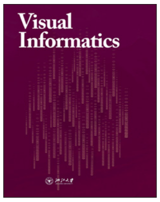
Visual Informatics 2 (2018) 191–197

Contents lists available at ScienceDirect

Visual Informatics

journal homepage: www.elsevier.com/locate/visinf



Steering data quality with visual analytics: The complexity challenge

Shixia Liu [a,*], Gennady Andrienko [b,c], Yingcai Wu [d], Nan Cao [e], Liu Jiang [a], Conglei Shi [f], Yu-Shuen Wang [g], Seokhee Hong [h]

[a] *Tsinghua University, Beijing, China*
[b] *Fraunhofer Institute IAIS, Sankt-Augustin, Germany*
[c] *City, University of London, London, UK*
[d] *Zhejiang University, Zhejiang, China*
[e] *Tongji University, Shanghai, China*
[f] *Airbnb, San Francisco, CA, USA*
[g] *National Chiao-Tung University, Hsinchu, Taiwan*
[h] *University of Sydney, Sydney, Australia*

ARTICLE INFO

Article history:
Received 20 November 2018
Accepted 11 December 2018
Available online 22 December 2018

MSC:
00-01
99-00

Keywords:
Data quality management
Visual analytics
Data cleansing

ABSTRACT

Data quality management, especially data cleansing, has been extensively studied for many years in the areas of data management and visual analytics. In the paper, we first review and explore the relevant work from the research areas of data management, visual analytics and human-computer interaction. Then for different types of data such as multimedia data, textual data, trajectory data, and graph data, we summarize the common methods for improving data quality by leveraging data cleansing techniques at different analysis stages. Based on a thorough analysis, we propose a general visual analytics framework for interactively cleansing data. Finally, the challenges and opportunities are analyzed and discussed in the context of data and humans.

Contents

1. Introduction 191
2. Related work 192
3. Data types and their relationships 193
4. Examples of different data types 193
5. Analysis pipeline 194
6. Research challenges and opportunities 195
   6.1. Data complexity 195
   6.2. Human complexity 196
7. Conclusion 196
   Acknowledgment 196
   References 196

1. Introduction

With the increasing predominance of data-centric approaches to business, scientific, and engineering problems, data and its quality have become more and more important (Fan and Geerts, 2012; Liu et al., 2019; McCurdy et al., 2019; Song and Szafir, 2019). However, during the data collection and processing stage, some incomplete, inconsistent, duplicate, inaccurate or irreversibly transformed (e.g., sanitized for privacy preservation (Amiri, 2007; Domadiya and Rao, 2013; Modi et al., 2010; Oliveira and Zaïane, 2003)) data can be fused, which usually affects the further usage of data (e.g., lower the accuracy of the learning models) and can lead to loss of consumer trust and revenues. As a result, in the era of big data, one key issue on preparing and processing data is to ensure the quality and usability of data (data quality management), including detecting, removing, and correcting errors and inconsistencies in the data.

[*] Corresponding author.
E-mail address: shixia@tsinghua.edu.cn (S. Liu).
Peer review under responsibility of Zhejiang University and Zhejiang University Press.

https://doi.org/10.1016/j.visinf.2018.12.001
2468-502X/© 2018 Zhejiang University and Zhejiang University Press. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Data quality management has been studied for many years in the area of database and data management (Liu et al., 2019; Kwon et al., 2014). Its major goal is to efficiently detect and correct errors in the data. As an important part of data-driven analysis, data quality management takes over 30%–80% of the time and resource (Saha and Srivastava, 2014). Most mature works focus on tabular data, such as assessing the data quality (Kandel et al., 2012), interactive data cleansing (Raman and Hellerstein, 2001), and data wrangler (Kandel et al., 2011a). However, due to the increasing complexity of the data (e.g., multimedia data, texts, graphs, sequences and trajectories etc.) collected through a variety of ways, it is more and more challenging to effectively and accurately improve data quality. In most of the cases, domain knowledge of experts is important to guide for better performance of data quality management algorithms (El Bekri and Peinsipp-Byma, 2016). As a consequence, there has been a growing interest in recent years to study how to better combine user-guided methods with system-guided methods during the analysis, where information visualization and visual analytics are the important parts to achieve this goal (McCurdy et al., 2019; Song and Szafir, 2019; Liu et al., 2017; Choo and Liu, 2018; Gschwandtner et al., 2012).

Data cleansing is a widely used practice for effective data quality management. As a result, most existing data quality management efforts focus on data cleansing. In this paper, we first report the related work from different research areas, including data management, visual analytics, and human-computer interaction. Then for different types of data, we summarize the common methods for improving data quality by leveraging data cleansing techniques at different stages. In addition, a high-level abstraction of a framework on designing a visual analytic system for data cleansing is needed as a general guideline for the research in this direction. Thus, inspired by pipeline proposed in Van den Broeck et al. (2005), we develop a visual analytics framework, focusing on iteratively and progressively improving data quality from the screening stage, to diagnosis stage and the correction stage. Finally, we explore the research challenges and opportunities and align them with the our visual analytics framework, which we hope can better guide the future visual analytics research on data cleansing.

2. Related work

Researchers have been extensively studying a variety of data cleansing techniques to improve the data quality for the past twenty years. Most efforts are mainly from two research areas: data management and visual analytics.

