International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 14, Number 12 (2018), pp. 1087-1094

http://www.ripublication.com

Data Cleaning Using Context Dependent and Independent Attribute Correction

V.Umesh

Research Scholar Department of Computer Science,
Bharathiar University, Coimbatore, India.

Abstract

Data cleaning is a process that detects and removes the errors and inconsistency in the data in oder to improve the quality of the data. To have a high data quality, data quality problems has to be solved. Data quality problems exist in single and multiple source systems. Single source problems refers to the errors, inconsistency, missing values, uniqueness violation, duplicated records and refrential integrity violations. Multiple source problems are structural conflicts, naming conflicts, inconsistent timing and aggregating. This work aims at providing access to the accurate and consistent data by solving these data quality problems using data mining techniques.

Keywords: Data cleaning, data quality.

1. INTRODUCTION

Data mining is a process that is used to extract hidden patterns from the huge data set. The steps involved in data mining are selection Pre-processing, Transformation, Data Mining and Interpretation/Evaluation. One of the major steps is data pre-processing.

Data cleaning [1] (data cleansing or data scrubbing) is one of the sub steps in data pre-processing. The intention of data cleaning process is to detect errors and remove the errors and inconsistencies from the data in order to improve the quality of the data.

Data Cleaning [2] is a process used to determine inaccurate, incomplete, inconsistent and unreasonable data, then improving the quality of the data by correcting the detected errors. The sources of errors are lexical errors, syntactical errors, irregularities, duplicates and data entry anomalies.

Data cleaning [3] approaches involve several phases:

a) **Analyse the data:** In order to detect different kinds of errors and inconsistencies that are to be removed, we require a detailed data analysis. Metadata about the data properties should be obtained to detect data quality problems.
b) **Definition of data transformation workflow and mapping the rules:** Depending on the different data sources, their heterogeneity and the errors in the data, a huge number of data transformation and data cleaning steps may need to be executed.
c) **Verifying the data:** The efficiency and the correctness of transformation definitions should be tested and evaluated.
d) **Transformation:** Execution of the transformation steps is both loading of data and refreshing a data warehouse and also answering multiple queries on different sources.
e) **Backflow:** After all the errors are corrected, the cleaned data should also replace the dirty data in the original sources.

DATA QUALITY:

Data quality specifies a state of completeness, consistency, validity, accuracy and timeliness that makes data appropriate for a specific use. Poor data quality leads to loss of money and inaccurate decision.

The following are characteristics of data quality:

a) **Completeness:** It is the characteristic of having filled all the required values for the data fields.

b) **Validity:** It is a measure of degree of conformance of data values to its domain and business rules. This includes Domain values, ranges, reasonability test, primary key uniqueness, referential integrity.

c) **Accuracy:** It is a measure of the degree to which data agrees with data contained in an original source.

d) **Precision:** The domain value which specifies business should have correct precisions as per specifications.

e) **Non-duplication:** It is the degree to which there is a one-to-one correlation between records and the real world object or events being represented.

f) **Derivation Integrity:** It is the correctness with which two or more pieces of data are combined to create new data.

g) **Accessibility:** It is the characteristic of being able to access data as needed.

h) **Timeliness:** It is the relative availability of data to support a given process as the time changes.

2. MOTIVATION

Data plays a fundamental role in every software system. In particular, information system and decision support system depends on it more deeply. In today's environment, there is a need for more correct information for a better decision making. Data quality is a crucial factor in data warehouse creation and during data integration. Medical data mining has got a great potential for exploring the hidden patterns in the data sets of medical domain. This pattern has to be utilized for clinical diagnosis. But there exists a data quality issues such as duplicated records, missing values, inconsistencies, referential integrity violations, uniqueness violations and errors that need to be handled in order to have high data quality. This motivated the approach of data cleaning, which is a process of identifying and removing the errors and inconsistency from the data to improve the quality of the data.

3. DATA CLEANING PROBLEMS

Data Cleaning is one of the tasks performed during the integration of database, during the process of knowledge discovery in the databases, and also in the creation of the data warehouse.

[4] Presents the requirement to cleanse the data during the integration of data into the database from the heterogeneous data sources. It lists out the sources of erroneous data (i.e., lexical errors, syntactical errors, duplication, data entry anomalies and missing values). It also explains what happens if the errors in the data are not cleaned.

[5] Presents the classification of data quality problems as single source and multi-source problems. Further single source and multi-source problems are classified as schema level and instance level problems. Schema-level problems.

[6] Data quality problems are addressed at schema level by improving the schema design, schema translation and integration. Instance-level problem refers to the errors and inconsistencies in the data set that are not visible at schema level. These problems are the primary focus in data cleaning. This paper then outlines the major steps in data cleaning as data analysing, data transformation and verification and also emphasizes the need to recover schema level and instance-related data transformations in a cleaned integrated way. Fig 1 shows the classification of data quality problems.

