A Review of Data Cleaning Methods for Web Information System

Jinlin Wang[1], Xing Wang[1, *], Yuchen Yang[1], Hongli Zhang[1] and Binxing Fang[1]

Abstract: Web information system (WIS) is frequently-used and indispensable in daily social life. WIS provides information services in many scenarios, such as electronic commerce, communities, and edutainment. Data cleaning plays an essential role in various WIS scenarios to improve the quality of data service. In this paper, we present a review of the state-of-the-art methods for data cleaning in WIS. According to the characteristics of data cleaning, we extract the critical elements of WIS, such as interactive objects, application scenarios, and core technology, to classify the existing works. Then, after elaborating and analyzing each category, we summarize the descriptions and challenges of data cleaning methods with sub-elements such as data & user interaction, data quality rule, model, crowdsourcing, and privacy preservation. Finally, we analyze various types of problems and provide suggestions for future research on data cleaning in WIS from the technology and interactive perspective.

Keywords: Data cleaning, web information system, data quality rule, crowdsourcing, privacy preservation.

1 Introduction

Digital is the media content of the network in the era of extensive data, which exists as a form of database in WIS. In general, a data-intensive information system that users can access through a web browser can be considered as WIS. The database produces a large number of entities in WIS and integrates the activities of the data business to form related reports that support essential business decisions. The errors induced by data are often inevitable due to a variety of reasons while the data is processed. These errors often lead to mistakes in the application business reporting, which harm business decisions. Therefore, it is essential to maintain the quality of the database in WIS, and the concept of keeping high-quality data is called data cleaning [Ganti and Sarma (2013)].

Data cleaning, also called data cleansing or scrubbing, which means detecting and removing errors or inconsistencies from data to improve the quality of data. Data quality problems are present in single data collections, such as files and databases. When multiple data sources need to integrate, the need for data cleaning increases significantly because the sources often contain redundant data in different representations. The consolidation of different data representations and the elimination of duplicate

[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150006, China.

[*] Corresponding Author: Xing Wang. Email: wxhit@hit.edu.cn.

CMC. doi:10.32604/cmc.2020.08675

information become necessary to provide access to accurate and consistent data.

Data collection has become a ubiquitous function of large organizations' WIS shown in Fig. 1. It not only can be used to keep records but also supports a variety of data analysis tasks that are critical to organizational tasks [Hellerstein (2008)]. The results of the analysis might be severely distorted due to the presence of incorrect or inconsistent data, and it usually negates the potential benefits of the information-driven approach. We typically refer to the data that triggers this condition as dirty data. Dirty information becomes a key factor affecting the decision-making and analysis. Isolating, processing, and reusing dirty data in time become the meaning of data cleaning.

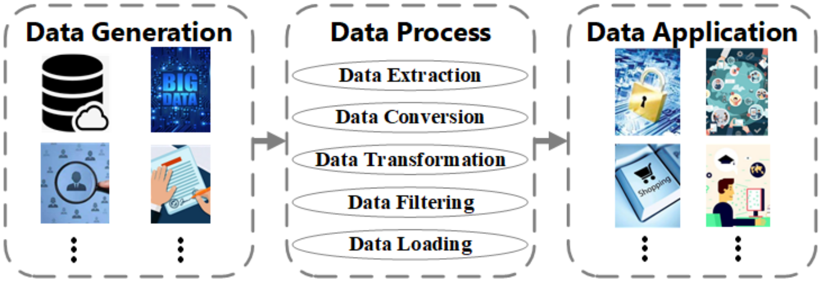

Figure 1: Aspects of WIS data cleaning procedure

In this paper, we provide a survey of data cleaning methods in WIS. These methods in recent works are classified according to the interaction objects, core technologies, and application scenarios in WIS. After discussing the advantages and limitations, we look ahead to the research questions and challenges of data cleaning in WIS and put forward recommendations on future research in this field.

This paper is organized an architectural description as follows. Section 2 introduces an overview of data cleaning in WIS, while a detailed analysis of existing data cleaning methods is presented in Section 3. Section 4 presents the research challenges of existing methods and provides suggestions for future research. Finally, the conclusions are presented in Section 5.

2 Overview of data cleaning in WIS

The presentation characterizes WIS to a broad audience of a large amount of data. In addition to a large amount of data, there exist several features of WIS, such as heterogeneous sources and high immediacy requirements. According to the methods about the application of WIS, we summarized several critical elements in the design and applicate of WIS.

2.1 Situation of data cleaning in WIS

With the development of web-based information service, WIS is widely used in the field of electronic commerce, communication, edutainment and entertainment, identity presentation, information services. Maintaining a high level of data quality is an essential topic in WIS research. Most present studies divide the WIS architecture into a logical and conceptual level based on business logic. The logical level leads to the database, and the conceptual level leads to the interface.

2.2 Critical elements of data cleaning in WIS

We have extracted the critical elements of data cleaning methods in WIS after summarizing and analyzing different research questions in the field of data cleaning, shown in Fig. 2.

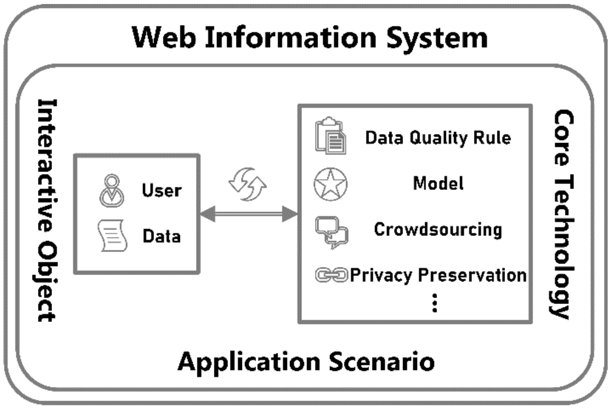

Figure 2: Critical elements of data cleaning methods in WIS

2.2.1 Application scenario

Data cleaning covers various types of business logic data as a necessary step in the data processing of WIS. Data cleaning depends on the different application characteristics in specific application scenarios, which makes data cleaning as an applied science with strong universality and versatility. The expected effect of data cleaning can be achieved according to local conditions in different application scenarios. In the scenarios of WIS, such as electronic commerce, community, education, entertainment, information services, the application effect of data cleaning performs excellent.

2.2.2 Interactive object

There is no doubt that the direct role of data cleaning in WIS is data. It is necessary to adjust the strategy according to the statistical characteristic of data and the users' demand indication in the process of cleaning the datasets. We refer to the process of adjustment as interacting and call the carrier that interacts with the data cleaning scheme as an interactive object. Data and user constitute the two essential parts of data cleaning interactive object.

2.2.3 Core technology

It is difficult to use a specific class of methods for unified processing since the WIS environment is complex, and data cleaning requirements are often in a wide variety. Core technology often embodies the irreplaceability of scientific methods.

Data cleaning uses models and algorithms to ensure the cleaning results and help users perform business decisions in WIS. Different core technologies often represent different research perspectives. Data quality rule, model, and crowdsourcing represent the most popular three typical research methods.

