Computers, Materials & Continua
DOI:10.32604/cmc.2021.011153
Article

Tech Science Press

Long-Term Preservation of Electronic Record Based on Digital Continuity in Smart Cities

Yongjun Ren[1,2], Kui Zhu[1,2], Yuqiu Gao[3], Jinyue Xia[4,*], Shi Zhou[1,2], Ruiguo Hu[1,2] and Xiujuan Feng[5]

[1]School of Computer and Software, Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science & Technology, Nanjing, 210044, China
[2]Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAEET), Nanjing University of Information Science & Technology, Nanjing, 210044, China
[3]Changwang School of Honors, Nanjing University of Information Science & Technology, Nanjing, 210044, China
[4]International Business Machines Corporation (IBM), USA
[5]School of Architectural and Surveying & Mapping Engineering, Jiangxi University of Science and Technology, Gan Zhou, 341000, China
[*]Corresponding Author: Jinyue Xia. Email: jinyue.xia@ibm.com
Received: 22 April 2020; Accepted: 28 September 2020

Abstract: Under the co-promotion of the wave of urbanization and the rise of data science, smart cities have become the new concept and new practice of urban development. Smart cities are the combination of information technology represented by the Internet of Things, cloud computing, mobile networks and big data, and urbanization. How to effectively achieve the long-term preservation of massive, heterogeneous, and multi-source digital electronic records in smart cities is a key issue that must be solved. Digital continuity can ensure the accessibility, integrity and availability of information. The quality management of electronic record, like the quality management of product, will run through every phase of the urban lifecycle. Based on data quality management, this paper constructs digital continuity of smart city electronic records. Furthermore, the work in this paper ensures the authenticity, integrity, availability and timeliness of electronic documents by quality management of electronic record. This paper elaborates on the overall technical architecture of electronic record, as well as the various technical means needed to protect its four characteristics.

Keywords: Smart city; electronic record; long term preservation; digital continuity

1 Introduction

Recently, the operation mode of modern cities and the living environment of urban residents have undergone fundamental changes as information and digital transformations [1–3]. Every aspect of city life, such as the economy, culture, traffic, entertainment and other, have been closely integrated with information technology. The cyberspace has become an integral part of urban life. The basic features of modern digital cities consist of a developed information infrastructure

and a wealth of digital applications. The rich achievements of urban information have brought great convenience to our daily life and laid the foundation for the further evolution of modern urban form [4,5].

Data-centered research methods and techniques have been widely used and recognized in various fields such as information, biology, energy, medicine and sociology. As a result, a large number of scientific research achievements [5–7] have been published. Data analysis and research methods are profoundly changing the working methods of traditional scientific exploration and naturally becoming an emerging model for the development of human science and technology and knowledge acquisition [8–10].

Driven by the wave of urbanization and the rise of data science, smart cities have become the next level of urbanization in the next generation [11]. Based on the new generation of information technology, smart city implements a thorough understanding of urban living environment, overall regulation and control of urban resources, coordination of all parts of the city through dynamic monitoring, analysis, integration and utilization of. Countries in the world, especially United States, Japan, South Korea and other developed countries, are actively carrying out relevant theoretical researches and technological explorations, exploring utilization of city's data resources and developments of urban intelligence applications. In addition, some corresponding pilots in cities have been carried out. In China, the development and construction of smart cities are actively being explored [10,12,13].

The smart city connects the city digitally through ubiquitous Internet of Things. Information data in smart city are the electronic record [14,15]. As we all know, the world generates more than imaginative big information data every day. In Beijing, the number of uses of bus cards is more than 40 million per day and 10 million for subway stations. Daily traffic data of Beijing traffic control center has increased to 30 gigabytes (GB), and storage volume to 20 terabytes (TB). National grid generates 510 TB of annual data (excluding video). Medical data like the CT image of a single patient is often up to two thousand megabytes (MB), and the amount of data has reached dozens of GB. Moreover, the massive space of 2 and 3 dimensional data in digital earth is growing rapidly, and it will reach TB and petabyte (PB) level respectively soon [16,17]. Civil aviation aircraft that is equipped with a large number of sensors generates 20 TB data for each flight every hour. The flight from London to New York could generate 640 TB data. The data of these engine states are monitored through a satellite to the engine company during the flight [18,19]. Nowadays, outpatient of hospitals in the large cities of China exceeds more than the thousands of cases every day. The number of outpatients in the country is up to billions each year, and the hospitalized patients have reached two hundred million. According to relevant regulations of medical industry, data of patients usually need to be saved more than 50a, and large medical electronic record will reach the level of exabytes (EB) [20–24].

Hence, in the process of building smart cities, it is bound to produce a huge amount of electronic records. The preservation of these electronic records has also become an important research topic. This paper puts the angle of view on the informatization of the city. The content focuses on the research hotspot of electronic records preservation in smart city. It drives the principle of digital continuity for long-term preservation, and gives a concrete technical framework based on data quality theory in the paper.

2 Related Works

2.1 Technology System of Smart City

Technology system is the top design of the technology research in smart city. It is very important because it guides the direction of technology development, defines the connotation and extension of research work, and optimizes the allocation and distribution of existing research resources.

An early technology system research of smart city is initiated by IBM researchers. Technical functions of Smarter CitiesTM, proposed by IBM [25–27], emphasize the importance of service and infrastructure as the center of smart cities. In the literature [28], from the point of view of integrated intelligence to understand the framework of the city, a smart city's initial framework was proposed. This framework believes that smart cities should integrate government, residential communities, economy, infrastructure and natural environment from the perspectives of policy, organization and technology. Literature [29] summarizes some early smart city technology frameworks and argues that the essential elements of smart cities include people, institutions and technologies. They can be subdivided into digital cities, smart cities, wireless cities, information city and so on. The research on technology systems of smart city is mainly based on the basic elements of urban construction. However, there is not much discussion on the role of data science and technology in the smart city.

