Hindawi

Research Article

Feature Extraction of Museum Big Data Text Information Based on the Similarity Mapping Algorithm

Zhe Yang ⓘ, Huiqin Wang, Qixuan Tang, Ting Wang, Shaowen Wang, and Yulei Kong

School of Management, Xi'an University of Architecture and Technology, Xi'an, Shaanxi, China

Correspondence should be addressed to Zhe Yang; zheyang1986@xauat.edu.cn

Under big data, a large number of features, as well as their complex data types, make traditional feature extraction and knowledge reasoning unable to adapt to new conditions. To solve these problems, this study proposes a museum big data feature extraction method based on a similarity mapping algorithm. Under the museum big data analysis, the museum big data text information is collected through web crawler technology. The web crawler is used to index the content of websites all across the Internet so that the museum websites can appear in search engine results and the collected text information is denoised and smoothed by a Gaussian filter to construct the processed text information set mapping matrix. The semantic similarity is computed according to the text word concept. Based on the calculation results, through word frequency and document probability inverse document frequency weight, the museum big data text information features are extracted. Simulation results show that the proposed method has high accuracy and short extraction time. Through the comparative analysis, it can be realized that this method not only solves the problems existing in traditional methods but also lays a foundation for the analysis of museum massive data.

1. Introduction

The problem of information overload has gotten increasingly critical as the Internet, particularly the mobile Internet, and e-commerce have grown in popularity. The way data is processed and maintained has changed dramatically as a result of big data issues. Big data is a challenge nowadays not only in terms of volume but also in terms of high-speed generation. Big data quality and validity differ from source to source, making processing difficult [1]. From a technical point of view, the five main characteristics of big data are summarized as 5V characteristics including large volume, diversity, velocity timeliness, veracity accuracy, and final value discovery. In terms of application, financial marketing, accurate advertising, selection and placement of goods in the retail industry, traffic and transportation status prediction, intelligent doctor service in the medical and health industry, and other big data application fields have all touched people's lives. Demand for big data applications encourages technological advancement, which in turn feeds back into societal development [2].

The Internet's development has tremendously influenced and moulded our modern social life attitude, as well as people's basic way of living standards [3]. As a central "cultural landmark" of the city where it is located, the traditional museum is a nonprofit permanent institution that serves society and its development and is open to the public. It is to collect, protect, and study cultural relics and disseminate and display the material and intangible cultural heritage of human beings and the human environment for public education, scholars' research, and appreciation. In recent years, domestic museums have successively carried out the construction of smart museums and accumulated a lot of valuable data. However, converting museum data into data assets is a key problem. The museum generates massive multisource heterogeneous data in the process of daily operation and maintenance [4]. However, generally, there are scattered, disordered, and difficult problems in the current situation of museum data. Scattered data resources, lack of high-quality cataloging and labeling information, and difficult data management lead to low data utilization. To transform it into data assets, it is essential to apply data

management and analysis techniques to extract the characteristics of museum big data text information, transform the data into knowledge, and sublimate from knowledge to wisdom, to provide support for mining data value and business innovation, and improve the management service level of the museum. Therefore, studying the feature extraction of museum big data text information is a pressing issue [5].