Table 1: Summary of data cleaning methods in WIS: nonfunctional requirements

Critical Element	Category	Reference	Scalability	Availability	Popularity	Privacy	Service-based	Distributed
Interactive Object	Data Interaction	[Berti-Equille, Dasu and Srivastava (2011)]	PL	S	M	NS	NI	N
		[Geerts, Mecca and Papotti et al. (2013)]	PL	S	H	NS	N	N
	User Interaction	[Fan, Li, Ma et al. (2011)]	AL	S	M	NS	Y	NI
		[He, Veltri, Santoro et al. (2016)]	AL	NI	L	NI	N	N
		[Bergman, Milo, Novgorodov et al. (2015)]	AL	S	M	NI	NI	N
		[Raman and Hellerstein (2001)]	AL	S	H	NS	NI	Y
Application Scenario	RFID and Sensor	[Jeffery, Alonso, Franklin et al. (2006)]	PL	S	H	NI	NI	NI
		[Jeffery, Garofalakis, Franklin et al. (2006)]	AL	S	H	NI	NI	NI
	ETL Process	[Bhattacharjee, Chatterjee, Shaw et al. (2014)]	PL	NI	L	NS	N	NI
		[Lettner, Stumptner and Bokesch (2014)	PL	NI	L	NS	NI	Y
	Other Scenarios	[Qu, Wang, Wang et al. (2016)]	AL	S	L	NS	NI	Y
		[Ong and Ismail (2014)]	AL	S	L	NS	NI	N
Core Technology	Data Quality Rule	[Chu, Ilyas and Papotti (2013)]	PL	S	H	NS	NI	NI
		[Dallachiesa, Ebaid, Eldawy et al. (2013)]	PL	S	H	NS	NI	Y
	Model	[Krishnan, Wang, Franklin et al. (2016)]	PL	S	L	NI	N	NI
		[Mayfield, Neville and Prabhakar (2010)]	PL	S	H	NI	NI	NI
		[Krishnan, Wang, Franklin et al. (2015)]	AL	S	M	NI	N	Y
	Crowdsourcing	[Tong, Cao, Zhang et al. (2014)]	AL	NI	M	N	Y	NI
		[Chu, Morcos, Ilyas et al. (2015)]	PL	S	H	N	N	Y
		[Haas, Krishnan, Wang et al. (2015)]	AL	NI	M	NI	Y	Y
	Privacy Preservation	[Ge, Ilyas, He et al. (2017)]	PL	S	L	S	Y	N
		[Krishnan, Wang, Franklin et al. (2016)]	PL	S	L	S	Y	N
		[Dong, Liu and Wang (2014)]	PL	S	L	S	Y	NI
		Legend	Processing Layer (PL) Application Layer (AL)	Supported (S) Not Supported (NS) No Information (NI)	Highly (H) Moderately (M) Less (L)	Supported (S) Not Supported (NS) No Information (NI)	Yes (Y) Not (N) No Information (NI)	Yes (Y) Not (N) No Information (NI)

3 Existing data cleaning methods in WIS

In this chapter, we classify the existing data cleaning methods based on the critical elements of the WIS mentioned above, shown in Tabs. 1-2.

3.1 Application scenario

According to existing research [Schewe and Thalheim (2019)], WIS has four main application scenarios: electronic commerce, communities and group, edutainment, and information services. In this paper, we respectively enumerate the main scenario of application.

Table 2: Summary of data cleaning methods in WIS: features and results

Critical Element	Category	Reference	Extractive features	Effect results
Interactive Object	Data Interaction	[Berti-Equille, Dasu and Srivastava (2011)]	**(1)** For definition, detection, and cleaning of complex, multi-type data glitches **(2)** A statistically rigorous methodology for evaluating and scoring glitches	DEC-PD constitutes 61% of the BQC strategies with Resource-Driven methods for the top 30% of glitches contributing 33%
		[Geerts, Mecca and Papotti et al. (2013)]	**(1)** A uniform framework **(2)** Propose a new semantics for repairs, and a chase-based algorithm	LLUNATIC-FR-S1 cost manager scales nicely and had better performances than some of the main memory implementations
	User Interaction	[Fan, Li, Ma et al. (2011)]	**(1)** Based on master data, editing rules, and certain regions. **(2)** Demonstrate a list of facilities	Unspecified
		[He, Veltri, Santoro et al. (2016)]	**(1)** Use SQL update queries as the language to repair data **(2)** Describe novel multi-hop search algorithms	Experiments confirm that Falcon can recover from these errors, at the price of more user interactions
		[Bergman, Milo, Novgorodov et al. (2015)]	**(1)** A novel query-oriented system for cleaning data with oracles **(2)** Present heuristic algorithms that interact with oracle crowds	90% of the errors are fixed within another hour, and the whole experiment completed within 3.5 hours, identifying all errors
		[Raman and Hellerstein (2001)]	**(1)** Integrate transformation and discrepancy detection tightly **(2)** Infer structures for data values in terms of user-defined domains automatically	Unspecified
Application Scenario	RFID and Sensor	[Jeffery, Alonso, Franklin et al. (2006)]	**(1)** Produced by physical receptor devices. **(2)** A declarative query processing tool with a pipelined design	ESP can correctly indicate that a person is in the room 92% of the time
		[Jeffery, Garofalakis, Franklin et al. (2006)]	**(1)** The first declarative, adaptive smoothing filter for RFID data cleaning. **(2)** Adapt the smoothing window size in a principled manner	SMURF is substantially easier to deploy and maintain and provide more reliable data
	ETL Process	[Bhattacharjee, Chatterjee, Shaw et al. (2014)]	**(1)** Check metadata in addition to various existing data cleaning algorithm **(2)** Emphasize on the citizen database system to make it errorless	After all steps are done, approximately 3% error remained on the table
		[Lettner, Stumptner and Bokesch (2014)]	**(1)** Correct corrupted data (semi-)automatically according to user-defined rules **(2)** Can be attached to an ETL process by defining "snapshot points"	Unspecified
	Other Scenarios	[Qu, Wang, Wang et al. (2016)]	**(1)** Based on Spark **(2)** Use exponential weighting moving mean value	The experimental result shows that the accuracy of identification can reach above 90%
		[Ong and Ismail (2014)]	**(1)** Create a weblog cleaning algorithm for web intrusion detection **(2)** Based on these five weblog attributes	The proposed algorithm managed to clean up 153372 entries which carried a percentage of reduction 40.41
Core Technology	Data Quality Rule	[Chu, Ilyas and Papotti (2013)]	**(1)** Let users specify quality rules using denial constraints with ad-hoc predicates **(2)** Exploit the interaction of the heterogeneous constraints	The holistic approach outperforms previous algorithms in terms of quality and efficiency of the repair
		[Dallachiesa, Ebaid, Eldawy et al. (2013)]	**(1)** Allow the users to specify multiple types of data quality rules **(2)** Allow cleaning algorithms to cope with multiple rules holistically	NADEEF can achieve better accuracy than existing methods
	Model	[Krishnan, Wang, Franklin et al. (2016)]	**(1)** Support convex loss models **(2)** Prioritize cleaning records likely to affect the results	ActiveClean can improve model accuracy by up-to 2.5x for the same amount of data cleaned
		[Mayfield, Neville and Prabhakar (2010)]	**(1)** Based on belief propagation and relational dependency networks **(2)** Include an efficient approximate inference algorithm	ERACER achieves accuracy comparable to a baseline statistical method using Bayesian networks with exact inference

	[Krishnan, Wang, Franklin et al. (2015)]	**(1)** Use the clean sample to estimate aggregate query results **(2)** Explore an outlier indexing technique	SVC is applicable for a wide variety of materialized views with high accuracy of 99%
Crowdsourcing	[Tong, Cao, Zhang et al. (2014)]	**(1)** For cleaning multi-version data on the Web **(2)** Blend active and passive crowdsourcing methods	Unspecified
	[Chu, Morcos, Ilyas et al. (2015)]	**(1)** A knowledgebase and crowd-powered data cleaning system **(2)** Generate top-k possible repairs for incorrect data	Katara usually achieves higher precision due to its use of knowledge bases and experts
	[Haas, Krishnan, Wang et al. (2015)]	**(1)** Support the iterative development and optimization of data cleaning workflows **(2)** Separate logical operations from physical implementations	Unspecified
Privacy Preservation	[Ge, Ilyas, He et al. (2017)]	**(1)** Allow data cleaning workflows while ensuring differential privacy **(2)** The privacy engine translates each query into a differentially private mechanism	DPClean can achieve high cleaning quality while ensuring a reasonable privacy loss.
	[Krishnan, Wang, Franklin et al. (2016)]	**(1)** A technique for creating private datasets of numerical and discrete-valued attributes **(2)** Maintain a bipartite graph relating dirty values to clean values	PrivateClean can be inverted to select maximal privacy levels given some constraint on query accuracy
	[Dong, Liu and Wang (2014)]	**(1)** Focus on data deduplication as the primary data cleaning task **(2)** Design two efficient privacy-preserving data-deduplication methods	The precision varies around 75%, and the recall is around 80%

3.1.1 Electronic commerce

The web information system in electronic commerce mainly has two purposes. The first one is managing the purchase and sale records. With the help of web information system, centralized management can be realized in large-scale electronic commerce. The second purpose is offering a better user experience for customers. Using the web information system can extremely promote the efficiency of customer service.