In recent years, academia has begun to pay close attention to the important role of data science in smart cities. Hence, some corresponding technical systems are making progresses. As shown in Fig. 1, Pan et al. [30] elaborated the technical framework of smart city based on trajectory data analysis and mining on smart communications Symposium of IEEE Communication Magazine in 2013. This framework divides the technology system in smart city based on trajectory data into three levels: trace, knowledge and application. The framework systematically describes the overall technical route to data-centric smart city technology from a strategic perspective.

Zheng et al. [31] of Microsoft Asia Research proposed a framework of urban computing technology with four-layer feedback structure in the topic of city computing in CCF Communications in 2013. The technology system divides the technical framework of urban computing into 4 levels: urban perception and data capture, urban data management, urban data analysis, and service provision. The feature of this framework is to introduce a feedback loop of service providing layer to the actual implementation, which further considers the influence of smart city technology on urban life.

2.2 Digital Continuity

As a valuable national resource and commercial asset, digital information has driven all governments to deal with the proper preservation and long-term use of them. Rationale behind digital continuity plan can be traced back to the record continuum theory. The theory of record continuum was first proposed in the 1950s by Ian Maclean, a famous Australian archivist and the first director of National Archives. From 1980 to 1990, the theory has been evolved and scholars reinterpreted and further developed the records continuum management model. Then it became a fundamental part of Australian electronic record and digital information management in principle. The Australian Document Management Standards, published in 1996, interprets records continuum as "a coherent approach to management of the entire process from the formation of documents (including the design of document management systems) to the management of documents as archives." The theory of records continuum breaks the stage distinction of document

custody. By constructing a multidimensional coordinate system containing 4 dimensions, 4 major axes and 16 focal points, the paper describes the increasingly vague boundaries of life movement and and the obvious connection of electronic records. It also reveals the continuity and integrality of the file movement in time and space. Lastly, it provides a scientific theoretical basis for the digital continuity plan [32,33].

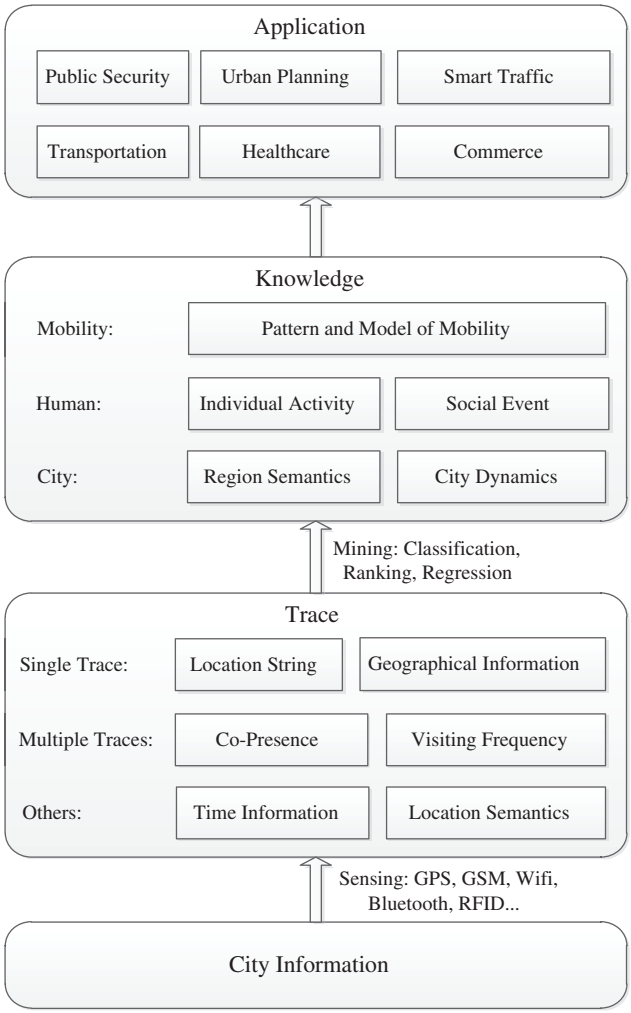

Figure 1: Framework of trace for smart cities

As shown in Fig. 2, Digital continuity is a strategy for the long-term preservation of digital information. It emphasizes the importance of the guidance of government agencies and business groups. As a way to manage and preserve digital information, digital continuity strengthens the management and preservation of native electronic records from different perspectives such as meeting the needs of institutions, governments and communities, the goal of ensuring the availability of digital information as needed [34–36]. Digital continuity ensures that the digital information is complete, available, and the time of storage cannot exceed the period for which it is required to be kept. Among them, usability is the ultimate goal of digital continuity planning.

More specifically, they are: 1. When you need to find it; 2. When you need to open it; 3. To deal with it the way you need it; 4. To understand that what it is and what it is to express? 5. Can trust what it says.

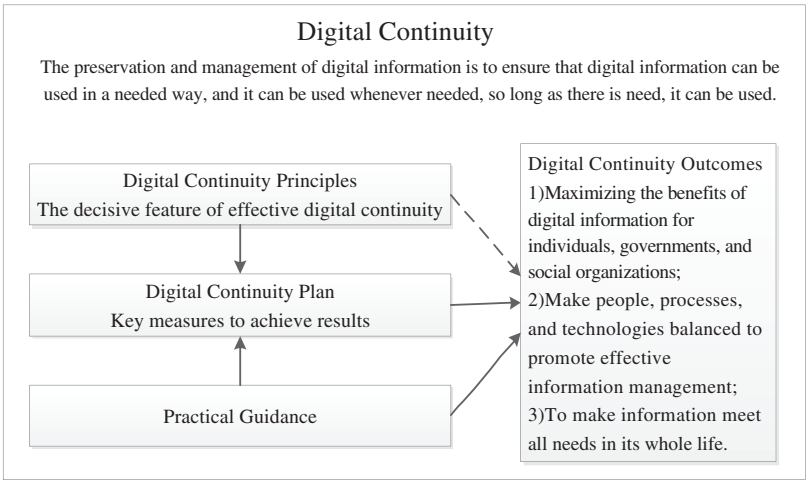

Figure 2: Basic elements and relationship of digital continuity

3 Problem Statement

3.1 Types of Electronic Record in Smart City

In the city, the information infrastructure not only provides information services, but also accumulates a large amount of urban dynamic electronic records. These records are numerous and contain the following common types.

