The Procedure For Employing Data Mining Techniques To Clear Up The Contaminated Data

Navneet Singh Chauhan[1], Dr. Ajay Singh[2], KapilKumar[3]
[1, 2] Dept of CSE
[2]Associate Professor & HoD
[3]Assistant Professor, Dept of Computer Application
[1, 2] Bhagwant Institute of Technology, Muzaffarnagar
[3]Shri Ram College, Muzaffarnagar

Abstract- *The primary concern in excellent information management is data quality. Issues of Data Quality management might arise in any part of an information system. For businesses, it has long been unclear whether they should use data analytics to determine whether their data is unclean or if they should clean up their data before using it. Data cleaning provides a solution to these issues. It is the procedure used to identify data that is erroneous, lacking, or irrational, and then enhance the quality by fixing any mistakes and omissions that are found. In general, data cleaning lowers errors and raises the calibre of the data. Although it can be a time-consuming and laborious procedure, data inaccuracies must be corrected and erroneous information must be removed. One important method for cleansing data is data mining. One method for finding valuable information in data is data mining. A new method called "data quality mining" uses data mining techniques to find and fix issues with data quality in big datasets. From data collections, data mining automatically extracts intrinsic and hidden information.*

I. INTRODUCTION

Businesses struggle to find good or clean data. For businesses, the question of whether to use data analytics to determine whether their data is dirty or to clean up their data first remains unanswered. Which came first, the chicken or the egg? is the question that remains. There isn't a definitive response. Without high-quality data input, businesses risk analysis paralysis, and they can never have clean data without analytics to assist them find data issues.

In the information industry, a vast amount of data is available, and this number is growing daily. Before statisticians used terms like data fishing and data dredging, the term "data mining" was first used in 1990. Finding useful information in huge data sets, or "Big Data," is the main goal of data mining. Since data mining is the process of extracting knowledge from data, we may also define it this way. These days, data mining is utilised in a wide range of contemporary applications, including fraud detection, corporate analysis and risk management, market analysis and management, sports, astrology, science exploration, and Internet web surfing assistance. Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large amounts of data to find patterns for big data. In this paper we provide an overview of Data Quality Problems, Dirty Data and different methods to clean.

Businesses have been gathering a lot of data from many sources to create their own "Data Lakes," hoping to improve their data bank, which is now their most important asset. The majority of the time, errors in data are introduced during the data gathering and collecting process. Examples of these errors include typos, missing values, redundant data, inconsistent entries for the same real-world entity, outliers, and business rule violations. According to a Kaggle 2017 assessment on the state of data science and machine learning, the most frequent obstacle faced by data workers is the issue of unclean data (Kaggle 2017).

Not surprisingly, developing effective and efficient data cleaning solutions is a challenging venue and is rich with deep theoretical and engineering problems. There are number of surveys and published books on different aspects of data quality and data cleaning. Rahm and Do (Rahm and Do 2000) give a classification of different types of errors that can happen in an Extract-Transform-Load (ETL) process, and survey the tools available for cleaning data in an ETL process; Bertossi (Bertossi 2011) provides complexity results for repairing inconsistent data, and performing consistent query answering on inconsistent data; Hellerstein (Hellerstein 2008) focuses on cleaning quantitative data, such as integers and floating points, using mainly statistical outlier detection techniques; Fan and Geerts (Fan and Geerts 2012) discuss the use of data quality rules in data consistency, data currency, and data completeness, how different aspects of data quality issues might interact; Dasu and Johnson (Dasu and Johnson 2003) summarize how techniques in exploratory data mining can be integrated with data quality management; Ilyas and Chu (Ilyas et al 2015) provide taxonomies and example

algorithms of qualitative error detection and repairing techniques; and Ganti and Das Sarma (Ganti and Sarma 2013) focus on an operator-centric approach for developing a data cleaning solution, which involves the development of customizable operators that could be used as building blocks for developing common solutions.