3.1.2 Communities and group

Communities and groups are designed to gather registered members who share the same interest, demands, or conviction. Commonly, communities are built on sharing experience, interests, or state of mind with other members. The primary function of WISs is to provide identity service, including but not limited to the identity recognition, the interest record, membership management.

3.1.3 Edutainment and entertainment

The principal intention of entertainment WISs is to provide fun for the users in the shape of games, videos. Edutainment is a compound word. It is also called educational entertainment, which means media designed to educate through entertainment. This kind of media includes content intended to educated but has incidental value. Edutainment is applied to audio and video, film and television, healthcare. Since the early 1990s, there has been a surge in interest in developing entertainment software, and WIS has begun to be used in the entertainment industry [Okan (2003)].

3.1.4 Information services

Information services is the most significant and crucial application scenario. The information services system is designed to provide information of all kinds, accommodation, news, social services. Under normal circumstance, the intentions of web-based information system is designed by the audiences of websites. That means the observing to audiences of the websites and user profiles are indispensable to the design of information services systems. In standard scenarios, the user enter the sites, search for information in need, and shut down the system. However, in realistic circumstances, the expectations of users can be complicated and unpredictable, which turns the system building a tough task.

3.2 Interactive object

3.2.1 Data interaction

It is an interesting research problem to integrate large-scale multi-source heterogeneous data. In 2015, Liu et al. [Liu, Kumar and Thomas (2015)] researched this problem. The work aims to identify identical or similar objects in WIS, link these associated objects, and then effectively clear and combine data. The method proposes structural and descriptive metadata for datasets and can identify the correlated data items and sets in the subsequent data cleaning process. The work determines the relations among objects according to the relation between data context and user mode. The experiment proves that an effective multi-source linkage can be structured through data context and user mode.

Volkovs et al. [Volkovs, Chiang, Szlichta et al. (2014)] designed a continuous data cleaning framework that could be applied in the dynamic data and constraint environment in 2014. The continuous data cleaning framework put forward adapts to naturally evolved data and constraints, can be applied in the dynamic data environment. This framework helps users better respond to inconsistency between data and constraints and minimize the problem of error spreading. Based on data and constraints as evidences, the method considers the repair behaviors that users selected and applied. Through experiment evaluation, the technology of the method achieved high prediction precision and high-quality repair effect. The work applied a series of data statistics display methods to predict the validity of specific repair types in WIS.

Quantitative data cleaning (QDC) measures, quantifies, and corrects data quality problems with the statistics and other analysis technologies. At present, QDC methods are exclusively used to solve separate category of data errors. However, different types of errors often occur concurrently in intricate forms. To solve the disadvantages of the existing QDC method, Bertiequille et al. [Berti-Equille, Dasu and Srivastava (2011)] presented DEC, an iterative framework which is oriented to exploring and cleaning complex data faults. The work defines different types of complex faults, develops heuristic data driving strategies, and applies error patterns and joint distribution to make quantitative data cleaning. The system completes the link detection and cleaning process through iteration, expert feedback, and dealing with correlated problems to solve some disadvantages of the current method. The system selects the best cleaning strategy and display effect for a candidate data set according to a strict statistical basis. QDC demonstrates the accuracy and scalability of the method and seems more effective than traditional strategies.

In the field of data cleaning, few universal algorithms are available to solve database repair problems which involve and contain different constraints and select the preferred value strategy. Geerts et al. [Geerts, Mecca, Papotti et al. (2013)] explained a uniform framework to solve the problem, a DBMS based prototype framework named LLUNATIC. LLUNATIC pays special attention to the implementation and execution of scalable programs in the parallel environment, which quickly generates a vast amount of repair targets. The paper forms a higher- level universality while being compatible with previous data cleaning methods. The system classifies the critical problem of data cleaning, namely, the problem of complexity of weight options and repair algorithms in the data quality solutions. Finally, the experiment also proves the excellent scalability and application effect of the system in terms of data cleaning algorithm through accumulating the previous industry experience.

3.2.2 User interaction

Fan et al. [Fan, Li, Ma et al. (2011)] built a data cleaning system named CerFix, which corrected the tuple determinately when data inputs. This method is based on master data, editing rules, and determined region. After verifying the attributes of given input tuples, the editing rules give other attributes to repair data correctly. The work defines the determinate regions as a group of attributes and ensures to repair the entire tuple after verification. Cerfix verifies the determined area in the WIS and monitors the data for the determined repair of the input tuple. It verifies the repair of the minimum quantity of attributes through guiding users and audits and displays the repaired attributes and sources of correct values.

He et al. [He, Veltri, Santoro et al. (2016)] proposed Falcon, which is an interactive data cleaning system and uses SQL update query as the repair language. The system is independent of the pre-defined data quality rules. On the contrary, Falcon encourages users to explore and solve possible problems in the process of data identification. Through guiding user update, the system finds the possible SQL update query, which can be used to repair data. The work plans to convert problems into search problems in lattice space, and further trims the search space and effectively maintains the lattice by applying novel search strategies and some optimization technologies. The experiment proves that Falcon can effectively communicate with users.

Many automatic data cleaning tools have been developed to solve the problem of inconsistency of a database. Bergman et al. [Bergman, Milo, Novgorodov et al. (2015)] provided QOCO in 2015, an Oracle-based data query and cleaning system. QOCO can remove incorrect query result tuples through editing the underlying database, which is accomplished by the communication between experts, which are transformed by Oracle Cloud. The work argues and demonstrates that the problem of the minimum interaction of Oracle Cloud to remove incorrect query result tuples is an NP-hard problem. QOCO implements a heuristic algorithm which interacts with Oracle Cloud-based on it.

Raman et al. [Raman and Hellerstein (2001)] introduced Potter’s Wheel, an interactive data cleaning system which is oriented to integration transformation and difference detection in WIS. Potter’s Wheel allows users to establish transformation and precise data by adding transformation when differences are detected. Potter’s Wheel automatically infers the data value structure according to the user-defined domains and examines whether there exists the situation of constraint violation based on it. To analyze character strings with the structure of user-defined domains makes Potter's Wheel form universal and scalable different detection mechanism. The defined domains provide a strong foundation for execution of decomposition and transformation via example values.

3.3 Core technology

3.3.1 Regulation of data quality

Data quality rules need to combine substantial and useful researches. Previous work focused on specific research forms, such as function dependency (FD), conditional function dependency (CFDs), and matching dependency (MD) relation. Many of them remain on the level of data flow processing and are unrelated to each other. Responding to this situation, Chu et al. explained a uniform framework [Chu, Ilyas and Papotti (2013)]. This framework makes users apply individual predicates to designate negative quality constraints and codes the interaction of various constraints in the hypergraph to show the conflicts. The conflict repair method, which is oriented to the overall global view, allows the framework to compute the automatic repair quality better. The method outperforms previous algorithms in terms of repair quality and efficiency through experiments on real datasets in WIS.

Bohannon et al. [Bohannon, Fan, Geerts et al. (2007)] proposed a kind of constraint named conditional function dependency and studied its application in data cleaning. Unlike traditional function dependency (TFD), CFDs aims to capture data consistency through binding semantic relativity. The Armstrong Axiom is used to infer and make consistency analysis on FD to prove the validity of CFD. As CFD allows data binding, there are a lot of independent constraints in datasheets, which increases the difficulty in detecting constraint violations. The paper proves constraints of CFD on the theoretical level and infers the constraint-based method of improving data quality.