Scholars have extensively studied the application of feature extraction for museum big data. Zhang and Jin [6] proposed a feature extraction method of museum big data text information based on "feature dimensionality reduction" using principal component analysis (PCA) and TOPSIS method. The keywords reflecting the theme of museum big data text are extracted as text feature words to realize museum big data text feature extraction. Xiaonan and Jiwei [7] introduced an adaptive feature extraction method of Museum big data text information regression analysis. Support vector regression was employed to separate the museum big data text information feature area, eliminate the redundant text information boundary, and extract the museum big data text information feature. However, the accuracy of the proposed method for feature extraction was low, resulting in the poor effect of text information feature extraction. The authors in [8] developed a short text feature extraction method integrating word cooccurrence distance and category information. The words in all categories did sort in descending order according to the weight value, and the first k words are selected as a new feature word item set. The experimental results show that this method can effectively improve the effect of feature extraction of short text. Xianlun et al. [9] proposed a short text feature extraction and classification method based on a serial-parallel convolution gate cyclic neural network. The serial-parallel convolution gate cyclic neural network was used to remove the pooling operation in the convolution layer, retain the temporal structure and position information of text data, and extract the multivariate feature combination of words with the serial-parallel convolution structure. The effects of network super parameter selection and convolution layer structure on classification accuracy are simulated and analyzed to verify the effectiveness of this method in extracting short text features. Heidorn and Wei [10] described the information properties of museum specimen labels and machine learning tools to automatically extract Darwin core and other features from these labels processed through optical character recognition. However, the abovementioned methods consume a long time for text feature extraction, resulting in low efficiency of text feature extraction. Because of the problems existing in the above methods, this study proposes a feature extraction method of museum big data text information based on a similarity mapping algorithm. Through the analysis of simulation experiments, it is verified that this method can extract the features of museum big data text information quickly and accurately, solve the problems existing in the traditional methods, and lay the foundation for the digital construction of the museum.

The rest of the paper is organized as: Section 2 provides analyses of the existing status of the museum and provides an overview of the application of big data techniques for the analysis of museum big data. Section 3 illustrates the proposed feature extraction method in detail. The results are given in Section 4, and the conclusion is presented in Section 5.

2. Museum Big Data Analysis

Museum, from traditional museum to today's smart digital museum, digitizes the information. Users can obtain exhibit information through a variety of channels, such as microblog, WeChat platform, or the museum's official website [11], to deepen the audience's understanding of the museum and the audience's further understanding of the exhibits. However, there are many museum exhibits, and it is impossible for users to take the initiative to query and obtain relevant knowledge one by one. Therefore, the concept of "smart museum" was born, so that museum visitors are no longer in the position of passively receiving information. Participants can recommend appropriate exhibits in the exhibition hall they are interested in based on historical information records of museum visitors' participation in the display. The formation process of the wisdom museum is shown in Figure 1. "Big data" refers to a dataset with a large volume and many data categories, and such a dataset can no longer be captured, managed, and processed with traditional database tools. When it comes to big data in museums, the first thing that comes to mind is collecting data, or even more so, audience data. [12]. Based on the principle of "coarse rather than fine" and guided by the function of the museum, the museum data can be divided into collection data, management data, communication data, and audience data. The collected data include cataloging data, detection data, research data, and storage and use data of the collection body.

The management data is generated according to the business behavior of the museum including the data generated by the daily management process of the museum, the data gathered by holding various activities, and the data formed by contacting all social parties. Similarly, the communication data are composed of various communication activities of the museum, digital communication tools, and their feedback mechanism, and the audience data is accumulated based on audience behavior [13]. Whether the data may become the driving force of museum work and the booster for the museum to leap to a new level depends mainly on the application of museum big data today, with the increasing attention devoted to the power of data.

3. Feature Extraction of Museum Big Data Text Information

3.1. Museum Big Data Text Information Collection. A web crawler, often known as a spider or a search engine bot, downloads and indexes content from all across the web. The purpose of a bot like this is to learn what (nearly) every web page on the Internet is about so that information may be obtained when needed. Crawling is the technical term for automatically accessing a website and getting data via a
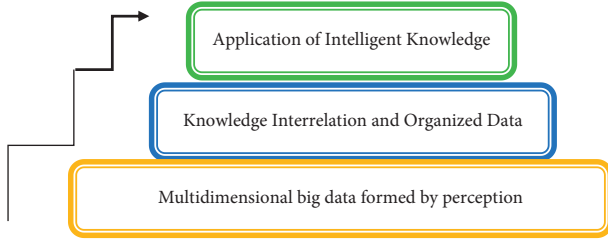
FIGURE 1: The process from data to wisdom.

software program. Web crawler is essentially a computer application that can automatically obtain web pages according to a certain routine; that is, it can automatically download web pages under relevant websites and store these web page information locally or in the cloud in a certain way. In this study, the museum's big data text information is collected through web crawler technology to obtain the required text information [14].

