Chapter

Use Data Mining Cleansing to Prepare Data for Strategic Decisions

Mawande Sikibi

Abstract

Pre-processing data on the dataset is often neglected, but it is an important step in the data mining process. Analyzing data that has not been carefully screened for such challenges can produce misleading results. Thus, the representation and quality of data are first and foremost before running an analysis. In this paper, the sources of data collection to remove errors are identified and presented. The data mining cleaning and its methods are discussed. Data preparation has become a ubiquitous function of production organizations – for record-keeping and strategical making in supporting various data analysis tasks critical to the organizational mission. Despite the importance of data collection, data quality remains a pervasive and thorny challenge in almost any production organization. The presence of incorrect or inconsistent data can significantly distort the results of analyses, often negating the potential benefits of strategical making driven approaches. This tool has removed and eliminated errors, duplications, and inconsistent records on the datasets.

Keywords: Data, Data cleaning, Data collection, Data mining, Data preparation, Data collection, Data quality, Messy data

1. Introduction

Time has changed for the production organizations who believe keeping messy data saves their day. This messy data is in the dataset, which is stored in databases, repositories, and data warehouses. Massive amounts of data are available on their resources for the organization to influence their strategic decision. Data collected from various resources are messy, and this affects the quality of the data result. Data preparation offers a better data quality, which will help the organizations yearly, making most existing methods no longer suitable for messy data.

The growing enthusiasm of messy data on the dataset for data-driven strategic decision-making has created the importance of preparing data for future use over the years. The rapid growth of messy data drives new opportunities for the organization and processing the quality of the data by cleaning and preparing data becomes essential for analysts and users. Unfortunately, this could be handled correctly as reliable data could lead to a misguided strategic decision.

Data mining is no longer a new research field [1]. It aims to prepare data to improve data quality before processing by identifying and removing errors and inconsistencies on the dataset [2]. Data mining can pull data to prepare it to inform organization strategic decisions. However, preparing data can be used before for specific organizational purposes.

Data mining could be added to a single application to pull anomalies within a large dataset. Utilizing the software arranges data in the large dataset to develop efficient organizational strategies. Data mining software is a user-friendly interface that allows organizational analysts and users who may not be technically advanced to execute data preparation in data mining [3]. Putting this capability in the hands of the non-technical user allows responding to data quality issues quickly.

Data preparation is the feature within data mining; it has immeasurable value working with data [4]. Utilizing the software will begin to embed within the organization. Data mining software is available on the market for an organization to use their data in the dataset. Thus, markets are different from a decade ago due to rapid change in the world economy and technological advancement. This technology is popular with marketers because it allows analysts and users to make smart strategic decisions. It enables better development of market strategies for competitive advantage ahead amongst organizations. As vendors continue to introduce solutions, the marketing strategy improves the data quality of the dataset stored in their resources. With data mining, analysts and users can access the dataset in preparation for it to be available for future use.

2. Objectives

Understanding the better contribution of data mining makes to the dataset. In addressing this, an attempt of the following should be met.

- To discover errors and inconsistencies in the dataset for data preparation.
- Minimizes the errors and inconsistencies on the dataset.
- Utilize the use of datasets stored in their various resources.

This paper aims to develop a process data mining capability undertake on the dataset. The literature review considers current knowledge contributions to this topic towards these paper objectives.

3. Literature review

Data preparation corrects inconsistent data in the dataset to prepare quality data [5]. Research indicates that data preparation in data mining formulates a workflow process covering steps to prepare data [6]. However, some research suggested that data preparation begins with data collection to check data quality [7]. This paper aims to demonstrate the evolution of collecting data into preparation steps to influence data quality. The paper examines the data preparation in data mining processes through data collection.

3.1 Data collection

Data mining is often described as an add-on software in checking the data quality in the dataset by searching through the large amount of data stored in databases, repositories, and data warehouses. The data stored is believed to be too messy, inconsistent, and have errors; it is unclear information to analysts and users, make it difficult to be ready to be used for its specific purposes [8]. Overloaded data limit analysts and users; thus, software such as data mining is developed to solve this challenge through automation.