Recently as the declarative rule of data cleaning and entity resolution, the matching dependency relation has been introduced. In cases where the values are sufficiently similar, it is necessary to identify the value of a particular attribute in two tuples to perform a matching dependency of the database instance. It is assumed that there are matching functions that make two attribute values equal. Bertossi et al. [Bertossi, Kolahi and Lakshmanan (2013)] used the cleaning instance process of the matching dependency relation. The work introduces the lattice structure into the attribute domain by applying the matching function and provides the order distribution with semantic allocation on the level of instance. The work is theoretically complete, makes an extending study on clean instances in the method of polynomial-time approximation, and computes more efficient and accurate approximate value for complex conditions.

Data cleaning technologies are often dependent on some quality rules to identify violation tuples, and then use some repair algorithms to repair these violations in WIS. The rules relating to business logic can be defined in specific target reports generated by transforming several data sources. These reports need to be updated repeatedly whenever data source get changed. This situation quickly makes violations detected in the reports delink with actual error sources spatially and temporally. In this case, repair reports become helpless to prevent target violations. Chalamalla et al. [Chalamalla, Ilyas, Ouzzani et al. (2014)] proposed a system to solve the coupling problem, and the system defines the quality rules in the output transformer and estimates the explanation of output errors. The system describes these errors on the level of target and executes solving measures on the level of source. Meanwhile, the method provides scalable technical detection of data errors. The work uses the TPC-H benchmark to verify the validity of the system from different scenes and quality rule categories.

Regarding to the automatic monitoring and the end-to-end solution to repairing violations of various and individual data quality constraints, Dallachiesa et al. [Dallachiesa, Ebaid, Eldawy et al. (2013)] presented NADEEF, a scalable and generalized and easily deployed data cleaning platform. The system contains a programming interface and core to realize universality and scalability. The programming interface allows users to designate various types of data quality rules and define data error uniformly. Then the interface realizes data repair through editing codes based on the realization of pre-defined categories. Many types of data quality rules can be expressed by the programming interface of NADEEF, such as function dependency (FD), matching dependency, and ETL rules. The system realizes the user interface in the form of the black box, and the core provides corresponding algorithms to detect and difference wrong and clean data. This work verifies the validity of the NADEEF experimentally through real datasets.

3.3.2 Model

Krishnan et al. [Krishnan, Wang, Franklin et al. (2016)] proposed a progressive data cleaning method of ActiveClean. The method applies the convex loss models (e.g., SVM, linear regression), and preferentially processes the records which influence results through building the user model structure. Progressive data cleaning is highly sensitive to sample size and error sparsity. The key to the ActiveClean prediction model is the convex loss model, which can be trained and cleaned simultaneously. Therefore, the convergence and error scope can be effectively guaranteed. The experimental results show that the optimization can significantly lower data cleaning costs when errors are sparse, and the cleaning budget is small. ActiveClean has more precise effects than active sampling and active learning model. In the same year, Krishnan et al. [Krishnan, Haas, Franklin et al. (2016)] raised another three main questions in data cleaning problems in both technical and organizational. The three questions are the iterative nature of data cleaning, the lack of rigor in evaluating the correctness of data cleaning, and the disconnect between the analysts and the infrastructure engineers who design the cleaning pipelines. Based on this, they conclude by presenting several recommendations for future work and envision an interactive data cleaning system that accounts for the observed challenges.

In the application programs, some missing values cannot be explicitly expressed but potential valid data values. Hua et al. [Hua and Pei (2007)] expressed the missing values as disguised missing data, which severely damage the data analysis quality. The missing values also cause deviations and missing results in the subsequent assumption tests, correlation analysis, and regression analysis. A model is built for the disguised missing data distribution and presents the heuristic algorithm with embedded unbiased samples to solve the issue of hidden missing data cleaning. An effective and efficient method is developed to identify commonly used disguised values and capture the subjects of disguised missing data. The experiments of real WIS datasets prove that the research method is authentic and valid. The work processes large-scale datasets in an effective and scalable manner.

Though the integrity constraint and other safety measures have been incorporated into DBMS, real databases often contain syntax and semantic errors. Mayfield et al. [Mayfield, Neville and Prabhakar (2010)] provided ERACER, an iterative statistical system framework which infers the information of missing values and automatically detects the category of errors. The method is based on the faith transmission and relationship dependence network. SQLs and user-defined functions are used in the standard DBMS to realize an effective approximate inference algorithm. The system executes reasoning and cleaning tasks in an integrated way and accurately infers correct values with the contraction technology under the circumstance with dirty data. The work applies multiple combinations and real datasets to evaluate the framework method, and the framework realizes the precision and baseline statistical method and makes precision reasoning with the Bayes network. The framework enjoys more extensive applicability than the Bayes network considering the complex circular dependency relations.

The materialized view is used to store pre-computed results and widely applied in promoting the rapid query of large-scale datasets from WIS. When new records are added efficiently, the batch-to-batch errors cause the materialized view to be increasingly outdated. Then the errors lead to an increasing quantity of errors, disappearances, and redundant records, which influence the query results. Krishnan et al. [Krishnan, Wang, Franklin et al. (2015)] provided SVC to solve the issue from the perspective of data cleaning. SVC can effectively eliminate the records from the outdated materialized view and uses clean samples to estimate the overall query results. This work considers the issue that outdated view is maintained in form as data cleaning. SVC builds a model to obtain accurate query results in the sampling data cleaning method, which significantly lowers the data computation cost. The author applies the TPC-D standard and a real video distribution application program to generate data sets and uses the evaluation system method. The experiment verifies that cleaning materialized views has higher efficiency and more accurate results than maintaining complete views.

3.3.3 Crowdsourcing

Mozafari et al. [Mozafari, Sarkar, Franklin et al. (2014)], designed the integrated machine learning algorithm with a combination of crowdsourcing database and the algorithm combines the accuracy of human labels and the speed and cost-benefit of machine learning classifier. The active learning method, as the optimization strategy for the crowdsourcing database labeling task, can reduce the problem scale through package and allows crowdsourcing application to label bigger datasets. The active learning algorithm of the crowdsourcing database requires universality, scalability, and usability. The work completes designing the algorithm by using the non-parameter theoretical derivation and verifies the validity of the method through testing the datasets of Amazon and UCI.

Through some network information, integration systems try to keep the latest updated version. The multi-version data often contain wrong and invalid information due to the errors in data integration or update delay. Tong et al. [Tong, Cao, Zhang et al. (2014)] provided the CrowdCleaner system, an intelligent data cleaning system based on the passive crowdsourcing strategy. This system finds wrong versions according to the package error reports and assigns tasks to artificial intelligence task manager. CrowdCleaner identifies which package gives the right answers in a practical packaging method and discovers the most reliable users according to the answer records. The system combines active and passive crowdsourcing methods to repair multi-version data errors.

Chu et al. [Chu, Morcos, Ilyas et al. (2015)] presented Katara, a data cleaning system based on knowledge bases and crowdsourcing. Katara gives tables, knowledge bases, interpretive tables, and semantic databases to identify whether the data are correct and generates the most probable incorrect datasets for reference. Firstly, through discovering and verifying the model of the datasheet, Katara builds the relationship between the dirty database and feasible knowledge base in WIS. Then, all tuples from databases are verified in the tablet mode. The work experimentally shows that Katara can be applied in various types of datasets and knowledge bases in order to efficiently clean data and provide accurate results.

