Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 161 (2019) 731–738

Procedia
Computer Science

www.elsevier.com/locate/procedia

The Fifth Information Systems International Conference 2019

A Review on Data Cleansing Methods for Big Data

Fakhitah Ridzuan, Wan Mohd Nazmee Wan Zainon*

School of Computer Science, Universiti Sains Malaysia, Penang, 11800 Malaysia

Abstract

各種資源收集到的資料往往存在髒資料問題

Massive amounts of data are available for the organization which will influence their business decision. Data collected from the various resources are dirty and this will affect the accuracy of prediction result. Data cleansing offers a better data quality which will be a great help for the organization to make sure their data is ready for the analyzing phase. However, the amount of data collected by the organizations has been increasing every year, which is making most of the existing methods no longer suitable for big data. Data cleansing process mainly consists of identifying the errors, detecting the errors and corrects them. Despite the data need to be analyzed quickly, the data cleansing process is complex and time-consuming in order to make sure the cleansed data have a better quality of data. The importance of domain expert in data cleansing process is undeniable as verification and validation are the main concerns on the cleansed data. This paper reviews the data cleansing process, the challenge of data cleansing for big data and the available data cleansing methods.

本文回顧資料清洗過程、大數據下資料清洗面臨的挑戰和方法

Keywords: data cleansing; big data; data quality

1. Introduction

The growing enthusiasm for data-driven decision-making has created the importance of accurate and precise prediction over the previous years. The rapid growth of the data drives new opportunity for business and the process of analyzing the data quickly become more essential. Unfortunately, the data must be handled correctly as unreliable information could lead to a misguided decision. Data cleansing, or sometimes called data cleaning is no longer a new research field. It aims to improve the quality of data by identifying and removing errors and inconsistencies [1].

識別並消除錯誤與不一致來提升數據品質

* Corresponding author. Tel.: +60-46-534-63; fax: +60-46-573-335.
E-mail address: nazmee@usm.my

10.1016/j.procs.2019.11.177

根據"Price Waterhouse Coopers"於2001年進行的調查，599家公司有75趴因為資料品質而遭受損失

Incomplete information will generate uncertainties during data analysis and this must be managed in the data cleansing stage. Errors or missing values in the dataset will produce a different result and may affect the business decision. The data must be accurate to avoid losses, problems and additional cost due to poor quality of data. For example, according to the “Price Waterhouse Coopers” survey conducted in 2001, 75% of 599 companies have suffered losses due to data quality issue [2]. Since these businesses rely on data like customer relationship management and supply chain management, therefore it is important for them to have excellent quality data in order to achieve more precise and useful result. Quality data only can be produced by data cleansing as the data collected from the various sources might be dirty [3].

進行精準分析之前，需要確保資料品質

Data quality can be defined as the fitness of data to fulfill the business requirement. It is achieved through people, technology and processes. It ensures compliance and consistency particularly when data from different databases are combined. Without proper data quality management, even a minor error might cause revenue loss, process inefficiency and failure to comply with the industry and government regulations [4]. Thus, data quality and data cleansing always linked together as ensuring data quality is critical and necessary before sharpening of analytic focus can occur[5].

Data cleansing is an operation that is performed on the existing data to remove anomalies and obtain the data collection which is an accurate and unique representation of the mini world [6]. It involves eliminating the errors, resolving inconsistencies and transforming the data into a uniform format [7]. With the vast amount of data collected, manual data cleansing is almost impossible as it is time-consuming and prone to errors. Data cleansing process is complex and consists of several stages which include specifying the quality rules, detecting data error and repairing the error [8]. This paper aims at reviewing the available data cleansing methods specifically for big data. Since data cleansing framework needs to meet data quality criteria and fulfill big data characteristics, therefore this paper will identify the data cleansing challenge in big data. Data cleansing methods will be explained in brief along with the weaknesses and strengths of each method.