Web crawler mainly includes three modules: crawling request module, link analysis, and the processing module, and web information analysis and a processing module. Among them, the crawling request module is responsible for the connection request and the request for relevant network protocols required when the web crawler crawls the big data text information of the museum. The main task of the link analysis and processing module is to analyze and process the collected links according to rules. To further improve the efficiency of the web crawler, the content detection module is used to process museum big data text information [15]. The data collection process of museum big data text information based on web crawler is shown in Figure 2.

$$f(x) = M_{i-1}\frac{(x_i - x)^3}{6h_i} + M_i\frac{(x - x_{i-1})^3}{6h_i} + (x_i - x)\left(\frac{y_i - 1}{h_i} - \frac{h_i M_{i-1}}{6}\right) + (x - x_{i-1})\left(\frac{y_i}{h_i} - \frac{h_i M_i}{6}\right), x \in [x_{i-1}, x_i]. \qquad (1)$$

The noise in museum big data text information will affect the information feature extraction results, so noise preprocessing is carried out. Gaussian filter is used to denoise and smooth the initial big data text information. The weight smoothing filter curve is obtained based on the Gaussian kernel function [20]. The Gaussian kernel function is expressed as

$$\mathrm{gauss}(x, \eta) = e^{-x^2/2\eta^2}, \qquad (2)$$

where η is the width parameter.

Reducing the complexity of morphological transformation of data mining based on information transformation can help museum big data text information obtain the same information matrix position and prevent excessive contrast weight of large and small initial value domain attributes [21].

The maximum and minimum values of the attribute vector S are expressed as $S_{\max}, S_{\min}$, respectively. Normalize the vector value to meet the interval $[S_1, S_2]$ and obtain

3.2. Information Preprocessing. Due to the noise included in the collected museum's big data text information, the collected museum's big data text information is preprocessed. The denoising process is used to remove the additive noise while retaining as much as possible the important features. The preprocessing mainly covers information selection, cleaning, and transformation. Selection refers to the selection and extraction of information attributes. The cleaning process refers to missing information and performing denoising and smoothing [15].

Aiming at the missing value processing of museum big data text information, polynomials are established according to the basic information to smoothly estimate the function of redundant points of basic information [16].

Let there be equidistant sampling points of $n + 1$ functions $f(x) = y$ in the interval $[x, y]$, and the value of sampling point x_i is $P_i(x_i, y_i)(i = 0, 1, \ldots, n)$. If function $S(x)$ belongs to $S(x_i) = y_i$, then $S''(x)$ is continuously differentiable, and function $S(x)$ is the cubic spline interpolation function of the function $f(x)$ in the interval, which is a cubic polynomial in subinterval $[x_i, x_{i+1}]$ [17].

Let the first and second derivatives of the function $f(x)$ at $x = x_i$ be $f'(x_i) = m_i$ and $f''(x_i) = M_i$, respectively, and the function $f(x)$ in the subinterval is all second-order continuous differentiable to obtain the following function. $f''(x) = M_{i-1}x_i - x/h_i + M_i x - x_{i-1}/h_i$, where $h_i = x_i - x_{i-1}$ [18]. The second derivative $f''(x_i)$ is quadratically integrated to obtain a cubic spline, obtain the full-scale constructor and the interpolation value, and realize the missing value supplement [19].

$$S_i^* = \frac{S_i - S_{\min}}{S_{\max} - S_{\min}}(S_1, S_2) + S_1, \qquad (3)$$

where S_i, S_i^* represents the values before and after the normalization of the ith attribute vector [22].