In the area of data management, researchers have developed a number of approaches to checking, repairing, and correcting inconsistencies and errors in the data. Existing efforts can be classified into three categories: rule-based detection methods for cleaning data based by a set of rules (Abedjan et al., 2015; Fan et al., 2012; Geerts et al., 2014; Khayyat et al., 2015), quantitative error detection methods for discovering and resolving outliers and glitches in the data (Dasu and Loh, 2012; Prokoshyna et al., 2015; Vartak et al., 2015; Wu and Madden, 2013), and record linkage and deduplication methods for detecting duplicate data items (Elmagarmid et al., 2007; Stonebraker et al., 2013). Recently, Abedjan et al. (2016) conducted a comprehensive evaluation to analyze the performance of existing algorithms on four common types of data errors, including outliers, duplicates, rule violations, and pattern violations. These error types are relatively general and can be applied beyond tabular data. However, these works do not provide an end-to-end data cleansing pipeline. In order to enable effective and efficient data cleansing practices, several frameworks have been developed. For example, Florescuand modeled the cleansing application as a graph-based data transformation, which can be applied to SQL-based database (Florescuand, 2000). Gill and Lee developed a distributed data cleansing framework specific for data streams (Gill and Lee, 2015). Due to the limited usage scope, these frameworks cannot be applied to a general data cleansing application. Broeck et al., proposed a three-stage framework on data cleansing, either manually or automatically. The framework categorizes the whole process into three stages, screening stage, diagnosis stage, and correction stage (Van den Broeck et al., 2005). For each stage, the key problems are identified. The major feature of this framework is that it covers the whole analysis workflow well, from the raw data exploration to actual error correcting.

In most real-world cases, the data cleansing process cannot be done fully automatically due to the ambiguity of the errors and the need of human knowledge to verify the cleansing results. To effectively loop human into the data cleansing process, visual analytics researchers have developed several works focusing on interactive data cleansing. These works in most cases aim to solve some specific tasks for certain types of data, mostly structured table representations. Krishnan et al. (2016) designed ActiveClean to interactively cleanse the data for statistical modeling. Profiler (Kandel et al., 2012) was designed to interactively detect and visually summarize the outliers from data. von Zernichow and Roman (2017) presented a prototype system for visual data profiling, with a focus on discovering and correcting missing values and outliers in tabular data. Wrangler (Kandel et al., 2011b) targets at interactively creating data transformation scripts. Guo et al. (2011) later extended Wrangler to a mixed-initiative system by integrating a proactive recommendation model, which suggests applicable data transforms to users for a more guided exploration of the transformation space. Both tools (Profiler and Wrangler) only support tabular data cleansing. Beyond these works, Kandel et al. (2011a) further summarized the research direction on how visualizations and interaction techniques can help data wrangling. The aforementioned works greatly demonstrate the usefulness and effectiveness of visual analytics techniques in helping improve the data quality, however, it is not easy to apply these techniques to other types of data or different cleansing tasks.

While tabular data is the predominant focus of data cleansing researchers, we also note some efforts on visually cleaning time-oriented data. Gschwandtner et al. (2012) derived a taxonomy of quality issues with time-oriented data, and envision the need of visualization tools for analyzing the quality issues with human in the loop. Following this line, they later proposed TimeCleanser (Gschwandtner et al., 2014), an interactive approach specifically for cleansing time-oriented data. The approach includes several syntax checks that are common for tabular data, including time checks (valid temporal range, consistent interval lengths, missing time points or intervals), time-oriented value checks (e.g., identifying values that do not change for a long time), and consistency about multiple data sets (same temporal range, resolution etc.) Though TimeCleanser demonstrates its effectiveness on correcting data, it is less flexible to support reasoning about underlying causes to the quality issues detected. In light of this, Arbesser et al. (2017) designed Visplause, an interactive visualization system to facilitate the inspection of the quality of multiple time series. In particular, the system mainly integrates the meta information of data to provide a hierarchical overview that aggregates the results of data quality checks at different levels of detail. This enables a flexible semantic reasoning about the data quality. Gschwandtner and Erhart (2018) presented "Know Your Enemy", a visual analytics approach to the identification of issues in time series data. Compared with prior work, they more adequately tailored the visualization design to the time-oriented characteristics of data (e.g., providing the temporal contexts). We also note some recent works that focus on cleaning the time-oriented data that has specialized characteristics in some particular application domains. For example, Schulz et al. (2015) proposed a visual cleansing tool tailored to the low-level eye-tracking data. Dixit et al. (2018) develop an interactive approach specifically for correcting event orderings in process logs.

3. Data types and their relationships

The majority of the existing data cleansing tools focus on structured table data. The word 'structured' is the key here, as the structure of the table suggests how to assess and improve the quality of the data. Table data representation assumes that values in each column have the similar formatting. Respectively, data cleansing procedures check data types, formatting, and help users to perform basic data cleaning operations such as fixing typos (e.g., by checking spelling of text values or ensuring consistency in using decimal dots and commas in numeric values), modifying representations of dates and times, computing statistics for numeric values, identifying and interactively inspecting outliers, resolving encoding rules (e.g., 999 for ***no_data***) etc. Special methods exist for detecting missing values and replacing them with plausible values taking into account values in other table columns. Modern tools like Tableau support cleansing of multiple related tables. For example, values in the connecting columns of the two tables can be checked for consistency and, if some values do not match, they can be modified interactively.

