AutoTest: Automation to Test Tabular Data Quality

Süleyman Eken, Ahmet Sayar
Kocaeli University, Department of Computer Engineering, Kocaeli, Turkey
{suleyman.eken, ahmet.sayar}@kocaeli.edu.tr

Kürşat Topçuoğlu
Turkcell Technology R&D, Kocaeli, Turkey
kursattopcuoglu@gmail.com

Abstract—Data can take various forms and types such as numbers, symbols, words, images, and graphics. All these data are significant resources in companies’ operating process. So, if data has any quality problems (e.g., poor schema design, data entry errors, misspelling, inconsistency, and etc.), it will very risky for companies. On the other hand, high quality data can increase opportunities dependent on performance issues. This study introduces an automation tool enabling data quality tests of data warehouse applications. Also with this tool, historical changes of a dataset are analyzed with linear regression algorithm and, thus, outlier variables of data trend are reported to application users. The efficiency of the proposed automation tool is also tested and results are evaluated.

Index Terms—data quality, data integrity, test automation, data warehouse, data statistics, data filtering

I. INTRODUCTION

Data is a strategic asset and assurance of data quality is becoming a business necessity in today’s organizations. Companies have begun to find promising solutions enhancing the data quality. Data owned by a company might be obtained from a variety of sources such as web, databases, and external resources. If management is not done properly, money, time and other physical resources will not be used efficiently. Moreover, collected data will be bulk of rubbish rather than useful and serviceable data. Also, Global Data Management Survey stated that 75% of the companies have serious difficulties in profit/loss because of poor data quality [1]. According to TDWI (The Data Warehousing Institute) research, lost is about USD $600 billion in a year due to poor data quality. Businesses depend on an increasing number of data intensive applications to power their strategic growth and operational efficiencies, when the quality of the data in these applications is bad it can lead to poor decision making or a serious breakdown in business operations and processes.

Incorrect or inconsistent data causes misinterpretation of the results and this situation causes misdirected users and companies. Data quality analysis is done by taking samples from the data sources by test engineers in many companies. A more comprehensive analysis of the data may be tedious and time consuming for employees due to the size of it. Investigations show that most of the problems come from overlooking incompatibility in data, not in data sources. Considering this evaluation, we developed an automation tool to clean inconsistent data from data sources and report outlier variables. So, test engineers or data quality experts can detect extraordinary data easily.

The rest of the article is structured as follows. In second section, data quality criteria’s and quality problems are discussed. In the third section, relevant works on data quality tools are proposed. In fourth section, basic technologies to develop the proposed tool are given, and system architecture is discussed. Last section draws a conclusion and presents future works.

II. DATA QUALITY DIMENSIONS AND PROBLEMS

A. Data Quality Dimensions

A data quality dimension is an aspect or feature of information and a way to classify information and data quality needs. Dimensions are used to define, measure, and manage the quality of the data and information.

- In order to improve information quality, there must be a way to measure it.
- There is no industry standard for the types of data quality dimensions.

Reason of different quality dimensions is that different tools, techniques, and processes are used to assess, measure, and manage the various dimensions of quality. High quality data need to pass a set of quality tests. Data quality criteria offer a way for measuring and managing qualities of data [2]. Some dimensions of criteria are listed in Table I:

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>The degree of conformity of a measure to a standard or a true value [3].</td>
</tr>
<tr>
<td>Completeness</td>
<td>The degree to which all required measures are known [4].</td>
</tr>
</tbody>
</table>
III. RELATED WORKS

Some tools have been developed in data quality and data cleansing era. One of them is WizSame created by WizSoft Company. It is an innovative data quality and data cleansing software for discovering duplicate records based on the user criteria and revealing similar records suspected as being duplicate. The user of WizSame tool determines the matching criteria by defining for each field whether it is identical, similar or ignored. Also the user may define several conditions connected by the AND or OR operators to be added to the matching criteria [9]. WizSoft also developed software called WizWhy. It is a data-mining software tool that automatically reveals the if-then and if-and-only-if rules in data, and on the basis of these rules, it summarizes the data, points out interesting phenomena in the data, reveals the main patterns, points out cases deviating from the rules and issues predictions for new cases [10].

Another tool is SAS Data Quality. It approaches to data quality from every angle, including data standardization, de-duplication and data correction. It enables users to establish data hierarchies and create reference data definitions, thus, users can have more control over their business information [11].

In the literature, some authors address methodology of assuring data quality. O-Hoon et al.,[12] propose quality assurance ontology to evaluate unexpected business rules and meaning of data value. To extract evaluate rules for data quality, they use ontology that has meanings of each word in itself. They gain the relationship among word in ontology, and then make SQL to evaluate data accuracy, especially focused on data meaning. Same authors propose MDRDP (Metadata Registry based on Data Profiling) to minimize the time and human resource for analyzing and extracting metadata as criteria standard for data profiling. Metadata Registry can guarantee the quality of metadata so that results of quality evaluation would improve [13]. Rehman and Esichaikul [14] focus on one of the major issue of data cleansing i.e. “duplicate record detection” which arises when the data is collected from various sources. Developed prototype which shows that adaptive duplicate detection algorithm is the optimal solution for the problem of duplicate record detection. For approximate matching of data records, string matching algorithms (recursive algorithm with word base and recursive algorithm with character base) have been implemented.

