International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 12 | Dec 2018 www.irjet.net p-ISSN: 2395-0072

A Review of Data Cleaning and its Current Approaches

Madhushree N

MTech Student, Department of Computer Science and Engineering, New Horizon College of Engineering, Bangalore, Karnataka, India

---

Abstract - *Nowadays data is being used everywhere, each and every field from research centers to academics, management and administration use large amount of data. Data is very important and valuable resource. The genuine use of high quality of data could help pupils who can make proper analysis, prediction and decisions. How much ever effort we try to put in collecting a good finite data set errors are prone and are penetrating into the data set. The major challenge for us is dirtiness in a data. Thus data cleaning is an important and iterative step as a preprocessing technique in data mining. Data cleaning is the process of detecting or removing a corrupted, inaccurate, inconsistent data from available raw data. Data cleaning may be complicated, time consuming and expensive but it is important and necessary step for mining. This paper makes an attempt to explain the various method and related works in data cleaning.*

***Key Words*: Data cleaning, Data quality, Noise removal, Outlier, Mining.**

1. INTRODUCTION

Analysis and detecting the dirty data is one of the challenges. If the detection leads to a failure it can result in accurate values and improper decision. Over the past few years many techniques have emerged in cleaning the raw data. The anomalies in common data include inconsistency where data among the resources available in the data with different meaning, data entry errors such as spelling mistakes inconsistent in the format of data, incomplete and incorrect attribute values. Data usually which is incomplete, inaccurate is known as dirty data.

Many technique have emerged efficiently in recent years which include fundamental services which can be used in data cleaning like token formation, attribute selection, using of any clustering algorithm, usage of similarity functions, eliminate and merge function etc. has been used.

The main objection or intention of data cleaning is to reduce the time and complexity of mining process and simultaneously increase the quality of data in a data warehouse. In existing world there are wide varieties of fields, where various types of data or similar kind of data are collected from different resources to single place. Then there also exists lot of mishaps that can be due to various improper measures in terms of missing data, noise, inconsistencies. The sample of data which doesn't contain the complete set of information about the data. This is why a heterogeneous data set which is collected from different sources is very difficult to analyze. If data available is with less quality then mining techniques like analysis of data, pattern recognition and decisions are not optimally possible. Solutions which are through wrong results that may lead to wrong decision and conclusions. It is essential and important to obtain quality data and correct sample of data in prior applying in mining techniques to get required useful information. Uncleansed data may leads to wrong interpretation results.

1.1 APROACHES IN DATA CLEANING

These days data is being used everywhere, each and every field from research centers to academics, management and administration use large amount of data. Data is very important and valuable resource. The genuine use of high quality of data could help pupils who can make proper analysis, prediction and decisions. How much ever effort we try to put in collecting a good finite data set errors are prone and are penetrating into the data set. The above represented is the general approach in data mining the following represented below give a brief note on data cleaning process.

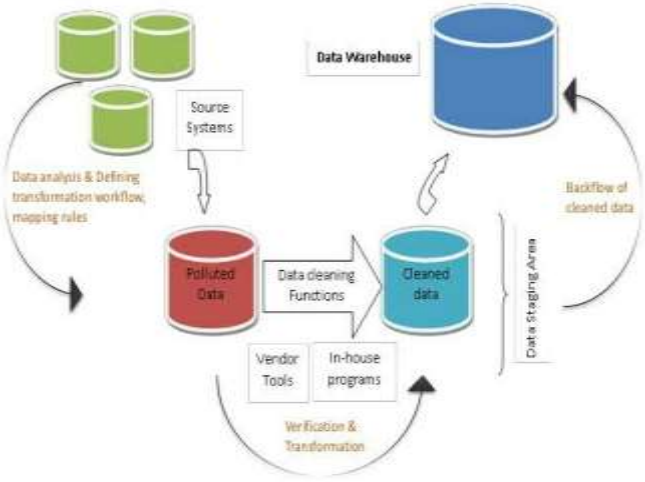

Fig -1: **Data cleaning process**

In general data cleaning involves several phases:

Data analysis: From the initial set of raw data one has to detect inconsistent and erroneous data are to be removed. Thus analysis of data is required. Analysis of data plays an important role and is used in gaining metadata about the properties of data and can detect the data quality problems.

Definition of transformation workflow and mapping rules: From the number of data sources, a huge number of data transformation and data cleaning steps has to be executed. Sometime schema translation is used to map sources to the obtained data model. The schema related data transformation model as well as the data cleaning steps should be specified by declarative query and a mapping language.

Verification: The effectiveness and true information of transformation definition and workflow should be tested and evaluated.

Transformation: In transformation steps the process of execution is either by running the ETL (Extract Transformation and Loading) for loading a data warehouse or at the time of answering the queries on the multiple sources present.

Back flow of cleaned data: After the removal of errors from the raw data, the cleansed data should replace the original set of dirty raw data in order to provide the improved quality data and also avoid the redoing the cleaning for the future extraction of data.