The data mining software uses recognition technologies and statistical techniques to clean messy data and discover the possible rule to govern data in databases, repositories, and data warehouses. Data mining considers the process that requires goals and objectives to be specified [9]. Once the intended goals met, it is necessary to determine what data is collected or available. However, before data is used, data preparation is performed, making data ready for its purposes.

The concept that strategic or effective decisions are based on appropriate data is not new. Finding the correct data for strategic decisions began 30 years ago [10]. During the late 1960s, organizations create reports from production sensors into databases, repositories, and data warehouses. These resources stored data to retrieved and manipulate to produce constructive reports containing information to meet specific strategic decision needs.

In the 1980s, analysts and users began to need data more frequently and to be more individualized. Thus, the organizations started to request data in the resources. Later in the 1990s, analysts and users required immediate access to be more detailed information. This meant to correlate with production and strategic decisions processes. It has helped the analysts and users extract its data from databases, repositories, and data warehouses.

The analysts and users began to realize the need for more tools to prepare data for future uses. Additionally, the organizations recognized the accumulated amount of data; thus, new tools to prepare data before meeting their needs. Such tools enabled the system to search for any possible errors and inconsistencies in the dataset. Data mining software was the first developed to help analysts and users to find quality data from a voluminous amount of data. Because the massive volume of data keeps rapidly growing, preparation methods are urgently needed. Therefore, data mining has become an increasingly important research field [11].

3.2 Data cleansing

Data cleansing is an operation within data mining software that can be performed on the existing data to remove anomalies and obtain the data collection. It involves removing the errors, inconsistencies and transform data into a uniform format in the dataset [12]. With the amount of data collected, manual data cleansing for preparation is impossible as it is time-consuming and prone to errors. The data cleansing process consists of several stages: detecting data errors and repairing the data errors [13]. Although, it is thought of as a tedious exercise. However, establish a process and template for the data cleansing process gives assurance that the method applied is correct. Hence, data cleansing focuses on errors beyond small technical variations and constitutes a significant shift within [14].

Data cleansing based on the knowledge of technical errors expects normal values on the dataset. Missing values may be due to interruptions of the data flow. Hence,

predefined rules for dealing with errors and true missing and extreme values are part of better practice. However, it is more efficient to detect the errors by active searching for them on the dataset in a planned way. Lack of data through data cleansing will arise if the analysts and users do not fully understand a dataset, including skips and filters [14].

Moore and McCabe [15] emphasized the serious strategic decision error would endure if the data quality were poor, leading to low data utilization efficiency. Although data cleansing follows data collection, data thoroughly checked for errors, and other inconsistencies are corrected for future use [16]. Although the importance of data-handling procedure is being underlined in better clinical practice and data management guidelines, gaps in knowledge about optimal data handling methodologies and standard of quality data are still present [14].

Detecting and correcting corrupted or inaccurate records help to meet standard quality data from the dataset. Find the incorrect, inaccurate, or irrelevant parts of the data, replace, modify, and delete coarse data [14]. The reality of the matter, data cannot always be used as it is and needs preparation to be used. Achieving higher preparation data quality during a data cleansing process is required to remove anomalies. Thus, the data cleansing process can be defined as assessing data's correctness and improving it. Therefore, enhancing data quality, pre-processing data mining techniques are used to understand the data and make it more easily accessible.

3.3 Data validation

Data validation is described as the process of ensuring data has undergone cleaning to ensure that it is both correct and useful. Although, it intended to provide a guarantee for the fitness and consistency of data in the dataset. Failure or omission in data validation can lead to data corruption. Catching data early on the dataset is important as it helps debug the roots of the cause and roll back in the working state [17]. Moreover, it is important to rely on mechanisms specific to data validation rather than on the detection of second-order effects.

Errors are bound to happen during the data collection process, while data is seldom 100% correct. Data validation helps to minimize erroneous data from the dataset. Data validation rules help organizations follow standards that make it efficient to work with data. Although, duplication data provide challenges to many organizations. Factors that cause the duplication of data are the data entry of machines and operators from production to capture data. An organization needs a powerful matching solution to overcome this challenge of duplicating records to ensure clean and usable data.