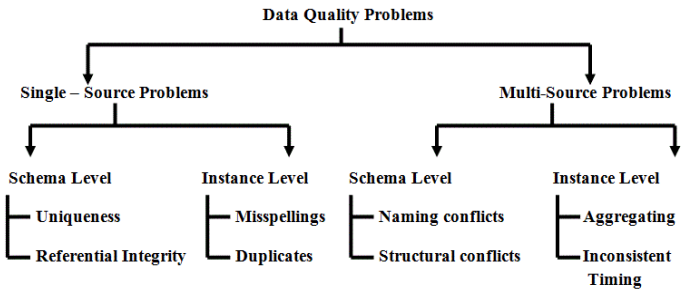

Fig 1: Categorization of data quality problems in data sources

[7] Presents the usage of medical database for the Management of Patients with Implanted Pacemakers. Here data cleaning is done for the purpose of Knowledge Discovery in the database. In this paper they have verified the medical database, repaired the broken references between the data fields, recreated the missing rows and performed some data type conversions by converting dates to the number of days.

[8] Presents various approaches in data cleaning for detecting and eliminating the duplicates. These approaches can either detect the exact duplicates or approximate duplicates.

[9] Detection of approximate duplicates is the most challenging task in data cleaning. Here there are two categories of data cleaning algorithm one is domain dependent that requires domain knowledge for detecting the duplicates and the second method is domain independent which does not require domain knowledge. A Sorted neighbourhood method is the only domain dependent approach whereas Token based data cleansing method, Priority queue method and Duplicate record elimination using external merge/purge sort are domain independent algorithms.

[10] Investigates a missing attribute value which is one of the forms of data incompleteness. The most common approaches used for computing the missing values are: use the most common attribute value, use attribute mean value, use some global constant, fill missing value manually, treat missing value as a special value, ignore the records or use probable value to fill the missing values.

[12] Presents a Regression methodology to predict the missing values. This paper concentrates mainly on filling the missing values by using most probable method called "Regression algorithm". A regression is a statistical analysis for assessing the association between two variables. Regression is used to find the relationship between two variables. Regression method can be used to find the missing values and fill it in the database without changing it manually. The missing value can be calculated using Regression Equation y = a + b*x where b = slope and a = Intercept.

$$\mathbf{a} = \frac{(\Sigma y)(\Sigma x^2)-(\Sigma x)(\Sigma xy)}{n(\Sigma x^2)-(\Sigma x)^2}$$

$$\mathbf{b} = \frac{n(\Sigma xy)-(\Sigma x)(\Sigma y)}{n(\Sigma x^2)-(\Sigma x)^2}$$

Where x and y are variables.

b = is the slope of the regression line.

a = is the intercept point of the regression line on the y axis.

N = is the Number of values or elements.

X = is First Score.

Y = is Second Score.

ΣXY = is Sum of the product of first and Second Scores.

ΣX = is Sum of First Scores.

ΣY = is Sum of Second Scores.

ΣX2 = is Sum of square First Scores.

[13] Estimates the missing values by using Lagrange Interpolation method. In a dataset if one of the attributes is depending on the other then by using known values we can compute the unknown values. Interpolation is a process of finding the unknown values from a known value. By taking the values of (xi, yi), where i ranges from 0, 1.....n of any function Y = f(x), the process of estimating the values of 'y', for any intermediate value of 'x' is called interpolation. The missing values can be calculated using Lagrange Interpolation as follows:

$$P(x) = \sum_{j=1}^{n} Pj\ (x)$$

$$Pj(x) = yj \prod_{\substack{k=1 \\ k \neq j}}^{n} \frac{x - xk}{xj - xk}$$

[14] Presents an attribute correction and data standardization techniques for finding errors. For attribute correction, two methods of data mining techniques are used i.e., clustering techniques for Context-independent correction and association rule for Context-dependent correction. In Context-independent, the attribute values are corrected without regard to the other attributes in the data set. This algorithm uses the concept of reference data set. Reference data is the data that occurs most frequently in the data set. This method uses clustering concept with the nearest neighbour algorithm. In Context-dependent, the attributes are examined and corrected by taking the consideration of values of other attributes within a given record. This method uses the concept of Apriori algorithm. In both the algorithms to find the distance between strings Levenshtein distance [18] is used. It is the number of text edit operations required between the strings

Algorithm ContextIndependent (C[0….n-1])

//Detects and correct the errors of the attribute chest pain type with the help of //reference dataset.

// Input: Database D, of medical datasets;

An array C [0…n-1] of chest pain type;

// Distance threshold, distThresh; Occurrence relation, occRel.

// Output: Corrected values of chest pain type.

Method:

1) for i ← 0 to n-1 do
2) value ← count the occurrence of each C[i]
3) if value > occRel then
4) add to reference dataset R[i]
5) for each chest pain type C_j in databasedo
6) distance←LevenshteinDistance(R[j],C[j)
7) if distance ≤ distThresh then
8) C[j] ← R[j] //replace C[j] with R[j]
9) return C[j]

Algorithm ContextDependent

(A[0….n-1], B[0….n-1], D[0….n-1],[0….n-1], C [0….n-1])

// Detects and correct the errors by considering values of other attributes like age, //blood pressure, cholesterol, heart beat.

