Journal of Ambient Intelligence and Humanized Computing (2023) 14:99–111
https://doi.org/10.1007/s12652-021-03590-2

ORIGINAL RESEARCH

Data cleansing mechanisms and approaches for big data analytics: a systematic study

Mehdi Hosseinzadeh[1] · Elham Azhir[2] · Omed Hassan Ahmed[3] · Marwan Yassin Ghafour[4] · Sarkar Hasan Ahmed[5] · Amir Masoud Rahmani[6] · Bay Vo[7]

Received: 1 October 2020 / Accepted: 28 October 2021 / Published online: 17 November 2021

Abstract

With the evolution of new technologies, the production of digital data is constantly growing. It is thus necessary to develop data management strategies in order to handle the large-scale datasets. The data gathered through different sources, such as sensor networks, social media, business transactions, etc. is inherently uncertain due to noise, missing values, inconsistencies and other problems that impact the quality of big data analytics. One of the key challenges in this context is to detect and repair dirty data, i.e. data cleansing, and various techniques have been presented to solve this issue. However, to the best of our knowledge, there has not been any comprehensive review of data cleansing techniques for big data analytics. As such, a comprehensive and systematic study on the state-of-the-art mechanisms within the scope of the big data cleansing is done in this survey. Therefore, five categories to review these mechanisms are considered, which are machine learning-based, sample-based, expert-based, rule-based, and framework-based mechanisms. A number of articles are reviewed in each category. Furthermore, this paper denotes the advantages and disadvantages of the chosen data cleansing techniques and discusses the related parameters, comparing them in terms of scalability, efficiency, accuracy, and usability. Finally, some suggestions for further work are provided to improve the big data cleansing mechanisms in the future.

Keywords Data cleansing · Big data · Data quality · Methods · Review

✉ Mehdi Hosseinzadeh
mehdi@gachon.ac.kr

Bay Vo
vd.bay@hutech.edu.vn

[1] Pattern Recognition and Machine Learning Lab, Gachon University, 1342 Seongnamdaero, Sujeonggu, Seongnam 13120, Republic of Korea

[2] Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

[3] Department of Information Technology, College of Science and Technology, University of Human Development, Sulaymaniyah, Iraq

[4] Department of Computer Science, College of Science, University of Halabja, Halabja, Iraq

[5] Network Department, Sulaimani Polytechnic University, Sulaymaniyah, Iraq

[6] Future Technology Research Center, National Yunlin University of Science and Technology, Douliou, Yunlin 64002, Taiwan

[7] Faculty of Information Technology, Ho Chi Minh City University of Technology (HUTECH), Ho Chi Minh City, Vietnam

1 Introduction

Nowadays, an enormous amount of data is being produced at unprecedented scales through heterogeneous sources like smartphones, social networks, sensors, logs, etc. (Klein 2017; Oussous et al. 2018). This is due to the growing trend of many technologies such as the Internet of Things (IoT) (Khorshed et al. 2015; Romero et al. 2016), cloud computing, and smart devices. The fast data growth creates novel opportunities for businesses, and this data explosion has led to a new term, big data (Manyika et al. 2011; Gantz and Reinsel 2012; Chang and Grady 2015). In this context, organizations need to collect, store and analyze data for strategic business decision-making, leading to the collection of valuable knowledge. However, there are substantial difficulties to gain valuable knowledge from huge volumes of data. The data has to be managed properly as inaccurate data may lead to wrong decisions. Data cleansing process tries to enhance the quality of data by detecting and repairing mistakes.

Incomplete, incorrect, inaccurate, and irrelevant records in a database will cause uncertainties during data analysis, and these have to be handled within the data cleansing process (Ridzuan and Zainon 2019). The purpose of data cleansing is to remove (correct) the errors, resolve inconsistencies, and convert the data into a uniform format to achieve accurate data collection. Due to the enormous amount of data, manual cleansing takes a long time and is prone to errors, and traditional data cleansing systems cannot be scaled very easily. Moreover, big data is generated from various sources and in multiple formats. As such, various mistakes such as missing values, duplication, and inconsistent values may occur within the huge volumes of data, and traditional data cleansing techniques are not appropriate for cleansing such dirty data (Ridzuan and Zainon 2019).

There are some literatures reviews on big data cleansing techniques (Chu et al. 2016; Hariharakrishnan et al. 2017; Jesmeen et al. 2018; Hariri et al. 2019; Ridzuan and Zainon 2019). For example, the major challenges faced in cleansing big data were discussed in (Jesmeen et al. 2018; Ridzuan and Zainon 2019). The authors presented a basic overview of various cleansing techniques for big data analytics. Ridzuan and Zainon (2019) categorized the investigated techniques based on traditional data cleansing and data cleansing for big data. However, only a few papers were discussed. Furthermore, studying the future works, study selection process, and an analysis of the related qualitative parameters have some limitations. Artificial intelligence (AI) techniques offer faster, more accurate, and scalable results in big data analytics as compared to traditional techniques. A detailed classification of big data cleansing methods also presented based on the machine learning algorithms in (Chu et al. 2016; Hariharakrishnan et al. 2017; Jesmeen et al. 2018; Hariri et al. 2019). Chu et al. (2016) reviewed various machine learning techniques and statistical methods for the qualitative repairing and cleansing process. Furthermore, emerging trends in data cleansing research were also highlighted in this survey. However, the paper doesn't include the newly proposed big data cleansing mechanisms. Furthermore, the article choosing mechanism is not defined and the qualitative parameters should have been checked in these papers. Also, different data mining techniques were compared by Hariharakrishnan et al. (2017). The authors presented a basic overview of several methods for data cleansing, such as filters, imputation, and wrappers is in this paper. This paper reviewed the efficiency of these data cleansing techniques in terms of blank spaces and missing values. Moreover, the features of the presented methods were investigated in this paper. However, few studies were addressed in their survey and the mechanism of choosing studies should be in the center of attention. Furthermore, Hariri et al. (2019) reviewed various AI techniques with regard to handling uncertainties in big data analytics. This paper focuses only on AI techniques for big data cleansing. Also, there are some gaps in studying the qualitative parameters, and article selection process. Finally, in (2018), the authors reviewed and compared various data quality management tools to clean messy data. Data quality issues which may occur in big data processing alongside data quality criteria were also discussed in this survey paper. But, few papers were reviewed, the article selection mechanism in not offered, and the qualitative parameters were not investigated.

