International Advance Journal of Engineering, Science and Management (IAJESM)
ISSN -2393-8048, January-June 2018, Submitted in May 2018, iajesm2014@gmail.com

# Review of Literature on Principle and Methods of Data Cleaning for Removing Erroneous Data from Database

Sri Sharanabasappa Raikoti, Assistant Professor, Department of Computer Science, Government Degree College Yadgir, Karnataka, India, Email: sr.raikoti@gmail.com

## Abstract

This paper includes the detail study of many techniques that have been found useful and important during the time of data cleaning process as well as consolidating the data quality in databases. For Instance, existing classification of dirty data types from the literature will be reviewed to present the multiple dirty data types observed in different data sources. Data cleaning methods, Data quality, data quality dimensions are reviewed in this chapter. They provide the foundation of development of the proposed data cleaning framework.

**Keywords: Review of Literature, Principle and Methods, Data Cleaning, Removing Erroneous Data from Database**

## INTRODUCTION

Data cleansing and transformation are important tasks in many of the domains of IT industry and as well as database management system, information collection systems, data warehousing etc. The existing data cleaning system is time consuming and traumatic longterm running non interactive operations, poor coupling between analysis and transformation, and complex transformation interfaces that often require user programming. There are so many interactive system for data cleaning and transformation such as Potter's Wheel architecture which cleans the data by removing the difference and inconsistency i.e. discrepancies as it encounters in database. One can check the result of transformation by graphical operations or through examples. Potter's Wheel architecture focuses on flat file data, tabular data, and nested data formats like XML. Incorrect and inconsistent data can gives rise to pseudo conclusion on misdirected investments on both public and private scales. Suppose the government may want to analyze population census figure to decide which region require further spending on infrastructure and service. In this scenario it is important exact and accurate data to avoid erroneous and fiscal decision. In the business world, incorrect data can be costly. Many companies use customer information database that record data like address, cell number, official address, official phone number and many others in case of incorrect information, the company suffer the various losses including money loss as well as customer loss. Data cleansing differs from data validation such as validation means rejecting the data at the time data entry rather than batches of data where as data cleansing refers to identifying incomplete, incorrect, inaccurate, irrelevant data from the database and modifying, removing or deleting these data from database. Actual process of data cleansing involves removing typographical errors or validating or correcting values against the known list of entries. The validation may be strict such as rejecting address without pin code, avoid fuzzy records that have partially match existing known records.

In most of the enterprises it is observed that they do not pay adequate attention to the occurrences of dirty data and have not applied useful methodologies to ensure high quality data for their application. Lack of awareness about the dirty data and its impacts on quality of data is one of the major reasons behind the dirty data occurrences and its negative impact on the quality of data. Therefore, in order to improve the data quality, it is necessary to understand wide variety of dirty data that may exist within the data source as well as how to deal with them.

### Review of Literature:

According to (Mueller, Freytag) Anomalies and impurities in data causes adverse effect in effective data utilization, degrades the performance and productivity. Some of the problems caused due to data anomalies are as follows - For Ex: - In retail marketing database an incorrect database of prices may cause billions of overcharges per annum to the consumer. Principle of data quality (Chapman 2005) stressed over preventing the errors rather than detecting and cleaning, as it is efficient and affordable to prevent the error then locate it and correct it later. On contrary of implanting lot of data constraints (i.e. primary key, foreign key etc.) error will still occur and therefore data validation and data cleansing cannot be ignored.

Error detection, validation and cleaning plays vital role especially with legacy data such as museum and herbarium data collected over 300 years ago). Hence error prevention and data cleansing both is essential component in organizations data management policy. While huge body of research deals with schema translation and schema integration, data cleaning have received only little attention in the research community. A number of authors focused on the problem of data cleaning and they suggest different algorithms to clean dirty data. According to Wejje Wei, Mingwei Zang (2008) In this paper, a data cleaning method is based on association rules is proposed. The new method adjust the basic business rules provided by the experts with association rules mined from the multi data sources and generates the advanced business rules for every data source. Using these method, time is saved and accuracy of data cleaning is improved. As per the Applied Brain and Vision Science- data cleaning Algorithm (2012) paper by Rajshree Y. Patil, The algorithm is designed for cleaning the EEG resulting from the brain function of stationary behavior such as an eyes-open or eyes closed data collection paradigm, This algorithm assumes that useful information in contained in the EEG data are stationary. That is , it assumes that there is very little change in the on-going statistic of the signals of interest contained in the EEG data. Hence, this algorithm is optimal for removing momentary artifacts in EEG collected while a participant is in eyes-opened or eyes-closed state. According to Helena Galhardas, Daniela Florescu, Dennis Sasha(2012) This paper presents the language. An execution model and algorithms that enables users to express data cleanings specifications declaratively and perform cleaning efficiently. They use as an example a set of bibliographic references used to construct the Citesser Web site. They propose the model to clean textual records so that meaningful queries can be performed. According to Arup Kumar Bhattacharjee, Atanu Mallick, Arnab Dey, Sananda Bandyopadhya Data Cleaning in Text Files (2013), In this paper, we mentioned about preventing as well as detecting & cleaning errors in primary biological collection of database, Data cleaning in Text files using ETL (Extract Transform Load) model along with the set of algorithms to correct errors such as alphanumeric errors, invalid gender, invalid ID pattern and Redundant ID errors, Hackers will try to submit harmful data to web server in order to crack server security and exploit it to vulnerability Data Validation and Data cleansing are two main methods to two defend server attacks against the bad data such as Code Injection, SQL Injections, Buffer Overflows.