IV. SYSTEM ARCHITECTURE

In the proposed system, data is processed or analyzed with appropriate statistical techniques. The proposed AutoTest tool takes into account five different statistics analysis:

- Table statistics: In this category, invariant variable values of a table are stored. The number of rows and columns in the table can be examples of this category.
- Column statistics: In this category, statistics give information about meaningful values within column and identifying the column. These statistics include

<table>
<thead>
<tr>
<th>Data quality problem</th>
<th>Category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-source problem</td>
<td>Schema level</td>
<td>Lack of integrity constraints, poor schema design, uniqueness constraints, referential integrity</td>
</tr>
<tr>
<td></td>
<td>Instance level</td>
<td>Data entry errors, misspelling, redundancy duplicates, contradictory values</td>
</tr>
<tr>
<td>Multi-source problems</td>
<td>Schema level</td>
<td>Heterogeneous data models and schema design, naming conflicts</td>
</tr>
<tr>
<td></td>
<td>Instance level</td>
<td>Overlapping contradicting and inconsistency data, inconsistent aggregating, inconsistent timing</td>
</tr>
</tbody>
</table>
standard deviation and mean values of the variables in the column and the number of unique values of the column.

- Value frequency statistics: In this category, statistics give information about how many times the values of the column repeat and if values of the column are metric, sum of these metric values are in this category. For instance, if a table contains prepaid and postpaid payment type column values, sum of these metric values are considered in this category.

- Value format statistics: In this category, statistics give information about how many times the values of the column are repeated according to the format of column values.

- Value distribution statistics: In this category, statistics give information about classification of numerical values of columns, according to minimum and maximum values.

Logic functions of AutoTest have been developed by using the PL/SQL database programming language on Oracle 11G Release 2 Enterprise Edition database management system [15]. Application Express (APEX) [16] is used to provide interface of AutoTest. It is integrated to Oracle database and it helps with developing user interfaces quickly. The login page interface for the proposed AutoTest is given in Fig. 1.

After logging on, the users first see the application home page (see Fig. 2). Working tasks/active jobs with name, owner, state and starting date can be seen on this page. Also, histories of application jobs are presented to the user in a table. The users can obtain some information such as how long successful jobs took to complete and for what reasons jobs failed. Moreover, a graphic shows the percentage of failed and successful jobs.

Value of frequency statistics information is presented to the user as shown in Fig. 3. Users can filter out a table and inspect the distribution of its values visually. In addition to this, the users can compare value frequency statistics, format statistics and distribution statistics of any two tables.
The users can examine changes in table’s variables over time. This process is done with linear regression approach. After this approach, trend lines of variables will occur. When analysis is complete, outlier value are sent to users by email messages (see Fig. 4). Depending on the quantity of deviation, an alert is offered to the users at three different levels (warning, error, critical).

![Figure 4. Trend analysis results reported to the user](image)

V. CONCLUSION

Data collections are very crucial for organizations’ successful decision making and their businesses in general. However, there are many kinds of data quality problems stemming from many reasons and indirectly effecting or reducing the profits of the businesses. In this paper, we proposed AutoTest Tool to overcome some of these problems. AutoTest enables data quality tests of data warehouse applications, analyze historical changes of a dataset with linear regression algorithm and reports outlier variables of data trend to the user of application. In the future, we plan to integrate artificial intelligent into our system for less user interaction.

REFERENCES

Süleyman Eken did his B.Sc. in Computer Engineering from Karadeniz Technical University in 2009 and M.Sc. in Computer Engineering from Kocaeli University in 2012. He is currently working towards Ph.D. degree in Computer Engineering from Kocaeli University, Turkey. Also, he is currently a Research Assistant of Computer Engineering Department at Kocaeli University in Turkey. His current research interests include Satellite Image Processing, Remote Sensing, WEB-GIS and Spatial Databases.

Kürşat Topçuoglu was born in Istanbul, Turkey, in 1990. He received the B.Sc. in Computer Engineering from Kocaeli University (Kocaeli, Turkey). Also, he is student at Industrial Engineering in Kocaeli University and software developer at Turkcell Technology R&D. His current research interests include data warehouses, data mining, and data analysis.

Ahmet Sayar has received his B.Sc. in Management Engineering from Istanbul Technical University (Istanbul, Turkey). He has received his M.Sc. in Computer Science from Syracuse University (NY, USA) and his PhD degree in Computer Science from Indiana University (IN, USA). He has worked at Los Alamos National Laboratory (New Mexico, USA) and Community Grids Laboratory (IN, USA) as a researcher. He is currently an Assistant Professor at Kocaeli University Computer Engineering Department in Turkey. His current research interests include Distributed Systems (such as Big Data Analytics and Distributed File Systems), Remote Sensing, WEB-GIS and Exploratory Spatial Data Analysis.