Haas et al. [Haas, Krishnan, Wang et al. (2015)] proposed Wisteria, a data cleaning system aiming to support iterative development and process optimization. The system realizes the separation of logical computation on the physical level and makes optimization according to the selected physical implementation after analyst feedback. The work outlines the system structure and key technologies and puts forward an instance to display how Wisteria improves iterative data analysis and cleaning. For the entire system structure, optimizer is the key and completes the interaction among recommendation, commands, queries, and results through various components.

3.3.4 Privacy preservation

Improving data quality in mining the value of large heterogeneous datasets has become a common challenge for business organizations. Existing data cleaning solutions focus on rapidly resolving data inconsistencies. Owing to the proliferation of sensitive, confidential user information, a few have explored privacy issues with data cleaning. Huang et al. [Huang and Chiang (2015)] proposed a privacy-conscious data cleaning framework designed to address data inconsistencies while minimizing the amount of disclosed information. Their research proposes a set of data disclosure operations in conjunction with the cleaning process and presents two technical measures for correcting privacy losses and data utility with inconsistent data. Based on considering the information sensitivity of data values during the data cleaning process, Chiang et al. [Chiang and Gairola (2018)] proposed a data cleaning framework called InforClean. InforClean allows data analysts to define the data quality rules that the data should satisfy, and errors are defined as violations of these rules. This research quantifies privacy as the information contained in each attribute value and minimizes the total amount of information disclosed. InforClean proposes a pragmatic, information-theoretic data cleaning framework that aims to safeguard data values with high information content from being disclosed during data cleaning.

Existing research systems assume that cleaning professionals can access the data to adjust the cleaning process. However, in many cases, privacy constraints do not allow unlimited access to the data, thus affecting the effectiveness of cleaning adjustments. Ge et al. [Ge, Ilyas, He et al. (2017)] proposed DPClean, which becomes the first system to adjust the data cleaning workflow while ensuring differential privacy, to solve this problem. In DPClean, cleaning users can construct fault-tolerant aggregated count query sequences to clean without accessing sensitive data. The privacy engine transforms each query into a differentiated answer that matches the specified tolerance and allows the data owner to understand the overall privacy loss. Their work achieves high-quality cleaning by ensuring reasonable privacy. Differential privacy analysis can ensure the privacy of users while retaining the main data features. However, most privacy mechanisms assume that the underlying dataset is clean. Krishnan et al. [Krishnan, Wang, Franklin et al. (2016)] proposed a data cleaning framework called PrivateClean and discussed the relationship between data cleanup and differential privacy. PrivateClean focuses on the effect of privacy on the accuracy of subsequent aggregated queries. This framework also amplifies certain types of errors in datasets and can be used to adjust privacy. This research uses functional dependencies, anomalous filtering, and inconsistent attributes to validate the bias due to the interaction between cleaning and privacy. Han et al. [Han, Chen, Zhang et al. (2018)] established the differential privacy cleaning model HRR, which is based on the contradiction generated by the function dependency, correct the contradictory data, and use the indistinguishability between the correction results to protect the data privacy. In this model, the local differential privacy mechanism is added in the process of data cleaning. While simplifying the data pre-processing process, this model achieves a balance between data availability and security.

The research on sensitive pattern hiding is of considerable significance to the privacy preservation of data mining. The existing hiding sensitive mode algorithm was initially designed for a static database and cannot adequately handle incremental datasets. Chen [Chen (2014)] designed the victim selection strategy with the least edge effect based on the delicate patterns and proposed the privacy preservation algorithm in the incremental update database to hide the sensitive mode under incremental environment. The privacy preservation algorithm is proposed in the incremental update database. The example analysis and experimental results verify the correctness, validity, and scalability of the proposed method.

Existing data cleaning methods focus on data repair with minimized cost functions or updates. However, these techniques do not consider basic data privacy requirements in real sensitive information datasets. Huang et al. [Huang, Gairola, Huang et al. (2016)] proposed a data cleaning system combined with privacy preservation through functional dependency and disclosure of sensitive values in the cleaning process. This system includes modules for evaluating key indicators during maintenance searches and addresses multi-objective optimization problems to identify repairs that balance the trade-off between privacy and utility. This research highlights the features of privacy-protected data remediation, customized data sanitation, and data privacy requirements using two real datasets and differentiates repair recommendations through a visual summary.

The data-cleaning-as-a-service (DCaS) enables users to outsource their data cleaning needs to third-party service providers that are highly capable of computing [Dong, Liu and Wang (2014)]. This option raises a few security issues in specific computing environments, and one of the salient issues is private information protection of customers in outsourced data. Dong et al. [Dong, Liu and Wang (2014)] used deduplication as the primary data cleanup task and designed two efficient methods for duplicate privacy-protected data de-duplication for DCaS. After analyzing the robustness of the two methods to the knowledge of attack and coding algorithm using auxiliary frequency distribution, this study experimentally proved the efficiency and effectiveness of the privacy preservation method.

PACAS is a privacy-aware data based on DCaS framework [Huang, Milani and Chiang (2018)], which facilitates communication between the client and the service provider. This communication is via a data pricing scheme where clients issue queries and the service provider returns clear answers for a price while protecting the data. This framework is proposed as a practical privacy model in such interactive settings called (X, Y, L)-anonymity. This model extends existing data publishing techniques to consider the data semantics while protecting sensitive values. The evaluation over real data shows that PACAS effectively safeguards semantically related sensitive values and provides improved accuracy over existing privacy-aware cleaning techniques.

4 Discussion of problems and further research

Existing data cleaning methods in WIS are presented in the previous section. For a more comprehensive discussion, we analyze various types of problems in existing methods and then provide relevant suggestions for further research.

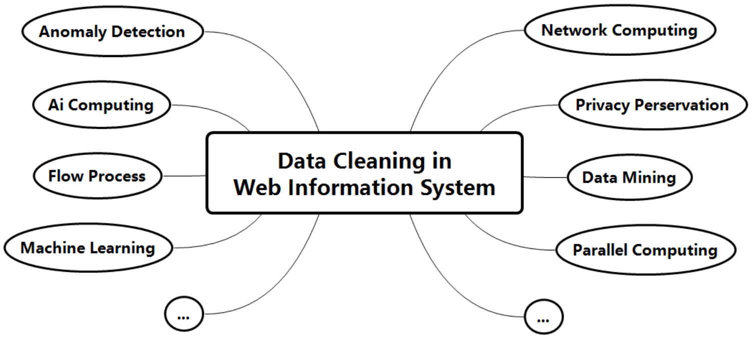

Figure 3: Technology combined with data cleaning methods in WIS

4.1 Exist problems of data cleaning methods in WIS

With the era of WIS, academic research increases due to the diverse needs of the data cleaning business. Data cleaning methods have been applied to many other areas of computer technology shown in Fig. 3. The diversity of research has led to the existence of specific problems in all aspects of the existing methods.

4.1.1 Application scenario

Data cleaning applies to different scenarios of WIS data processing, including RFID and sensor and ETL. Scenarios often provide the data a few additional unique properties in the procedure of data quality improvement. Data cleaning methods and application scenarios become closely related.

RFID and sensors are in the sensing layer of WIS. In the data processing of RFID and sensor, transmission inefficiencies, delays, noise, and other problems often occur. Existing data cleaning studies on RFID and sensor mainly focus on proposing noise and invalid data processing methods. By summarizing the existing methods [Chen, Yu, Gu et al. (2011); Jiang, Xiao, Wang et al. (2011); Ziekow, Ivantysynova and Günter (2011); Ali, Pissinou and Makki (2012); Fan, Wu and Lin (2012); Zhang, Kalasapudi and Tang (2016)], we find that the research in this area can still be improved in terms of cleaning accuracy. This research will achieve excellent results for RFID and sensors to adjust data business logic according to the data cleaning result.