Data validation checks the accuracy and data quality of source data, usually performed before processing the data. It can be seen as a form of data cleansing. Data validation ensures that the data is complete (no blanks or empty values), unique (includes different values that are not repeated), and the values that range consistent with the expectations. When moving and merging data, it is important to ensure that data from different sources and repositories conform to organizational rules and not become corrupted due to inconsistencies in type or context. Data validation is a general term and can be performed on any data. However, including data within a single application, such as Microsoft Excel, or merging simple data within a single data store.

The data validation process is a significant aspect of filtering the large dataset and improving the overall process's efficiency. However, every technique or process consists of benefits and challenges; therefore, it is crucial to have a complete

acknowledgement. Data handling can be easier if analysts and users adapt this technique with the appropriate process, then data validation can provide the best outcome possible for data. Data validation can be broken down into the following categories: data completeness and data consistency.

3.3.1 Data integrity

Data integrity refers to the integrity of the data. However, for the data to be valid, there should not be any gaps or missing information for data to be truly complete. Occasionally incomplete data is unusable, but it is usually used in the absence of information, leading to cost error and miscalculations.

An incomplete data is usually the result of unsuccessful data collection. This denotes the degree to which all required data are available in the dataset [18]. A measure of data completeness would be the percentage of missing data entries. However, the true goal of data completeness is not to have perfect 100% data. It ensures that data the essential to the purpose of validity. Therefore, it is a necessary component of the data quality framework and is closely related to validity and accuracy.

3.3.2 Data consistency

Data consistency means that there is consistency in the measurement of variables throughout the datasets. This becomes a concern, primarily when data aggregates from multiple sources. Discrepancies in data meanings between data sources can create inaccurate, unreliable datasets. Since the data inconsistency comes from the storage format, semantic expressions, and numerical values, a method of consistent quantification assesses the degree of data consistency quantitatively after defining the degree of consistency.

Data consistency could be the difference between great business success or failure. Data is the foundation for successful organizational strategic decisions, and inconsistent data can lead to misinformed business decisions. Organizations must ensure data consistency, especially when aggregating data from multiple internal or external sources without changing their structure, to be confident and successful in their strategic decision-making.

Data consistency checks that the data values of all instances of the application are the same. These data belong together and describe a specific process at a specific time, which means that the data remains unchanged during processing or transmission. Synchronization and protection measures help to ensure that data consistency during the multi-stage processing [19]. Data consistency is essential to the operation of programs, systems, applications, and databases. Locking measures prevent data from being altered by two applications simultaneously and ensure correct processing order. Controlling simultaneous operations and handling incomplete data are essential to maintain and restore data consistency in power failures.

3.4 Data preparation

Data preparation is the process of cleaning and transforming raw data before processing and analysis for future use. It is an important step before processing and often involves reformatting data, correcting data, and combining data sets to enrich data [20]. Its task is to blend, shape, clean, consolidate data into one file or data table to get it ready for analytics or other organizational purposes.

The data must be clean, formatted, and transformed into something digestible by data mining software to achieve the final preparation stage. These actual processes include a wide range of steps, such as consolidating or separating fields and columns, changing formats, deleting unnecessary or junk data, and making corrections to data.

In this literature review, several studies have used data preparation and data mining on the messy data on the dataset for future use, few studies on the quality data check. This is the gap in this paper, as it aims at reviewing the available data mining preparing methods for messy data. Since the data preparation framework needs to meet data quality criteria, using a quality dimension includes accuracy, completeness, timeliness, and consistency [21]. Quality data check is crucial because it automates data and provides information about the number of valid, missing, and mismatched values in each column. The result shows the quality data above each column in the dataset. A data mining software will help remove errors and inconsistencies in the dataset to meet quality data check percentage [22].

Quality data check on the dataset, it may be better to use a transformation. These quality data checks can create data quality rules which persist in checking columnar data against defined. Performing variety checks, transform data automatically show the effect of transformations on the overall quality of data. It can provide various services for the organization and only with high-quality data and achieve the top-service in the organization [13].