Non-table data requires special methods for cleansing that take into account the specifics of the data type. For example, preparation of textual data requires language detection, removal of too short documents, detection of duplicated content, fixing miss-spelled words and incorrect punctuation, just to name a few operations. It is important to have a possibility to perform fast computations for a subset of the data (e.g., calculating the size of the vocabulary used in different documents or assessing the overall sentiment) for looking at the data set from different perspectives and thus selecting appropriate documents for analysis.

For more complex data it is necessary to exploit their structures for developing appropriate data wrangling tools. For example, spatial event data (Andrienko and Andrienko, 2005) may be considered as a special type of table data that includes columns with temporal and spatial references. Each event is described by its type ('*what happened*'), location ('*where happened*'), time ('*when happened*') and attributes (e.g., event magnitude, actors, etc.). Event types and attributes can be treated as regular columns, applying traditional methods for validating their quality. Spatial and temporal references of the events can be used for attaching additional data sources (that describe locations and times) for checking if events are plausible at given locations and times. For example, traffic events are expected to happen on roads but not in lakes, and visiting restaurant events are highly unlikely at night times.

Another representative example is the graph data. Generally, a graph can be represented using two connected tables describing graph vertices and edges. Both vertices and edges can be described by their attributes. Moreover, additional attributes of vertices and edges can be computed from the topology of the graph. Graph centrality measures can be used for identifying outliers and disconnected subgraphs. Respectively, the data cleansing tools for graphs need to take into account the structure of the graph data and potentially useful additional information that can be derived from the graph data.

Understanding the specifics and structure of the data is essential for designing appropriate data transformations. Let us consider an example of trajectories of moving objects (Andrienko et al., 2013). Each position is a spatial event that can be described by a reference to the moving object *id*, time stamp *t*, its coordinates *x* and *y*, and possible attributes: *id, t, x, y, attributes*. Sequences of events for the same moving object can be integrated into a trajectory (Fig. 1). Such integration allows computation of derived attributes such as displacement, time difference, speed estimate etc. These derived attributes can be used for extracting secondary events from trajectories (e.g., stops) and dividing trajectories into smaller subsets (e.g., trips between stops). Both trajectories and events can be aggregated by areas and links between areas, creating spatial time series referring to locations and links between them, respectively. Further events (e.g., extreme values) can be extracted from such derived spatial time series. This example demonstrates how data structure defines possible and potentially useful transformations. Such transformations can be used for looking at the data from a different perspective, thus facilitating the data cleansing process, identifying and eventually fixing data problems.

4. Examples of different data types

In this section we consider the specifics of steering data quality for different types of data beyond regular data tables.

Multimedia. Multimedia covers many different types of data, which had been widely used for communication after cameras and recorders were invented. In this study, we focus on images, audios, and videos because they are the most popular. Typically, a pixel in images represents the color of a small area. While many pixels are arranged to form an image, which describe the appearance of a large area, they can be used to convey visual information. Compared to an image, a video has an additional time coordinate, in which each time span contains an image. Therefore, a video can show dynamic visual changes over time. In terms of audios, each sample in a time span describes the frequency and magnitude of energy that would be heard by humans. Note that the most important characteristic of the above-mentioned data is that each sample/pixel is meaningless, whereas the integration of them is not. Therefore, there are no point anomalies, but contextual and collective anomalies in multimedia data. Cleansing abnormal multimedia data demands semantics and can only be achieved by complex algorithms and human guidance.

Textual data. Textual data consists of bag-of-words representation in the form of sentences, paragraphs, documents, and topics. Typical examples such as news articles, interviews, emails, field notes, as well as text descriptions from feeds and social media sources, are essential to communicating information about people, events, and activities, sharing findings, knowledge, and best practices within a community, as well as to connecting people and propelling the organization forward (Liu et al., 2012).

In the era of big data, with the ever-increasing size of a document corpus, it is simply not possible in many cases for people to quickly locate key information or derive insights from such large amounts of textual data. Generally, text analysis tasks include information retrieval, cluster/topic analysis, natural language process, classification, outlier analysis, etc. (Liu et al., 2018).

To better perform the aforementioned tasks, the quality of textual data must be guaranteed. Examples of quality issues of textual data are duplicate documents that are intentionally rephrased, irrelevant content such as an advertisement in an news article, documents with mixed topics that are hard to disentangle for a specific application (e.g., information retrieval), multilingual documents, as well as inconsistent documents with varying document lengths, diverse writing styles, and different writing quality (e.g., news articles and tweets).