By observing the procedure of data cleaning in WIS, we find that ETL is a crucial application scenario in the aspect of data process. Existing works focus on quality changes in the data conversion process [Cheng (2009) ; Esuli and Sebastiani (2009); Lettner, Stumptner and Bokesch (2014); Ong and Ismail (2014); Mok, Mok and Cheung (2015); Ding, Lin and Li (2016); Mykland and Zhang (2016); Deng and Wang (2017)]. Owing to the diversity of data processing requirements, unifying the quantitative standards of data quality has always been difficult. Existing data cleaning methods have poor portability and evaluability in the application environment of ETL.

4.1.2 Interactive object

The process of data cleaning must interact with entity objects to improve data quality. In the interactive process, WIS data quality is mainly affected by data and users.

After summarizing the works in this research area [Berti-Equille, Dasu and Srivastava (2011); Müller, Freytag and Leser (2012); Geerts, Mecca and Papotti et al. (2013); Volkovs, Chiang, Szlichta et al. (2014); Liu, Kumar and Thomas (2015)], we find that traditional data cleaning methods focus on data interaction. Identifying and correcting data errors are necessary for data quality improvement. Owing to the emergence of data errors that change the statistical properties and interdependence, data interaction studies are carried out to investigate this phenomenon from qualitative and quantitative aspects. Correcting data errors leads to uneven data quality improvement in different data quality evaluation systems. In the absence of a unified data quality evaluation system, continuous research is necessary to obtain the ideal data cleaning effect.

In the process of WIS data cleaning, the importance of user interaction rises with the increasing demand for customization. With rising demand for customization, the importance of user interaction is increasing. Existing works provide customized services that combine the actual needs of users with processes [Raman and Hellerstein (2001); Kalashnikov and Mehrotra (2006); Christen (2009); Fan, Ma, Tang et al. (2011); Galhardas, Lopes and Santos (2011); Bergman, Milo, Novgorodov et al. (2015); He, Veltri, Santoro et al. (2016)]. Herein, we discuss a particular case where data cleaning is ruined with the occurrence of user interactions with evident business logic errors. The extent of instructions that should be followed is an exciting research direction.

4.1.3 Core technology

Data cleaning is oriented toward different interactive objects in various application scenarios of WIS. The typical examples are data quality rule, model, crowdsourcing, and privacy preservation.

Data quality rules can be divided into integrity constraint and duplicate record detection according to the type of data errors. Integrity constraints contain FD and negative constrains. Numerous existing works show that integrity constraints and pattern rules can be used for data repair after clarifying the type of error [Bohannon, Fan, Geerts et al. (2007); Richards and Davies (2012); Bertossi, Kolahi and Lakshmanan (2013); Chu, Ilyas and Papotti (2013); Dallachiesa, Ebaid, Eldawy et al. (2013); Terracina, Martello and Leone (2013); Chalamalla, Ilyas, Ouzzani et al. (2014); Dara, Satyanarayana and Govardhan (2016); Salem and Abdo (2016); Zhang, Szabo and Sheng (2016); Hazen, Weigel, Ezell et al. (2017); Berghe and Gaeveren (2017)]. Specific problems can be solved by single data quality rule. In the discussion on multi-role applications, we can assume that two sets of data are in line with one kind of data quality rule. However, the conclusion often disobeys common sense when merging the two sets. Subtle deviations are frequently observed in the high-dimensional statistical analysis, in which the Simpson’s paradox played a significant role. Therefore, the industry must conduct follow-up research to solve this difficult problem.

Although large-scale data in WIS can help train complex learning models, data errors also simultaneously affect the reliability of model training in the field of model application. Recent studies use robust technology and high- dimensional model to continually optimize the method effect in response to this problem [Limas, Meré, Ascacibar et al. (2004); Hua and Pei (2007); Arumugam and Devadas (2010); Mayfield, Neville and Prabhakar (2010); Pehwa (2013); Krishnan, Wang, Franklin et al. (2015); Laufer and Schwieger (2015); Chen, Ouyang and Jiang(2016); Gueta and Carmel (2016); Krishnan, Wang, Franklin et al. (2016); Merino, Caballero, Rivas et al. (2016); Vetrò, Canova, Torchiano et al. (2016); Zhang, Zhang, Liang et al. (2017)]. The common problem with the model applications of data cleaning is that the model requires training to learn the characteristics of datasets. However, data cleaning and model training is not as tacit as imagined because dirty data always exist, which is missing without processing and produces results. It has more difficulty for model-based data cleaning methods to perform better in terms of accuracy.

Crowdsourcing is widely used in data cleaning. Recent studies have used active learning to solve problems due to the difficulty in measuring clustering. The basic idea of active learning is to use human input as a label and obtain the most meaningful label using the supervised learning methods [Mozafari, Sarkar, Franklin et al. (2014); Tong, Cao, Zhang et al. (2014); Ye, Wang, Li et al. (2014); Chu, Morcos, Ilyas et al. (2015); Haas, Krishnan, Wang et al. (2015)]. Although it looks perfect, crowdsourcing needs more labor costs than traditional data cleaning methods. Finding the most significant label from the process of data cleaning under the cost constraints of WIS remains a popular topic in this research field.

After summarizing the research on WIS data cleaning [Babar, Mahalle, Stango et al. (2010); Dong, Liu and Wang (2014); Ukil, Bandyopadhyay and Pal (2014); Huang and

Chiang (2015); Huang, Gairola, Huang et al. (2016); Krishnan, Wang, Franklin et al. (2016); Ge, Ilyas, He et al. (2017)], which has been combined with privacy preservation, we find that ensuring the confidentiality of data cleaning is a top priority. A few research works exist in this field, mostly from the two aspects of data privacy and application scene characteristics. In the specific privacy data cleaning process, current research works provide a patch plan for privacy preservation. Specific constraints control the scalability of each study to meet the growing demand for WIS privacy data cleaning.

4.2 Suggestions for further research

After presenting the current problems of existing methods, a few suggestions for future data cleaning research in WIS are as follows. Reasonable response to large datasets The processing data scale is gradually increasing in the era of big data. The datasets of most methods in the field of data cleaning research often presently exist in the primary storage. Exploring the use of auxiliary storage and parallel processing and modifying the system framework with the appearance of large datasets in WIS are feasible.

4.2.1 Reduce parallel processing

Numerous redundancies often occur in data cleaning while processing data in the WIS environment. Dividing the entire processing task into sub-tasks and merging them in the same situation while minimizing parallel processing in different situations are feasible solutions.

4.2.2 Technology for wider applications

Unstructured document text plays an irreplaceable role in WIS. The data cleaning methods for such data have not been sufficiently explored. The increase in RFIDs and sensors leads to additional streaming data with the increase in WIS. Carrying out distributed qualitative data cleaning research based on streaming data will be the future research hotspot.

4.2.3 User-friendly interaction

The data attributes and processing results are perceived with user interaction, whereas an instructive role is crucial in user intervention, such as crowdsourcing. Although no uniform technology is available for different data cleaning frameworks, user-friendly interaction is necessary for data cleaning framework design in terms of utility and effectiveness.

4.2.4 Extensible privacy cleaning framework

Owing to the confidentiality of WIS privacy data, its cleaning work is limited by factors, such as processing scenarios and permission management. With the growing demand for WIS privacy data cleaning, the emergence of a privacy data cleaning framework that accommodates additional scenarios and data types in WIS is necessary.

5 Conclusion

Users interact with each other via an information communication system in WIS. As WIS offers the possibility to information dissemination, the data brings practical significance to the business logic process during generating and disseminating as the relevant carrier of the content. Data cleaning plays an essential role in a variety of application scenarios by numbers of core technologies with different interactive objects in WIS to improve the quality of data. Although there are still some related problems, the data cleaning has brought a good application prospect for the high-quality data flow of WIS. We surveyed the current situation of data cleaning field and discussed the critical elements of data cleaning methods in WIS in this paper. The existing cleaning methods are classified and introduced, and the problems in each type are analyzed. We have discussed some viable solutions in response to the challenges in the field of data cleaning. Data cleaning technology is expected to be more efficient and accurate in playing its role in WIS for the future.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Ali, B. Q.; Pissinou, N.; Makki, K. (2012): Belief based data cleaning for wireless sensor networks. *Wireless Communications and Mobile Computing*, vol. 12, no. 5, pp. 406-419.