3.3. Semantic Similarity Calculation Based on the Similarity Mapping Algorithm. Semantic similarity, or semantic textual similarity, is a task that scores the relationship between texts or documents using a defined metric. For the museum big data text, to facilitate the computer to understand and process, the text needs to be mapped into a numerical vector space. Text vectorization has the advantages of fast processing speed for a large amount of data, so it is widely used in museum big data text processing. Traditional text vectorization methods, such as calculation methods, determine the weight value according to the number of words and phrases appearing in the text and the number of words appearing in the whole text set. Because hot words are typical in hot event discovery, higher weight values should be given when calculating the weight values of these words [23]. In
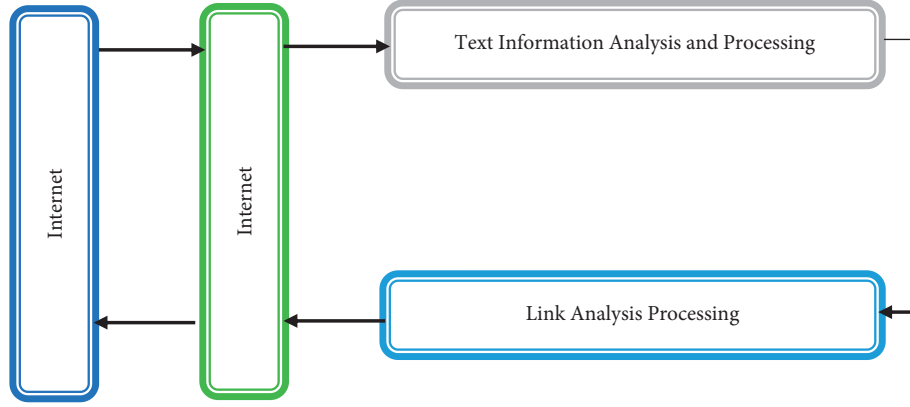
FIGURE 2: The collection process of museum big data text information based on web crawler.

addition, the position of words in the museum big data text is different, and their importance is also different. The influence of position on the calculation of word weight should be considered. Therefore, the calculation method of the weight value of words in the museum big data text can comprehensively consider multiple factors. This study uses the similarity mapping algorithm to calculate the semantic similarity of the preprocessed museum big data text information [24]. The process is described as follows.

Suppose we set the numerical mapping of museum big data text information as follows:

$$f = [d(c_1), d(c_2), \ldots, d(c_n)]. \tag{4}$$

The text is mapped into a vector s-dimensional space, where the element $d(c_n)$ represents the weight value of the nth word c_n in the museum big data text information f [25]. We collect the text information in a period in the museum big data to form a text set, which can be represented by a matrix:

$$F(t_i) = \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix}$$

$$= \begin{bmatrix} d_1(c_1) & \ldots & d_1(c_m) \\ \vdots & \ddots & \vdots \\ d_n(c_1) & \ldots & d_n(c_m) \end{bmatrix}, \tag{5}$$

where $F(t_i)$ represents the mapping matrix of the museum big data text information collection collected in the period t_i, where $d_n(c_m)$ represents the weight value of the m-th word of the n museum big data text information [26].

Due to the huge amount of museum big data text information and many words involved, this study introduces the concept of word heat value, and the measure of word burst can be determined by its heat value [27]. Suppose that in the period t_i, the number of texts containing the word c_m is a, the number of texts without c_m is c, the number of texts containing c_m before the period t_i is b, and the number of texts without c_m is f. According to the distribution of word c_m s in different time periods, the abruptness of word c_m s can be obtained. Since the distribution of c_m in period t_i

satisfies the binomial distribution, the ρ^2 value of the binomial distribution of discrete events can be obtained according to the statistical formula of the binomial distribution. The probability density function of ρ^2 distribution is computed as

$$p_k(x) = \frac{(1/2)^{m/2}}{m/2} x^{m/2-1} F(t_i). \tag{6}$$

The ρ^2 value of the binomial distribution is used as the scale value of word heat, and $p_b(c_m)$ is expressed as

$$p_b(c_m) = \frac{(a\,d - bc)^2 p_k(x)}{(a+c)(a+b)(b+c)(b+d)}. \tag{7}$$

Semantic similarity is a value used to measure the matching degree of semantic information of museum big data text. In this representation method, words are represented by concepts, and the similarity between concepts can fully express the similarity between words. The similarity between concepts is expressed by calculating the distance between concepts [28]. The weighted sum of the shortest path lengths between multiple attributes of two concepts can represent the distance calculation formula of the concept, and the result is the similarity between the two concepts.