4. Methodology

This chapter aims to provide the research methodology roadmap designed to meet the objectives of this paper. It is important to select an appropriate method to ensure the accuracy, validity, and quality of data and findings. This chapter shows the method chosen, the tools used to extract data and data analysis. Hence, the phenomenological concept is focused on preparing data and reference [23]. A research method refers to how data can be collected and analyzed, such as data analysis software.

This paper used ethnography as the researcher was directly involved in preparing messy data on the dataset. Ethnography is usually described as participant observation, and this was where the researcher became actively involved, demonstrating the data preparation.

A single case approach was chosen for this paper to be the suitable method for executing data preparation into a single organization. It was not done to represent other same organization using data mining analysis. It was using the quantitative and qualitative method to explore data preparation. It began with a data collection approach to the analysis of data preparation. Although, it may be possible to generalize this paper.

The company set the principles of ethics, which was honored by the researcher. The company was informed that participating in this demonstration was voluntary and would not impact the company's brand. Ensuring anonymity, the paper removed some information that would be manipulation to favor the competitors [24]. Thus, the name of the organization is referred to as company A. Public information that could have damaged the company authenticity that could result in negative was removed.

4.1 Company description

Company A is one of the leading companies in producing steels. This company is situated in Alberton, where most of the production industries are built. It has a

history of making several sheets of steel at a high rate. It increases the data in the dataset, not only proper data but also messy or dirty data. The company was selected due to its nature of producing a high number of products. Therefore, it was suitable for this research, which is dealing with data.

4.2 Data collection

Data collection is the method of gathering observations or collecting information using standard validated techniques. It is important to collect data to understand what can be done using it. Data collection consisted of two sources, which is primary and secondary data. Primary data refers to raw data collected. Secondary data is data that is already collected. Therefore, this paper selected secondary as company A already collected its data using sensors embedded in their machines into databases, repositories, and data warehouses.

The researcher extracted the dataset from the repository of company A based on the experience obtained through training in extracting data. This potential skill has helped the researcher to use data mining tool for preparing data. This was done during the period month of February and March 2021. Datasets were sent by company A to the researcher as the active participant in preparing data datasets due to the coronavirus pandemic. The datasets that were sent contained the machine, alarm data, and sensor data.

4.3 Data analysis

Data analysis is the process of systematically applying statistical and technique to evaluate data. According to [25], this type of research whereby data gathered is categorized into themes and sub-themes. Analysis helps data collected being reduced and simplified while at the same time producing results that may then measure using quantitative techniques. Moreover, the analysis provides the ability to the researcher to structure the qualitative data to satisfy the accomplishment of the paper objectives. The researcher installed a data mining tool as an "Add-on" to the Microsoft Excel spreadsheet. Microsoft Excel is a powerful tool for handling large data [26]. It consists of a grid with columns and rows that store data from resources of data. Data mining employed to arrange and remove inconsistencies that were on the datasets. Data mining was performed into a Microsoft Excel spreadsheet to prepare data for its readiness to be used for specific purposes. The resources were used in this paper are computer, Microsoft Excel spreadsheet, and data mining tool.

5. Results and discussion

This section describes the findings, and the overall discussion represents the datasets with data cleansing preparation. These three datasets were obtained from the data repository, and **Table 1** represents the excel files dataset before using the data mining tool. Machine data file contain 30 000 records in the dataset. It contains 10% missing values and 7 duplicate records. The alarm data file contains 45 000 records. It contains 25% missing values and 28 duplicated records in the dataset. Finally, the sensor data file contains 100 000 records in the dataset. It contains 45% missing values and 100 duplicated records. These files format was using Microsoft Excel as the technique to use datasets.

Filename	No. of records	No. of fields	Missing value	Duplicated records
Machine data	30 000	1	10%	7
Alarm data	45 000	1	25%	28
Sensor data	100 000	1	45%	100

Table 1.
Raw data.

Filename	No. of records	No. of fields	Missing value	Duplicated records
Machine data	26 993	1	0%	0
Alarm data	33 722	1	0%	0
Sensor data	54 900	1	0%	0

Table 2.
Data mining uses.