3.1.1 Map and Electronic Record of Interest Point

Streets and buildings are the basic components of a city. Digital map is the a way to show the urban architecture. The electronic record of interest point is the basic information of each functional unit in city.

3.1.2 Electronic Record of Passenger Flow

The electronic record of commuters commuting by different means of transportations is called electronic record of passenger flow. The record contains a lot of urban activity information and can be heavily used in urban functional analysis, population flow monitoring, urban transportation system evaluation, human behavior research, urban transportation economics research and other fields.

3.1.3 Electronic Record of Location Service

Location service is a new type of network service in the mobile Internet era. The electronic record collects through location services provides a well-defined geographical location and coordinates as well as the semantic characteristics of traditional Web services. We can treat the electronic record of location service as a deep description and complement of interest point records. Compared with simple urban geographic data, the electronic record of location ser-

vice contains a large amount of semantic information, and can help people understand the urban operation.

3.1.4 Electronic Record of Video Surveillance

Video surveillance technology has been widely used in traffic management, community security, security protection systems and other parts of urban life. Moreover, making full use of these video records can revisist the history of city life, which has great theoretical and practical value.

3.1.5 Electronic Record of Environment and Meteorology

Electronic record of meteorology was paid much attention to by Urban Science in the early days. In recent years, with the environmental and health issues such as the urban environmental record represented by air quality. An important feature of urban environmental and meteorological record is its low sampling density of geography and temporality. How to achieve fine, high-precision environment and weather record and analyze them is an important challenge.

3.1.6 Social Activity Record

Social activity record is an essential ingredient for in-depth understanding and analysis of urban social behavior. The electronic record of urban social activities includes urban population record, household registration, medical and health care, energy consumption and other social dynamic record. Because of the feature of urban social activities record, the records are vulnerable to the influence of industry segmentation, and often isolated from each other. Breaking the industry fragmentation and realizing the integration of multi-source and heterogeneous urban record are the primary task for the deep use of urban social activity record.

3.2 Characteristics of Electronic Record in Smart City

The characteristics of electronic record in smart city are described as follows.

3.2.1 Spatial and Temporal Characteristics

The electronic record of spatial structure based on map is a basic organization of urban record. The urban fast-paced lifestyle also makes urban record very sensitive to changes in the temporal dimension. Therefore, multi-dimension characteristics with spatial and temporal dimension become an important feature for urban record. In space, according to the size of urban geography, urban record has different scales of space span. According to the time of generation, the changes and distributions of urban record are time- dependent. Therefore, in the analysis and application of urban record, we not only need to consider the evolution characteristics of record but also make full use of the record in the two dimensions of time and space. This makes a great demand for the utilization of urban record.

3.2.2 Multi-Scale and Multi Granularity

To study and utilize urban record, we need to consider the influence of record scale and granularity on record characteristics, in addition to multi dimensions, such as time and space. According to the scale of the city, the cities can be divided into small cities, medium-sized cities, large cities and super large cities and so on. Geographically, the description of urban record varies from several street blocks to thousands of square kilometers. On the time scale, the coverage time of city record can be short to some event monitoring, and can grow to the hundred year city vicissitude. In the geographical sampling granularity, the record can be as accurate as a few meters, and can also be the same as the environment record as units with counties, regions, and even provinces. In time granularity, the record can be sampled based on the clock of sampling

device, storage and transmission capacity, computing speed and other factors. In the time and space multi-dimensional condition, the efficient processing of large scale and multi granularity record is one of the key technical problems that must be solved effectively by the urban record.

3.2.3 Pluralism and Isomerism

There are many types and sources of urban record, i.e., the diversity of electronic record. These different sources of urban record differ greatly in the aspects of structure, organization, dimension, scale and granularity, namely the heterogeneity of electronic record. The application services of smart cities require that these heterogeneous data organically should integrate and acquire new knowledge by mining the correlation and interaction between active record. It is a common challenge for the academia and industry to explore the intelligent city in order to analyze the heterogeneous record.

3.3 Challenge of Electronic Record Preservation in Smart City

In the development of smart city, the sources of electronic record had changed from a single source to multi sources. The value of record is also translated into multiple uses. The functionality of record is shifting from the structure level to the organizational level. Moreover, due to electronic record from multiple sources, it is challenging to achieve quality assurance with authenticity, reliability, integrity, and. There is a growing need to ensure quality of electronic record.

In addition, stakeholders are increasingly focusing on changes and reliability issues of electronic record. In general, for applications services of smart city, sensors fall into 3 broad categories. First, the quantitative sensors, which are used to observe environmental variables, monitor security events and obtain flow information. The second is the video sensor, which plays an important role in traffic monitoring and monitoring in city. The third is the position sensor, which is used to record the time and location of the object during the moving process.

The quantitative sensors are usually deployed over fixed positions to collect information about soil, hydrology, and the atmosphere at regular intervals. They are important approaches to perceive the natural environment and its changes in city. Numerical record is often observed by quantitative sensors, such as temperature, barometric pressure, water level, and so on. They can also be processed categorical record, such as the output of PM2.5 observations can be the following 6 categories: excellent, good, mild pollution, moderate pollution, severe pollution and severe pollution. The quantitative sensors are usually used together with the sensor observation service, and the Insert Observation method is exploited to embed the observation record to the service system. Video sensors are usually deployed in fixed positions, with basic functions such as video photography and static image capture. Obviously, video record has both spatial and temporal properties. Position sensors are utilized and usually obtain spatial-temporal sequence record [37]. Moreover, the position sensor and the environment sensor are integrated on the moving object, and the location and the environment monitoring records can be obtained at the same time.