If $\text{sim}(Q_1, Q_2)$ represents the similarity of two concepts Q_1, Q_2 and $\text{dis}(Q_1, Q_2)$ represents the distance between two concepts, the semantic similarity between concepts is calculated as

$$\text{sim}(Q_1, Q_2) = \frac{\varepsilon p_b(c_m)}{\text{dis}(Q_1, Q_2) + \varepsilon}, \tag{8}$$

where ε is a constant that can be adjusted, which is the conceptual distance value when the similarity is half [29].

3.4. Feature Extraction of Museum Big Data Text Information. According to the semantic similarity calculated above, the text features of museum big data are extracted. Text information feature extraction is measured according to the weight of words in the text. Each featured item has a feature weight in its text, that is, feature weight, which expresses the importance of the featured item to the text. It shows the ability of this feature item to distinguish the text from other

texts. The factors affecting the weight of the featured item generally include word frequency, document probability, and inverse document frequency (IDF) weight [30].

3.4.1. Word Frequency. Term frequency (TF) refers to the total number of times a word is in a museum's big data text. The number of words in the museum's big data text information can represent its relevance to the text. The higher the frequency is, the more qualified it is to represent the museum's big data text information [31]. Feature selection through word frequency and deleting words that appear very few times in the museum big data text information can reduce the number of feature words in the museum big data text information and reduce the spatial dimension. However, in some cases, words with high word frequency do not have great information value, and words with low frequency have an important amount of museum big data text information. Word frequency is an important factor in feature weight calculation. It will be considered in many cases in the process of feature extraction. Generally, the more the word segmentation appears in the museum big data text information, the higher the position of the word segmentation in the museum big data text information is [32]. The word frequency is computed as follows:

$$W_{TF} = r_{c_m} \text{sim} (Q_1, Q_2), \tag{9}$$

where r_{c_m} refers to the number of times that word c_m appears in the text and W_{TF} refers to the word frequency of the word.

3.4.2. Document Probability. Document frequency refers to the number of documents with this word in the museum big data text information dataset. The smaller the document frequency of the word, the less information of the museum big data text with the word, indicating that the word can distinguish the text from other texts [32]. If the document frequency is particularly high, it indicates that this word appears in a large number of collections and is not representative of museum big data text information. Similarly, if the document frequency is particularly low, this indicates that the frequency of the word in the whole dataset is very small, and it does not exist in the text set. Therefore, the document frequency of each word is usually calculated, and the words with particularly high and low values are removed.

$$W_{im} = UG_{im}, \tag{10}$$

where W_{im} is the number of times UG_{im} m and feature items that appear in the i th text.

3.4.3. Inverse Document Frequency Weight. The reciprocal of document frequency is the inverse document frequency. It represents the characteristics of a word and describes the layout of the word in the dataset. The calculation method is as follows:

$$q_m = \log\left(\frac{M}{s_e + 0.01}\right), \tag{11}$$

where M is the total number of documents in the text set and s_e is the amount of museum big data text information with

the mth word in the text set. If the amount of museum big data text information with a certain word is less, the word is likely to fully express the museum big data text information. If a word appears in a large number in the text sets, it is of no value to the big data text information of the museum.

*3.4.4. TF * IDF.* TF * IDF fully focuses on word frequency and inverse document frequency. If a word appears frequently in a museum's big data text information, but not in the whole museum's big data text information dataset, it indicates that the word can express the content of the museum's big data text information. TF reflects the occurrence degree of this word in the big data text information of a museum, and IDF reflects the incidence degree of this word in the whole text set. To sum up, TF-IDF is expressed as

$$E(f_i) = q_m * \log\left(\frac{M}{m_j + 0.01}\right). \tag{12}$$

3.4.5. Part of Speech. Generally, the notional words in the museum's big data text information, such as verbs, nouns, and adjectives, cover more sentence information. The quantifiers, conjunctions, adverbs, and other function words in the museum big data text information are not of great information value to express the content of the museum big data text. In notional words, nouns, verbs, and adjectives play a decisive role in expressing the meaning of the text.