2. LITERATURE SURVEY

Current method of data cleaning: Nowadays many methods have been emerged in data in data cleaning, many approached methods refers to ETL (Extraction Transformation and Loading) tools. A variety of data cleaning methods have been proposed which also include outlier, noise removal from the raw data.

Constraint based data repairing: This method mainly include two important steps, firstly it identifies well defined constraints and set that the data has to follow these constraints. Next using one existent constraint in data find the another constraint from the data base.

Statistical method for data repairing: This method mainly focus or relies on data imputation, de duplication and error detection. Data imputation infer mainly the missing values.

Hong Liu, Ashwin Kumar T K, Xiaofei Hou proposed an interactive approach in data cleaning, which improves the data quality. For a big data there is no need to use all data only a sample of small subset of useful data is required. The goal of association is to manipulate the original raw data into relevant data, after obtaining there is also required to analyze the data. This paper provides a uniform frame work that unifies the association, which can make use of data cleaning and also provide enriched data information. Many algorithms like Markov's, iterative inductance cutting algorithms, metadata generation algorithm are used which improves the data quality.

Xiuyuan Liu, Aizhang Guo, TaoSun proposed an data cleaning technology based on handoop distribution platform and outlier technology for cleaning data attribute. It is mainly applied to mass data which mainly improves speed and accuracy of cleaning. Determining of inconsistencies and repeatability of values are clearly analyzed and solutions are proposed in this method. Algorithms such as outlier algorithm which determines outlier and inconsistent data from the data set. Density based outlier algorithms usually used for global outliers and also distributed outlier mining algorithms are used in the proposed paper.

Dr. Deepali Virmani, Preeti Arrova, Ekta Sethi proposed an important approach in data cleaning which uses the mechanism variegated data swabbing an efficient algorithm which calibre of raw data set by removing the incorrect, inconsistencies, duplicate, inaccurate values from the provided structured data . Spell checker algorithm is applied to this proposed work in order to check the mistakes from the data set as well as misspellings from the raw data. The suggested method also provide an efficient and good result while comparing in terms of space, accuracy of data and execution.

He Liu Jing Qi Chen, Fupeng Huang proposed a data cleaning solution for electric power sensor data which is combination of data cleaning and cleaning frame work. The proposed method detects the outliers on the basis of K means clustering algorithm. Effectiveness of a data can be evaluated on a data set. This method improves the data quality, strong, scalability and provides a high performance. This method which provide solution to outlier detection. Clustering algorithm is proposed in order to deal with outliers.

Joeri Rammeraere, Floris Geerts Bart Goethals proposed a dynamic cleanliness for the data quality which is static. They have introduced a forbidden item set has enhanced properties of the life measure and given an algorithm to mine low-lift the forbidden item set. The algorithm which is provided is much efficient and can capture the consistencies with high precision of dirtiness in data. This method also developed as an efficient repair algorithm that after the repair in a data set no new consistencies or inaccurate values can be found. Experimental results demonstrate that there is a high quality in repairs.

3. CONCLUSIONS

Data cleaning as specified is a mandatory part in data mining. The cleaning approaches depends upon the type of data which one has to clean and based on that necessary tool or method are to be applied. Each tool has its own desired features and function depending upon the data we can use the necessary tools and clean the data.

Noise, outliers, inconsistencies data, duplicate data, missing values can be addressed by data cleaning. In this survey a

brief analysis of existing techniques of data cleaning is mainly focused in this paper. Different kind of data cleaning techniques are presented in this survey paper. Each method described can be used to identify type of errors present in a data set. The technique mainly depends on the type of data available.

Future research may include the wider review and fine investigation of various methods in data cleaning.

REFERENCES

[1] Hong Liu, Ashwin Kumar T K, Xiaofei Hou 'A cleaning frame work for big data an interactive method approach for data cleaning' 2016 IEEE second international conference on big data computing services and application.

[2] Joeri Rammeraere , Floris Geerts Bart Goethals 'cleaning data with forbidden item sets' 2017 IEEE 33rd international conference on data engineering

[3] Xiuyuan Liu, Aizhang Guo,TaoSun 'Application of handoop based distribution data cleaning technology in periodical meta data integration' 2017 10th international symposium on computational intelligence and design

[4] Dr. Deepali Virmani, Preeti Arrova, Ekta Sethi 'Variegated data swabbing an improved purge approach for data cleaning'2017 IEEE

[5] He Liu Jing Qi Chen , Fupeng Huang 'An electric power sensor data oriented data cleaning solutions' 2017 14th international symposium on pervasive systems, algorithms and networks and 2017 11th international conference on frontier of computer science and technology and 2017 third international symposium of creative computing

[6] Rajashree Y. Patil 'A review of data cleaning algorithms for data warehouse systems' (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (5) , 2012,5212 - 5214