This brief overview shows that a complete study of the existing data cleansing methods for big data together with their categorization were not provided in these studies. The limitations of the previous studies can be summarized as follows: the newly proposed mechanisms, especially from 2018, 2019, and 2020, are not provided; an organized structure is not presented; a logical taxonomy is not provided in most of these papers; only a few articles are studied in a number of these papers; the qualitative parameters are not provided; and directions for future works have not been discussed well. This survey is thus a first attempt to completely and systematically investigate the data cleansing problem in the context of big data. In this survey, the existing research on big data cleansing is classified into five main categories, including machine learning-based, sample-based, expert-based, rule-based and framework-based mechanisms. The big data cleansing techniques are described along with the strengths and weaknesses of each technique. The four most significant qualitative parameters are defined to assess each data cleansing method and compare it with previous methods with a view to finding the best data cleansing algorithm, as follows:

- **Scalability**: The data cleansing mechanism's ability to detect and repair dirty data often involves processing large and different types of datasets without compromising data quality.
- **Efficiency**: It means the ratio of the method to the overall time and cost need.
- **Data quality and accuracy**: This is detected with various parameters like data incompleteness, out of range values, obsolete data, and values in the wrong field.
- **Usability**: Does the mechanism allow ease-of-use via a user-friendly programming interface?

The main contributions of this paper are as follows:

1. Giving a brief description of the problems in big data cleansing.
2. Presenting a systematic review of the present data cleansing techniques and their constraints with relation to big data.

3. Providing a classification for crucial methods in the field of big data cleansing.
4. Outlining important areas for enhancing the use of data cleansing methods for big data analytics within future research.

The rest of this paper is structured as follows: the background will be discussed in Sect. 2. The mechanisms for selecting papers are introduced in Sect. 3. The intended classification for the chosen big data cleansing studies is described in Sect. 4. The investigated studies will be compared and discussed in Sects. 5 and 6. Finally, some open issues and the conclusion are provided in Sects. 7 and 8, respectively.

2 Background

Data cleansing is an essential task for creating quality data analytics. The need for cleansing data increases along with the size of big data, as the sources might have problems such as incomplete data, redundant data or incompatible data formats. The iterative process of data cleansing is shown in Fig. 1.

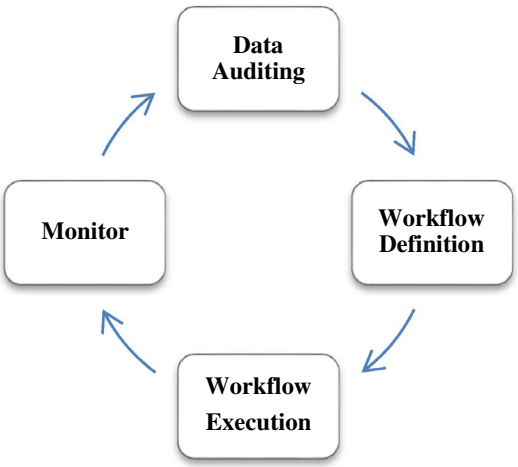


Fig. 1 Data cleansing process (Müller and Freytag 2005; Ridzuan and Zainon 2019)

The auditing step is used to evaluate the data to discover the kinds of anomalies that reduce the data quality. Next, detection and elimination of anomalies is executed by a series of tasks on the data within the data cleansing workflow definition step. Appropriate methods are selected to detect and remove the existing anomalies. The correctness and effectiveness of the workflow is additionally examined and evaluated within the second step. The third step is the workflow execution stage. The data cleansing workflow is applied to the tuples in the data collection after specification and verification of its correctness. After executing the cleansing workflow, the outcomes are inspected to again verify the correctness of the required operations. In the monitor step, we examine the results and carry out exception handling for the tuples not corrected within the actual processing.

Data scientists spend most of their time cleansing datasets with huge data volumes (Ridzuan and Zainon 2019). This can be a labor-intensive and time-consuming process, and because of big data's characteristics, the complexity of the data cleansing algorithm will increase (Cappiello et al. 2018).

3 Research methodology

This paper will use the systematic literature review (SLR) guidelines proposed by Kitchenham (2004) to study the current literature associated with big data cleansing mechanisms. Conducting a comprehensive and thorough search of the literature for appropriate articles is the main step to provide a systematic review. Also, a criteria-based selection of relevant studies is performed based on the search factors such as search terms and the publication year (Zhang and He 2021).

The article selection process is shown in Fig. 2. In the first stage, Google Scholar and electronic databases such as ACM, ScienceDirect, Springerlink, and IEEExplore are used to find the primary studies based on the keyword search. Selecting the appropriate search keywords to identify related papers from journals and conference papers is important in SLR. The following keywords are searched for the period 2010–2021:

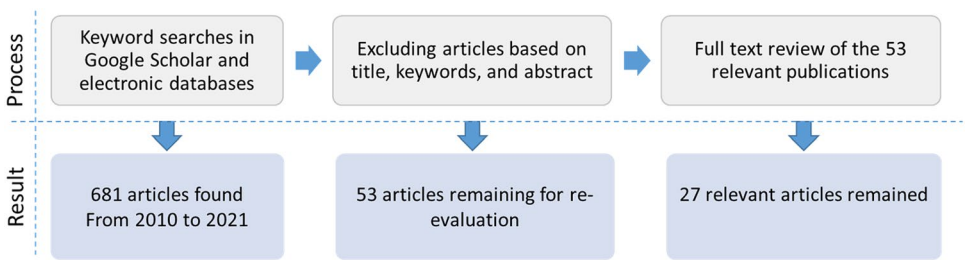


Fig. 2 The process of choosing relevant studies

- Data cleaning
- Data cleansing
- Big data cleaning
- Big data cleansing
- Dirty data cleansing
- Dirty data cleaning
- Entity resolution
- Data deduplication
- Data quality
- Big data quality

Masters theses and doctoral dissertations, book chapters, and non-English articles were excluded from the study. Therefore, we found 681 articles using this keyword search strategy. Some study selection criteria are considered within the next steps to select high-quality studies. Titles, abstracts, and keywords were reviewed to choose the articles for the next stage of the process. Therefore, 53 articles remained for re-evaluation. In stage 3, a review of the full text of the chosen studies from the second stage was performed to validate the relevance of these studies. Finally, a total of 27 articles were identified and reviewed.

Figure 3 shows the distribution of the chosen articles by publication dates and their publishers, including Springer, ACM, Elsevier, IEEE, and Scientific Research. The chosen publications showing a trend of growth since 2016.

The distribution of the articles by different publishers is shown in Fig. 4. The highest number of articles is related to ACM with 11 articles (41%) and the lowest number of articles is related to Scientific Research with one article (4%).