Trajectories of moving objects. Trajectory data is a sequence of events that correspond to recorded time-referenced positions of moving objects (Andrienko et al., 2013). While each particular record has limited value, many applications require consideration of large collections of such position records. However, even simple tasks like detection of duplicates become difficult due to the complex structure of the data. For position records, a duplicate is a repeating combination of the moving object identity and temporal reference. If the positions and attributes are equal, this is a case of duplicate records that need to be eliminated. If positions or

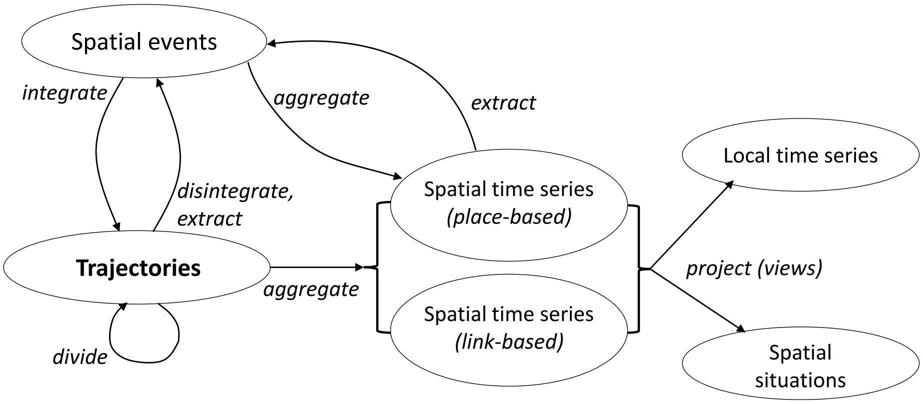


Fig. 1. Potentially possible transformations of trajectory data.

attributes differ, sophisticated conflict resolution procedures are needed.

Paper (Andrienko et al., 2016) analyses in detail the properties of trajectory data (properties of moving objects, space and time) and related properties of data collection procedures. These properties are used for identifying potential problems that may appear in trajectory data sets, which are considered in respect to all components of the data structure (moving objects, space and time). In addition, derived attributes (e.g., speed or acceleration) and possible transformations (e.g., aggregating trajectories into occupancy indicators for visited areas and counts of moves between the areas) enable looking at data from different perspectives (see Fig. 1).

Graph data. A graph $G = (V, E)$ consists of a vertex set V and an edge set E, where a vertex represents an entity and an edge between two vertices represents a relationship between them. Both vertices and edges can have multiple *attributes*, such as numbers, texts, categorical data and images. An edge set defines a *topology* structure of the graph, and there can be more than one edge sets for a given vertex set, which represent multiple relationships between the entities.

For example, a vertex v can represent a student, where v has multiple attributes such as student id number, enrolled degree, year, home address, phone number, and photo id. Similarly, an edge e can have numbers such as weights, texts, time stamps and directions. For more complex data, an edge set E can be changing over time, i.e., change the values of their attributes as well as the topology of a graph. The student set can have multiple relationships, such as facebook friends, instagram friends, and tennis friends etc.

Therefore, the data cleansing for graph data need to take into account these attributes for vertices and edges, as well as the topology of the graph, defined by the edge set. For example, based on the types of attributes of vertices and edges (such as numbers, texts and images) and the topology of the graph, one need to use a variety of techniques for processing quality issues of graph data.

Common issues. Besides the distinctive quality issues from different data types, there are also common ones shared by all kinds of data. For example, missing and inaccurate values are universal problems. It inevitably introduces uncertainty in the process of data transformation and analysis. Common approaches to tackling these issues include deleting missing records, interpolations and uncertainty modeling. Outliers also generally exist in the data. They reflect unusual data patterns, which often hampers the extraction of the main trend of the data. To handle this issue, many outlier detection methods (e.g., density-based approach) have been proposed in the context of different data types. We also find that duplicates and conflicts are issues often encountered in data quality management. Generally, they lead to an incorrect data distribution that mistakenly gives more weights to some particular data, and results in inconsistency in the knowledge generated from data. Additionally, scalability is an important and common problem in steering data quality. Almost all approaches for improving data quality, become less effective when scaled to a large collection of data.

5. Analysis pipeline

Based on the *Screening* → *Diagnosis* → *Correction* framework (Van den Broeck et al., 2005), we propose a visual analytics framework for analyzing and improving data quality. The goal of the proposed framework is to help a user (e.g., an analyzer) find the potential problems of the data to be analyzed/visualized and provide efficient and convenient method to improve the data quality under the users' supervision via their domain knowledge and experiences. To achieve the goal, our framework, as shown in Fig. 2, is designed in a three-layer/module structure with (1) a data layer shown at the top and (2) a visualization layer shown at the bottom connected by an interaction layer show in the middle. These modules are respectively designed for (1) discovering the data insights (for tackling data complexities); (2) intuitive data representation and interpretation (for tackling data complexities); as well as (3) easy data exploration and analysis with human in the loop (for tackling human complexities).

Specifically, the data layer takes various types of data (e.g., raw data, metadata, or analysis results) collected from different sources or produced at different analysis stages as the input. It then preprocesses or analyzes the input data for retrieving data samples, discovering data uncertainty, revealing hidden patterns, or uncovering outliers.

Based on the preprocessing results, three interaction modules: screening, diagnosis, and correction, are designed in a row as shown in the interaction layer. Through screening, a user is able to choose to summarize and illustrate the overview, statistic features, and data patterns (e.g., trend and cluster) via intuitive visual representations. After that, the user can further make a diagnosis of the data to find out the potential problems (e.g., missing values, duplications, pattern/constraint violations, inconsistencies) that affect the data quality. Finally, a user can interactively correct the detected problems within the data.