II. PROBLEMS OF DIRTY DATA

It might sound a bit abrupt, but clean data is a myth. If data is dirty, so is everyone else's. Enterprises or individuals are more than dependent on data these days, and it is not going to change in coming years. They need to collect data in order to analyze it, which necessarily will not be 100% clean, pristine, or perfect in nature. Nearly all companies face the challenge of dirty data in the form of a lot of duplicates, incorrect fields, and missing values. This happens due to omnichannel data influx, followed by hundreds, if not thousands, of employees wrestling and torturing that data to derive professional outcomes and insights. Don't forget that even the best of the data has that tendency to decay in few weeks. Because the data is time relevant. The Figure (1) shows the fundamental reasons of Program Data Consolidation Inefficiency. (1) conformance, uniformity, density and uniqueness. Data quality problems are raised in industry and academic areas. Gathering data from heterogeneous, analyzing it for further usage are important and remarkable aspect of the data quality problems. i.e. Data Integration Single source and multiple sources are classifications of data quality problems. Single source problems occur in a single database whereas multiple source problems occur whenever data integrate from two or more sources. e.g. Overlapped data and differences in entity names.

Outlier Detection:

Figure 1

Nowadays, not only the public or private organizations but every one of us understands the value of Data. Data is the key to improve business, productivity, decision making and to support top-level decisions as well as strategy management for the organizations. However, as organizations

Outlier detection as the name suggests refers to the act of finding values lying outside the range or domain probably and is an quantitative error detection task. While definition of an outlier depends on the application, there are some commonly used definitions, such as "an outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins 1980). E.g., for a company whose employees' salaries are mostly around 100K, an employee with a salary of 10K can be considered to be an outlying record. Multiple surveys have been done and number of articles have been published to summarize different definitions of outliers, and algorithms for detecting them (Hodge and Austin 2004; Chawla and Sun 2006; Aggarwal 2013). In general, if we see that the outlier detection techniques can be categorized into three broad categories: statistics-based, distance-based, and model-based. Statistics-based outlier detection techniques assume that the normal data points would appear in high probability regions of a stochastic model, while outliers would occur in the low probability regions of a stochastic model (Grubbs 1969). They can often provide a statistical interpretation for discovered outliers, or a score/confidence interval for a data point being an outlier, rather than making a binary decision. Distance-based outlier detection techniques often define a distance between data points, which is used for defining a normal behavior, for example, normal data point should be close to a lot of other data points, and data points that deviate from such normal behavior are declared outliers (Knorr and Ng 1998). A major advantage of distance-based techniques is that they are unsupervised in nature and do not make any assumptions regarding the generative distribution for the data. Instead, they are purely data driven. Model-based outlier detection techniques first learn a classifier model from a set of labeled data points, and then apply the trained classifier to a test data point to determine whether it is an outlier (De Stefano et al 2000). Model-based approaches assume that a classifier can be trained to distinguish between the normal data points and the anomalous data points using the given feature space. They label data points as outliers if none of the learned models classify them as normal points.

III. DATA QUALITY PROBLEMS

Data quality is a root issue in many areas mainly in the pattern discovery. As it is very much clear that Data Quality problem may raise wrong output based on analysis of Dirty Data. If data quality satisfies a quality criteria and the data is treated as high quality data. Data quality criteria are accuracy, integrity, completeness, validity, consistency, schema begin to create integrated data warehouses for decision support, the resulting Data Quality (DQ) problems

become painfully clear. A study by the Meta Group revealed that 41% of the Data Warehouse projects fail, mainly due to insufficient DQ, leading to wrong decisions. The quality of the input data greatly influences the quality of the results.

The concept of DQ is vast not limited to a specific thing. It has different definitions and interpretations. It is essentially analyzed and discussed by two communities: Database and Management. The First one studies DQ from technical point of view, while the second one is also concerned with other aspects or dimensions (e.g. accessibility, believability, relevancy, interpretability, objectivity) involved in DQ.