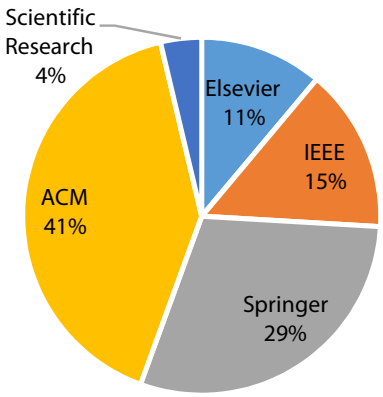


Fig. 4 Percentage of the number of articles in each electronic database

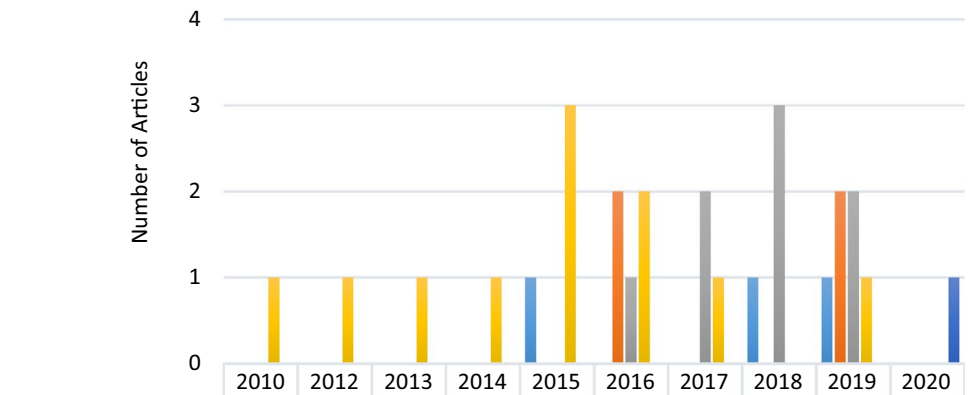
4 Review of big data cleansing mechanisms

A total of 27 articles will thus be investigated during this part, and their benefits and drawbacks will be described. Figure 5 indicates the taxonomy of the data cleansing techniques. The proposed taxonomy has five categories, including machine learning-based, sample-based, expert-based, rule-based, and framework-based techniques. Furthermore, the studies are compared regarding the scalability, efficiency, accuracy, and usability as explained in Sect. 1.

4.1 Machine learning-based mechanisms

Traditional preprocessing techniques cannot be applied efficiently to large datasets. Current machine learning

Fig. 3 Distribution of selected studies by publication dates and publishers



	2010	2012	2013	2014	2015	2016	2017	2018	2019	2020
Elsevier					1			1	1	
IEEE						2			2	
Springer						1	2	3	2	
ACM	1	1	1	1	3	2	1		1	
Scientific Research										1

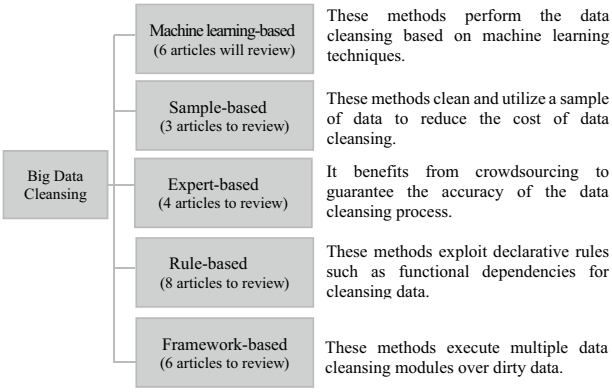


Fig. 5 Big data cleansing techniques taxonomy

techniques help to create models for predicting, thus enabling an automated system. These techniques aim to enhance the data cleaner by integrating machine learning techniques. The selected machine learning-based mechanisms are described in the following paragraphs.

Yakout et al. (2013) presented a novel approach, named SCARE (SCalable Automatic REpairing), combining machine learning and likelihood techniques to cleanse a dirty dataset. It learns the correlations from the correct data to predict the most appropriate replacement values. SCARE relies upon the probabilistic principles to produce predictions for more than one attribute at a time. The scalability and accuracy of replacement values are the two main advantages of the proposed approach. Scalability is guaranteed using horizontal data partitioning. Furthermore, SCARE also uses data partitioning for parallel computing. SCARE can scale well in terms of computational expense.

Mayfield et al. (2010) developed ERACER, an iterative statistical framework to infer the information of missing values and correct such errors automatically. The proposed method is based on the belief propagation and relationship dependence network. SQL and user- defined functions are applied to develop ERACER in standard database management systems (DBMS). In ERACER, a shrinkage technique has been developed to execute the inference and cleansing tasks in an integrated way. The authors did several trials to assess the functionality of the proposed ERACER method. The outcomes show that ERACER can provide more accurate inferences than the baseline statistical method using Bayesian networks. This mechanism offers high efficiency and scalability according to the experimental results.

Data deduplication is the task of identifying duplicate records to improve data quality. Kolb et al. (2012) have presented a data cleansing tool, named Dedoop (Deduplication with Hadoop), to remove all duplicates in large datasets. Dedoop benefits from machine learning techniques for the automatic generation of match classifiers. Hadoop-MapReduce is a scalable and distributed processing engine in the cloud environment. The authors used the map-reduce technique to parallelize the data deduplication operations, which can decrease the execution time. Dedoop also supports several advanced load balancing strategies to achieve high efficiency.

Label noise is an important issue in classification problems. However, traditional noise removal approaches have problems coping with a large quantity of data. It is thus necessary to develop new noise removal methods for big data problems. García-Gil et al. (2019) presented a novel big data preprocessing framework based on Apache Spark to remove noisy data composed of two ensemble methods. The first is a single base learning algorithm (Random Forest), named Homogeneous Ensemble for Big Data (HME-BD). The second is a heterogeneous ensemble, namely Heterogeneous Ensemble for Big Data (HTE-BD), that employs three types of classifiers: Random Forest, Logistic Regression and K-Nearest Neighbors (KNN). The results show that the presented framework is able to get a clean dataset more efficiently. The results also indicate that the performance and scalability of the proposed framework is satisfactory for removing noise from any big data classification problem.

De et al. (2016) introduced a new system, BayesWipe, for modifying individual attribute values in relational databases. The presented method uses both Bayesian generative and statistical error models. BayesWipe, employs an error model to automatically identify the data mistakes and repair them without clean master data. The Bayesian inference assigns probabilities to the various possible clean replacement values. To perform the large-scale data cleansing so as to meet the scalability, BayesWipe is implemented using a two-stage MapReduce model. The experiments validate the efficiency and quality of the BayesWipe method.