Arumugam, G.; Devadas, T. J. (2010): Object oriented intelligent multi-agent system data cleaning architecture to clean email data. *International Journal of Computer Applications*, vol. 9, no. 8, pp. 34-44.

Babar, S.; Mahalle, P.; Stango, A.; Prasad, N.; Prasad, R. (2010): Proposed security model and threat taxonomy for the Internet of Things (IoT). *International Conference on Network Security and Applications*, pp. 420-429.

Berghe, S. V.; Gaeveren, K. V. (2017): Data quality assessment and improvement: a vrije universiteit brussel case study. *Procedia Computer Science*, vol. 106, pp. 32-38.

Bergman, M.; Milo, T.; Novgorodov, S.; Tan, W. C. (2015): Query-oriented data cleaning with oracles. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 1199-1214.

Berti-Equille, L.; Dasu, T.; Srivastava, D. (2011): Discovery of complex glitch patterns: a novel approach to quantitative data cleaning. *IEEE International Conference on Data Engineering*, pp. 733-744.

Bertossi, L.; Kolahi, S.; Lakshmanan, L. V. S. (2013): Data cleaning and query answering with matching dependencies and matching functions. *Theory of Computing Systems*, vol. 52, no. 3, pp. 441-482.

Bhattacharjee, A. K.; Chatterjee, P.; Shaw, M. P.; Chakraborty, M. (2014): ETL based cleaning on database. *International Journal of Computer Applications*, vol. 105, no. 8, pp. 34-40.

Bohannon, P.; Fan, W.; Geerts, F.; Jia, X.; Kementsietsidis, A. (2007): Conditional functional dependencies for data cleaning. *IEEE 23rd International Conference on Data Engineering*, pp. 746-755.

Chalamalla, A.; Ilyas, I. F.; Ouzzani, M.; Papotti, P. (2014): Descriptive and prescriptive data cleaning. *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pp. 445-456.

Chen, H.; Ouyang, Y.; Jiang, **W.** (2016): An optimized data integration model based on reverse cleaning for heterogeneous multi-media data. *Multimedia Tools and Applications*, vol. 75, no. 23, pp. 15571-15586.

Chen, M.; Yu, G.; Gu, Y.; Jia, Z.; Wang, Y. (2011): An efficient method for cleaning dirty-events over uncertain data in WSNs. *Journal of Computer Science and Technology*, vol. 26, no. 6, pp. 942-953.

Chen, W. (2014): Privacy preservation for the incremental updating database. *Pattern Recognition and Artificial Intelligence*, vol. 27, no. 7, pp. 638-645.

Cheng, R. (2009): Querying and cleaning uncertain data. *International Workshop on Quality of Context*, pp. 41-52.

Chiang, F.; Gairola D. (2018): InfoClean: protecting sensitive information in data cleaning. *Journal of Data and Information Quality*, vol. 9, no. 4, pp. 1-22.

Christen, P. (2009): Development and user experiences of an open source data cleaning, deduplication and record linkage system. *ACM SIGKDD Explorations*, vol. 11, no. 1, pp. 39-48.

Chu, X.; Ilyas, I. F.; Papotti, P. (2013): Holistic data cleaning: putting violations into context. *IEEE 29th International Conference on Data Engineering*, pp. 458-469.

Chu, X.; Morcos, J.; Ilyas, I. F.; Ouzzani, M.; Papotti, P. et al. (2015): Katara: a data cleaning system powered by knowledge bases and crowdsourcing. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 1247-1261.

Dallachiesa, M.; Ebaid, A.; Eldawy, A.; Elmagarmid, A.; Ilyas, I. F. et al. (2013): NADEEF: a commodity data cleaning system. *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pp. 541-552.

Dara, R.; Satyanarayana, C.; Govardhan, **A.;** (2016): A novel approach for data cleaning by selecting the optimal data to fill the missing values for maintaining reliable data warehouse. *International Journal of Modern Education and Computer Science*, vol. 8, no. 5, pp. 64-70.

Deng, W.; Wang, G. (2017): A novel water quality data analysis framework based on time-series data mining. *Journal of Environmental Management*, vol. 196, no. 7, pp. 365-375.

Ding, Y.; Lin, H.; Li, R. (2016): Change semantic constrained online data cleaning method for real-time observational data stream. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. XLI-B2, no. 6, pp. 177-183.

Dong, B.; Liu, R.; Wang, W. H. (2014): Prada: privacy-preserving data-deduplication-as-a-service. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 1559-1568.

Esuli, A.; Sebastiani, F. (2009): Training data cleaning for text classification. *2nd International Conference on the Theory of Information Retrieval*, pp. 29-41.

Fan, H.; Wu, Q.; Lin, Y. (2012): Behavior-based cleaning for unreliable RFID data sets. *Sensors*, vol. 12, no. 8, pp. 10196-10207.

Fan, W.; Li, J.; Ma, S.; Tang, N.; Yu, W. (2011): Cerfix: a system for cleaning data with certain fixes. *Proceedings of the VLDB Endowment*, vol. 4, no. 12, pp. 1375-1378.

Fan, W.; Ma, S.; Tang, N.; Yu, W. (2011): Interaction between record matching and data repairing. *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, pp. 469-480.

Galhardas, H.; Lopes, A.; Santos, E. (2011): Support for user involvement in data cleaning. *International Conference on Data Warehousing and Knowledge Discovery*, pp. 136-151.

Ganti, V.; Sarma, A. D. (2013): Data cleaning: a practical perspective. *Synthesis Lectures on Data Management*, vol. 5, no. 3, pp. 1-85.

Ge, C.; Ilyas, I. F.; He, X.; Machanavajjhala, A. (2017): Private exploration primitives for data cleaning. https://arxiv.org/pdf/1712.10266.pdf.

Geerts, F.; Mecca, G.; Papotti P.; Santoro, D. (2013): The LLUNATIC data-cleaning framework. *Proceedings of the VLDB Endowment*, vol. 6, no. 9, pp. 625-636.

Gueta, T.; Carmel, Y. (2016): Quantifying the value of user-level data cleaning for big data: a case study using mammal distribution models. *Ecological Informatics*, vol. 34, pp. 139-145.

Haas, D.; Krishnan, S.; Wang, J.; Franklin, M. J.; Wu, E. (2015): Wisteria: nurturing scalable data cleaning infrastructure. *Proceedings of the VLDB Endowment*, vol. 8, no. 12, pp. 2004-2007.

Han, Q.; Chen, Q.; Zhang L.; Zhang, K. (2018): HRR: a data cleaning approach preserving local differential privacy. *International Journal of Distributed Sensor Networks*, vol. 14, no. 12.

Hazen, B. T.; Weigel, F. K.; Ezell, J. D.; Boehmke, B. C.; Bradley, R. V. (2017): Toward understanding outcomes associated with data quality improvement. *International Journal of Production Economics*, vol. 193, pp. 737-747.

He, J.; Veltri, E.; Santoro, D.; Li, G.; Mecca, G. et al. (2016): Interactive and deterministic data cleaning. *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, pp. 893-907.

Hellerstein, J. M. (2008): Quantitative data cleaning for large databases. *United Nations Economic Commission for Europe*.

Hua, M.; Pei, J. (2007): Cleaning disguised missing data: a heuristic approach. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 950-958.

Huang, D.; Gairola, D.; Huang, Y.; Zheng, Z.; Chiang, F. (2016): PARC: privacy-aware data cleaning. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2433-2436.

Huang, Y.; Chiang, F. (2015): Towards a unified framework for data cleaning and data privacy. *International Conference on Web Information Systems Engineering*, pp. 359-365.

