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

Fauziah Abdul Rahman, Rahimah Kassim , Zirawani Baharum , Helmi Adly Mohd Noor, Norhaidah Abu Haris

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 (DC). DC is a process used to determine inaccurate, incomplete or unreasonable data and then improve the quality through correcting of detected errors and omissions. Various process of DC have been discussed in the previous studies, but there is no standard or formalized the DC process. The Domain Driven Data Mining (DDDM) is one of the KDD methodology often used for this purpose. This paper review and emphasize the important of DC in data preparation. The future works was also being highlight.***

Index Terms: Data cleaning; Data Mining; 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. In other words, it deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. 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 in data preparation are identified as the most important phase in Domain driven Data Mining (DDDM) 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.

Revised Manuscript Received on September 22, 2019.

Fauziah Abdul Rahman, Universiti Kuala Lumpur-Malaysian Institute of Information Technology, 50250 Kuala Lumpur, Malaysia *fauziahar@unikl.edu.my, pojiah78@gmail.com
Rahimah Kassim, Universiti Kuala Lumpur-Malaysian Institute of Information Technology, 50250 Kuala Lumpur, Malaysia.
Zirawani Baharum, Universiti Kuala Lumpur-Malaysian Institute of Information Technology, 50250 Kuala Lumpur, Malaysia
Helmi Adly Mohd Noor, Universiti Kuala Lumpur-Malaysian Institute of Information Technology, 50250 Kuala Lumpur, Malaysia
Norhaidah Abu Haris, Universiti Kuala Lumpur-Malaysian Institute of Information Technology, 50250 Kuala Lumpur, Malaysia

II. DATA CLEANING

A. Data Cleaning issues

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. 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 due to the time required for an expert to cleanse it manually. 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. 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. 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.

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.

DDDM was the second 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 DDDM is used for discovering interesting information in data validate by the domain knowledge and 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 Fig. 1. 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 Fig. 1, the second 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.

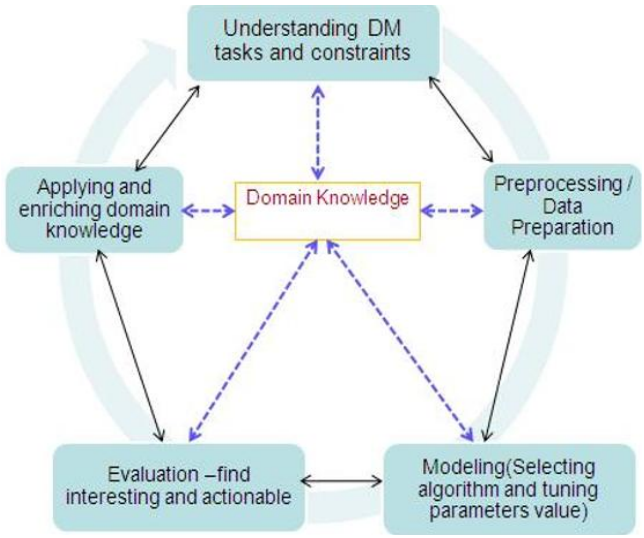

Fig 1: Domain driven data mining methodology (DDDM)

TABLE 1: PAST RESEARCH ON DC PROCESS

References	DC Process
Data preprocessing for prediction of rerecirculating water chemistry faults (Qiang et al., 2010)	Steps in DC: Step 1: Data classification classify the data in case study based on categories Step 2: 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
Discovering knowledge in data (Larose, 2014)	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
E-clean: A data cleaning framework for patient data (Mohamed et al., 2011)	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 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 2 categories: single-source problems or multiple-source problems. The examples of single-source problems are the redundancy and duplicate. The multiple-source problems are referring to the contradicting or overlapping and inconsistence data. DC framework proposed is ETL process model: 1) E: Extract stage 2) Transform stage 3) Load stage
Attribute correction-data cleaning using association rule and clustering method (Kumar and Chadrasekaran, 2011)	DC framework proposed: 1) Raw data 2) Pre-processing 3) Selection of attributes 4) New data 5) Selection of DM techniques (association rule and clustering) 5) Clean data

III. APPLICATION ON DATA MINING (DM) ON MISSING VALUE (MV)

Over the years, a great deal of attention has been paid to resolving the problems of Missing Value (MV). Referring to the Table 2, 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. Analyzing such enormous data using conventional technique is mind boggling task for the company.

TABLE 2: PAST RESEARCH ON DM METHODS IN MV

References	DM Methods
A novel two-phase method for the classification of incomplete data (Qu et al., 2009)	Conducting classification on incomplete data without applying deletion or imputation. a) Feature deletion b) Imputation 1. EM algorithm 2. Generalized EM 3. 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.T-R, Two-phase method 4. Feature deletion Result: Two phase method was the best
Attribute correction -data cleaning using association rule and clustering method (Kumar and Chadrasekaran, 2011)	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 correction 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.
Learning with missing or incomplete data (Gabrys, 2009)	i) Highlight the benefit of using General Fuzzy Min-Max (GFMM) algorithms for clustering and classification that support incomplete datasets. ii) dataset: Pattern recognition
Fuzzy belief pattern classification of incomplete data (Chou et al., 2007)	i) Suggested by combining FCM and Dempster-Shafer theory because FCM cannot directly treat the missing data. ii) dataset: Breast cancer from UCI database iii) Show improvement of classification accuracy in both experiments in both databases.