Oussous et al. (2018) proposed a configurable approach, Rule Assembler, to classify duplicate records on the basis of confidence scores computed by logical rules. Various heuristics are employed to turn the results produced with the aid of logical rules into confidence scores. The proposed approach estimates the confidence scores concerning the duplicity of record pairs. The calculated scores are then combined by an aggregation function. Finally, the aggregated confidence scores are used to classify similar records. The results show that the Rule Assembler improves the quality of the data compared to state-of-the-art methods. Moreover, the proposed Rule Assembler has high efficiency and a lower time of entity resolution compared to the basic assembly approach.

Table 1 illustrates the summary of the reviewed techniques along with their major strengths and weaknesses.

Table 1 Summary of the data cleansing in machine learning-based mechanisms

Paper	Technique	Advantage	Weakness
Yakout et al. (2013)	Learn the correlations from the correct data using machine learning and likelihood methods	• High scalability • High accuracy • High efficiency	• Low usability
Mayfield et al. (2010)	A statistical framework to infer the missing data and correct such errors automatically	• High scalability • High accuracy • High efficiency	• Low usability
Kolb et al. (2012)	Automatic generation of match classifiers using machine learning techniques	• High scalability • High efficiency	• Low usability • Low accuracy
García-Gil et al. (2019)	Ensemble methods for noise elimination	• High scalability • High efficiency	• Low usability • High complexity • Low accuracy
De et al. (2016)	Correct the attribute values using Bayesian generative and statistical error models	• High scalability • High accuracy • High efficiency	• Low usability • High complexity
Oussous et al. (2018)	Duplicate record detection using configurable assembly of classification rules	• High accuracy • High efficiency	• Low usability • Low scalability

4.2 Sample-based mechanisms

In emerging big data scenarios, it is difficult to get high-quality data within a reasonable time frame. Sample-based methods have been developed to address the problem of data volume, when the estimated results might be valid. The selected mechanisms regarding sample-based data cleansing are discussed as follows.

Salloum et al. (2019) proposed a sampling-based method for exploring and cleansing huge datasets using small computing clusters. The random sample partition (RSP) is a data model represents a data set as a collection of disjoint distributed partitions of data, named RSP blocks. The RSP blocks are employed to perform statistical inference and estimation, detection of errors, and data cleansing tasks. Henceforth, approximate outcome for the whole data set is calculated with a particular confidence interval. The experimental results reveal that the estimated outcomes from RSP-Explore converges quickly to true values.

Wang et al. (2014) proposed a novel framework, named SampleClean, for efficient and accurate processing of queries on large datasets. In their proposed framework, the cleansing process is executed on a small subset of the data. The cleaned sample is then been used to process aggregate queries. The experiments show that the query outcomes of the proposed algorithm are identical to the outcomes of the entire clean dataset. The results also confirm the time efficiency of the presented method.

The existing cleansing methods require lots of completely clean data to train the predictive model. However, it is impractical to obtain this. Ding and Qin (2018) thus proposed a new model updating algorithm. The algorithm combines the data cleansing method with the conjugate gradient algorithm. Instead of cleaning the entire data, the initial model is updated incrementally by cleaning samples at each iteration. Furthermore, the authors presented a cluster descent sampling algorithm for model convergence acceleration. The trial outcomes reveal the accuracy of the model compared with that of the training raw data directly in the model. However, it suffers from high time demands and complexity.

The selected sample-based mechanisms are reviewed in this part. The advantages, and disadvantages of the investigated techniques are shown in Table 2.

Table 2 Summary of the data cleansing in sample-based mechanisms

Paper	Technique	Advantage	Weakness
Salloum et al. (2019)	Clean noisy data based on RSP data model	• High scalability • High accuracy • High efficiency	• Low usability
Wang et al. (2014)	Applied the cleansing process on a small subset of the data for accurate and fast query processing	• High accuracy • High efficiency	• Low usability • Low scalability
Ding and Qin (2018)	Interactive data cleansing	• High scalability • High accuracy	• Low efficiency • High complexity

4.3 Expert-based mechanisms

Data cleansing is the process of identifying and repairing inaccurate records from a dataset. There are many data cleansing techniques to enhance data quality. However, it is difficult to control the accuracy of the data cleansing process without verifying it via experts. The data cleansing mechanisms that use crowdsourcing platforms are discussed below.

Duplicate data can have a negative effect on different areas of an organization's operations. Saberi et al. (2019) proposed a crowdsourcing-based framework to help in duplicate detection, named DedupCrowd. A statistical quality control method is developed to control the performance of the crowd, and members with poor performance may be evicted from the crowdsourcing process. The proposed system evaluates the quality of the crowd's work to guarantee the quality of the automated process.

Social media data needs to be curated before being used for deeper analytics. A data curation pipeline, named CrowdCorrect, was proposed by Beheshti et al. (2018) that exploits the hybrid combination of machine and human-driven functionality to curate and clean social media data. The procedures of automatic feature extraction, correction and enrichment are performed in the first step. The method next uses task-based crowdsourcing platforms to recognize and correct the unmodified items of the first step. The knowledge of experts is employed in the last step to recognize and correct the unmodified items of two previous steps. The results show that CrowdCorrect significantly enhances the quality of extracted knowledge in comparison with the classical curation pipelines.

Chu et al. (2015) proposed a novel end-to-end big data cleansing system, named KATARA, which uses crowdsourcing and trustworthy knowledge-bases (KBs) to discover correct and wrong data tuples and produces probable repairs for the latter. Crowdsourced workers define the target table and the reference KB. First, KATARA finds out the table patterns to map the table to a KB. Next, data tuples will be interpreted as correct or wrong by cooperation between the KB and humans. For the wrong tuples, the top-k possible mapping is obtained from the KB and analyzed by domain experts. This method tries to generate accurate repairs by relying on KBs and humans.

Evaluating the accuracy of the algorithms used in the data cleansing process is an important issue. Ilyas (2016) introduced an effective data cleansing life-cycle with continuous evaluation. The article proposes a decoupling between error-detection and error-correction within a continuous data cleansing life-cycle with humans in the loop. Domain experts are involved in evaluating the effectiveness of the cleansing algorithms at the report and analytics level.

The advantages and disadvantages of the studied expert-based data cleansing techniques are provided in Table 3.