Huang, Y.; Milani, M.; Chiang, F. (2018): PACAS: privacy-aware, data cleaning-as-a-service. *2018 IEEE International Conference on Big Data*, pp. 1023-1030.

Jeffery, S. R.; Alonso, G.; Franklin, M. J.; Hong, W.; Widom, J. (2006): A pipelined framework for online cleaning of sensor data streams. *Proceedings of the 22nd International Conference on Data Engineering*, pp. 140.

Jeffery, S. R.; Garofalakis, M.; Franklin, M. J. (2006): Adaptive cleaning for RFID data streams. *Proceedings of the VLDB Endowment*, pp. 163-174.

Jiang, T.; Xiao, Y.; Wang, X.; Li, Y. (2011): Leveraging communication information among readers for RFID data cleaning. *International Conference on Web-Age Information Management*, pp. 201-213.

Kalashnikov, D. V.; Mehrotra, **S.** (2006): Domain-independent data cleaning via analysis of entity-relationship graph. *ACM Transactions on Database Systems*, vol. 31, no. 2, pp. 716-767.

Krishnan, S.; Franklin, M. J.; Goldberg, **K.; Wang, J.; Wu, E.** (2016): ActiveClean: an interactive data cleaning framework for modern machine learning. *Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data*, pp. 2117-2120.

Krishnan, S.; Haas, D.; Franklin, M. J.; Wu, E. (2016): Towards reliable interactive data cleaning: a user survey and recommendations. *Proceedings of the Workshop on Human-In-the-Loop Data Analytics, Co-Located with SIGMOD 2016*, pp. 9.

Krishnan, S.; Wang, J.; Franklin, M. J.; Goldberg, K.; Kraska, T. (2015): Stale view cleaning: getting fresh answers from stale materialized views. *Proceedings of the VLDB Endowment*, vol. 8, no. 12, pp. 1370-1381.

Krishnan, S.; Wang, J.; Franklin, M. J.; Goldberg, K.; Kraska, T. (2016): Privateclean: data cleaning and differential privacy. *Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data*, pp. 937-951.

Laufer, R.; Schwieger, V. (2015): Modeling data quality using artificial neural networks. *1st International Workshop on the Quality of Geodetic Observation and Monitoring Systems*, pp. 3-8.

Lettner, C.; Stumptner, R.; Bokesch, K. (2014): An approach on ETL attached data quality management. *International Conference on Data Warehousing and Knowledge Discovery*, pp. 1-8.

Li, J.; Feng, Y.; Yu, T.; Liu, J.; Zhu, L. et al. (2014): A data mining system for herbal formula compatibility analysis. *Frontier and Future Development of Information Technology in Medicine and Education*. Springer Berlin Heidelberg.

Limas, M. C.; Meré, J. B.; Ascacibar, F. J.; González, E. P. (2004): Outlier detection and data cleaning in multivariate non-normal samples: the PAELLA algorithm. *Data Mining and Knowledge Discovery*, vol. 9, no. 2, pp. 171-187.

Liu, H.; Kumar, T. A.; Thomas, J. P. (2015): Cleaning framework for big data-object identification and linkage. *2015 IEEE International Congress on Big Data*, pp. 215-221.

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

Merino, J.; Caballero, I.; Rivas, B.; Serrano, M.; Piattini, M. (2016): A data quality in use model for big data. *Future Generation Computer Systems*, vol. 63, pp. 123-130.

Mok, R. V.; Mok, W. Y.; Cheung, K. Y. (2015): A security price data cleaning technique: Reynold's decomposition approach. *Information Management and Big Data*, Springer Berlin Heidelberg.

Mozafari, B.; Sarkar, P.; Franklin, M.; Jordan, M.; Madden, **S.** (2014): Scaling up crowd-sourcing to very large datasets: a case for active learning. *Proceedings of the VLDB Endowment*, vol. 8, no. 2, pp. 125-136.

Müller, H.; Freytag, J.; Leser, U. (2012): Improving data quality by source analysis. *Journal of Data and Information Quality*, vol. 2, no. 4, pp. 15.

Mykland, P. A.; Zhang, L. (2016): Between data cleaning and inference: pre-averaging and robust estimators of the efficient price. *Journal of Econometrics*, vol. 194, no. 2, pp. 242-262.

Okan, Z. (2003): Edutainment: Is learning at risk?. *British Journal of Educational Technology*, vol. 34, no. 3, pp. 255-264.

Ong, Y. C.; Ismail, Z. (2014): Enhanced web log cleaning algorithm for web intrusion detection. *Recent Advances in Information and Communication Technology*, Springer Berlin Heidelberg.

Pehwa, P.; Arora, R.; Thakur, G. (2013): An efficient algorithm for data cleaning. *International Journal of Knowledge-Based Organizations*, vol. 1, pp. 56-71.

Qu, Z.; Wang, Y.; Wang, C.; Qu, N.; Yan, **J.** (2016): A data cleaning model for electric power big data based on spark framework. *International Journal of Database Theory and Application*, vol. 9, no. 3, pp. 137-150.

Raman, V.; Hellerstein, J. M. (2001): Potter's wheel: an interactive data cleaning system. *Proceedings of the VLDB Endowment*, pp. 381-390.

Richards, K.; Davies, N. (2012): Cleaning data: guess the olympian. *Teaching Statistics*, vol. 34, no. 1, pp. 31-37.

Salem, R.; Abdo, A. (2017): Fixing rules for data cleaning based on conditional functional dependency. *Future Computing and Informatics Journal*, vol. 1, no. 4, pp. 10-26.

Schewe, K.; Thalheim, B. (2019): *Design and Development of Web Information Systems*. Springer Berlin Heidelberg.

Terracina, G.; Martello, A.; Leone, N. (2013): Logic-based techniques for data cleaning: an application to the italian national healthcare system. *International Conference on Logic Programming and Nonmonotonic Reasoning*, pp. 524-529.

Tong, Y.; Cao, C. C.; Zhang, C. J.; Li, Y.; Chen, L. (2014): CrowdCleaner: data cleaning for multi-version data on the web via crowdsourcing. *IEEE 30th International Conference on Data Engineering*, pp. 1182-1185.

Ukil, A.; Bandyopadhyay, S.; Pal, A. (2014): Sensitivity inspector: detecting privacy in smart energy applications. *IEEE Symposium on Computers and Communication*, pp. 1-6.

Vetrò, A.; Canova, L.; Torchiano, M.; Minotas, C. O.; Iemma, R. et al. (2016): Open data quality measurement framework: definition and application to open government data. *Government Information Quarterly*, vol. 33, no. 2, pp. 325-337.

Volkovs, M.; Chiang, F.; Szlichta, J.; Miller, R. J. (2014): Continuous data cleaning. *IEEE 30th International Conference on Data Engineering*, pp. 244-255.

Ye, C.; Wang, H.; Li, K.; Chen, Q.; Chen, J. et al. (2014): CrowdCleaner: a data cleaning system based on crowdsourcing. *Asia-Pacific Web Conference*, pp. 657-661.

Zhang, C.; Kalasapudi, V. S.; Tang, **P.** (2016): Rapid data quality oriented laser scan planning for dynamic construction environments. *Advanced Engineering Informatics*, vol. 30, no. 2, pp. 218-232.

Zhang, J.; Zhang, S.; Liang, J.; Tian, B.; Hou, Z. et al. (2017): Photovoltaic generation data cleaning method based on approximately periodic time series. *IOP Conference Series: Earth and Environmental Science*.

Zhang, Y.; Szabo, C.; Sheng, Q. Z. (2016): Reduce or remove: individual sensor reliability profiling and data cleaning. *Intelligent Data Analysis*, vol. 20, no. 5, pp. 979-995.

Ziekow, H.; Ivantysynova, L.; Günter, O. (2011): RFID data cleaning for shop floor applications. *Unique Radio Innovation for the 21st Century,* Springer, pp. 143-160.