// Input: Database D, of medical datasets; An array A [0…n-1] of age;

// An array B [0…n-1] of blood pressure; An array D [0…n-1] of cholesterol;

// An array H [0…n-1] of heart beat rate; An array C [0…n-1] of chest pain type;

// Minimum support, minSup; Distance threshold, distThresh.

// Output: Corrected values of chest pain type.

Method:

1) L_1 ← {frequent items}
2) for ($k=0; L_k \neq \phi; k++$) do
3) C_{k+1} ← Candidates generated from L_k
4) for each transaction t in D do

5) $c_t \leftarrow$ subset (C_k,t)

6) for each candidate c $\in$ C_t c.count++

7) $L_k \leftarrow$ {c $\in$ C_k | c. count $\geq$ minSup}

8) return $\cup_k L_k$

9) for each frequent itemset i do

10) for each subset s of i do

11) conf $\leftarrow$ support(i) / support(i - s)

12) if conf $\geq$ minConf then

13) Output the rule (i-s) $\rightarrow$ s

14) for each chest pain type C_j in database do

15) distance$\leftarrow$LevenshteinDistance(S[j],C[j)

16) if distance $\leq$ distThresh then

17) C[j] $\leftarrow$ S[j]

18) return C[j]

IV CONCLUSION

Data or the information that are used in the medical field is one of the most important assets. High data quality is one of the most important requirements for taking successful decisions over the medical datasets. In order to provide high data quality for these datasets we have used different data mining techniques. Data quality issues such duplicates, missing values and errors are solved using these techniques making the data efficient and accurate.

Our proposed work implemented the algorithm to find duplicates, missing values and errors in the single source data. Further as a future work, multi-source problems such as naming and structural conflicts can be solved by collecting the data from different sources.

REFERENCES

[1]. KDnuggets Polls. “Data Preparation Part in Data Mining Projects”, Sep30-Oct-12, 2003. http://www.kdnuggets.com/polls/2003/data_preparation.htm.

[2]. Sweety Patel,“Requirement to cleanse DATA and why is data cleansing in Business Application?” International Journal of Engineering Research and Applications, Vol. 2, Issue 3, May-Jun 2012.

[3]. Erhard Rahm, Hong Hai Do. “Data Cleaning: Problems and Current Approaches”. IEEE Data Engineering Bulletin, 2000, 23(4):3-13.

[4]. Adam Widera, Michal Widera, Daniel Feige,“ Data Cleaning on Medical Data Sets” Journal of Medical Informatics and Technologies,Vol. 8,ISSN 1642-6037, Jun e 2008.

[5]. Dr. Payal Pahwa, Rashmi Chhabra, “Domain Dependent and Independent Data Cleansing Techniques”, International Journal of Computer Science and Telecommunications, Vol. 2, Issue 3, September 2011.

[6]. M. Bilenko and R. J. Mooney, “Adaptive duplicate detection using learnable string similarity measures”. ACM SIGKDD, 39-48, 2003.

[7]. Monge, A. E. “Matching Algorithm within a Duplicate Detection System.” IEEE Techn. Bulletin Data Engineering 23 (4), 2000.

[8]. H.H. Shahri; S.H. Shahri , “Eliminating Duplicates in Information Integration: An Adaptive, Extensible Framework, Intelligent Systems”, IEEE, Volume 21, Issue 5, Sept.-Oct. 2006 Page(s):63 – 71.

[9]. Jerzy, W.Grzymala-Busse1 “MingComparison of Several Approaches to Missing Attribute Values in Data Mining Techniques” Journal of Computer Science ISSN 1549-363 Science Publications.

[10]. Z. Mahesh Kumar , R. Manjula “Regression model approach to predict missing values” International Journal of Computer Science & Engineering Technology , ISSN : 2229-3345 Vol. 3 No. 4 April 2012.

[11]. L.Sunitha, Dr M.BalRaju , J.Sasikiran “Estimation of Missing Values Using Lagrange Interpolation Technique” International Journal of Advanced Research in Computer Engineering & Technology, Volume 2, Issue 4, April 2013.

[12]. R. Kavitha Kumar and Dr. RM. Chadrasekaran “Attribute Correction – Data Cleaning using Association rule ” International Journal of Data Mining & Knowledge Management Process ,Vol.1, No.2, March 2011.

[13]. Lukasz Ciszak, “Application of clustering and Association Methods in data cleaning”, 978-83-60810-14-9/08, 2008 IEEE.

[14]. W. Cohen, P. Ravi Kumar, S. Fienberg “A Comparison of String Metrics for Name-matching Tasks” in Proceedings of the IJCAI-2003.