4.4 Rule-based mechanisms

Rule-based mechanisms are designed to resolve data inconsistencies problems by enforcing data quality rules (i.e., integrity constraints). The resolution of data inconsistencies focuses on the detection of violations of data quality rules. The newly generated data conforms to the given rules. Eight selected rule-based mechanisms are discussed here, as follows.

Wang et al. (2016) presented a parallel big data cleansing system, named Cleanix, to handle mixed errors. Cleanix supports four types of data cleansing tasks, namely detection and correction of abnormal values, fill-in the missing data, elimination of duplicate records, and conflict resolution. These data repairing tasks are performed in parallel. Cleanix combines data from various sources and cleans them on a shared-nothing architecture. Scalability and unification are the two main features of Cleanix. Furthermore, it does not need any domain expert due to its easy-to-use graphical user interface (GUI). Users are enabled to define their cleansing rules to resolve the mistakes by using a web interface.

Table 3 Summary of the data cleansing in expert-based mechanisms

Paper	Technique	Advantage	Weakness
Saberi et al. (2019)	Monitor the performance of the crowd in duplicate detection	• High accuracy	• Low usability • Low scalability • Low efficiency
Beheshti et al. (2018)	Exploits the hybrid combination of machine and human-driven functionality to curate and clean social media data	• High accuracy	• Low usability • Low scalability • Low efficiency
Chu et al. (2015)	Uses knowledge-bases and crowdsourcing to recognize correct and wrong data and produces top-k possible repairs for the wrong data	• High accuracy	• Low usability • Low scalability • Low efficiency
Ilyas (2016)	An effective crowd-based data cleaning life-cycle with continuous evaluation	• High accuracy	• Low usability • Low scalability • Low efficiency

Khayyat et al. (2015) introduced BigDansing, a rule-based big data cleansing system, to adjust dataflows for error detection. BigDansing provides a user-friendly interface that enables users to define the logic data quality rules. The logical plans are later translated into optimal physical plans. It thus supports a wide range of quality rules by abstracting the process of rule specification, and high efficiency is achieved through some physical optimizations. Furthermore, by supporting the scalability of current parallel computing frameworks, BigDansing can be scaled to large datasets.

Data quality rules have also been used to describe correct data instances. Satish and Kavya (2017) introduced a novel big data cleansing method based on data quality rules with high accuracy. In this regard, functional and inclusion dependencies are exploited to detect and repair data quality problems. The authors developed a novel hybrid algorithm, Cuckoo Search Optimization with Gravitational Search algorithm (CSO-GSA), which is developed to efficiently identify and correct the errors in the records received from data sources before delivering them. Some data quality factors, such as consistency, currency, deduplication and completeness, are considered as objective functions in this work. The outcomes reveal the time efficiency and accuracy of the proposed method for error detection and correction in large datasets.

Functional dependencies (FD) are widely used in data cleansing. Currently, several tools are available for rule mining. However, existing discovery methods have usually ignored the time dimension. Abedjan et al. (2015) proposed a new system for the discovery of approximate temporal functional dependencies over noisy web data. The system uses association measures and outlier detection methods for the discovery of the rules, along with an aggressive repair of the data within the mining step itself. The results indicate that high-quality data with lower execution time can be generated using the proposed system.

Ding and Cao (2016) presented a data cleansing method for large spatio-temporal data. The temporal range is find out by using time-based clustering with no previous knowledge, which ensures consistent timestamps. Moreover, invalid records are found and corrected using a rule-based filtering module, which guarantees legal spatio-temporal relationships. The presented method is implemented using the map-reduce programming model. The results indicate the efficiency of the parallelization of the proposed method. Furthermore, accuracy and high scalability are two benefits of the method.

Mezzanzanica et al. (2015) presented an iterative domain-independent technique called Multidimensional Robust Data Quality Analysis to enhance quality of data. The data quality tasks of the knowledge discovery in database (KDD) process can be represented as a model checking problem. The consistency check of the proposed technique is implemented by using the UPMurphi model checking tool. The cleansing process improvement is iteratively achieved by recognizing the issues to be handled. The results show the accuracy of the proposed model-based technique. Furthermore, it facilitates the management of constraints through the model design. However, it has disadvantages such as high time and wasteful consumption of resources.

Some irrelevant data can be identified and removed using the data cleansing operation to organize data for analytical processes. Martinez-Mosquera et al. (2017) proposed a comparative data cleansing technique to discard a number of fields from the security log files before their analysis. The proposed technique is based on Fellegi-Sunter probabilistic approach to compare the original log files and final user reports. Furthermore, the authors use Levenshtein Distance to recognize similar strings (Fellegi and Sunter 1969; Luján-Mora and Palomar 2001a, b). However, despite the simplicity of the proposed method, efficient automated selection of matched and non-matched records is considered as a key challenge for future work.

Data cleansing is also applicable in medical imaging, where huge quantities of data need to be processed in order to extract valuable information. For example, Godinho et al. (2019) proposed an extract, transform, load (ETL) framework for analyzing clinical records. The proposed framework improves the quality of the repositories by using a rule system and essential data cleansing functionalities. The charts and reports can be customized and less experienced users can use it easily. It can also index medical data across distributed repositories. Therefore, low time and high scalability are two benefits of the proposed method.

The strengths and weaknesses of the studied rule-based techniques are provided in Table 4.

4.5 Framework-based mechanisms

The framework-based mechanisms are based on a combination of cleansing tasks to improve the data quality and reduce redundancy tasks during the process of data cleansing. In this section, some framework-based mechanisms will be reviewed.

NoSQL is a novel technology that is developed to address the limitations of relational databases like unlimited scalability and continuous availability. Ramzan et al. (2019) proposed an approach to migrate the data from relational to NoSQL databases. This consists of two main modules, called data transformation and data cleansing. The aim of data transformation is to utilize model transformation techniques to transform a relational database into an Oracle NoSQL database. Next, quality data may be generated through the data cleansing phase. It identifies the errors using clustering and duplicate detection techniques, and then applies the

Table 4 Summary of the data cleansing in rule-based mechanisms

Paper	Technique	Advantage	Weakness
Wang et al. (2016)	A parallel rule-based big data cleansing system that supports four types of data cleaning tasks	• High scalability • High efficiency • High usability	• –
Khayyat et al. (2015)	A rule-based big data cleansing system, to adjust dataflows for error detection	• High scalability • High efficiency • High usability	• –
Satish and Kavya (2017)	Uses functional and inclusion dependencies to detect and repair data quality problems	• High accuracy • High efficiency	• Low scalability • Low usability
Abedjan et al. (2015)	Temporal rules discovery for web data cleaning	• High accuracy • High efficiency	• Low scalability • Low usability
Ding and Cao (2016)	Rule-based filtering for massive spatio-temporal data cleaning	• High scalability • High efficiency • High accuracy	• Low usability
Mezzanzanica et al. (2015)	Model-based evaluation of data quality activities	• High accuracy • High usability	• Low efficiency
Martinez-Mosquera et al. (2017)	Fellegi-Sunter theory for record linkage	• High usability	• Low efficiency • Low accuracy • Low scalability
Godinho et al. (2019)	An ETL framework over medical imaging repositories	• High usability • High efficiency • High scalability	• –

corrections. High accuracy and efficiency are considered as the main advantages of this approach.

