International Journal of Advances in Computer Science and Its Applications– IJCSIA
Volume 4: Issue 2 *[ISSN: 2250-3765]*
Publication Date : 25 June 2014

Data Cleaning in Knowledge Discovery Database (KDD)-Data Mining

Fauziah Abdul Rahman
Faculty of Technical Foundation,
Universiti Kuala Lumpur
Malaysian Institute of Industrial Technology (MITEC)

Mohammad Ishak Desa
Faculty of Computing
Universiti Teknologi Malaysia
Johor, Malaysia

Antoni Wibowo
Faculty of Computing
Universiti Teknologi Malaysia
Johor, Malaysia

Norhaidah Abu Haris
Faculty of Software Engineering
Universiti Kuala Lumpur,
Malaysia

Abstract— ***Data quality is a main issue in quality information management. Data quality problems occur anywhere in information systems. These problems are solved by data cleaning. Data cleaning (DC)is a process used to determine inaccurate, incomplete or unreasonable data and then improve the quality through correcting of detected errors and omissions. Generally data cleaning reduces errors and improves the data quality. It is well known that the process of correcting errors in data and eliminating bad records are time consuming and involve a tedious process but it cannot be ignored. Various process of DC have been discussed in the previous studies, but there's no standard or formalized the DC process. Knowledge Discovery Database (KDD) is a tool that enables one to intelligently analyze and explore extensive data for effective decision making. The Cross-Industry Standard Process for Data Mining (CRISP-DM) is one of the KDD methodology often used for this purpose. This paper review and emphasize the important of DC in data preparation. The wrong analysis will probably turn out to be expensive failures. The future works was also being highlighted.***

Keywords—***Data Cleaning, Data Mining, DC Process, Missing Value***

I. Introduction

Data cleaning (DC), also called data cleansing or scrubbing, includes operations that correct bad data, filter some bad data out of the data set, and filter out data that are too detailed for use in the mode [1]. In other words it deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data [2]. Data Quality(DQ) problems are present in single data collections, such as files and databases, for example, due to misspellings during data entry, missing information or an invalid data. This is because the sources often contain redundant data in different representation. In order to provide access to accurate and consistent data, consolidation of different data representations and elimination of duplicate or missing information become necessary. While a huge body of research deals with schema translation and schema integration, DC process has received only little attention in the research community. A number of authors focused on the problem of duplicate identification and elimination as well as on data mining approaches in DC, only a little have been known on the DC process. DC processes are identified as the most important phase in CRISP-KDD Data Mining (DM)to determine data quality that reflects the end results. Few research efforts have been carried out in these steps compared to data mining, suggesting that DC process should be formalize are needed.

Fauziah Abdul Rahman

Universiti Kuala Lumpur
Malaysia

II. Data Cleaning (DC)

A. *DC Current Issue*

When multiple data sources need to be integrated, the need for data cleaning increases significantly. For example in data warehouses, federated database systems or global web-based information systems, the need for DC increases significantly. The continuously refresh huge amounts of data from these variety of sources also allowed the probability of "dirty data" is high. If it is used for decision making, then it will cause a wrong conclusion. Due to the wide range of possible data inconsistencies and the sheer data volume, DC is considered to be one of the biggest problems in data warehousing [3].

In the current practices in DC process, the involvement of domain expert is very important, because the detection and correction of anomalies requires detailed domain knowledge. DC is therefore described as semi-automatic but it should be as automatic as possible because of the large amount of data that usually is be processed and because of the time required for an expert to cleanse it manually [1]. The ability for comprehensive and successful DC is limited by the available knowledge and information necessary to detect and correct anomalies in data. So far only a little research has appeared on DC, although the large number of tools indicates both the importance and difficulty of the cleaning problem [4][1].

Another issue in DC is there is no common description about the objectives and extend of comprehensive in DC. Additionally DC is a term without a clear or settled definition [5]. There is no formalize of DC process and most authors of peer-reviewed journal articles go to great lengths to describe their study, the research methods, the sample, the statistical analyses used, results, and conclusions based on those results. However, few seem to mention DC which can include screening for extreme scores, missing data, normality and a little have been known on the DC process specifically on the DC process using data from the real world [9].

B. *Past Literatures on DC Process*

DC is a process used to determine inaccurate, incomplete or unreasonable data and then improve the quality through correcting of detecting errors and omissions. Generally DC reduces errors and improves the data quality [4]. CRISP-KDD was the first generation of KDD where Data Mining (DM) process attached together in the KDD life cycle to ensure a discover knowledge can meet the business requirements. However, correcting errors in data and eliminating bad records can be a time consuming and tedious process but it cannot be ignored. DC process in the CRISP-KDD is used for discovering interesting information in data by applying DM techniques to identify and recover data quality problems in large databases. It is important to understand each phase before implementing the DM process as Figure 1 below. Nowadays researchers with strong industrial engagement realized the need from DM to KDD to deliver useful knowledge in the business decision making.