指定品質規則、檢測資料錯誤、修正錯誤等多個階段

2. Methodology

傳統資料清洗 大數據資料清洗 兩大類 2013 到 2019 年間文獻回顧

In order to review the current trends, articles describing or defining data cleansing, its challenges and methods were considered. The methodology was categorized into two phases; the first one focused on investigating data cleansing issue in big data. Data cleansing articles focusing on big data and data quality were selected. The second phase involved scrutiny of the references in resulting articles, which allowed the discovery of methods and specific authors related to the search objective. The results were classified into two groups which are traditional data cleansing and data cleansing for big data. For data cleansing methods for big data, articles whose publication date does not exceed seven years old from 2013 to 2019, were considered in order to obtain updated literature review.

3. Problem statement

Dirty and noisy data is common to big data, but the traditional way to handle dirty data may not be easily adaptable to a large dataset. Dirty data is defined as inaccurate, inconsistent and incomplete due to the error found within the dataset. Big data are typically described as ill-conditioned because of the amount of time and resources needed to cleanse the data [9]. Big data usually characterized into five main dimensions called 5V’s which are value, volume, variety, velocity, and veracity. The volume often makes up for the lack of quality data, many forms of data, quality, and accuracy are less controllable [10]. The most definition only focuses on the size of big data, but the variety and veracity also important in data cleansing.

Variety refers to the type of data that can be processed which can consist of structured or unstructured data. Challenges like incompatible data formats, incomplete data, non-aligned data structures, and inconsistent data can affect the analysis result. Veracity refers to trustworthiness in the data to make a decision. This proves that it is important to acquire the right correlation between the attributes for the business future [11]. Ensuring the accuracy and relevancy of data will drive the business a step ahead from their competitors. According to L'Heureux et al. [9], veracity is not only about the reliability of the analyzed data, but also the reliability of the data source. Data are being gathered enormously but the means and method to gather the data can introduce uncertainty which will affect the veracity of the dataset.

這篇論文提及，雖然大量資料，但收集方式or方法可能影響資料的表現，進而影響真實性。

特別提及大數據的特性
多樣性 Variety: 可處理的數據類型(data can be processed which can consist of structured or instructured data.)
真實性 Veracity : 可信度。

數據品質可以透過四個因素: (用於評估真實性，而非數量 速度和多樣性)
準確性
完整性
即時性
一致性

When processing even larger datasets, handling a sophisticated mechanism to discover errors or managing large arbitrary errors, the overhead of data cleansing may reach up to more than 60% of the data scientists' time [12]. Despite there are various tools being introduced for data cleansing, data scientist still finds it's time-consuming. This janitorial task is important as the data needs to be cleaned, labeled and enriched before it is used for the analysis. There is no longer an issue with the shortage of data; instead, a new problem arises to get good training data. Besides, access to quality data is the main issue faced by a data scientist to complete their work. Data quality experts estimate that a business spends about 40 to 50% of the budget for the data cleansing process as it is time-consuming, labor-intensive and tedious processes [13].

Data quality can be measured using a quality dimension which includes accuracy, completeness, timeliness, and consistency. This data quality dimension is evaluated to address the veracity of big data. However, due to big data volume, velocity and variety, the complexity of the data quality algorithm become more complex [14].

3.1. Data cleansing process

As data increasingly used to support organizational activities and drive the business decision, poor quality data may negatively affect organizational effectiveness and efficiency. Data quality is the main concern faced by most of the organization. In fact, this issue rises due to improper maintenance and will indirectly generate inconsistency in the database [15]. Data quality issue is one of the obstacles to effectively use the data as dirty data may lead to a false decision. It can provide various services for the organization and only with high quality of data, they able to achieve the top service in the organization [16].

Data cleansing process consist of five phases; (1) data analysis, (2) definition of transformation workflow and mapping rule, (3) verification, (4) transformation and (5) backflow of cleaned data. Fig. 1 shows the data cleansing process.