Tae et al. (2019) also developed a novel whole data quality framework, named MLClean. The proposed framework consists of data cleansing, unfairness mitigation, and data sanitization. The data sanitization and cleansing components are executed together, followed by the mitigation component. Machine learning techniques are also included in MLClean. The results show the efficiency of MLClean in preprocessing due to eliminating redundant operations on the data. However, it has some disadvantages such as low scalability.

Wang et al. (2017) proposed a MapReduce-based system, named CleanCloud, for cleaning big data in the cloud. CleanCloud provides multiple types of data cleansing methods, including record linkage, data imputation, and rule-based error discovery. The system is offered as a service on the cloud. The cleansing tasks work on different machines in parallel to reduce the execution time, and scalability, efficiency and a user-friendly interface are considered as the main features of CleanCloud.

MapReduce-based frameworks are often developed to perform data cleansing with high scalability. However, because of the lack of effective design of MapReduce-based frameworks, there is some redundant tasks within the data cleansing procedure, which leads to lower performance, and some operations may be conducted multiple times on the same files. Lian et al. (2020) proposed a novel data cleansing technique that is based on merging redundant computations. The results show that the system runtime is substantially reduced. Moreover, fill-in the missing data, entity recognition, and restoration of inconsistent data modules of data cleansing are also improved in this paper.

Liu et al. (2016) presented a big data cleansing framework to enhance data quality. Three kinds of data errors are considered in this paper: value error, condition error, and missing value. The framework seamlessly unifies the association and repairing operations and supports context patterns, usage patterns, metadata, and repairing rules. The proposed system produces a data association report that may benefit the later data cleansing process. The results reveal the effectiveness of the proposed framework in terms of data quality.

Wahyudi et al. (2018) introduced a new model based on process patterns to enhance data quality. A proven series of activities are included within the proposed process model to enhance the quality of data considering a particular context, a certain objective, and a selected set of initial conditions. In this paper, a number of patterns are proposed to solve the big data quality issues. However, the patterns are derived based on a single case study in a specific context, and the method also leads to more complex flowcharts and models that need more analysis time.

A side-by-side comparison of the framework-based techniques along with their main advantage and weakness are revealed in Table 5.

Table 5 Summary of the data cleansing in framework-based mechanisms

Paper	Technique	Advantage	Weakness
Ramzan et al. (2019)	A novel approach for migration to NoSQL databases	• High accuracy • High efficiency	• Low usability
Tae et al. (2019)	A novel data quality framework consists of data cleaning, unfairness mitigation, and data sanitization modules	• High efficiency	• Low usability • Low scalability
Wang et al. (2017)	A MapReduce-based system provides multiple types of data cleansing tasks	• High scalability • High efficiency • High usability	• –
Lian et al. (2020)	Uses task merging to improve big data cleansing	• High efficiency • High scalability	• Low usability
Liu et al. (2016)	A big data cleansing framework supported context patterns, usage patterns, metadata, and repairing rules	• High accuracy • High scalability	• Low efficiency • Low usability
Wahyudi et al. (2018)	A big data cleansing framework supporting context patterns	• High accuracy • High usability	• Low scalability • Low efficiency

5 Results and comparison

As described in the previous sections, machine learning-based, sample-based, expert-based, rule-based, and framework-based are five different types of big data cleansing mechanisms. The machine learning-based mechanisms use learning mechanisms to adapt the automated decisions. These mechanisms have high complexity and overhead compared to other approaches. They also suffer from a high possibility of error. The sample-based mechanisms focus on cleaning and using the small subset of data to obtain the accurate results. These methods focus on efficiency. The expert-based mechanisms enhance the data quality using expert knowledge. The expert-based mechanisms have high accuracy. The rule-based mechanisms attempt to clean dirty datasets by matching against the rules of the knowledge base. The key benefit of rule-based mechanisms is their relative simplicity of development. Although coverage for diverse scenarios is lower, whatever scenarios are covered by the rule-based systems will provide high accuracy. The framework-based mechanisms have high flexibility, and can have more advantages because they use two or more algorithms in combination.

We assess the parameters which have an effect on the big data cleansing approaches. Table 6 gives the key features of the reviewed data cleansing techniques, which include scalability, efficiency, data quality and accuracy, and usability. The highest advantage is revealed by three stars and 1 star indicates the lowest capability. As can be seen in Fig. 6, the previous researchers have emphasized on scalability in 25% of the studies, efficiency in 28%, accuracy in 29%, and usability in 18%. The usability has the lowest score and accuracy has the highest score according to Table 6; Fig. 6. The accuracy has thus attracted more attention in the literature.

6 Discussion

This paper aims to investigate data cleansing in big data. Therefore, five categories are considered to review these mechanisms, which are machine learning-based, sample-based, expert-based, rule-based, and framework-based mechanisms. A total of 27 articles were identified and reviewed.

In the review of machine learning-based mechanisms, it can be seen that this method focused on solving noisy data problems by using machine learning techniques. The important factors that have improved with all of the machine learning-based techniques are efficiency and scalability. However, model complexity and probability of error are two main disadvantages of these approaches. The sample-based mechanisms address the problem of data volume for cleansing huge datasets. The key parameter that has improved with all of the sample-based techniques is efficiency. The expert-based mechanisms require human intervention to control the accuracy of the data cleansing process. The expert-based data cleansing techniques have high accuracy, but also have limitations in efficiency and scalability. The rule-based mechanisms require a set of rules that have been defined by experts. The major problems of rule-based mechanisms are the problems encountered during knowledge acquisition, and adaptability. Moreover, these techniques cannot be applied for huge amounts of data. Also, the framework-based mechanisms have high flexibility, and can have more advantages because they are based on a combination of two or more algorithms.