In the above analysis procedure, visualization plays an important role in supporting data interpretation and decision making. In the framework, two types of visualization designs are required: (1) the visualization designed for illustrating and summarizing the data with the goal of showing overview, patterns, distributions, and constraints of the data, thus handling the complexities associated with data; (2) the visualization designed for data error correction with the goal of identifying missing data, outliers, duplicates, pattern/constraint violations, and data inconsistencies, which aims at tackling a variety of complexities related to human.

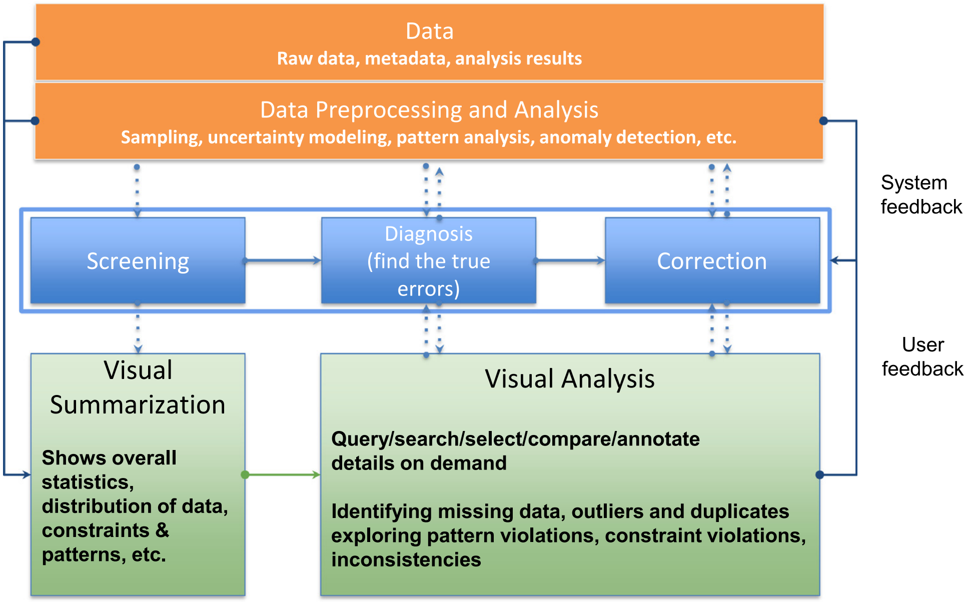

Data
Raw data, metadata, analysis results
Data Preprocessing and Analysis
Sampling, uncertainty modeling, pattern analysis, anomaly detection, etc.
Screening
Diagnosis (find the true errors)
Correction
System feedback
User feedback
Visual Summarization
Shows overall statistics, distribution of data, constraints & patterns, etc.
Visual Analysis
Query/search/select/compare/annotate details on demand
Identifying missing data, outliers and duplicates exploring pattern violations, constraint violations, inconsistencies

Fig. 2. Visual analytics framework for steering data quality, which is based on the data cleansing framework developed in Van den Broeck et al. (2005) (Screening → Diagnosis → Correction).

6. Research challenges and opportunities

6.1. Data complexity

Multimedia. Quality management for multimedia is challenging. Users have to apply different methods to handle images, audio recordings, and videos because they are of different nature. In addition, caused by different reasons, anomalies of these media can be classified into low level and high level categories. On one hand, low level anomalies are often caused by data transmission, compression, or fault of a machine. These anomalies, such as blank images, white noise audios, duplicated frames in a video, can be detected by simple rules. However, on the other hand, high-level anomalies are difficult to detect because of requiring semantics. Specifically, each small segment of multimedia data, such as a pixel in images or a sample in audios, is meaningless, but the integration of them is not. To determine whether an image or a video is anomalous, users have to apply image processing methods to assess visual quality, or object detection techniques to obtain semantics. Otherwise, anomalies such as overexposure, underexposure, and serious object occlusions cannot be detected. Similarly, speech recognition and emotion identification are often used to understand semantics in an audio, and signal processing techniques are used to identify low quality audios such as containing load but irrelevant background speech.

Textual data. Although textual data is widely used in many lines of work, data quality problems for such type of unstructured data remain largely unexplored. This is because, due to the unstructured nature of textual documents, quality management for textual data is challenging. First, textual data often contains several data fields and mixes the useful information with irrelevant information. As a result, one key challenge is how to retrieve interactively the useful content and remove the noisy information. For example, a web page is usually a mixture of many types of information, such as main textual content, advertising panels, navigation bars, copyright blocks, images, etc. In real-world applications, only part of information, typically the main textual content, is useful and the rest is treated as noise. Therefore, it is important to remove the irrelevant information, which is still a hot research topic in the area of information retrieval. Second, a text corpora may contain text strings of different distributions, such as different lengths and language usages. For example, news articles and formal publications are usually long and consist of sentences with grammatical rules, while tweets or micro-blogs are short, noisy, and with limited context information. This makes it impractical to use a unified text mining model to analyze them together. As a result, another challenge is how to effectively improve the quality of a text corpora with inconsistent data distributions.

