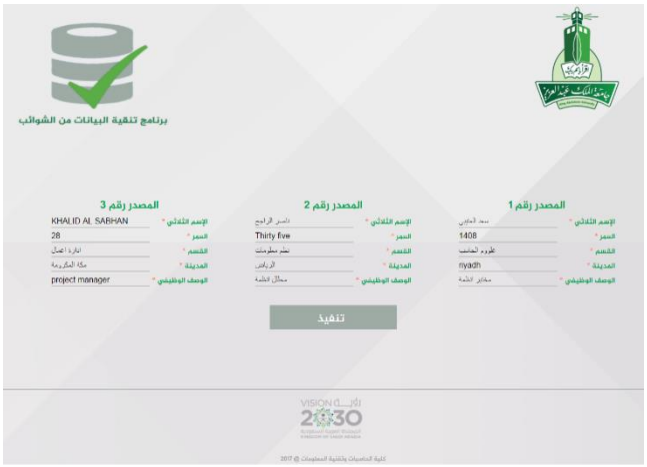

Figure 3. Dirty data

The system will take the data and trying to analyze it and detect the errors. The error type will be handling with specific algorithm. See the result figure 4.

Figure 4. Result

6. CONCLUSON

Data cleaning is a primary key for any data wherehouse system success, it is the main factor to achieve a perfect data quality, which leads to accurate and correct statistics for the business decision making people. As a result of that the researches for the data cleaning are increasing significantly. I think that on the next couple of years researches will appear on data cleaning, although the large number of tools indicates both the importance and difficulty of the cleaning problem. We see several topics deserving further research, and it will be a very hot topic for the researchers.

7. REFERENCES

[1] Hamad M-M, Jihad A-A. An Enhanced Technique to Clean Data in the Data Warehouse. In Developments in E-systems Engineering.IEEE. 2011;306-311.

[2] A. Paul, et-al "HADCLEAN: A Hybrid Approach to Data Cleaning in Data Warehouses", Information Retrieval & Knowledge Management (CAMP), 2012 International Conference, IEEE, 2012.

[3] L. He, Z. Zhang, et-al, "An Efficient Data Cleaning Algorithm Based on Attributes Selection," th International Conference on Computer Sciences and Convergence Information Technology (ICCIT), 2014.

[4] T. Ohanekwu, C.I. Ezeife "A Token-Based Data Cleaning Technique for Data Warehouse Systems", 2003.

[5] Sh. Taneja, et-al, "DFT: A Novel Algorithm for Data Cleansing", International Journal of Computer Science and Information Technologies, 2014.

[6] Agusthiyar.R, K. Narashiman, "A Simplified Framework for Data Cleaning and Information Retrieval in Multiple Data Source Problems", International Journal of Innovative Research in Science, 2014.

[7] Xu Chu, et-al, "KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing", SIGMOD'15, 2015.

[8] P. Pahwa, et-al, "Alliance Rules for Data Warehouse Cleansing" in IEEE International Conference on Signal Processing Systems, 2009.

[9] J. Tamilselvi,V. Saravanan et-al "Unified Framework and Sequential Data Cleaning Approach for a Data warehouse", IJCSNS International Journal of Computer Science and Network Security, 2008.

[10] M. Ashwini, S. Kolkur, "Hybrid Technique for Data Cleaning", International Journal of Computer Applications, 2014

[11] Spence Green, et-al, "NP Subject Detection in Verb-Initial Arabic Clauses", Third Workshop on Computational Approaches to Arabic Script-based Languages, 2009

[12] Z. Chen and M. Cafarella. Integrating spreadsheet data via accurate and low-effort extraction. In KDD. ACM, 2014.

[13] I. F. Ilyas and X. Chu. Trends in cleaning relational data: Consistency and deduplication. Foundations and Trends in Databases, 5(4):281–393, 2015.

[14]] D. Haas, S. Krishnan, J. Wang, M. J. Franklin, and E. Wu. Wisteria: Nurturing scalable data cleaning infrastructure. PVLDB, 8(12), 2015.

[15] M. Volkovs, F. Chiang, J. Szlichta, and R. J. Miller. Continuous data cleaning. In ICDE, pages 244–255, 2014.

[16] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth. Preserving statistical validity in adaptive data analysis. In STOC, pages 117–126, 2015.

[17] M. Bergman, T. Milo, S. Novgorodov, and W. C. Tan. Query-oriented data cleaning with oracles. In SIGMOD, 2015.