Because electronic record comes from a variety of sensors, preservation of them is facing enormous challenges to ensure reliable record as evidence.

3.4 Electronic Record Preservation Based on Digital Continuity

Digital continuity is the useful method to protect the long-term use of the information. Specifically, it refers to the maintenance and management of digital information to ensure that

it can be used now and in the future when needed. Thus, the preservation of electronic record based on digital continuity will guarantee their quality. If the quality of electronic record used is not good enough and it will probably have seriously negative impacts on subsequent data processing [38–40]. The quality management of electronic record, like the quality management of product, needs to run through every phase of the data life cycle. According to the theory of data quality, the preservation of electronic record can be shown in Fig. 3.

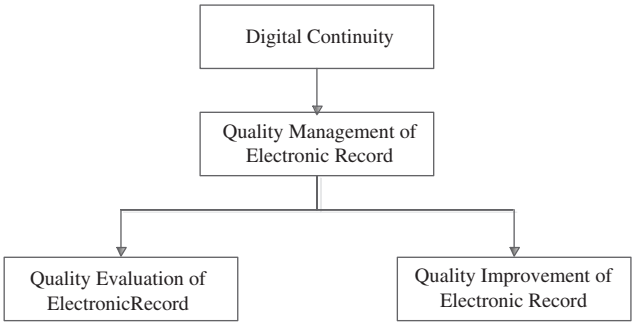

Figure 3: The preservation of electronic record based on digital continuity

The quality management of electronic record concerns the following research topics [41,42].

Quality requirements. The quality requirements of electronic record enable users to obtain specific quality record, and they are the indicators that need to be appended to the electronic record. Thus, they are the measurements of the quality of electronic record. The quality requirements of electronic record are generally expressed in the form of quality parameters. These common quality parameters include source credibility, availability, timeliness, and stability of electronic record.

Quality evaluation. The quality evaluation of electronic record should include at least two basic evaluation indicators. One is credibility of electronic record; and the other one is the usability of electronic record. In general, the former contains accuracy, completeness, consistency, validity, uniqueness of electronic record. The latter includes time and stability of electronic record.

Quality improvement. The purpose of electronic record quality improvement is to decrease the rate of defective record. The quality problem of electronic record can be divided into pattern level and instance level. The quality problem of pattern level is primarily the patterns design of electronic record, which exists at higher levels. The methods of pattern selection, matching and optimization can be chosen to resolve the issues. And the higher management and work models can be also reconstructed. The examples of the quality problems of instance level include data duplication, data missing, abnormal data, logical errors and inconsistent data, and so on.

Quality management system. The research of quality management system of electronic record mainly focuses on two aspects. On the one hand, in order to emphasize the integrality of quality management, the concept, principles and methods of total quality management have been introduced into quality management of electronic record. On the other hand, the process maturity model has been introduced into the quality management system due to the importance of process management.

To ensure digital continuity, it is necessary to have an accurate examination of the quality status of electronic record. Thus, recording the digital information and processing is recommended by the digital continuity plan. The recordings are created, captured and used during the business

management process. Moreover, the plan requires a clear responsibility, obligations, conditions, costs, benefits of digital information management.

4 Electronic Record Preservation Based on Quality Management

4.1 Quality Evaluation of Electronic Record

Quality evaluation and monitoring is one of the most critical problems for quality management of electronic record. It is generally believed that record quality is a concept of hierarchical classification. And each quality class is eventually broken down into specific record quality dimensions. The core of quality evaluation for electronic record is how to evaluate each dimension concretely. The evaluation methods are divided into two categories: qualitative strategies and quantitative strategies. To analyze the "good" or "bad" from a qualitative perspective is an important performance metric of evaluation. Due to the lack of objectivity and reproducibility of qualitative analysis, quantitative assessment technology has become an alternative method. Based on data quality theory, the quality evaluation of electronic record mainly examines the accuracy, integrity, consistency and timeliness of electronic record, which is shown in the following Fig. 4.

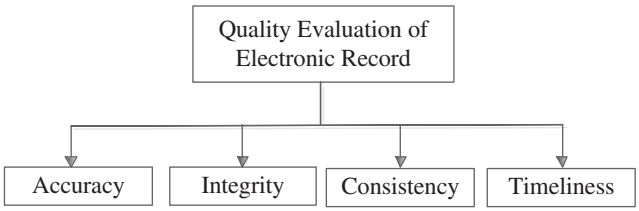

Figure 4: Quality evaluation of electronic record

4.1.1 Accuracy Evaluation of Electronic Record

Accuracy refers to the degree to which the record provides the correct and objective representation of the resource to be described. The method of accuracy measurement is to calculate the semantic distance between the information from user records and the information obtained by the same user from the resource itself. The distance indicates that the matching accuracy of the content of record and the content of the resource itself. The computation of semantic distance, in the field of information retrieval to calculate the similarity of vector space model between two texts, can be used as a reference. Two multidimensional vectors are created based on the different words contained in the text field of the record and record of the described resource. The value of each dimension in the vector of the described resource equals the relative frequency that the dimension counterpart appears in the text of the described resource. Similarly, the electronic record corresponding to the vector is constructed. The distance between the two vectors can be computed by the most commonly used cosine function. Thus, the semantic distance between the record and the described resource is obtained, namely the accuracy of electronic record. The formula is as follows:

$$RQ_{au} = \sum_{i=1}^{n} (\sigma_i \times \xi_i) \Big/ \sqrt{\sum_{i=1}^{n} \sigma_i^2 \times \sum_{i=1}^{n} \xi_i^2} \tag{1}$$

where σ_i and ξ_i are the frequency of the first i value in the record and electronic records of the described resource relatively. n is the total number of different values in two records. In practical applications, because resource and record creators may not be the same person, they may use different records when expressing the same semantic. Therefore, it is necessary to adopt the semantic analysis technique of artificial intelligence to reduce the interference of multiple records phenomenon on the accuracy value, and reduce the dimension and the amount of calculation as well. The retrieval systems of different subject areas have already constructed useful relative thesaurus and thesaurus, which can provide reference for such semantic analysis.