Trajectory data. Quality management for trajectories is challenging. Users need to understand the nature of the data and of the phenomenon they represent for assessing the data quality properly and, subsequently, validating the data correctly. Concerning the phenomenon, it is necessary to distinguish different modes of movement (e.g., separate walking from biking or using public transportation, and then use different constraints concerning the physics of movement: speed, acceleration, inertia) and take into account the context of movement (e.g., multiple cars on a road must follow common direction). Concerning the data, it is necessary to take into account the data collection procedure (e.g., positions collected every minute, every 20m during straight movement, by performing certain activities such as making a call).

It is necessary to understand the coverage properties of a data set under inspection, ensuring that it corresponds to analysis tasks. Often data sets are limited in spatial extent, causing complete trajectories or their parts being absent in data. Sometimes data collection is impossible in specific conditions (e.g., positioning device does not work in tunnels or indoor). Proper temporal coverage is essential, too. Another coverage aspect that needs to be considered is population: for example, it is dangerous to make conclusions about mobility of elderly people based on data derived from positions of social media activities that are performed mostly by youngsters. Another example in vehicle traffic: projecting mobility patterns of public buses onto individual cars is doubtful.

Graph data. Similarly to other data types, quality management for graph data is challenging due to the complexity of the data. In addition to the challenges for various types of attributes, the topology structure of a graph adds more challenges. Users need to apply a variety of techniques for missing values, duplicates, uncertainty, and outlier detection to handle different types of attributes (such as numbers, texts and images) and the topology of the graph. Furthermore, multiple relationships change over time (i.e., dynamic graphs) adding more challenges.

For example, users may need to compute statistics about the topology of a graph, such as the density, diameter, clustering coefficient, connectivity and average neighbor degree, as well as the property testing on the topology structure of a graph, such as testing whether a given graph is a tree (i.e., no cycle), a planar graph (i.e., can be drawn in the plane without edge crossings), or a Directed Acyclic Graph (DAG).

In addition to the attribute-based outlier detections, users need to perform topology-based pattern detections. Examples include finding high frequency patterns such as *motifs*, a small subgraph consists of three or four vertices, small cycle patterns such as *triads* (i.e., triangles), and other special subgraphs such as paths, trees, stars, and complete subgraphs. Algorithms for finding such special topological patterns are complex with high runtime complexity.

Users may need to analyze topology-based constraints. For example, a tree has $n - 1$ edges, and a planar graph can have at most $3n-6$ edges, where n is the number of vertices. Others include domain specific constraints; for example a family tree should not have a cycle, which represents a blood marriage.

Common challenges and opportunities. In addition to the aforementioned data-type-specific research challenges and opportunities, there are also some common ones. First, existing visual data cleansing methods cannot be scalable to large-scale datasets. One potential solution to handle large-scale datasets is to sample only a small subset of the whole training set. The challenge here is how to develop effective sampling methods that can both keep the data density and preserve important data such as influential points, outliers, and exceptions. Second, there is a lack of effective quality metrics to measure the quality of different types of data such as textual data, images, videos, graph data, and trajectory data. As a result, a potential research opportunity is to develop quality metrics from data content and evaluate them within specific usage contexts. In real-world applications, the analyst often needs to examine multiple types of data and correct the errors among them. Accordingly, the third challenge is designing an integrated interface to visually illustrate the distributions of different types of data.

6.2. *Human complexity*

Several challenges will arise when human intelligence is integrated into automatic data cleaning pipeline. We classify these challenges into three categories and identify the associated opportunities as follows.

- **Lack of domain knowledge.** Better integration of human domain knowledge plays an important role in steering data quality. However, sufficient knowledge or expertise regarding new types of data or new data sets might not be always available. To overcome such a challenge, it is important to explore how to tackle integration and calibration of insufficient or incomplete knowledge and expertise. One important direction is to design progressive or collaborative visual interfaces that enable crowdsourcing, such that users without sufficient knowledge can gradually gain more knowledge or seek support of other users with sufficient knowledge. It would also be interesting to create a visual recommendation mechanism for providing necessary automation in data cleaning, especially when users lack domain knowledge.
- **Limitations of perception/cognition.** Prior psychological studies have revealed limitations on visual perception and cognition, such as restricted field of view in perception (Creem-Regehr et al., 2005) and limited working memory in cognition (Baddeley, 2003). These limitations can directly influence the way in which human perceive and understand the world. For complicated types of data, it is challenging to design a visual analytics system while keeping the complexity of the system within the limitations. One worthy research direction is to explore a mixed initiative mechanism which seamlessly integrates system initiative guidance and user initiative guidance for better human machine intelligence, such that perception or cognition limitations of users can be largely addressed.
- **Difficulty in understanding uncertainty and its implications.** Uncertainty might arise in any stage of a data cleaning process, and propagate in subsequent stages (Wu et al., 2012). Misunderstanding the uncertainty and its implications would result in erroneous decisions and low-quality data. However, understanding the uncertainty and its implications would be generally difficult without a proper visual guidance. To circumvent the problem, it is highly necessary to model and visualize the uncertainty in data cleaning, such that users can make informed decisions during the data cleaning process.

7. Conclusion

Data quality is of crucial importance to a wide variety of real-world applications. In this paper, we review and summarize research efforts on steering data quality, with a focus on data cleansing, a widely-used technique for effective data quality management. First, we summarize the relevant work from different research fields, including data management, visual analytics and human computer interaction. Then, for different types of data, we discuss the common methods for improving data quality by leveraging data cleansing techniques. Building upon the existing analysis pipeline of data cleansing by Van den Broeck et al. (2005), we further propose a visual analytics framework for iteratively and progressively improving data quality from the screening, diagnosis and correction stage. Finally, we analyze the research challenges and opportunities in the context of data and human complexities, which we believe are critical for future research on visual data cleansing.