However, Referring to Figure 1, the third phase is data preparation which consists of DC process is the most difficult and time-consuming element in KDD process. To perform DC process there are several of steps implemented in the previous studies as in Table 1. Most of the studies determined that handling missing data are needed to solve at the early stage in DC process. However, the DC process was based on the type of data set. Therefore, few research efforts have been carried out in these steps compared to data mining, suggesting that DC process should be formalize are needed.

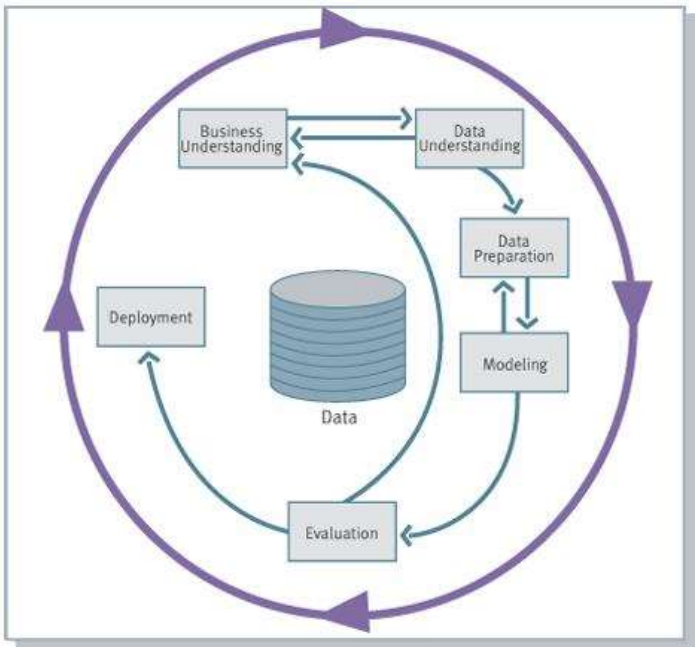

Figure 1. Cross-Industry Standard Process for Data Mining Methodology (CRISP-DM)

TABLE I. PAST RESEARCH ON DC PROCESS

Title	DC Process
1. Data Preprocessing for prediction of recirculating water chemistry fault By: Gao Qiang and et. el (2010)	Steps in DC: Step 1: Data classification- classify the data in case study based on categories Step2: Data Cleaning - Eliminate the noise of data preprocessing - Replace the missing data with mean of the value - Delete the redundancy data - Delete the duplicate data Step 3: Data Transformation Step 4: Data reduction using PCA method
1.Discovering Knowledge in Data By: Daniel T.Larose (2005)	-This paper emphasized the used of Exploratory Data Analysis (EDA). -The EDA process is based on CRISP-DM KDD Methodology. The steps are: a) Handling Missing Data b) Identifying Misclassification c) Identifying Outliers using Graphical and Numerical Methods d) Data Transformation: Numerical transformation and Categorical transformation.
1. E-Clean: A Data Cleaning Framework for Patient	-This paper emphasizes the process involve in DC: a) DC should detect and remove all major errors and inconsistency in the database. b) DC should perform mapping and

Data By: Hasimah Hj Mohamed and et. El. (2011)	merging function. DC should be able to let a user to insert valid value for each newly created attribute. The data quality issues can be divided into categories: single-source problems or multiple-source problems. The examples of single-source problems are the redundancy and duplicate. The multiple-source problems are reffering to the contradicting or overlapping and inconsistence data . - DC Framework proposed is ETL Process Model: 1) E: Extract stge 2)Transform stage 3)Load stage
2. Attribute Correction-Data Cleaning using Association Rule and Clustering Method. By: R.KAVITHA KUMAR and et.el. (2011)	DC Framework proposed: 1) Raw data 2) Pre-processing 3) Selection of attributes 4) New Data 5)Selection of DM techniques (Association Rule &Clustering) 5) Clean Data

III. Application of DM in Missing Value

A. *Missing Value*

Over the years, a great deal of attention has been paid to resolving the problems of Missing Value (MV). Referring to the Table II, however, the selection of DM methods in MV was based on the type of data set in case study. Some methods are Classification and Regression Trees (CART), Genetic Algorithm, Association Rule and k-Nearest Neighbor. As far the author concerned the DM methods has not been tested on other data sets such as vehicle maintenance data sets. It is important for Malaysia as most logistics companies have main operation activities of the land transportation involving tankers and cargo trailers are transportation of palm oil, dry cargo, palm fruit, latex and courier. These transportation vehicles are the most contributing costs of operations and maintenance. Some of the companies are using a system in their operations however unfortunately, they are not able to use the data in making a decision making. Aanalysing such enormous data using conventional technique is mind boggling task for the company.

