Evaluating the Quality of databases Instances based on Completeness and Accuracy of Data

Mohammed N. Al_Khwlani[1], Mariam Shmsan[2], Nadia Al-akhram[2]

[1]Amran University
[2]UST University

Summary
Data quality for database represents the most important of the essential things for any organization where these organizations depend on their databases for providing the required information in any situation or time. In addition, data quality directly influences every day important decisions that made on all management levels based on the data stored within databases. However, real-world databases often contain both syntactic and semantic errors that cause several problems and damages to the organization. This paper proposes an approach to evaluate the quality of databases instances for treating and cleaning the errors. The evaluation process is based on several criteria for data quality such as completeness and correctness. These criteria are measured by counting the problems of data items for each relation in databases in order to identify the relation that has the most number of problems. Detecting the noisy data values in this paper is made by the clustering approach, especially using the K-means algorithm. The proposed approach was applied on several samples of databases from different drivers of ODBC using an application system that is designed by java beans for this purpose.

Key words:
data quality, database, clustering, noisy data

1. Introduction

Although the databases community has produced a large amount of research on integrity constraints and other safety measures to maintain and ensure the quality of information stored in relational databases, real-world databases often still contain a non-trivial number of errors. These errors, both syntactic and semantic, are generally subtle mistakes, which are difficult or even impossible to express (and detect) using the general types of constraints available in the modern databases systems. Data quality plays an important role in a wide variety of information systems applications.

Evaluating the quality of databases aims to help the organization in treating their data constantly. This paper proposes an approach to evaluate the quality of databases instances for cleaning the errors. This evaluation process is based on several criteria of data quality such as completeness and correctness. These criteria are measured by counting the problems of data items for each relation in databases in order to identify the relations that contain the most problems of data. Detecting the noisy data values in this paper is made by the clustering approach, especially using the K-means algorithm.

2. Definitions

The following paragraph includes several definitions of terms that relate to data quality.

Data Quality: is defined as fitness for use; meaning it is strongly influenced by contextual factors [10].

Completeness: deals with to ensure is all the requisite information available? Are some data values missing, or in an unusable state?[2]

Accuracy: Do data objects accurately represent the “real world” values they are expected to model? Incorrect spellings of product or person names, addresses, and even untimely or not current data can impact operational and analytical applications [2].

3. Related work

[11] performed the first survey on the quality of relational schemas in open-source software. The overall quality results in this work are worse than envisioned at the beginning of the study. [10] proposes a framework for data quality assessment based on the contextual factors management level and decision process phase which allows a decision maker to relate data quality dimensions to values of features of the information demand. Based on this individual assignment of dimensions, the user can access situation specific information regarding data quality. The key features of [1] method are that (1) it uses an efficient approximate inference algorithm that is easily implemented in standard DBMSs and scales well to large databases sizes, and (2) it uses shrinkage and joint inference to accurately infer correct values even in the presence of both missing and corrupt values. [2] has attempted to collect all possible causes of data quality problems that may exist at all the phases of data warehouse. The researches of this study reached to that the most common type of problems that are manifested in literature of data quality are: lack of standardization of data, non standardization of formats, heterogeneity of data

Manuscript received January 5, 2013
Manuscript revised January 20, 2013

sources, Non-Compliance of data in data sources with the standards, missing data, and inconsistent data across the sources. The main objective of [4] is clustering the outliers in the numerical domain and sequence of data set based on the leader algorithm instead of reclustering the entire sliding window /data set by calculating the threshold with the average method and maximal resemblance of leaders. This is more efficient than using reclustering the sliding window. [5] discusses different methods to impute the missing values. [6] proposes the notion of data quality constraint that may be associated to any of the intermediate relations produced by data transformations in a data cleaning graph (DCG). This work equip a data cleaning graph with data quality constraints to help users identifying the points of the graph and the records that need their attention and manual data repairs for representing the way users can provide the feedback required to manually clean some data items. [7] proposes a new progress on the basis of spectral clustering to analyze the structure of a given dataset and find anomaly data points in the dataset. [8] focuses on outlier detection in health data sets such as Pima Indians Diabetes data set and Breast Cancer Wisconsin data set using partitioning clustering algorithms. The algorithms used in this research work are PAM, CLARA AND CLARANS and a new clustering algorithm ECLARANS is proposed for detecting outliers. In order to find the best clustering algorithm for outlier detection several performance measures are used. The experimental results show that the outlier detection accuracy is very good in the proposed ECLARANS clustering algorithm compared to the existing algorithms. [9] presents an approach called ERACER. This approach is an iterative statistical framework for inferring missing information and correcting such errors automatically.