According to Lin Li Data Cleaning in Database Application (2013), In these paper they suggested five different methods of data cleaning as well as main activities addressed by these data cleaning approaches, special features associated with the five approaches are detailed in above table According to Kazi Shah Nawaz Ripon, Ashiqur Rahman, G.M. Atiqur Rahman (2013). In this paper, they propose novel domain independent techniques for better reconciling the similar duplicate records. They also introduced new ideas for making similarduplicate detection algorithm faster and more efficient. According to Payal Pahwa, Rajiv Arora, Garima Thakur, In this paper, they address issues related to detection and correction of duplicate records. Also it analyses data quality and various factors that degrade it. A brief analysis of existing techniques is discussed, pointing out its major limitations. And a new framework is proposed that is an improvement over existing system. According to R.Kavitakumar, Dr. RM. Chandrasekaran, In this paper they designed two algorithms using data mining techniques to correct the attribute without external reference. One is Context-dependent attribute correction and other is Contextindependent attribute correction.

According to Shubi Anand, Rinkle Rani, World Wide Web is a monolithic repository of web pages that provides internet users with heaps of information. With growth in number and complexity of websites, the size of web has become massively large. This paper emphasizes on Web Usage Mining process and makes an exploration in the field of data cleaning. According to Mohammad, H.H., They developed a system uses the extract, transform and load model as the system main process model to serve as guidelines for the implementation of the system. Besides that, parsing techniques is also used for identification of dirty data. Here they selected K-nearest Neighbor algorithm for the data cleaning. According to Chris Mayfield, Jennifer Neville, Sunil Prabhakar, They presented ERACER, an iterative statistical

framework for inferring missing information and correcting such errors automatically. Their approach is based on belief propagation and relational dependency networks, and includes an efficient approximate inference algorithm that is easily implemented in standard DBMS using SQL and user defining functions.

Big data is often described by five characteristics, namely volume, velocity, variety, veracity, and value Han et al. (2016), which differentiate it from traditional data. Volume refers to the size, scale, quantity, and magnitude of the data that have been generated, the size of which can exceed hundreds of terabytes. Velocity refers to the arrival speed of data, which results in the accumulation of enormous datasets within very short periods. Variety refers to the different formats of data, namely structured data, semi-structured data, and unstructured data. Structured data have a standardized format and a well-defined structure and generally reside in relational databases to represent the relationship between entities such as tables. Semi-structured data are not as rigid as structured data; however, they have several elements that are similar, being organized hierarchically, although it cannot be verified whether the tabular structure is associated with relational databases. Unstructured data, which are growing at a faster rate than structured and semi-structured data, have no easily identifiable structure and they cannot be stored in any logical form Erl et al. (2016). Veracity refers to the validity or quality of data. Value refers to the usefulness of data for decision making. Big data's characteristics increase the difficulty of the cleaning process compared with that for normal data due to the heterogeneous data, the volume, and speed of arrival data, which reveal a limitation of using the normal data cleaning method. However, the normal data cleaning methods can be considered as a baseline for developing cleaning methods able to cope with big data's characteristics.

**Reference:**

Christidis, L. & Boles, W.E. 1994. Taxonomy and Species of Birds of Australia and its Territories. Royal Australasian Ornithologists Union, Melbourne. 112 pp.

Clarke, K.C. 2002. Getting Started with Geographic Information Systems, 4th edn. Upper Saddle River, NJ, USA: Prentice Hall. 352 pp.

Flower dew, R., 1991. Spatial Data Integration. pp. 375-387 in: Maguire D.J., Good child M.F. and Rhind D.W. (eds) Geographical Information Systems Vol. 1, Principals: Longman Scientific and Technical.

Blakers, M., Davies, S.J.J.F. and Reilly, P.N. 1984. The Atlas of Australian Birds. Melbourne: Melbourne University Press.

Erl, T.; Khattak, W.; Buhler, P. Big Data Fundamentals: Concepts, Drivers & Techniques; Prentice Hall Press: Upper Saddle River, UJ, USA, 2016.

Kolajo, T.; Daramola, O.; Adebiyi, A. Big data stream analysis: A systematic literature review. J. Big Data 2017, 6, 47.

Han, J.; Pei, J.; Tong, H. Data Mining: Concepts and Techniques; Morgan kaufmann: Burlington, MA, USA, 2016.

Ridzuan, F.; Zainon, W.M.N.W. A review on data cleansing methods for big data. Procedia Comput. Sci. 2017, 161, 731–738. [CrossRef]