4.1.2 Integrity Evaluation of Electronic Record

The integrity of electronic record requires that record to contain a comprehensive description of the target's resources. Whether or not it can describe the resources in a comprehensive way is closely related to the record specifications and the guidelines for the use of the record library. A field of the same record specification is an essential attribute in a digital resource library, and may be optional in another digital resource library. Another factor is that the filed also affects integrity is the type of resource. Traditional bibliographic record is easier to have higher integrity than record of digital resources. The record of the latter changes with the use of target resources, which increases the difficulty of maintaining high integrity. The most straightforward way to quantify integrity is to calculate the number of non-empty fields except for record specifications and resource types, which are the two most important factors. The formula is as follows:

$$RQ_{ite} = \sum_{i=1}^{n} \delta(i) / n \tag{2}$$

RQ_{ite} is called simple integrity. If the first i field is null in the formula, then $\delta(i) = 0$, otherwise $\delta(i) = 1$. n is the total number of different values. Each field in the above formula is equivalent to assessing integrity. In fact, the integrity of some records is more important than other records for the discovery, acquisition, and utilization of resources. Thus, a better way to measure integrity is to assign weights to each field, then we can calculate the number of non-empty fields, and finally get a value for the integrity of the record. The formula is as follows:

$$RQ_{wite} = \sum_{i=1}^{n} (\mu_i \times \delta(i)) \Big/ \sum_{i=1}^{n} \mu_i \tag{3}$$

where μ_i is the weight of the first i field. The range of values for RQ_{wite} is [0, 1]. When the fields with non-zero weight in a record are not empty, the value is 1; otherwise if all are empty, then 0.

Because the application service of record is different and the function should be satisfied and the provided service by record are different, the weight value of the same field may be different in different application environments. This means that each of the different application services needs to set up a set of weighting values for itself to reflect the extent to which the links between the record fields and the functions to be implemented. To determine the weight value of the main consideration is the adopted fields, namely the weight of a field value will change with the utilized frequency by the user. The more users use the field, the greater the weight.

4.1.3 Consistency Evaluation of Electronic Record

Consistency is mainly examined by record specifications, application guidelines. The common damage consistency is definied as follows: (1) Fields contained in electronic record that are

not defined by the specified record specification; (2) There is no required specified field in the record specification; (3) Some fields are not taken from the controlled vocabulary in the record specification; (4) An application guide that does not follow the record specification and integrates multiple values in the controlled vocabulary into one or some fields. At present, many record repositories use XML as the grammatical cornerstone. Thus, parsing the syntax of records is carried out by a XML parser. Then, statistics are drawn on records that violate record application rules for above mentioned four situations. Specific quantitative calculation can use formula as follows:

$$RQ_{cs} = 1 - \left(\sum_{i=1}^{n} \chi_i / n \right) \tag{4}$$

where χ_i is that record follow the first i rule. If the record meets rule i, then $\chi_i = 0$; otherwise $\chi_i = 1$; n is the number of rules in the metadata specification and in the guide, which is used in the digital resource library.

4.1.4 Timeliness Evaluation of Electronic Record

The determination of timeliness is one of the key issues to ensure the freshness of electronic record. The discovery and repair of the timeliness of electronic record determines the timeliness of the electronic record. Users are required to provide the timeliness of the electronic record, when users use electronic record.

Due to lack of effective maintenance and record integration, these timestamps are often unavailable or inaccurate even if there are timestamps. Moreover, the different attributes of the same record vary with time, resulting in different temporal characteristics of the different attributes of the same record. In addition, it is not easy to keep the corresponding timestamp for each attribute of each record. Therefore, a main challenge for record validation is that there is no complete, accurate, or usable timestamp in the record set. If the time stamp is missing or imprecise, it is difficult to determine the timeliness of the record. Some redundant records and time constraints can be used for the timeliness of record. Redundant records indicate that there are multiple records describing the same entity in the record set. Time constraint is the semantic information of record, which can help to recover the partial temporal relation of different attribute values of the same entity.

4.2 Quality Improvement of Electronic Record Independent of Knowledge

The technology of record quality improvement mainly involves two aspects: instance and model. Data cleaning is the main technology to improve the quality of electronic record. It focuses on instance level issues. By detecting and eliminating errors and inconsistencies in the record, data cleaning also improves quality of electronic record. The purpose of cleaning is mainly to improve quality of electronic record from the point of view of instance layer. The specific methods are shown in the following Fig. 5.

4.2.1 Duplicate Detection

In reality, entities often have multiple representations in different record sources. In the process of record integration, it is necessary to determine whether the expression in different record sources represents the same entity. So the problems can be solved by databases and artificial intelligence.

The same or similar records may represent the same entities, so the detection of similar duplicate records of structured data can be an issue. Fortunately, similar duplicate record recognition can be done using the following techniques: sorting and merging, building index, information identification based on context, identification based on record characteristics and identification based on specific knowledge. The detection is described in the following Fig. 6.