The following Table 1 illustrates the basic DQ problems.

Table 1 – E.g. Basic data quality problems

DirtyData	DQ Problem
Cus1 = (name="Santosh Singh"...) Cus2=(name="S.Singh"…)	DuplicatedRecords
Country="SOUTHINDIA"...	MisfieldedValues
Age=00	MissingValues
PIN=825413 City=Hazaribag	Violated AttributeDependencies
Gender=Q	IllegalValues
Name="Santosh01-01-2020"	Multiple Values insinglecolumn

Figure 2 presents the well known typical model of data organization:(i) data is stored in multiple data sources; (ii) a data source is composed of several relations and relationships are established among them; (iii) a single relation is made up of several tuples; and (iv) a tuple is composed by a predefined number of attributes. This model results in a hierarchy of four levels of granularity: multiple data sources; multiple relations; single relation; and attribute/tuple.

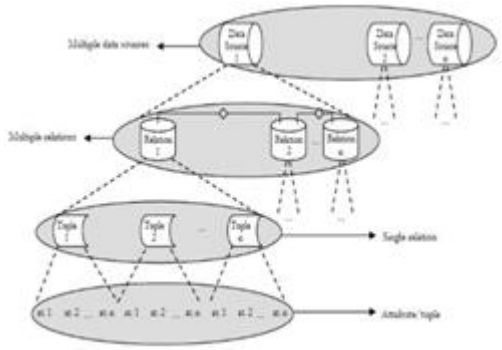


Figure 2: Typical model of Data Organization

With the advent of data socialization and data democratization, many organizations are organizing, sharing and making available the information in an efficient manner to all the employees. While most organizations are profiting by the liberal usage of such mine of information at their employees' fingertips, others are facing problems with the quality of data being used by them.

As most organizations also look at implementing systems with artificial intelligence or connecting their business via internet of things, this becomes especially important.

Business analysts determine market trends, performance data, and even present insights to executives that will help direct the future of the company. And as the world becomes even more data-driven, it is vitally important for business and data analysts to have the right data, in the right form, at the right time so they can turn it into insight.

The basic model that a company follows when implementing data socialization is:

However, many times, business analysts end up spending the majority of their time focused on data quality. This is a problem because data preparation and management isn't the business analyst's' primary responsibility. But they also don't need to depend on IT to do it for them either.

Some of the most common data quality-related issues faced by analysts and organizations in general are:

1. Duplicates: Multiple copies of same record.
2. Incomplete Data: Many times the data has not been entered correctly so it not gives proper information or message due to missing variables.
3. Inconsistent Format: If the data is not entered in proper format then it will be problematic for the software to analyze the data and to produce correct result.
4. Accessibility: The information the most of the data scientist use or the system analyst use to create, evaluate, theorize and predict the results and end products often gets lost.
5. System Upgrade: Every time when the system gets updated or the hardware required to be updated there are chances of information getting lost or corrupt.
6. Data purging and storage: With the management level in an organization there are chances that locally saved

information can be deleted either knowingly or un-knowingly. So, saving the data or the information safely is the important task.

IV. DATA CLEANING

Data cleaning is applied with comprehension in the different areas of data processing and maintenance. Data profiling examines the data available in an existing source and collects statistics information about that data. It is an application of data analysis technique. It determines actual content, structure and quality of the data

Data profiling gives overall idea about the database which are obliging for data cleaning to perform their work efficiently. Data cleaning is essential to maintain the data warehouse; it deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. It is applied in the field of data warehousing when several databases are merged. How data cleaning is important task in data warehousing is described in. Records referring to the same entity are represented in different formats in the different data sets or are represented erroneously. Thus, duplicate records will appear in the merged database. The issue is to identify and eliminate these duplicate records. This is called Merge/purge problem.