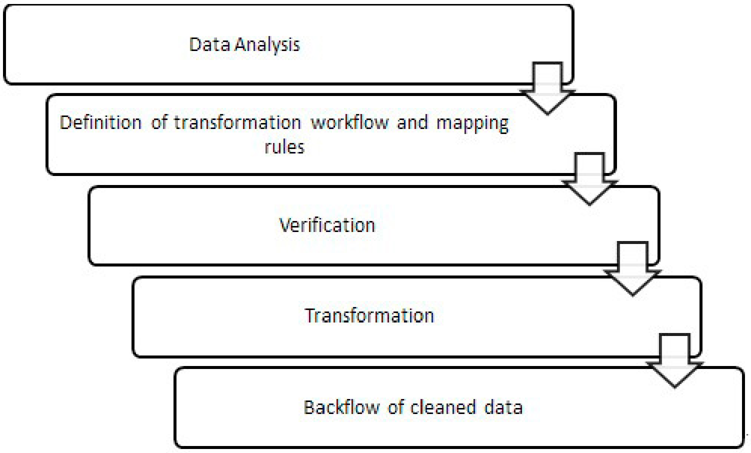

Fig. 1. Data cleansing process [1].

The first step in data cleansing is analyzing the data to identify the errors and inconsistencies that occurred in the database. In other words, this phase is called data auditing where this phase will find all types of anomalies inside the database. Besides, metadata about the data properties will be obtained through data analysis to detect data quality problems. There are two approaches in data analysis which are data profiling and data mining. Data profiling is an emphasis on the instance analysis of individual attributes. Meanwhile, data mining focus on discovering the specific data pattern in the large dataset. The result from the first step is the indication for each possible anomaly whether it occurs inside the database.

Next, transformation workflow defines the detection and elimination of anomalies performed by a sequence of operations on data. It is specified after data analysis to gain information about the existing anomalies. The number of transformation steps required depends on the number of data source, degree of heterogeneity and the ‘dirtiness’ of the data. To enable the automatic generation of the transformation code, the schema-related transformation, and the cleansing steps must be specified by a declarative query and mapping language. One of the main challenges in this phase is the workflow specification and the mapping rules which will be applied to the dirty data.

The third step is the verification stage. In this phase, the correctness and effectiveness of the transformation workflow are evaluated. This phase consists of multiple iterations to verify all the errors are being corrected and it requires the interaction with the domain experts. Since some errors are only visible after the transformation, thus a new cycle of analysis, design and verification process is needed.

After the data is verified and validated, the transformation steps will be executed to refresh the data in the data warehouse. The transformation process requires a large amount of metadata such as schema and instance level data characteristics, transformation mapping and workflow definitions. Detailed information about the transformation process must be recorded to support data quality. Finally, after all the errors have been removed, the dirty data should be replaced with the cleaned data.

3.2. Data cleansing challenges in Big Data

Various researches have been done throughout the years to find the ultimate data cleansing techniques in order to solve data quality problems. However, this is not an easy task as the amount of data is increasing every day and the current approach might no longer able to cope with this situation. Organizations are crammed with a huge volume of information and it is being collected every day at an unprecedented scale. This situation decreases the value of data collected and indirectly affects data quality and data analysis. Even though there is no standard volume of how big dataset need to be considered as big data [17], but various challenge related to the volume needs to be tackled. Besides, current data cleansing tools are not suitable for cleansing big data because none of the existing systems can scale out to thousands of machines in a shared-nothing manner [8, 18].

Data variety might be the biggest obstacle to effectively use the large volume of data for the analysis[19]. Different types of error like incompleteness, inconsistency, duplication, and value conflicting may co-exist in the big data and will affect the analysis result. Furthermore, the variety of constraints may detect the presence the errors, but fail to correct the errors and may introduce new errors when repairing the data [13]. Most existing solutions repair dirty databases by value modification follow constraint-based repairing approaches which search for a minimal change of the database to satisfy a predefined set of constraints.

Current data cleansing approaches also cannot ensure the accuracy of the repaired data and need domain expert [20]. The domain expert is important in the cleansing process to understand and implement quality rules [8, 21, 22] as well as verify the corrected data. However, human involvement in the cleansing process needs to be minimized as they are expensive to employ and limited [23]. Thus, many data cleansing solutions has been developed using highly domain-specific heuristic [7].