3.4.6. Location. Words appear in different positions in the museum of big data text information, and their contribution to the expression of text content is also very different. Especially in English texts, the first and last sentences of a paragraph are usually important places to express the central idea of the whole paragraph. The title, abstract, and conclusion are the top priorities to explain the content of the article. Short words and sentences represent the main content of an article. The first paragraph, the last paragraph, and the beginning and end of the paragraph are generally used to summarize the content information of the article. Therefore, this factor can also be considered in weight calculation.

3.4.7. Syntactic Structure. There are typically some rules in sentences. The sentences in the abstract, for example, essentially summarize the article's fundamental topic. Most of the declarative sentences in the text, while exclamatory sentences and interrogative sentences do not represent the positive meaning of the content of the article. In addition, the sentences behind words such as "sum up" and "in short" are used to summarize the main contents of paragraphs and articles, which are of a great reference value.

4. Simulation Experiment Analysis

4.1. Experimental Environment Setting. To verify the effectiveness of the feature extraction method of museum big

FIGURE 3: Wisdom museum.

TABLE 1: Hardware components.

Component	Description
CPU related parameters	Binuclear Intel@Xeon @E5-2650v3 series processor
Memory capacity and memory type	16g and above ddr4recc2133mhz
Maximum hard disk capacity and interface type	6tb sataisas72k3.5-inch enterprise class
Hadoop node	10

TABLE 2: Software components.

Component	Description
Operating system	Red hat enterprise linux 7
Programming development platform IDE	Eclipse-jee-mars-i-linux
JDK version	Jdk-8u65-linux-i586
Hadoop version	Hadoop-2.6.0-cdh55.
HBase version	HBase-1 0.3

data text information based on similarity mapping algorithm in practical application, a simulation experiment is carried out. The wisdom museum is shown in Figure 3.

The different hardware and software components used in the simulation experiment are shown in Tables 1 and 2, respectively.

4.2. Dataset. The dataset used in this study is obtained from the data center of a provincial museum, in which the basic data of 5000000 visitors are collected and the basic data of 1000 exhibits are used for experimental tests. During the experiment, 90% of the dataset is used as a training set, and the remaining 10% is used as a test set. The division of training and testing samples was divided randomly.

4.3. Experimental Comparative Analysis. To verify the effectiveness of the proposed museum big data text information feature extraction method based on similarity mapping algorithm, we compared it with the museum big data text information feature adaptive extraction method based on support vector machine regression proposed in [7] and the feature extraction method of museum big data text information based on word cooccurrence distance and category information presented in [8], respectively. The

comparative results for feature extraction accuracy of the three methods are shown in Figure 4.

It can be seen from Figure 4 that the accuracy of the proposed museum big data text information feature extraction method based on similarity mapping can reach 100%, which is higher than the adaptive feature extraction method of museum big data text information based on support vector machine regression proposed in [7] and the method based on word cooccurrence distance and category information given in [8], respectively.