A data mining tool was used as the result of the analysis. **Table 2** shows the importance of using data mining in removing errors and inconsistencies in records. The data mining tool in **Table 2** has removed machine data records decreases from 30 000 to 26 993. There was no missing value found on the dataset, with no duplication records. Alarm data records decreased from 45 000 to 33 722, with no missing values and duplicated records. Sensor data records decreased 100 000 to 54 900, with no missing values and duplicated records.

Missing values represent how efficient this tool in finding missing values of a file. Other features were whether this tool could find duplication, illegal values, merging the records and misspelling. Ease file format supported by these records and of use.

5.1 Discussion

This paper aims to investigate data cleansing in big data. Based on the available data cleansing methods discussed in the previous section, data cleansing for big data needs to be improvised and improved to cope with the massive amount of data. The traditional data cleaning method is important for developing the data cleaning framework for big data applications. In the review of Potter, this method only focused on solving data transformation challenges [13]. The Excel spreadsheet supports problems like duplicate record detection, and the user needs other approaches to deal with duplicate record detection problems [27].

Data mining can require manual and automatic procedures, but this approach focuses on duplication and missing elimination despite various data quality challenges in the dataset. Traditional data cleansing tools tend to solve only one data quality problem throughout the process and require human intervention to resolve data cleansing conflicts. In the big data era, the traditional data cleansing process is no longer acceptable as data needs to be cleansed and analyzed fast. The data is growing more complex as it may include structured data, semi-structured data, and unstructured data. The discussed methods focus only on structured data. However, existing methods have some limitations when working with dirty data. Data mining performs the computations of each stage as “local” in each Excel spreadsheet, and the data exchange is done at the stage boundaries by broadcast or hash partitioning.

6. Recommendations and conclusion

6.1 Recommendation

The chapter discusses the contribution of data mining cleaning on a dataset. This is achieved by discovering the errors and inconsistencies in the dataset and utilizing datasets stored in various resources. The authors discuss the importance of the management in organizations for attaching the vitality of data sourcing and strategic decision-making. The management must ensure that the correct, timely and accurate data is used in strategic decision-making to generate the ever-elusive competitive advantage. Furthermore, due to the key roles of the available data, big data has become a strategic resource. The data security required to be enhanced at all strategic decision-making levels to avoid unauthorized person (s) must be explored as future work.

6.2 Conclusion

Most organizations rely on data-driven decision making; therefore, the information system is closely related to business process management to leverage their processes for competitive advantage. Nowadays, the amount of data is constantly increasing, but the data quality is decreasing as much of the data collected is messy or dirty. There are various data cleansing approaches to solve this challenge, but data mining cleansing remains a tool to deal with the criteria of big data. Some of the approaches are not suitable for big data as there is a significant amount of data that needs to be processed simultaneously. Despite the availability of existing frameworks for data cleansing for big data, the value and veracity of the data are often disregarded while developing the approaches. Moreover, data mining is undeniably required to verify and validate the data before it can be subjected to an analysis process.

Acknowledgements

First, I would like to thank God for His blessing in completing this paper and my highest gratitude goes to my mentor for guarding me throughout this paper. Her patience on this paper was something I admired.

Also, thanks to seen and unseen hands that have given me direct and indirect help to finish this paper. Finally, thanks to my family who keeps encouraging through difficult time. Even if it was not fashionable to do so.

Declarations

I, Mawande Sikibi, hereby declare that this paper is wholly my work and has not been submitted anywhere else for academic credit either by myself or another person.

Author details

Mawande Sikibi
University of Johannesburg, South Africa

*Address all correspondence to: mawandesikibi@gmail.com

IntechOpen

References

[1] Adelman-McCarthy JK. VizieR Online Data Catalog: The SDSS Photometric Catalog, Release 8 (Adelman-McCarthy+, 2011). VizieR Online Data Catalog. 2011 Sep:II-306.

[2] Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data?.

[3] Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. Data science journal. 2015 May 22;14.

[4] Collis J, Hussey R. Business research: A practical guide for undergraduate and postgraduate students. Macmillan International Higher Education; 2013 Nov 29.

[5] Cong G, Fan W, Geerts F, Jia X, Ma S. Improving Data Quality: Consistency and Accuracy. InVLDB 2007 Sep 23 (Vol. 7, pp. 315-326).