4. Scope of the Research

This research concentrates on applying the two most important measurements of data quality: completeness and Accuracy. For completeness as it defined in above, it is based on the missing data that have been one major factor affecting data quality. The presence of missing data is a general and challenging problem in the data analysis field. Fortunately, missing data imputation techniques can be used to improve data quality [5]. The numerical values columns are only considered to detect the noisy data for evaluating the accuracy of databases instances with ignoring the auto number and derived values columns because these values are computed not inserted by users. Relational database only considered and ignoring the other types of databases.

5. Methodology of the research

This research was conducted by several steps as follows:

1. Adding the database that will be evaluated in the ODBC bridge.
2. Connecting the database with the proposed approach model.
3. Detecting and counting the missed values that is contained in each relation for the database .
4. Detecting and counting the noisy numeric data values for each relation using clustering approach.
5. Computing the percentages of missed values and noisy values for each relation.

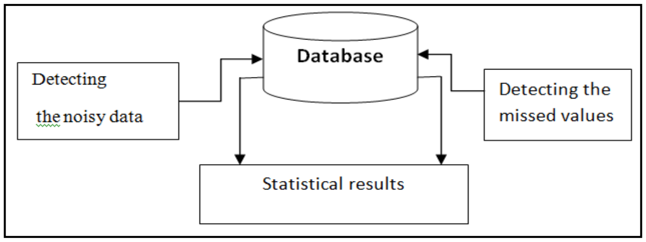
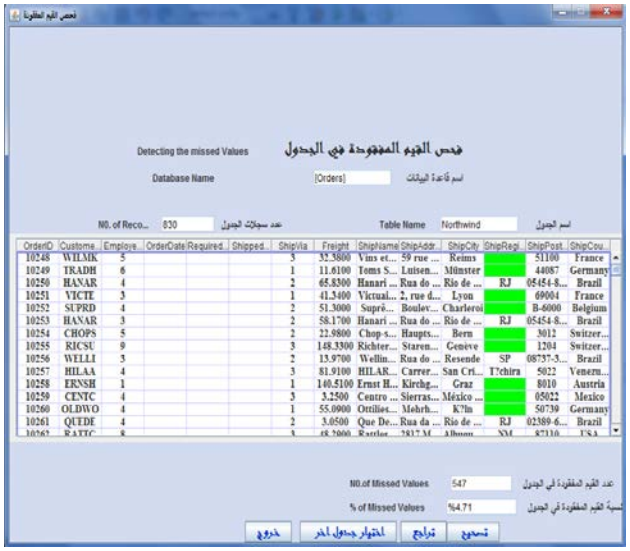


Figure 1 shows the architecture of the proposed approach.

6. Detecting the missed values

Detecting the missed values is based on scanning any attribute does not contain any values. The detection process is performed by comparing each attribute or cell for each table with null for text columns or with zero for numeric columns. The proposed approach after finishing of detection process shows the locations of missed values and statistics of these values as presented in Figure 1.


Detecting the missed Values فحص القيم المفقودة في الجدول

Database Name [Orders] اسم قاعدة البيانات

N0. of Reco... 830 عدد سجلات الجدول

Table Name Northwind اسم الجدول

OrderID	Custome...	Employe...	OrderDate	Required...	Shipped...	ShipVia	Freight	ShipName	ShipAddr...	ShipCity	ShipRegi...	ShipPost...	ShipCou...
10248	WILMK	5				3	32.3800	Vins et...	59 rue ...	Reims	🟩	51100	France
10249	TRADH	6				1	11.6100	Toms S...	Luisen...	Münster	🟩	44087	Germany
10250	HANAR	4				2	65.8300	Hanari ...	Rua do ...	Rio de ...	RJ	05454-8...	Brazil
10251	VICTE	3				1	41.3400	Victuai...	2, rue d...	Lyon	🟩	69004	France
10252	SUPRD	4				2	51.3000	Suprê...	Boulev...	Charleroi	🟩	B-6000	Belgium
10253	HANAR	3				2	58.1700	Hanari ...	Rua do ...	Rio de ...	RJ	05454-8...	Brazil
10254	CHOPS	5				2	22.9800	Chop-s...	Haupts...	Bern	🟩	3012	Switzer...
10255	RICSU	9				3	148.3300	Richter...	Staren...	Genève	🟩	1204	Switzer...
10256	WELLI	3				2	13.9700	Wellin...	Rua do ...	Resende	SP	08737-3...	Brazil
10257	HILAA	4				3	81.9100	HILAR...	Carrer...	San Cri...	T?chira	5022	Venezu...
10258	ERNSH	1				1	140.5100	Ernst H...	Kirchg...	Graz	🟩	8010	Austria
10259	CENTC	4				3	3.2500	Centro ...	Sierras...	México ...	🟩	05022	Mexico
10260	OLDWO	4				1	55.0900	Ottilies...	Mehrh...	K?ln	🟩	50739	Germany
10261	QUEDE	4				2	3.0500	Que De...	Rua da ...	Rio de ...	RJ	02389-6...	Brazil
10262	RATTC	8				3	48.2900	Rattler...	2817 M...	Albuqu...	NM	87110	USA