4. Data cleansing methods

A number of authors have proposed a solution to address data cleansing problems. It can be divided into traditional data cleansing and data cleansing for big data. Traditional data cleansing methods is called traditional because it is not suitable to handle a huge amount of data. Potter’s Wheel and Intelliclean are some of the examples of traditional data cleansing. Meanwhile, some of the methods are designed specifically for big data like Cleanix, SCARE, KATARA, and BigDansing. These methods are developed to address the issue arises when dealing with big data during the cleansing process.

4.1. Traditional data cleansing

Potter’s wheel is an interactive data cleansing system that integrates data transformation and error detection using a spreadsheet like an interface. According to Raman and Hellerstein [24], existing data cleansing tools are lack of interactivity where the transformation is done in the batch process and the user has to face long frustrating delays without any feedback. Besides, data often have many ‘nested discrepancies’ which hard to detect and more time is required for data transformation and discrepancy detection. Both data transformation and discrepancy detection need user effort, thus making the cleansing process become painful and prone to error. Potter's Wheel allows the user to define custom domains and the corresponding algorithms to enforce domain constraint [6]. Based on the given domain, the system will extrapolate appropriates structures for values in each column.

Intelliclean [25] is a knowledge-based approach where the main focus is on duplicate elimination. It was developed as a framework which provides a systematic approach for representation standardization, duplicate elimination, anomaly detection, and removal in dirty databases. The framework consists of three stages; (1) pre-processing stage, (2) processing stage and (3) validation and verification stage. During the pre-processing stage, data anomalies will be detected and cleaned. The output if this stage will be input to the processing stage. In the processing stage, there are four different rules which are duplicate identification rules, merge/purge rules, update rules, and alert rules. These rules are fed into an expert system engine for comparing the large collection of rules to a large collection of objects. The actions taken in these two stages will be logged for the verification and validation process. Human involvement is required in the final stage to verify the consistency and accuracy of the updates.

4.2. Data cleansing for Big Data

Wang, et al. [22] have designed and proposed Cleanix; a parallel big data cleansing system aims to solve the issue related to the volume and variety of big data. Four types of data quality issues are tackled by Cleanix which are abnormal value detection, incomplete data filling, deduplication, and conflict resolution. It is developed with the scalability, unification and usability features which enable Cleanix to perform data cleansing and data quality reporting task in parallel. Besides, it integrates various automated data repairing task into single parallel dataflow. The dataflow consists of four main stages; (1) read data, detect and correct abnormal data; (2) fill missing data; (3) broadcast the updated value in local gram; (4) solve deduplication and conflict resolution. This system also does not require any data cleansing expert because of the friendly and easy graphical user interface. Cleanix provides a web interface for the user to input the information of data source, parameters, and the rule selections. Users are allowed to select their own cleansing rules to solve the error found in the dataset.

On the other hand, Yakout, et al. [13] tried to use machine learning techniques and likelihood methods for the repairing and cleansing process. However, these techniques require accurate modeling of correlation between the databases attributes as some of the attributes of the same records may be dirty. SCARE (SCalable Automatic REpairing) is a systematic scalable framework that has a robust mechanism for horizontal data partitioning to ensure the scalability and enable parallel processing of data blocks. It is developed to address the issue on scalability and accuracy of replacement values by leveraging machine learning techniques for predicting better quality updates to repair dirty databases. SCARE offers a probabilistic principles technique to provide predictions for multiple attributes at a time. No constraint or editing rules is needed as it will analyze the data, learns the correlations from the correct data and takes advantages of them for predicting the most accurate replacement values.

Another method for data cleansing in big data is KATARA [23]. It is end-to-end data cleansing systems that use trustworthy knowledge-bases (KBs) and crowdsourcing for data cleansing. Chu, et al. [20] believed that integrity constraint, statistics and machine learning cannot ensure the accuracy of the repaired data. Thus, the authors introduced the presence of crowds as the main component in the data cleansing process along with KB. Crowdsourcing is needed to discover and verify table patterns, identify errors, and suggest possible fixes. The main functionalities are to interpret table semantics, identify correct and wrong data and generate top-k possible repairs for the wrong data. It is developed with the easy specification which enables the user to easily declare the target table and the reference KB. Besides, KATARA able to identify the top-K table pattern, validate the best pattern via crowdsourcing and annotate table with different categories. KATARA aims to produce accurate repairs by relying on KBs and domain expert. First, it will discover the table patterns to map the table to a KB. With table pattern, KATARA annotates tuples as either correct or incorrect by interleaving the KB and humans. For the incorrect tuples, the top-k mapping will be extracted from the KB and examined by humans.