From the literature, accuracy, efficiency, and scalability are three main issues considered by the authors when developing the data cleansing methods to deal with the volume and variety of data. However, less attention is given to address the usability problem. Therefore, extensive data profiling and cleansing, easy data mapping, and automation of

Table 6 An overview of the reviewed techniques and their main features scored from 1 (lowest) to 3 (highest) stars for scalability, efficiency, accuracy and usability

Paper	Scalability	Efficiency	Accuracy	Usability
Yakout et al. (2013)	***	***	***	*
Mayfield et al. (2010)	***	***	***	*
Kolb et al. (2012)	***	***	*	*
García-Gil et al. (2019)	***	***	*	*
De et al. (2016)	***	***	**	*
Oussous et al. (2018)	*	***	***	*
Salloum et al. (2019)	***	***	***	*
Wang et al. (2014)	*	***	***	*
Ding and Qin (2018)	***	*	***	*
Saberi et al. (2019)	*	*	***	*
Beheshti et al. (2018)	*	*	***	*
Chu et al. (2015)	*	*	***	*
Ilyas (2016)	*	*	***	*
Wang et al. (2016)	***	***	**	***
Khayyat et al. (2015)	***	***	**	***
Satish and Kavya (2017)	*	***	***	*
Abedjan et al. (2015)	*	***	***	*
Ding and Cao (2016)	***	***	**	*
Mezzanzanica et al. (2015)	**	*	***	***
Martinez-Mosquera et al. (2017)	*	*	*	***
Godinho et al. (2019)	***	***	**	***
Ramzan et al. (2019)	**	***	***	*
Tae et al. (2019)	*	***	**	*
Wang et al. (2017)	***	***	**	***
Lian et al. (2020)	***	***	**	*
Liu et al. (2016)	***	*	***	*
Wahyudi et al. (2018)	*	*	***	***

*indicates the lowest capability; **indicates the medium capability; ***indicates the highest capability

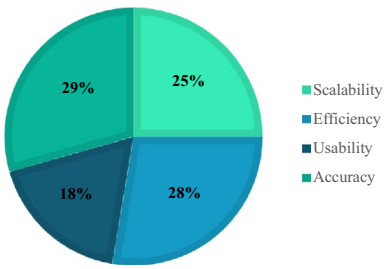


Fig. 6 Parameters considered in the selected articles

data cleansing workflow are key features that should be considered to solve usability problems in data cleansing tools.

7 Open issues

The data gathered in this paper explain the state-of-the-art in the field of big data cleansing mechanisms and approaches. However, these techniques have some limitations and challenges such as insufficient training and validation data, real-time data, huge volume, and unstructured data. This part highlights many key issues that are necessary to explore in future work.

7.1 Vast amount of unstructured data

Big data contains large amounts of semi-structured and unstructured data. Processing of large amounts of unstructured, incomplete, inconsistent, and imprecise data is a critical issue in the process of big data. Further research is desired with a focus on data quality problems for unstructured, and semi-structured data formats. Furthermore, this data cannot be processed by traditional data cleansing methods. Therefore, developing new artificial intelligence methods to clean big data is also needed in future works, which could apply novel methods such as Convolutional Neural Networks (CNNs) (LeCun and Bengio 1995; LeCun et al. 1998).

7.2 Data streams

Big data contains a large volume of data collected from social media, sensors, mobile devices, and the like. Various

techniques must be implemented to clean such large amounts of such data in real-time. The study of data cleansing techniques for handling distributed streams of data is very interesting for future studies.

7.3 Qualitative parameters and metrics

As reviewed in this article, numerous methods used to clean big data. The authors used different qualitative attributes to validate the proposed techniques. However, the study of data cleansing using the same real-world datasets, with the same methods and the same trial infrastructure and their evaluation using different quality attributes is very interesting. Also, reviewing the techniques in this study has revealed that big data cleansing requires to be extended to account for more qualitative parameters. For instance, accuracy and efficiency are considered in some mechanisms, while parameters such as scalability are ignored.

8 Conclusions

Nowadays, the ever growing size of data makes it necessary to further develop the process of data cleansing. Various data cleansing mechanisms are developed to resolve this issue, and this survey presents a systematic review of big data cleansing mechanisms. First, the data cleansing process is overviewed. Next, the issues of big data cleansing are discussed. This paper focused on data cleansing techniques which use humans, constraints, rules, or patterns to detect errors. Therefore, 27 studies are selected which are classified into five main groups, including machine learning-based, sample-based, expert-based, rule-based, and framework-based mechanisms. The advantages and disadvantages of each of these mechanisms have been investigated. The machine learning-based mechanisms improve data quality by learning from their mistakes but have high complexity in most of the reviewed studies. The sample-based mechanisms are used to improve the data cleansing efficiency while enhancing data quality. The expert-based mechanisms are mostly designed based on expert knowledge. The rule-based mechanisms need to define data quality rules for error detection. Finally, the framework-based mechanisms include different cleaning tasks such as detecting and eliminating data duplication.

The data collected in this survey helps to explain the state-of-the-art in the field of big data cleansing. This paper tries to conduct a detailed systematic review, but also has some limitations. It fails to study the data cleansing techniques that are available in various sources. Moreover, the studies which are not in the context of big data are not explored. However, the results will help researchers to develop more effective data cleansing methods in big data environments.

Funding None.

Data availability All data and results are reported in the paper.

Declarations

Conflict of interest There is no conflict of interest among authors.

Ethical approval The submitted work is original and has not been published elsewhere in any form or language.

Informed consent None.

References

Abedjan Z, Akcora CG, Ouzzani M, Papotti P, Stonebraker M (2015) Temporal rules discovery for web data cleaning. Proc VLDB Endow 9(4):336–347

Beheshti A, Vaghani K, Benatallah B, Tabebordbar A (2018) CrowdCorrect: a curation pipeline for social data cleansing and curation. International conference on advanced information systems engineering. Springer, Cham, pp 24–38

Cappiello C, Samá W, Vitali M (2018) Quality awareness for a successful big data exploitation. In: Proceedings of the 22nd International Database Engineering & Applications Symposium, pp 37-44

Chang WL, Grady N (2015) NIST big data interoperability framework: volume 1, big data definitions. No. special publication (NIST SP)-1500-1

Chu X, Morcos J, Ilyas IF, Ouzzani M, Papotti P, Tang N, Ye Y (2015) KATARA: reliable data cleaning with knowledge bases and crowdsourcing. Proc VLDB Endow 8(12):1952–1955

Chu X, Ilyas IF, Krishnan S, Wang J (2016) Data cleaning: overview and emerging challenges. In: Proceedings of the 2016 International Conference on Management of Data, pp 2201-2206