N0.of Missed Values 547 عدد القيم المفقودة في الجدول

% of Missed Values %4.71 نسبة القيم المفقودة في الجدول

Figure 1: Detection Results for Missed Values in a table

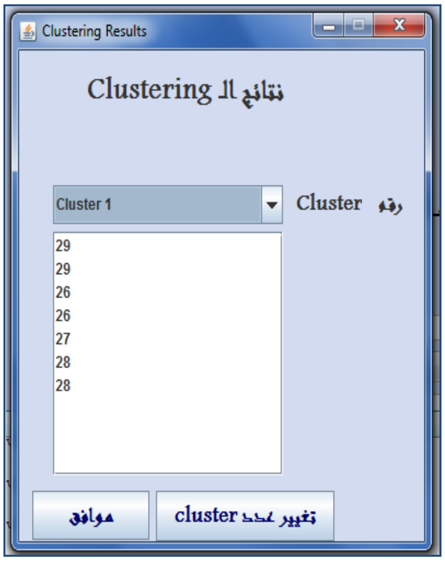
7. Detecting the noisy data

Detecting process for the noisy data considers the numerical data items for each column in the relation using the k-means algorithm of clustering approach. The algorithm detects the noisy data for overall database, table by table a way. Then in each table, the algorithm scans the numerical columns to detect the noisy data, column by column as shown in Figure 3.

7.1 K-means Algorithm

Let DB=database, R is a relation in databases, n is a number of relations, m is the number of columns C for each relation, V is a data value in a column, Nv is the number of values for each column C, Nc is the number of clusters.

```
1. For i=1 to n
2.    For j=1 to m
3.    Scan Ci
4.  Let Nc=2
5.  Choose randomly two centers from values of Ci.
6.  For k=1 to Nv
7.     Compute the distance for each Vk about the two centers.
8.    Compare the two distances for each Vk
9.     Add each Vk to the nearest center that will be as a cluster.
10.  Compute the average of values for each cluster.
11.   Let the two averages for the two clusters values as new centers.
12.   Compute the constant of centers by subtracting the old centers values from the new centers Values.
13  If the constant value is zero then goto step 14
       Else goto step 7.
14  Display the clusters values
```



نتائج ال Clustering

Cluster 1

Cluster رقم

29
29
26
26
27
28
28

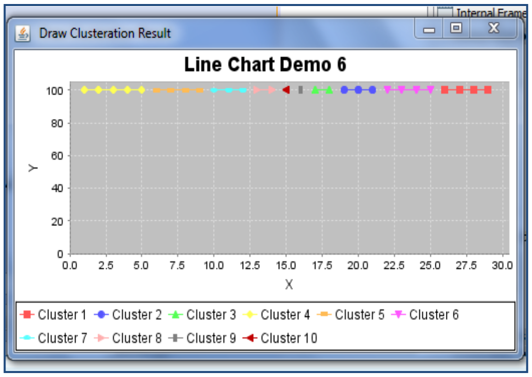


Figure 2: Output clusters by the proposed Approach

8. Data preparation

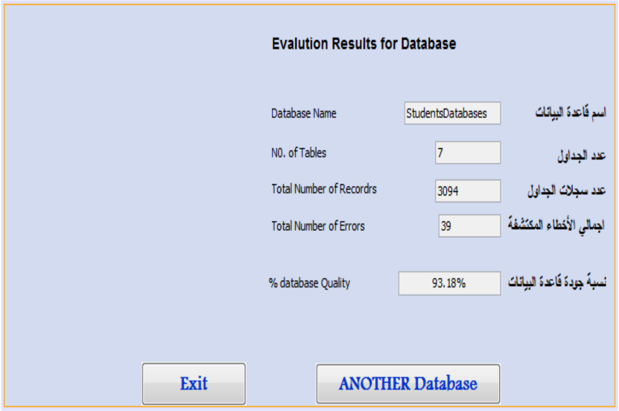
For testing the proposed approach, many databases with different drivers were created with several samples. These databases have different numbers of tables. Table 1 presents the databases that used for testing the proposed approach.