Khayyat, et al. [8] proposed an approach that focuses on the efficiency, scalability, and ease of use issue in data cleansing called BigDansing. It is developed to address the scalability and abstraction problems when designing a distributed data cleansing system. Rule specification allows users to specify dataflow for error detection and it will be abstracted into a logical plan. User can focus on the logic rules instead of the details on how to implement it. Besides, the authors presented a technique that able to translate the logical plan into an optimized physical plan. A major goal of BigDansing is to allow users to express a variety of data quality rules in a simple way. It also supports a large variety of data quality rules by abstracting the rule specification process and able to achieve high efficiency when cleansing datasets by performing a number of physical optimizations. Besides, it can scale to big datasets by fully leveraging the scalability of existing parallel data processing frameworks. Table 1 summarizes the data cleansing methods for big data.

Table 1. Data cleansing methods for big data.

Methods	Key Features	Execution Method	Approach
Cleanix	Scalability, unification, and usability	Parallel	Rule selection
SCARE	Scalability	Parallel	Machine learning technique
KATARA	Easy specification, pattern validation, data annotation	Sequential	Knowledge-base and crowdsourcing
BigDansing	Efficiency, scalability, and ease of use	Parallel	Rule specification

宗旨:探討大數據資料清理

Potter's Wheel
Intelliclean

5. Discussion

傳統資料清理工具僅能解決單一類型數據品質問題，過程中仍須依賴人工介入

This study aims to explore data cleansing in big data. From the available data cleansing method discussed in the previous section, it is clear that the data cleansing for big data needs to be improvised and enhanced to cope with the vast amount of data. Traditional data cleansing method is important as the baseline to design the data cleansing framework for big data application. In the review of Potter's Wheel, it can be seen that this method only focused on solving data transformation problems only. Problems like duplicate record detection are not well supported by the system and user needs other approaches to deal with duplicate record detection problem [2]. Meanwhile, Intelliclean requires less human dependency but this approach only focuses on duplicate elimination despite there is various type of data quality problems occurs in the dataset. It also requires hand-coding when dealing with complex rules which decrease the degree of automation. Traditional data cleansing tools tend to solve one specific data quality issue only throughout the process and require human intervention to solve the data cleansing conflicts.

In the big data era, traditional data cleansing process no longer acceptable as the data need to be cleaned and analyzed at a fast pace. The data become more complex as it may comprise structured data, semi-structured data, and unstructured data. All the discussed methods focus more on the structured data only. However, the existing methods have some limitation when dealing with dirty data. Cleanix performs the computation of each stage is 'local' to every single machine and data exchange happen at the stage boundaries through broadcast or hash partitioning. Cleanix tries to reduce the need of domain expert by requiring the rules in stage 1 where user can select the rules or use any rules related to the dataset. Despite the quality, the report is shown at the end of the cleansing process, but the accuracy of the cleansed result is a doubt.

KATARA have similar intention as Cleanix where it aims to minimize the need for domain expert in the process. It actively involves the crowd together with the KBs for data cleansing, but there are some arguments on the KBs and the crowdsourcing. First, KBs might not be available to cover all the data. Besides, the crowd must be available in most of the process to make sure the data is accurate which making data cleansing is time-consuming. However, crowdsourcing can leads to uncertainty and noisy data because it makes use of human judgment to label the data [7]. The number of noisy labels will directly affect the veracity of the data. Even though crowdsourcing can substitute the presence of domain expert in the process, but it is still inefficient for large dataset [21].

On the other hand, Cleanix, SCARE, and BigDansing focus on the scalability issue faced in the data cleansing process. But SCARE and BigDansing do not require any domain expert or human in the cleansing process. SCARE admitted that domain expert should be involved to confirm the updates made on the dataset however no expert is a presence in the process. Instead, it requires a set of rules that have been specified by domain experts to cover the data. However, the process is expensive and the rules may fail to identify the correct fixes for the dirty dataset. Furthermore, the result of SCARE depends on the quality of training data in terms of its redundancy and a threshold ML parameter that is hard to set precisely [20].