Acknowledgment

This research was funded by National Key R&D Program of China (No. SQ2018YFB100002), the National Natural Science Foundation of China (No.s 61761136020, 61672308), Microsoft Research Asia, Fraunhofer Cluster of Excellence on "Cognitive Internet Technologies", EU through project Track&Know (grant agreement 780754), NSFC (61761136020), NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization (U1609217), Zhejiang Provincial Natural Science Foundation (LR18F020001), NSFC Grants 61602306, and Fundamental Research Funds for the Central Universities. The authors would like to thank the anonymous reviewers for their thoughtful comments.

References

Abedjan, Z., Akcora, C.G., Ouzzani, M., Papotti, P., Stonebraker, M., 2015. Temporal rules discovery for web data cleaning. Proc. Very Large Database Endow. 9 (4), 336–347.

Abedjan, Z., Chu, X., Deng, D., Fernandez, R.C., Ilyas, I.F., Ouzzani, M., Papotti, P., Stonebraker, M., Tang, N., 2016. Detecting Data Errors: Where are we and what needs to be done? Proc. Very Large Database Endow. 9 (12), 993–1004.

Amiri, A., 2007. Dare to share: Protecting sensitive knowledge with data sanitization. Decis. Support Syst. 43 (1), 181–191.

Andrienko, N., Andrienko, G., 2005. Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach. Springer-Verlag, Berlin, Heidelberg.

Andrienko, G., Andrienko, N., Bak, P., Keim, D., Wrobel, S., 2013. Visual Analytics of Movement. Springer Publishing Company, Incorporated.

Andrienko, G., Andrienko, N., Fuchs, G., 2016. Understanding movement data quality. J. Locat. Based Serv. 10 (1), 31–46.

Arbesser, C., Spechtenhauser, F., Mühlbacher, T., Piringer, H., 2017. Visplause: Visual data quality assessment of many time series using plausibility checks. IEEE Trans. Vis. Comput. Graphics 23 (1), 641–650.

Baddeley, A., 2003. Working memory: looking back and looking forward. Nat. Rev. Neurosci. 4, 829839.

Van den Broeck, J., Cunningham, S.A., Eeckels, R., Herbst, K., 2005. Data cleaning: detecting, diagnosing, and editing data abnormalities. Public Libr. Sci. Med. 2 (10), e267.

Choo, J., Liu, S., 2018. Visual analytics for explainable deep learning. IEEE Comput. Graph. Appl. 38 (4), 84–92.

Creem-Regehr, S.H., Willemsen, P., Gooch, A.A., Thompson, W.B., Creem-Regehr, S.H., Willemsen, P., Gooch, A.A., Thompson, W.B., 2005. The influence of restricted viewing conditions on egocentric distance perception: Implications for real and virtual indoor environments. Perception 34 (2), 191–204.

Dasu, T., Loh, J.M., 2012. Statistical distortion: Consequences of data cleaning. Proc. Very Large Database Endow. Endow. 5 (11), 1674–1683.

Dixit, P.M., Suriadi, S., Andrews, R., Wynn, M.T., ter Hofstede, A.H., Buijs, J.C., van der Aalst, W.M., 2018. Detection and interactive repair of event ordering imperfection in process logs. In: International Conference on Advanced Information Systems Engineering. Springer, pp. 274–290.

Domadiya, N.H., Rao, U.P., 2013. Hiding sensitive association rules to maintain privacy and data quality in database. In: IEEE International Conference on Advance Computing Conference. pp. 1306–1310.

El Bekri, N., Peinsipp-Byma, E., 2016. Assuring data quality by placing the user in the loop. In: International Conference on Computational Science and Computational Intelligence. pp. 468–471.

Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S., 2007. Duplicate record detection: A survey. IEEE Trans. Knowledge Data Eng. 19 (1), 1–16.

Fan, W., Geerts, F., 2012. Foundations of Data Quality Management. Morgan & Claypool Publishers.

Fan, W., Li, J., Ma, S., Tang, N., Yu, W., 2012. Towards certain fixes with editing rules and master data. Very Large Database J. 21 (2), 213–238.

Florescuand, D., 2000. An extensible framework for data cleaning. In: IEEE International Conference on Data Engineering. 312–312.

Geerts, F., Mecca, G., Papotti, P., Santoro, D., 2014. Mapping and cleaning. In: IEEE International Conference on Data Engineering. pp. 232–243.

Gill, S., Lee, B., 2015. A framework for distributed cleaning of data streams. Procedia Comput. Sci. 52, 1186–1191.

Gschwandtner, T., Aigner, W., Miksch, S., Gärtner, J., Kriglstein, S., Pohl, M., Suchy, N., 2014. TimeCleanser: A visual analytics approach for data cleansing of time-oriented data. In: International Conference on Knowledge Technologies and Data-driven Business. 18:1–18:8.

Gschwandtner, T., Erhart, O., 2018. Know your enemy: Identifying quality problems of time series data. In: IEEE Pacific Visualization Symposium. pp. 205–214.