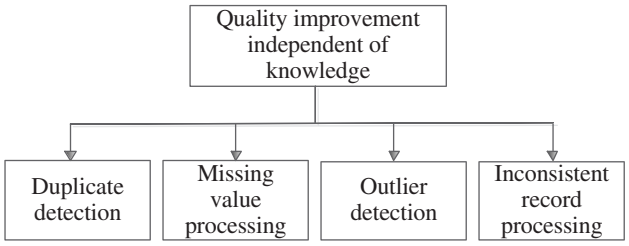

Figure 5: Quality improvement of electronic record independent of knowledge

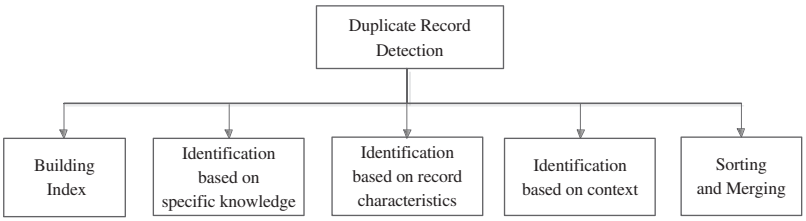

Figure 6: Detection method of duplication record

The method of sorting and merging is a generic identification algorithm, which does not rely on specific application domains. A number of attribute strings selected by the user can be sorted as keys. A fixed size sliding window clustering is used to identify similar duplicate records. In addition, multiple sorting and merging are performed according to different attributes, and a priority queue is used instead of a fixed size sliding window to cluster. The two methods may cause high frequency I/O generated by sorting outside the strings. Thus, they are very resource intensive. In addition, because the string sort is too sensitive to the character and word position, it does not guarantee the duplicate records similar in adjacent positions. As a result, the subsequent clustering operation will not identify duplicate records.

The way to build indexes using the R tree is as follows. First, a number of strings are selected as axes, and each record is calculated according to these axes. Second, the coordinates in multidimensional space are computed. The R tree is then used for multi-dimensional similarity connection to implement the recognition of similar duplicate records. Because the dimension disaster determines that the dimension cannot be too high, this method is not universal. Additionally, data cleaning methods can also be performed under the conditions of a clean reference table. The basic idea is to establish an error tolerance index on the reference table. Then the most appropriate clean record can be quickly found based on this index. The record is then replaced with the complete cleaning of the input record.

In the duplicate record recognition, it is often invalid to determine the duplicate records only by the content of the data. Contextual information can be used to eliminate suspicious links.

For example, for a similar duplicate record that cannot be verified, contextual similarity can be computed to determine the underlying information hidden in other records. The idea of using context related information provides a good way to identify difficult and duplicated records.

In reality, information systems are oriented to a specific application and associated with specific business rules. Therefore, similar duplicate records can be identified by establishing rules based on domain specific knowledge. The method can achieve good recognition accuracy according to the business rules. However, the main problem of this method is that in order to identify similar duplicate records, rules library in the corresponding fields must be established. What this means is that it requires higher knowledge of the field, and the workload of manual definition is relatively large. Another approach is using fuzzy duplication record recognition method which is based on the compact set feature of fuzzy duplication. In other words, duplicate records are more immediate than non-duplicate records. Moreover, the local neighbors of repeated records are sparse features. Thus, similar duplicate records are identified.

4.2.2 Missing Value Processing

The actual record set often suffers from missing value and have a great impact on the analysis results. To solve the problem, the methods of single imputation and multiple imputations can be used. The single imputation method constructs a single substitution value for missing values. The representive methods of filling include average or intermediate filling, regression filling, maximum expectation filling, and nearest filling method and so on. In the nearest filling method, the corresponding variable value is used as the filling value. But the single value filling method often cannot reflect the uncertainty of the original data set, and the drawback is that it can lead to great deviation. On the other hand, the multiple imputation method uses multiple values to populate, and analyzes them and leads to composite results. In the multiple imputation method, though, the common methods include trend score and prediction matching method. The advantage of the methods is that the relationship between variables can be maintained by simulating the distribution of missing record; while the disadvantage is that the computation is complex.

There are also corresponding filling methods for record containing vacancy values. In a method for filling missing value, the log models and log linear models are utilized to fill the missing values of data cubes in a metric set of multidimensional record sets. Some constraints can also be used to fill in the missing data cube. In relational databases, the conditional table can be used to fill the missing record tables. Furthermore, k-means and Markov chain can be adopted to fill with the missing values. Missing value fills are primarily intended to prevent analysis bias, due to the value missing of considerable records. The method of filling has statistical significance for the filling of individual value.

4.2.3 Outlier Detection

Unexpected records are usually caused by two reasons: Inherent variability in record and measurements errors. Abnormal record can be discovered by data auditing in two aspects. The first step is record generalization, in which the distribution of record is generalized by mathematical statistics, and then the overall distribution characteristics of record are obtained automatically. The second step is to excavate a specific quality problem to discover the abnormal data. At first, the record is partitioned into layers based on the distance. Then, the statistics characteristics of each layer are presented. Based on the defined distance, the distance between the data points and the center distance in the layer is computed, which is utilized to determine the possible abnormalities.

Association rules mining can also be used to implement nominal type outlier record discovery. But this method cannot be directly applied to numerical attribute processing. The use of record auditing methods to discover the effect of abnormal record largely depends on whether data mining algorithms can accurately distinguish abnormal and non-abnormal record. Therefore, to generate simulated data, a decision tree algorithm is adopted to detect and validate specific data mining algorithms to detect bias record. In general, statistical models, distance based and offset based methods all can be used to detect abnormal record. Because record is often unclean, record generalization before specific mining algorithms are executed to help in exploratory mining. As a result, more unusual records are found.

4.2.4 Inconsistent Record Processing

Overlapping content is recorded and inconsistent data appears when the records come from multiple sensor nodes. How to obtain an ideal record from a number of inconsistent results is a frequently asked question. The inconsistency between records fall into two types: context independent conflict and context dependent conflict. The former conflict results from different records from different sources due to design and presentation factors inherent in different systems or applications. And the conflicts can be solved by data transformation rules. A context independent conflict, however, is an inconsistency caused by some external random occurrence. The solution of such problems generally requires manual intervention and specific methods. Each record value is evaluated from different performance parameters, i.e., feature, and the overall evaluation value is a linear combination of each feature evaluation value, and then the value is determined as the only correct one based on the evaluation value. Records can be sorted, fused, and cleaned by rules and other methods. In addition, an extended relational model that handles both legacy and inconsistent record can also handle inconsistent record.