Table 1: Databases for testing the proposed approach

Database	Driver
Moneychanger	Oracle
Students + test	MS SQL Server
Northwind + Market	Access
Test	MySQL

9. Implementation

This research was implemented using several tools. These tools include NetBeans IDE version 6.1, Open Database Connectivity (ODBC).



Evalution Results for Database

Database Name	StudentsDatabases	اسم قاعدة البيانات
N0. of Tables	7	عدد الجداول
Total Number of Recordrs	3094	عدد سجلات الجداول
Total Number of Errors	39	اجمالي الأخطاء المكتشفة
% database Quality	93.18%	نسبة جودة قاعدة البيانات

Figure 4: Evaluation Results for StudentsDatabase

10. Conclusion

The proposed approach in this study evaluates the databases instances based on two measurements of data quality: completeness and accuracy. The evaluation process is performed for each individual relation of database. The applied algorithms for this study scan the columns of relation to detect the missed and noisy data values. The proposed approach can shows the locations of missed values and noisy data in database relations to enable the user to correct them later. In addition, this approach identifies the relations that contain the most problems even the columns for each relation based on the percentages of these problems.

For Future work, there are several works can be proposed. First one is evaluating the data warehouse that constructed by integrating multiple databases. Second one, considering the noisy data for texts columns by the current proposed approach. Third one is extending the proposed approach to enable the user to clean and correct the errors of data.

References

[1] Chris Mayfield Jennifer Neville Sunil Prabhakar. A Statistical Method for Integrated Data Cleaning and Imputation.

[2] Ranjit Singh, Dr. Kawaljeet Singh. A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing. IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 2, May 2010.

[3] Rahime Belen. Detecting Disguised Missing Data. Master Thesis, the Middle East Technical University, February, 2009.

[4] N.Sudhakar& KVN Sunitha. Efficient Clustering Algorithm for Large Data Set. International Journal of Advanced Research in Computer Science and Software Engineering(ijarcsse), Volume 2, Issue 1, January 2012.

[5] Dinesh J. Prajapati, Jagruti H. Prajapati. Handling Missing Values: Application to University Data Set. International Journal of emerging trends in engineering and development Issue 1, Vol 1, August, 2011.

[6] Helena Galhardas1, Ant_onia Lopes2, and Emanuel Santos. Support for User Involvement in Data Cleaning.

[7] Hai LIN, Qingsheng ZHU. A Spectral Clustering-Based Dataset Structure Analysis and Outlier Detection Progress. Journal of Computational Information Systems 8: 1 (2012) 115–124.

[8] S.Vijayarani and S.Nithya. Article:An Efficient Clustering Algorithm for Outlier Detection. International Journal of Computer Applications 32(7):22-27, October 2011.

[9] Chris Mayfield, Jennifer Neville, Sunil Prabhakar. ERACER: A Database Approach for Statistical Inference and Data Cleaning. SIGMOD'10, June 6–11, 2010, Indianapolis, Indiana, USA.

[10] Daniel Poeppelmann & Christian Schultewolter. Towards a Data Quality Framework for Decision Support in a Multidimensional Context. International Journal of Business Intelligence Research, 3(1), 17-29, January-March 2012.

[11] Fabien Coelho, Alexandre Aillos, Samuel Pilot, and Shamil Valeev. On the Quality of Relational Database Schemas in Open-source Software. International Journal on Advanceds in Software, Vol. 4, No 3&4, 2011.

Mohammed N. Ahmed was born in Yemen 1968. He has got PhD degree in Database Systems, 2008, faculty of Computer Science and Information Systems, UPM, Malaysia. His Master degree was in Computer Science, UST, Yemen, 2001. His Bsc degree in computer Science, Mustanseria University, Baghdad, Iraq. Currently, Dr Mohammed is an assistant professor in Math Dept. at Amaran University, Yemen. Further, he is a head of computer Science in Future University, Sana'a Yemen.