Generally data cleaning is updating a record with cleaned data but serious cleaning involves decomposing and reassembling the data. Data transformation is essential for extracting data from legacy data formats and for Business – to - Business Enterprise data integration. Data cleaning is performed by domain expert because it is valuable in identifying and eliminating of anomalies. Anomaly is a property of data values it may causes the errors in measurements, lazy input habits, omission of data and redundancies. Anomalies basically classified into three types Syntactic - describes characteristic values and format. Semantic - hides data collection from a comprehensive and non- redundant representation. Coverage anomalies - reduce the amount of entities and their properties.

4.1 Data Cleaning Process:

i. Data Auditing: Auditing the data is done to find the types of anomalies contained within it. Statistical methods are used for auditing. Syntactical anomalies are detected using parsing. The results of auditing the data support the specification of integrity constraints and domain formats. Integrity constraints are depending on the application domain and are specified by domain expert. Each constraint is checked to identify possible violating tuples. For onetime data cleansing only those constraints that are violated within the given data collection has to be further regarded within the cleansing process.

ii. Workflow Specification: Multiple operations over the data are applied for Detection and elimination of common order problems. This is called the data cleansing workflow. It is specified after auditing the data to gain information about the existing anomalies in the data collection at hand. One of the main challenges in data cleansing insists in the specification of a cleansing workflow that is to be applied to the dirty data automatically eliminating all anomalies in the data.

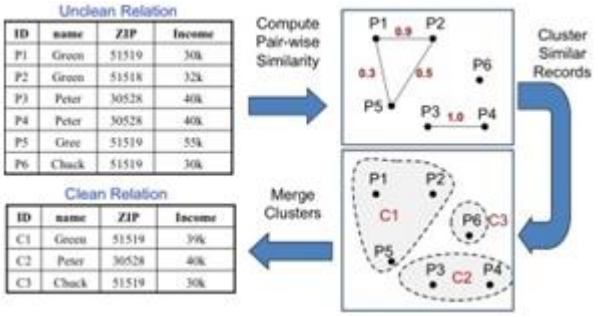
iii. Workflow Execution: The data cleaning workflow is executed after specification and verification of its correctness.

iv. Post-Processing/Control: After executing the cleansing workflow the results are checked to again verify the correctness of specified operations. Within the controlling step the tuples that could not be corrected initially are inspected intending to correct them manually.

Applications of Data Cleaning

Data cleaning is an important step in all types of data-driven analytics. Different data cleaning tasks target different types of errors. Applications of outlier detection include network intrusion detection, financial fraud detection, and abnormal medical condition detection. For example, in the context of computer networks, different kinds of data, such as operating system calls and network traffic, are collected in large volumes. Outlier detection can help with detecting possible intrusions and malicious activities in the collected data. Rule-based data cleaning can help clean any relational databases where data quality rules can be defined. The richer the semantics of the data is, the better rule-based data cleaning techniques are at detecting and repairing violations. Data transformations are used in a variety of tasks, and at different stages of the ETL life cycle. For example, before running a data integration project, transformations are often used to standardize data formats, to enforce standard patterns, or to trim long strings. Transformations are also used at the end of the ETL process, for example, to merge clusters of duplicate records, to find a unique representation for a cluster of records (aka golden record), or to prepare data to be consumed by analytics tools. Data transformations can also be seen as a tool for data repair in rule-based data cleaning, since it can be used to “transform” erroneous data. Duplicate records can occur due to many reasons. For example, a customer might be recorded multiple times in a customer database if the customer used different names at the time of purchase; a single item might be represented multiple times in an online shopping Website; and a record might appear multiple time after a data

integration project because that record had different representations in original data sources. Data deduplication targets specifically duplicate records, and resolves them.