Gschwandtner, T., Gärtner, J., Aigner, W., Miksch, S., 2012. A taxonomy of dirty time-oriented data. In: International Conference on Availability, Reliability, and Security. pp. 58–72.

Guo, P.J., Kandel, S., Hellerstein, J.M., Heer, J., 2011. Proactive wrangling: mixed-initiative end-user programming of data transformation scripts. In: ACM Symposium on User Interface Software and Technology. pp. 65–74.

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N.H., Weaver, C., Lee, B., Brodbeck, D., Buono, P., 2011a. Research directions in data wrangling: Visualizations and transformations for usable and credible data. Inf. Vis. 10 (4), 271–288.

Kandel, S., Paepcke, A., Hellerstein, J., Heer, J., 2011b. Wrangler: Interactive visual specification of data transformation scripts. In: ACM Special Interest Group on ComputerHuman Interaction. pp. 3363–3372.

Kandel, S., Parikh, R., Paepcke, A., Hellerstein, J.M., Heer, J., 2012. Profiler: Integrated statistical analysis and visualization for data quality assessment. In: Proceedings of the International Working Conference on Advanced Visual Interfaces. pp. 547–554.

Khayyat, Z., Ilyas, I.F., 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: ACM Special Interest Group on Management of Data. pp. 1215–1230.

Krishnan, S., Wang, J., Wu, E., Franklin, M.J., Goldberg, K., 2016. ActiveClean: interactive data cleaning for statistical modeling. Proc. Very Large Database Endow. 9 (12), 948–959.

Kwon, O., Lee, N., Shin, B., 2014. Data quality management, data usage experience and acquisition intention of big data analytics. Int. J. Inf. Manag. 34 (3), 387–394.

Liu, S., Chen, C., Lu, Y., Ouyang, F., Wang, B., 2019. An interactive method to improve crowdsourced annotations. IEEE Trans. Vis. Comput. Graph. 25 (1), 235–245.

Liu, S., Wang, X., Collins, C., Dou, W., Ouyang, F., El-Assady, M., Jiang, L., Keim, D., 2018. Bridging text visualization and mining: A task-driven survey. IEEE Trans. Vis. Comput. Graph. (in press).

Liu, S., Wang, X., Liu, M., Zhu, J., 2017. Towards better analysis of machine learning models: A visual analytics perspective. Vis. Inform. 1 (1), 48–56.

Liu, S., Zhou, M.X., Pan, S., Song, Y., Qian, W., Cai, W., Lian, X., 2012. TIARA: Interactive, topic-based visual text summarization and analysis. ACM Trans. Intell. Syst. Technol. 3 (2), 25.

McCurdy, N., Gerdes, J., Meyer, M., 2019. A framework for externalizing implicit error using visualization. IEEE Trans. Vis. Comput. Graph. 25 (1), 925–935.

Modi, C.N., Rao, U.P., Patel, D.R., 2010. Maintaining privacy and data quality in privacy preserving association rule mining. In: International Conference on Computing Communication and Networking Technologies. pp. 1–6.

Oliveira, S.R., Zaïane, O.R., 2003. Protecting sensitive knowledge by data sanitization. In: IEEE International Conference on Data Mining. pp. 613–616.

Prokoshyna, N., Szlichta, J., Chiang, F., Miller, R.J., Srivastava, D., 2015. Combining quantitative and logical data cleaning. Proc. Very Large Database Endow. Endow. 9 (4), 300–311.

Raman, V., Hellerstein, J.M., 2001. Potter's wheel: An interactive data cleaning system. In: Very Large Database Conference, Vol. 1. pp. 381–390.

Saha, B., Srivastava, D., 2014. Data quality: The other face of big data. In: IEEE International Conference on Data Engineering. pp. 1294–1297.

Schulz, C., Burch, M., Beck, F., Weiskopf, D., 2015. Visual data cleansing of low-level eye-tracking data. In: Eye Tracking and Visualization. pp. 199–216.

Song, H., Szafir, D.A., 2019. Where's my data? Evaluating visualizations with missing data. IEEE Trans. Vis. Comput. Graph. 25 (1), 914–924.

Stonebraker, M., Bruckner, D., Ilyas, I.F., Beskales, G., Cherniack, M., Zdonik, S.B., Pagan, A., Xu, S., 2013. Data curation at scale: The data tamer system. In: Conference on Innovative Data Systems Research.

Vartak, M., Rahman, S., Madden, S., Parameswaran, A., Polyzotis, N., 2015. SeeDB: efficient data-driven visualization recommendations to support visual analytics. Proc. Very Large Database Endow. 8 (13), 2182–2193.

Wu, E., Madden, S., 2013. Scorpion: Explaining away outliers in aggregate queries. Proc. Very Large Database Endow. 6 (8), 553–564.

Wu, Y., Yuan, G.-X., Ma, K.-L., 2012. Visualizing flow of uncertainty through analytical processes. IEEE Trans. Vis. Comput. Graphics 18 (12), 2526–2535.

von Zernichow, B.M., Roman, D., 2017. Usability of Visual Data Profiling in Data Cleaning and Transformation. In: OTM Confederated International Conferences "On the Move to Meaningful Internet Systems". pp. 480–496.