IV. CONCLUSION

In the real world scenarios, domain experts are slightly important for data validation in DDDM methodology. However, previously 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.

ACKNOWLEDGEMENT

The author is grateful to Universiti Kuala Lumpur-Malaysian Institute of Industrial Technology (UniKL-MITEC).

REFERENCES (APA FOR SOCIAL SCIENCES)

1. Cao L and Zhang C (2018). The evolution of KDD: Towards domain-driven data mining. International Journal of Pattern Recognition and Artificial Intelligence, 21(4): 677-692. https://wwwstaff.it.uts.edu.au/~lbcao/publication/dmba-dddm.pdf
2. Chou TS, Yen KK, An L, Pissinou N and Makki K (2015). Fuzzy belief pattern classification of incomplete data. Proceedings IEEE International Conference on Systems, Man and Cybernetics, pp. 535-540.
3. Erhard R and Hong HD (2000). Data cleaning: Problems and current approaches IEEE Data Engineering Bulletin, 23(4): 3-13. https://pdfs.semanticscholar.org/6046/770d1c3e08edfdd39bdb57fccaca84f5139c.pdf
4. Gabrys B (2009). Learning with missing or incomplete data. Proceedings International Conference on Image Analysis and Processing, pp. 1-4.
5. Jehn YW and Pi HC (2007). Managing valuable Taiwanese airline passengers using knowledge discovery in database techniques. Journal of Air Transport Management, 13(6): 362-370, https://www.researchgate.net/publication/222413068_Managing_valuable_Taiwanese_airline_passengers_using_knowledge_discovery_in_database_techniques
6. Kalaivany N, Jiuyong L and Andy K (2010). Data mining techniques for data cleaning. In Engineering Asset Lifecycle Management, D. Kiritsis, C. Emmanouilidis, A. Koronios and J. Mathew, eds., Springer, London, England, pp. 796-804. http://link.springer.com/chapter/10.1007/978-0-85729-320-6_91
7. Kumar R and Chadrasekaran RM (2017). Attribute correction-data cleaning using association rule and clustering method. International Journal of Data Mining and Knowledge Management Process, 1(2): 22-32.
8. Larose DT (2014). Discovering knowledge in data: an introduction to data mining. John Wiley and Sons, Hoboken, NJ.

9. Lily S and Cleopa JM (2010). Case-based analysis in user requirements modeling for knowledge construction. Journal of Information and Software Technology, 52(7): 770-777. http://centaur.reading.ac.uk/5737/
10. Mohamed HH, Kheng TL, Collin C and Lee OS. (2011). E-clean: A data cleaning framework for patient data. Proceedings 1st IEEE International Conference on Informatics and Computational Intelligence, pp. 63-68.
11. Müller H and Freytag JC (2005). Problems, methods and comprehensive data cleaning. Technical report, Humboldt University, Berlin, Germany. http://www.dbis.informatik.hu-berlin.de/fileadmin/research/papers/techreports/2003-hub_ib_164-mueller.pdf
12. Qiang G, Xinmin W, Chao D and Chenguang L (2010). Data preprocessing for prediction of rerecirculating water chemistry faults. Proceedings IEEE International Conference on Computer Application and System Modeling, pp. 553-556.
13. Qu X, Yuan B and Liu W (2009). A Novel Two-Phase Method for the Classification of Incomplete Data. Proceedings IEEE International Conference on Information Management, Innovation Management and Industrial Engineering, pp. 452-455).
14. Sang JL and and Siau K (2011). A review of data mining techniques. Industrial Management and Data Systems, 101(1): 41-46. http://dx.doi.org/10.1108/02635570110365989
15. P. Dempster, N. M. Larid, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, vol. 39, pp. 1-38, 1977.
16. X. Meng, D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm: A general framework," Biometrika, vol.80, pp. 267-278, 1993.
17. D. B. Rubin, Multiple Imputation for Non-Response in Surveys. New York: John Wiley and Sons, 1987

AUTHORS PROFILE

Fauziah Abdul Rahman, Department of Technical Foundation, Universiti Kuala Lumpur-Malaysian Institute of Industrial Technology, Persiaran Sinaran Ilmu, Bandar Seri Alam, 81750 Johor Bahru, Johor, Malaysia. *fauziahar@unikl.edu.my, pojiah78@gmail.com

Rahimah Kassim, Department of Technical Foundation, Universiti Kuala Lumpur-Malaysian Institute of Industrial Technology, Persiaran Sinaran Ilmu, Bandar Seri Alam, 81750 Johor Bahru, Johor, Malaysia. And Department of Industrial Logistics, Universiti Kuala Lumpur-Malaysian Institute of Industrial Technology, Persiaran Sinaran Ilmu, Bandar Seri Alam, 81750 Johor Bahru, Johor, Malaysia

Helmi Adly Mohd Noor, Department of Technical Foundation, Universiti Kuala Lumpur-Malaysian Institute of Industrial Technology, Persiaran Sinaran Ilmu, Bandar Seri Alam, 81750 Johor Bahru, Johor, Malaysia. And Department of Industrial Logistics, Universiti Kuala Lumpur-Malaysian Institute of Industrial Technology, Persiaran Sinaran Ilmu, Bandar Seri Alam, 81750 Johor Bahru, Johor, Malaysia.

Norhaidah Abu Haris, Department of Software Engineering, Universiti Kuala Lumpur-Malaysian Institute of Information Technology, 50250 Kuala Lumpur, Malaysia