Similar to SCARE and Cleanix, BigDansing also require a set of data quality rules for the cleansing process. Nevertheless, BigDansing allows users to express the data quality rules in a simple way without worrying about how to make the code distributed. The cleansing process is done in parallel to reduce waiting time, but if there are too many rules and constraint to follow, data quality rule optimization might be needed. Among all the approaches discussed, only KATARA used sequential execution process. However, the error detection in the KATARA is done in parallel due to the cluster of machines is interconnected by a fast network [26].

From the literature, it is evident that there is lack of research on data cleansing which focuses on big data criteria. Scalability is the main issue highlighted by the authors [8, 13, 22] when designing the data cleansing techniques to deal with the volume and variety of data. In addition, the process must be executed in parallel to cope with the velocity of the data. Data cleansing methods should be able to process the data while it is being generated without affecting the existing data. Furthermore, value and veracity of the data must be taken into consideration when evaluating the proposed methods. Data analytics is not about having the information known, but about discovering the predictive power behind the data collected. Thus, all five criteria of big data must be addressed when designing and developing data cleansing methods.

6. Conclusion

領域專家的參與不可或缺，因為專家在數據清理過程中負責驗證以及確認數據的準確性

Most of the organization depend on the data-driven decision making, thus information system is closely related to the business process management to leverage their processes for competitive advantage. Nowadays, the amount of data keeps increasing, but the quality of the data is decreasing as many of the data collected is dirty. Various data cleansing approaches are available to solve this issue but data cleansing remains as a challenge in order to cope with the criteria of big data. Some of the approaches are not suitable for big data as it has a significant amount of data that need to be processed at a time. Despite the availability of existing frameworks to address data cleansing for big data, but the value and veracity of the data often left out when designing the approaches. Besides, the need for domain expert in undeniable as an expert is needed to verify and validate the data before it can undergo an analysis process.

Acknowledgements

This work was supported by Universiti Sains Malaysia under RUI Grant Scheme (1001/PKOMP/8012206).

References

[1] Rahm, Erhard and Hong Hai Do. (2000) “Data Cleaning: Problems and Current Approaches.” *IEEE Bulletin of the Technical Committee on Data Engineering* **(23)**: 3-13.

[2] Li, Lin. (2012) “Data Quality and Data Cleaning in Database Applications.” [doctoral dissertation], School of Computing, Edinburgh Napier University.

[3] Someswararao, Chinta, J. Rajanikanth, V. Chandra Sekhar, and Bhadri Raju M. S. V. S. (2012) “Data Cleaning: A Framework for Robust Data Quality In Enterprise Data Warehouse.” *International Journal of Computer Science and Technology* **3** (**3**): 36-41.

[4] Saha, Barna, and Divesh Srivastava. (2014) “Data Quality: The other face of Big Data.” in *2014 IEEE 30th International Conference on Data Engineering*. pp. 1294-1297.

[5] Shneiderman, Ben, and Catherine Plaisant. (2015) “Sharpening Analytic Focus to Cope with Big Data Volume and Variety.” *IEEE Computer Graphics and Applications* **35** (**3**): 10-14.

[6] Müller, Heiko, and Johann-Christoph Freytag. (2003) *Problems, Methods, and Challenges in Comprehensive Data Cleansing,* Humboldt University Berlin.

[7] Gu, Randy Siran. (2010) “Data Cleaning Framework: An Extensible Approach to Data Cleaning.” [master’s thesis], University of Illinois, Urbana, Illinois.

[8] Khayyat, Zuhair, Ihab F. Ilyas, Alekh Jindal, Samuel Madden, Mourad Ouzzani, and Paolo Papotti. (2015) “BigDansing: A System for Big Data Cleansing”, in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, Melbourne, Victoria, Australia.