4.3 Quality Improvement of Electronic Record Based on Application Logic

The actual information system is designed to an application field. For a specific application, how to use automated methods to solve the data does not meet the business logic error, is a practical application of the problem. This kind of problem can be considered as data editing and imputation. The idea is to automate the process by setting up a rule system based on application dependent domain knowledge and then it can be processed by constructing mathematical model. The basic idea also includes making minimal changes to all the variables in the record to meet all editorial rules. For specific applications, explicit constraint rules are defined according to domain knowledge, and then the whole closed set of rules is calculated based on a certain mathematical method. Each record is automatically judged whether it violates the rules constraint. The advantage of this method is that the mathematical foundation is strict; the rules are automatically generated and relatively mature.

5 Discussion and Conclusion

In conclusion, data quality of electronic record in smart city is an important research aspect that has attracted great attention from industry and academia. So far, some research achievements have been focusing on the expression mechanism of usability, the theories and methods of high quality data acquisition, data error detection. Although the research on the long-term preservation of electronic records in smart cities are progressing, it is far from meeting the needs of practical applications, especially big data applications in smart cities. Considering the characteristics of electronic records in smart cities, there are still many challenges and we should carry out the following further research:

1) We need to study the theories and key techniques of data quality with multi-data set family as the object, including data quality expression mechanism of interrelated multi-dataset family, detection and repair of cross-related data errors in multiple data sets, etc. It is because the existing data quality researches mainly focus on a single data set, and do not consider the overall usability of the interrelated multi-data set family.

2) With respect to existing data quality calculation problems, its computational complexity and researches on solving algorithms still take "determine the Turing machine is solvable in polynomial time" as the standard of the problem resolvability, which could lead to the situation that many results are not suitable for the large data scale in smart cities (data size can be up to PB or even EB). Moreover, polynomial time algorithm (even linear time algorithm) is difficult to solve this calculation problem in the time. Therefore, to meet the requirement of large data quality in smart city, it is necessary to design sub-linear and logarithmic polynomial time algorithms to solve data computing problems.

Funding Statement: This work is supported by the NSFC (Nos. 61772280, 62072249), the AI recognition scoring system of weather map (No. SYCX202011), the national training programs of innovation and entrepreneurship for undergraduates (Nos. 201910300123Y, 202010300200), and the PAPD fund from NUIST. Jinyue Xia is the corresponding author.

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

References

[1] D. Li, Y. Yao and Z. Shao, "Big data in smart city," *Geomatics and Information Science of Wuhan University*, vol. 39, no. 6, pp. 631–640, 2014.

[2] Y. J. Ren, F. J. Zhu, S. P. Kumar, T. Wang, J. Wang *et al.*, "Data query mechanism based on hash computing power of blockchain in internet of things," *Sensors*, vol. 20, no. 1, pp. 207, 2020.

[3] Y. J. Ren, Y. Leng, F. J. Zhu, J. Wang and H. J. Kim, "Data storage mechanism based on blockchain with privacy protection in wireless body area network," *Sensors*, vol. 19, no. 10, pp. 2395, 2019.

[4] Q. Li, "From geomatics to urban informatics," *Geomatics and Information Science of Wuhan University*, vol. 42, no. 1, pp. 1–6, 2017.

[5] C. P. Ge, Z. Liu, J. Xia and L. M. Fang, "Revocable identity-based broadcast proxy re-encryption for data sharing in clouds," *IEEE Transactions on Dependable and Secure Computing*, 2019.

[6] Z. Chen, Z. Xu, Q. Li, W. Lu and Z. Xiong, "A novel framework of data sharing and fusion in smart city-SCLDF," *Journal of Computer Research and Development*, vol. 51, no. 2, pp. 290–301, 2014.

[7] C. P. Ge, S. Willy, Z. Liu, J. Xia, P. Szalachowski *et al.*, "Secure keyword search and data sharing mechanism for cloud computing," *IEEE Transactions on Dependable and Secure Computing*, 2020.

[8] L. Fang, C. Yin, L. Zhou, Y. Li, C. Su *et al.*, "A physiological and behavioral feature authentication scheme for medical cloud based on fuzzy-rough core vector machine," *Information Sciences*, vol. 507, pp. 143–160, 2020.

[9] L. M. Fang, Y. Li, X. Y. Yun, Z. Y. Wen, S. L. Ji *et al.*, "THP: A novel authentication scheme to prevent multiple attacks in SDN-based IoT network," *IEEE Internet of Things Journal*, 2019.

[10] Y. Lu and T. Feng, "Research on trusted DNP3-BAE protocol based on hash chain," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, pp. 893, 2018.

[11] L. Xiang, X. Wang and J. Gong, "A multi-level abstraction model for city sensing observation data," *Bulletin of Surveying and Mapping*, vol. 11, pp. 39–44, 2015.

[12] Y. Zhang, Y. Chen, B. Du, J. Pu and Z. Xiong, "Multimodal data fusion model for smart city," *Journal of Beijing University of Aeronautics and Astronautics*, vol. 42, no. 12, pp. 2683–2690, 2016.