De S, Hu Y, Meduri VV, Chen Y, Kambhampati S (2016) Bayeswipe: a scalable probabilistic framework for improving data quality. J Data Informn Qual (JDIQ) 8(1):1–30

Ding W, Cao Y (2016) A data cleaning method on massive spatio-temporal data. In: Proceedings of the Asia-Pacific Services Computing Conference, pp 173-182

Ding X, Qin S (2018) Iteratively modeling based cleansing interactively samples of big data. In: International Conference on Cloud Computing and Security, pp 601-612

Fellegi IP, Sunter AB (1969) A theory for record linkage. J Am Stat Assoc 64(328):1183–1210

Gantz J, Reinsel D (2012) The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. IDC iView 2007:1–16

García-Gil D, Luengo J, García S, Herrera F (2019) Enabling smart data: noise filtering in big data classification. Inf Sci 479:135–152

Godinho TM, Lebre R, Almeida JR, Costa C (2019) Etl framework for real-time business intelligence over medical imaging repositories. J Digit Imaging 32(5):870–879

Hariharakrishnan J, Mohanavalli S, Kumar KS (2017) Survey of preprocessing techniques for mining big data. In: 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), pp 1-5

Hariri RH, Fredericks EM, Bowers KM (2019) Uncertainty in big data analytics: survey, opportunities, and challenges. J Big Data 6(1):44

Ilyas IF (2016) Effective data cleaning with continuous evaluation. IEEE Data Eng Bull 39(2):38–46

Jesmeen M, Hossen J, Sayeed S, Ho C, Tawsif K, Rahman A, Arif E (2018) A survey on cleaning dirty data using machine learning paradigm for big data analytics. Indones J Electr Eng Comput Sci 10(3):1234–1243

Khayyat Z, Ilyas IF, Jindal A, Madden S, Ouzzani M, Papotti P, Quiané-Ruiz J-A, Tang N, Yin S (2015) Bigdansing: a system for big data cleansing. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp 1215-1230

Khorshed MT, Sharma NA, Kumar K, Prasad M, Ali AS, Xiang Y (2015) Integrating internet-of-things with the power of cloud computing and the intelligence of big data analytics—a three layered approach. In: 2015 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), pp 1-8

Kitchenham B (2004) Procedures for performing systematic reviews, vol 33. Keele, UK, pp 1–26

Klein S (2017) The world of big data and IoT. IoT solutions in Microsoft's azure IoT suite. Springer, New York, pp 3–13

Kolb L, Thor A, Rahm E (2012) Dedoop: efficient deduplication with hadoop. Proc VLDB Endow 5(12):1878–1881

LeCun Y, Bengio Y (1995) The handbook of brain theory and neural networks. Convolutional networks for images, speech, and time series. MIT press, Cambridge

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proce IEEE 86(11):2278–2324

Lian F, Fu M, Ju X (2020) An improvement of data cleaning method for grain big data processing using task merging. J Comput Commun 8(3):1–19

Liu H, Tk AK, Thomas JP, Hou X (2016) Cleaning framework for bigdata: an interactive approach for data cleaning. In: Proceedings of IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), pp 174-181

Luján-Mora S, Palomar M (2001a) Comparing string similarity measures for reducing inconsistency in integrating data from different sources. In: International Conference on Web-Age Information Management, pp 191-202

Luján-Mora S, Palomar M (2001b) Reducing inconsistency in integrating data from different sources. In: Proceedings 2001b International Database Engineering and Applications Symposium, pp 209-218

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Hung Byers A (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute, New York

Martinez-Mosquera D, Luján-Mora S, López G, Santos L (2017) Data cleaning technique for security logs based on Fellegi-Sunter theory. EuroSymposium on systems analysis and design. Springer, Cham, pp 3–12

Mayfield C, Neville J, Prabhakar S (2010) ERACER: a database approach for statistical inference and data cleaning. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pp 75-86

Mezzanzanica M, Boselli R, Cesarini M, Mercorio F (2015) A model-based evaluation of data quality activities in KDD. Inf Process Manag 51(2):144–166

Müller H, Freytag J-C (2005) Problems, methods, and challenges in comprehensive data cleansing. Professoren des Inst. Für Informatik

Oussous A, Benjelloun F-Z, Lahcen AA, Belfkih S (2018) Big data technologies: a survey. J King Saud Univ-Comput Inform Sci 30(4):431–448

Ramzan S, Bajwa IS, Ramzan B, Anwar W (2019) Intelligent data engineering for migration to NoSQL based secure environments. IEEE Access 7:69042–69057

Ridzuan F, Zainon WMNW (2019) A review on data cleansing methods for big data. Procedia Comput Sci 161:731–738

Romero CDG, Barriga JKD, Molano JIR (2016) Big data meaning in the architecture of IoT for smart cities. In: International Conference on Data Mining and Big Data, pp 457-465

Saberi M, Hussain OK, Chang E (2019) Quality management of workers in an in-house crowdsourcing-based framework for deduplication of organizations' databases. IEEE Access 7:90715–90730

Salloum S, Huang JZ, He Y (2019) Exploring and cleaning big data with random sample data blocks. J Big Data 6(1):45

Satish KR, Kavya N (2017) Hybrid optimization in big data: error detection and data repairing by big data cleaning using CSO-GSA. In: Proceedings of the International Cnference on Cognitive Computing and Information Processing, pp 258-273

Tae KH, Roh Y, Oh YH, Kim H, Whang SE (2019) Data cleaning for accurate, fair, and robust models: a big data-AI integration approach. In: Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning, pp 1-4

Wahyudi A, Kuk G, Janssen M (2018) A process pattern model for tackling and improving big data quality. Inform Syst Front 20(3):457–469

Wang J, Krishnan S, Franklin MJ, Goldberg K, Kraska T, Milo T (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, pp 469-480

Wang H, Li M, Bu Y, Li J, Gao H, Zhang J (2016) Cleanix: a parallel big data cleaning system. ACM SIGMOD Rec 44(4):35–40

Wang H, Ding X, Chen X, Li J, Gao H (2017) CleanCloud: cleaning big data on cloud. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp 2543-2546

Yakout M, Berti-Équille L, Elmagarmid AK (2013) Don't be scared: use scalable automatic repairing with maximal likelihood and bounded changes. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, pp 553-564

Zhang G, He B-J (2021) Towards green roof implementation: drivers, motivations, barriers and recommendations. Urban For Urban Green 58:126992

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com