[6] Gschwandtner T, Aigner W, Miksch S, Gärtner J, Kriglstein S, Pohl M, Suchy N. TimeCleanser: A visual analytics approach for data cleansing of time-oriented data. InProceedings of the 14th international conference on knowledge technologies and data-driven business 2014 Sep 16 (pp. 1-8).

[7] Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. arXiv preprint arXiv:1706.02216. 2017 Jun 7.

[8] Hellerstein JM. Quantitative data cleaning for large databases. United Nations Economic Commission for Europe (UNECE). 2008 Feb 27;25.

[9] Jones S, Pryor G, Whyte A. How to Develop Research Data Management Services-a guide for HEIs.

[10] Kandel S, Heer J, Plaisant C, Kennedy J, Van Ham F, Riche NH, Weaver C, Lee B, Brodbeck D, Buono P. Research directions in data wrangling: Visualizations and transformations for usable and credible data. Information Visualization. 2011 Oct;10(4):271-88.

[11] Kusiak A. Data mining: manufacturing and service applications. International Journal of Production Research. 2006 Sep 15;44(18-19):4175-91.

[12] Li D, Wang S, Li D. Spatial data mining. Berlin, Heidelberg: Springer Berlin Heidelberg; 2015.

[13] Liao X, Lochhead P, Nishihara R, Morikawa T, Kuchiba A, Yamauchi M, Imamura Y, Qian ZR, Baba Y, Shima K, Sun R. Aspirin use, tumor PIK3CA mutation, and colorectal-cancer survival. New England Journal of Medicine. 2012 Oct 25;367(17):1596-606.

[14] Mathew PS, Pillai AS. Big Data solutions in Healthcare: Problems and perspectives. In2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) 2015 Mar 19 (pp. 1-6). IEEE.

[15] Moore CA, McCabe ER. Utility of Population-based Birth Defects Surveillance for Monitoring the Health of Infants and as a Foundation for Etiologic Research. Birth defects research. Part A, Clinical and molecular teratology. 2015 Nov;103(11):895.

[16] Olson DL, Delen D. Advanced data mining techniques. Springer Science & Business Media; 2008.

[17] Polyzotis N, Zinkevich M, Roy S, Breck E, Whang S. Data validation for

machine learning. Proceedings of Machine Learning and Systems. 2019 Apr 15;1:334-47.

[18] Rahm E, Do HH. Data cleaning: Problems and current approaches. IEEE Data Eng. Bull. 2000 Dec;23(4):3-13.

[19] Rajkumar SV. Multiple myeloma: 2012 update on diagnosis, risk-stratification, and management. American journal of hematology. 2012 Jan;87(1):78-88.

[20] Raman V, Hellerstein JM. Potter's wheel: An interactive data cleaning system. InVLDB 2001 Sep 11 (Vol. 1, pp. 381-390).

[21] Ridzuan F, Zainon WM. A review on data cleansing methods for big data. Procedia Computer Science. 2019 Jan 1;161:731-8.

[22] Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med. 2005 Sep 6;2(10):e267.

[23] Vandecruys O, Martens D, Baesens B, Mues C, De Backer M, Haesen R. Mining software repositories for comprehensible software fault prediction models. Journal of Systems and software. 2008 May 1;81(5):823-39.

[24] Wang L, Jacques SL, Zheng L. MCML—Monte Carlo modeling of light transport in multi-layered tissues. Computer methods and programs in biomedicine. 1995 Jul 1;47(2):131-46.

[25] Wirth R, Hipp J. CRISP-DM: Towards a standard process model for data mining. InProceedings of the 4th international conference on the practical applications of knowledge discovery and data mining 2000 Apr 11 (Vol. 1). London, UK: Springer-Verlag.

[26] Wu X, Zhu X, Wu GQ, Ding W. Data mining with big data. IEEE transactions on knowledge and data engineering. 2013 Jun 26;26(1):97-107.

[27] Zhang S, Zhang C, Yang Q. Data preparation for data mining. Applied artificial intelligence. 2003 May 1;17(5-6):375-81.