[13] L. Gong, B. Yang, T. Xue, J. Chen and W. Wang, "Secure rational numbers equivalence test based on threshold cryptosystem with rational numbers," *Information Sciences*, vol. 466, pp. 44–54, 2018.
[14] Y. J. Ren, J. Shen, D. Liu, J. Wang and J. Kim, "Evidential quality preserving of electronic record in cloud storage," *Journal of Internet Technology*, vol. 17, no. 6, pp. 1125–1132, 2016.
[15] Y. Ren, J. Qi, Y. Cheng, J. Wang and J. Xia, "Digital continuity guarantee approach of electronic record based on data quality theory," *Computers, Materials & Continua*, vol. 63, no. 3, pp. 1471–1483, 2020.
[16] X. Liu, Z. Hu and S. Pan, "Control strategy for the number of replica in smart city cloud storage system," *Geomatics and Information Science of Wuhan University*, vol. 41, no. 9, pp. 1205–1210, 2016.
[17] W. Wan, J. Chen and S. Zhang, "A cluster correlation power analysis against double blinding exponentiation," *Journal of Information Security and Applications*, vol. 48, no. 10, 102357, 2019.
[18] Z. Lu and G. Chen, "Research on prediction model of flights alternate probability distribution," *Computer Engineering and Applications*, vol. 53, no. 20, pp. 259–264, 2017.
[19] J. Xu, Y. J. Zhang, K. Y. Fu and S. Peng, "SGX-based secure indexing system," *IEEE Access*, vol. 7, pp. 77923–77931, 2019.
[20] C. X. Wang, X. Shao, Z. Gao, C. X. Zhao and J. Gao, "Common network coding condition and traffic matching supported network coding aware routing for wireless multihop network," *International Journal of Distributed Sensor Networks*, vol. 15, no. 6, pp. 1–20, 2019.
[21] Y. C. Mao, J. H. Zhang, H. Qi and L. B. Wang, "DNN-MVL: DNN-multi-view-learning-based recover block missing data in a dam safety monitoring system," *Sensors*, vol. 19, no. 13, pp. 2895, 2019.
[22] W. Zhao, J. J. Liu, H. Z. Guo and T. Hara, "ETC-IOT: Edge-node-assisted transmitting for the cloud-centric internet of things," *IEEE Network*, vol. 32, no. 3, pp. 101–107, 2018.
[23] D. Li, J. Shan and Z. Shao, "Geomatics for smart cities-concept, key techniques, and application," *Geo-Spatial Information Science*, vol. 16, no. 3, pp. 13–24, 2013.
[24] W. Zhang, F. Y. Shih, S. Hu and M. Jian, "A visual secret sharing scheme based on improved local binary pattern," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 32, no. 6, 1850017, 2018.
[25] Y. Chen, J. Wang, R. Xia, Q. Zhang, Z. Cao *et al.*, "The visual object tracking algorithm research based on adaptive combination kernel," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 12, pp. 4855–4867, 2019.
[26] Y. J. Ren, J. Qi, Y. Liu, J. Wang and G. Kim, "Integrity verification mechanism of sensor data based on bilinear map accumulator," *ACM: Transactions on Internet Technology*, 2020.
[27] C. Harrison, B. Eckman and R. Hamilton, "Foundations for smart city," *IBM Journal of Research and Development*, vol. 54, no. 4, pp. 1–16, 2010.
[28] F. Victoria, "Stakeholders approach to smart cities: A survey on smart city definitions," in *Smart-CT 2016, LNCS 9704*, E. Alba et al. (Eds.), pp. 157–167, 2016.
[29] A. Hefnawy, T. Elhariri, A. Bouras, C. Cherifi, J. Robert *et al.*, "Lifecycle management in the smart city context: Smart parking use-case," in *PLM 2016, IFIP AICT 492*, R. Harik et al. (Eds.), pp. 631–641, 2016.
[30] G. Pan, G. Qi and W. Zhang, "Trace analysis and mining for smart cities: Issues, methods, and applications," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 120–126, 2013.
[31] Y. Zheng, "Urban computing and big data," *Communications of CCF*, vol. 9, no. 8, pp. 8–18, 2013.
[32] NAA: National Archives of Australia, "Digital continuity 2020 policy [EB/OL]," 2015. [Online] http://www.naa.gov.art/records-management/digital-transition-and-digital-continuity/digital-continuity-2020/index.aspx.
[33] Q. Xiao and L. Wu, "Study on digital continuity plan of national archives of Australia," *Information Resources Management*, vol. 2015, no. 4, pp. 19–23, 2015.
[34] X. An, "Accelerate the formulation of the government's digital continuous action plan to modernize the state's governance capacity," *China Archives*, vol. 2868, no. 3, 2016.

[35] X. J. Zhao, X. H. Zhang, P. Wang, S. L. Chen and Z. X. Sun, "A weighted frequent itemset mining algorithm for intelligent decision in smart systems," *IEEE Access*, vol. 6, pp. 29271–29282, 2018.

[36] W. Zhou and N. Zhang, "The panorama and inspiration of global digital continuity," *Information Studies: Theory and Application*, vol. 40, no. 3, pp. 138–142, 2017.

[37] G. S. Li, J. H. Yan, L. Chen, J. H. Wu, Q. Y. Lin *et al.*, "Energy consumption optimization with a delay threshold in cloud-fog cooperation computing," *IEEE Access*, vol. 7, pp. 159688–159697, 2019.

[38] Y. Qian, "Study on the long-term preservation standard of trusted electronic records in China," *Archives Science Bulletin*, vol. 3, pp. 75–79, 2014.

[39] G. S. Li, Y. C. Liu, J. H. Wu, D. D. Lin and S. S. Zhao, "Methods of resource scheduling based on optimized fuzzy clustering in fog computing," *Sensors*, vol. 19, no. 9, pp. 1–16, 2019.

[40] Y. Huang, "Research on the connotation and management of trusted electronic records," *Zhejiang Archives*, vol. 31, no. 5, pp. 12–15, 2014.

[41] J. Han, L. Xu and Y. Dong, "An overview of data quality research," *Computer Science*, vol. 35, no. 2, pp. 1–5, 2008.

[42] N. Zhang, C. Wang, Z. Liu and W. Wang, "Study on the evaluation strategy of electronic document authenticity based on digital continuity thought," *Archives Research*, vol. 6, pp. 69–72, 2015.