[9] L'Heureux, A., K. Grolinger, H. F. Elyamany, and M. A. M. Capretz. (2017) "Machine Learning with Big Data: Challenges and Approaches." *IEEE Access* (**5**): 7776-7797.

[10] Abdullah, Noraini, Saiful Adli Ismail, Siti Sophiayati, and Suriani Mohd Sam. (2015) "Data Quality in Big Data: A Review." *International Journal of Advance Soft Computing and Its Application* **7** (**3**):16-27.

[11] Hadi, Hiba Jasim, Ammar Hameed Shnain, Sarah Hadishaheed, and Azizah Haji Ahmad. (2015) "Big Data and Five V's Characteristics." *International Journal of Advances in Electronics and Computer Science* **2** (**1**): 16-23.

[12] Crowdflower. (2016) "Data Science Report." *Crowdflower*. Available from: https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf. [Accessed Jan., 28, 2018].

[13] Yakout, Mohamed, Laure Berti-Équille, and Ahmed K. Elmagarmid. (2013) "Don't be SCAREd: use SCalable Automatic REpairing with maximal likelihood and bounded changes", in *the Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, New York, New York, USA.

[14] Cappiello, Cinzia, Walter Samá, and Monica Vitali. (2018) "Quality Awareness for a Successful Big Data Exploitation", in *ACM International Conference Proceeding Series*. pp. 37-44.

[15] Cohen, Bevin, David K. Vawdrey, Jianfang Liu, David Caplan, E. Yoko Furuya, and Frederick W. Mis. (2015) "Challenges Associated With Using Large Data Sets for Quality Assessment and Research in Clinical Settings." *Policy, Politics & Nursing Practice* (**16**): 117-124.

[16] Sidi, Fatimah, Payam Hassany Shariat Panahy, Lilly Affendey, A. Jabar, Marzanah, Hamidah Ibrahim, and Aida Mustapha. (2012) "Data Quality: A Survey of Data Quality Dimensions." *International Conference on Information Retrieval & Knowledge Management*. pp. 300-304.

[17] Sonka, Steven. (2016) "Big Data Characteristics." *International Food and Agribusiness Management Review* **19** (**A**): 7-12.

[18] Wang, Hongzhi, Mingda Li, Yingyi Bu, Jianzhong Li, Hong Gao, and Jiacheng Zhang. (2015) "Cleanix: a Parallel Big Data Cleaning System." *ACM SIGMOD Record* **44** (**4**): 35-40.

[19] Swapnil, Walunj K., Anil H. Yadav, and Sonu Gupta. (2016) "Big Data: Characteristics, Challenges and Data Mining." *International Journal of Computer Applications*: 25-29.

[20] Chu, Xu, John Morcos, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, and Nan Tang. (2015) "KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing", in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, Melbourne, Victoria, Australia.

[21] Wang, Jiannan, Sanjay Krishnan, Michael J. Franklin, Ken Goldberg, Tim Kraska, and Tova Milo. (2014) "A Sample-And-Clean Framework for Fast and Accurate Query Processing on Dirty Data", in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, Snowbird, Utah, USA.

[22] Wang, Hongzhi, Mingda Li, Yingyi Bu, Jianzhong Li, Hong Gao, and Jiacheng Zhang. (2014) "Cleanix: A Big Data Cleaning Parfait", in *the Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, Shanghai, China.

[23] Chu, Xu, John Morcos, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Nan Tang, et al. (2015) "KATARA: Reliable Data Cleaning with Knowledge Bases and Crowdsourcing." *Proceedings of the VLDB Endowment* **8** (**12**): 1952-1955.

[24] Raman, Vijayshankar, and Joseph Hellerstein. (2001) "Potter's Wheel: An Interactive Data Cleaning System", in *Proceedings of the 27th International Conference on Very Large Data Bases*, Roma, Italy.

[25] Lee, Mong Li, Tok Wang Ling, and Wai Lup Low. (2000) "IntelliClean: A Knowledge-based Intelligent Data Cleaner", in *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, Massachusetts, USA.

[26] Chu, Xu. (2017) "Scalable and Holistic Qualitative Data Cleaning." [doctoral dissertation], University of Waterloo.