TABLE II. PAST RESEARCH ON DM METHODS IN MV

Title	DM Methods
1. A Novel Two-Phase Method for the Classification of Incomplete Data. (2009) By: Xiuyun Qu, Bo Yuan, Wenhuang Liu	- conducting classification on incomplete data without applying deletion or imputation. a) Feature deletion b)Imputation 1. Dempster et al.: EM algorithm 2. Meng and Rubin: Generalized EM 3. Rubin: Multiple imputations 4. others: k-NN and kernel-based method. c) Learning with Missing data. Using classifiers. existing algorithms: Artificial Neural Network (ANN), C4.5 decision trees, Bayesian Networks (BN) Rough sets and Logistic regression algorithm.* the algorithm selected based on data properties. -Dataset: Income information - Based on the experiments, 4 methods used and compared: 1. Incomplete data, using C4.5 and BN 2. EM imputation **3. Tr-Methdod .Two-phase method** 4. Feature deletion -Result: Two phase method was the best
2. Attribute Correction: Data cleaning using Association rule and Clustering Methods (2011) By: Kumar, K.	AR-context-Dependent Correction means attribute values are corrected with regard not only to the reference data values it is most similar but also take consideration values of other attributes within a given record. -Dataset: Customer record a) Association Rule – Dependent corerction -Used Apriori algorithm which 2 parameters is used: a) Minsup- same name for the Apriori Algorithm used.

	b) DustTresh- minimum distance between the value of suspicious attribute and the proposed value being successor rule it violates in order to make correction. b) Clustering-Independent correction - The most-representative values may be the source of reference data. The values with low number of occurrences are noise or misspelled instance of the reference data.
3. Learning with Missing or Incomplete Data (2009) by: Bogdan Gabys (Springer –Verlag.	- highlight the benefit of using General Fuzzy Min-Max (GFMM) algorithms for clustering and classification that support incomplete datasets. -dataset: Pattern recognition
4. Fuzzy Belief Pattern Classification of Incomplete data By: Te-Shun Chou, Kang K. Yen, Liwei Aan, Niki Pissinou and Kia Makki (2007)	- suggested by combining FCM and Dempster- Shafer theory because FCM cannot directly treat the missing data. -dataset: Breast cancer from UCI database - Show improvement of classification accuracy in both experiments in both databases.

IV. Conclusion and Future Works

In the real world scenarios, domain experts are slightly important for data validation in CRISP-KDD methodology. Previous researchers have difficulty experienced in doing the existing DC process in term of long time DC process that produced an inaccurate results. Therefore, a formalize DC process that generate high data quality are critically needed specifically for the logistics company in Malaysia

In future, the researchers are planning to explore the current DM methods in other case study to make comparisons of DC process with accurate results. The other important DC processes such as duplicates and inconsistencies data also important instead of MV as future works.

Acknowledgment

We would like to express our thanks specifically to ASL company for their support of domain knowledge and data extraction and to Universiti Kuala Lumpur (UniKL) for the financial support to the first author to undertake a PhD research at Universiti Teknologi Malaysia (UTM).

References

[1] Heiko Müller, Johann-Christoph Freytag, Problems, Methods and Comprehensive Data Cleaning, Technical Report, 2003.

[2] Daniel T.Larose, Discovering Knowledge in Data:An Introduction on Data Mining.", Cresswell, 2005, pp. 27-65.

[3] Erhard Rahm, Hong Hai Do.., Data Cleaning: Problems and Current Approaches, IEEE Data(base) Engineering Bulletin - DEBU Journal, 2011, vol. 23(4), pp. 3-13.

[4] Erhard Rahm and Hong Hai Do., Data Cleaning: Problems and Current Approaches, Techn. Report, Dept. of Computer Science, Univ. of Leipzig,, 2003.

[5] Jehn-Yih Wong, Pi-Heng Chung., Managing valuable Taiwanese airline passengers using knowledge discovery in database techniques, Journal of Air Transport Management., 2007,Vol.. 13, pp.362–370.

[6] Kalaivany Natarajan, Jiuyong Li and Andy Koronios, Data Mining Techniques or Data Cleaning., In Proceedings of the 4th World Congress on Engineering Asset Management Athens,Greece, 28 – 30, September.

[7] R.Kavitha Kumar and et.el.., Attribute Correction-Data Cleaning using Association Rule and Cluestering Method, International Journal of Data Mining & Knowledge Management Process (IJDKP), March 2011,vol. 1(2), pp. 22-32.

[8] Sang Jun Lee and et. el., A Review of Data Mining Techniques", Industrial Management & Data Systems, 2001, vol.(1).., pp.41-46.

[9] Jehn-Yih Wong, Pi-Heng Chung., Managing valuable Taiwanese airline passengers using knowledge discovery in database techniques, Journal of Air Transport Management., 2007,Vol.. 13, pp.362–370.

[10] Hasimah Hj Mohamed and et. el., E-Clean: A Data Cleaning Framework for Patient Data, First International Conference on Informatics and Computational Intelligence., 2011.

[11] Müller, Heiko,Freytag, J.C.., Problems, Methods, and Challenges in Comprehensive Data Cleansing., Technical Report. Humboldt University Berlin, 2003.

[12] Cao Longbing and Zhang Chengqi, The Evolution of KDD: Towards Domain-Driven Data Mining,, International Journal of Pattern Recognition and Artificial Intelligence , 2007,vol. 21(4), pp.677-692.

About Author (s):

Fauziah Abdul Rahman is a lecturer at Universiti Kuala Lumpur in Johor, Malaysia. She is currently pursuing her PhD in Computer Science at a local public university in Malaysia.