V. CONCLUSION

Data cleaning is very necessary part of data mining. Data cleaning is a complicated process, and an end-to-end data cleaning solution usually involves many different cleaning sub-tasks. We have discussed techniques for tackling data cleaning tasks. There are still many challenges and opportunities in building practical data cleaning systems:

(1) the scale of data renders many data cleaning techniques insufficient. New cleaning solutions must adapt to growing datasets of the Big Data era, for example, by leveraging sampling techniques or distributed computation. (2) although there are existing research about involving humans to perform data deduplication, for example, through active learning (Tejada et al 2001; Sarawagi and Bhamidipaty 2002; Arasu et al 2010), involving humans in other data cleaning tasks, such as repairing IC violations, and taking user feedback in discovering of data quality rules, is yet to be explored; (3) a significant portion of data is residing in semi- structured formats (e.g., JSON) and un- structured formats (e.g., text documents). Data quality problems for semi-structured and unstructured data remain largely unexplored; and (4) there is significant concerns about data privacy as increasingly more individual data are collected by governments and enterprises. Data cleaning is by nature a task that requires examining and searching through raw data, which may be restricted in some domains including finance and medicine. How to perform most data cleaning tasks, while preserving data privacy, remains an open challenge. From the above study we can see that there are different types of problems in data cleaning. Data cleaning methods and approaches depend upon the type of data which we want to clean and according to that we apply particular methods. There are different types of tools present for Data Cleaning Next we will do the comparison of data cleaning tools and determine the best tool. Each tool has its own specific features and depending upon the data we can use the tool to clean data. In future work we can check other functionality of these tools and suggest own.

REFERENCES

[1] Orr, K. – "Data Quality and Systems Theory". Communications of the ACM, 41 (2). 1998. pp. 66-71.
Meta Group – Data Warehouse Scorecard. Meta Group, 1999.
[2] Sattler, K. and Schallehn, E. – "A Data Preparation Framework based on a Multi database Language". In Proceedings of International Database Engineering and Applications Symposium (IDEAS – 2001), Grenoble, France. IEEE Computer Society. 2001. pp. 219 – 228.
[3] Galhardas, H.; Florescu, D.; Shasha, D.; Simon, E. and Saita, C.-A. – "Data Cleaning: Language, Model, and Algorithms". In Proceedings of the Very Large databases Conference (VLDB). 2001.
[4] Pipino, L.; Lee, Y. and Wang, R. – "Data Quality Assessment". Communications of the ACM, 45 (4). 2002. pp. 211 – 218.
[5] Wand, Y. and Wang, R. – "Anchoring Data Quality Dimensions in Ontological Foundations". Communications of the ACM 39 (11). 1996. pp. 86-95.
[6] Li Lee Mong , Cleansing Data for Mining and Datawarehousing, school of computing National University of Singapore, 1999 .
[7] Rahm E. & Hai Do Hong, Data Cleaning: Problems and current approaches, IEEE Bulletin of the Technical Committee on Data Engineering, 2000
[8] Müller Heiko & Christoph Freytag Johann , Problems, Methods, and Challenges in Comprehensive Data Cleansing ,Humboldt-Universität zu Berlin zu Berlin,10099 Berlin, Germany.
[9] Devi Sapna & Kalia Arvind, Study of Data Cleaning and Comparision of Data Cleaning Tools, International Journal of Computer Science and Mobile Computing, Vol 4, Issue 3, March 2015, pp. 360 – 370.
https://analyticsindiamag.com/ accessed on 25.01.2020 at 12:20 pm.
[10] Abedjan Z, Morcos J, Ilyas IF, Ouzzani M, Papotti P, Stonebraker M (2016) Dataxformer: A robust transformation discovery system. In: Proc. 32nd Int. Conf. on Data Engineer- ing, pp 1134–1145.
[11] Aggarwal CC (2013) Outlier Analysis. Springer Arasu A, Gotz M, Kaushik R (2010) On active learning of record matching packages. In: Proc. ACM SIGMOD Int. Conf. on Management of Data, pp 783– 794
[12] Bertossi LE (2011) Database Repairing and Consistent Query Answering. Morgan & Claypool Publishers.