To further verify the performance of this method in terms of computation time, we compared all three methods. Figure 5 shows the comparative computation time for all three methods. The proposed feature extraction method of museum big data text information based on similarity mapping algorithm takes less than 5s to extract the feature of museum big data text information. The time taken by the method given in [7] is about 20s, and the method presented in [8] took about 30s. The computation time of the proposed feature extraction method of museum big data text information based on similarity mapping algorithm is the shortest as compared to the other two feature extraction methods. This confirms that the proposed feature extraction method museum big data analysis is superior to the other two methods in terms of recognition accuracy and computation time.
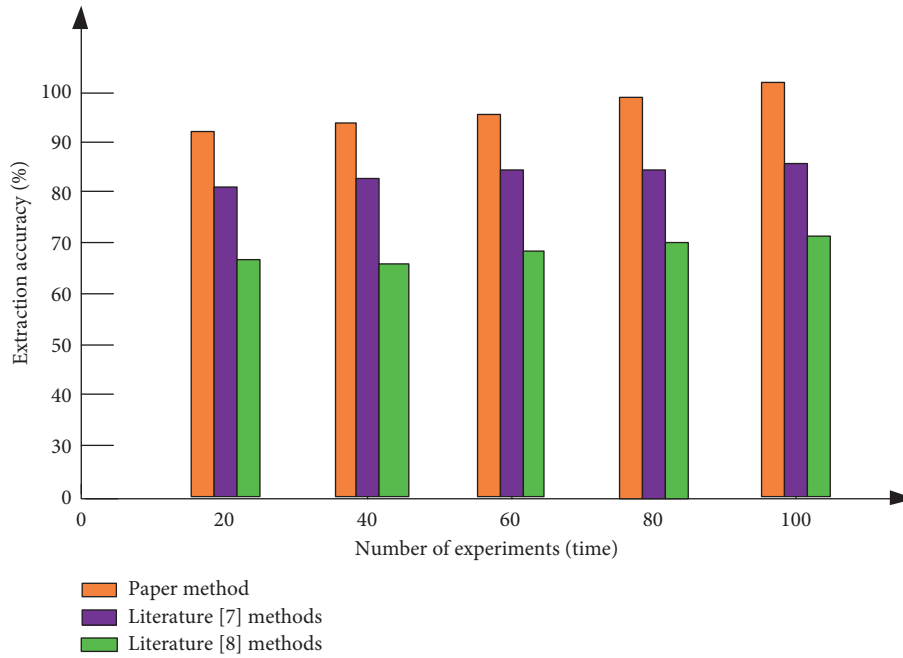
FIGURE 4: Comparison results of feature extraction accuracy of museum big data text information by three methods.
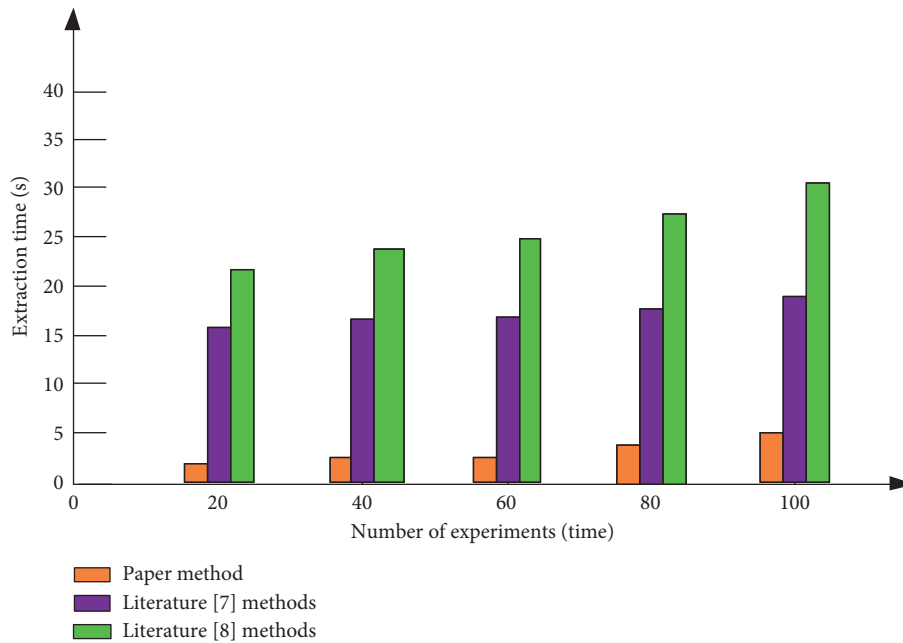

FIGURE 5: Comparison results of feature extraction time of museum big data text information by three methods.

5. Conclusion

The use of big data is an inevitable topic in the process of human beings moving from the Information technology era to the data technology age. The core of data technology is about data-driven innovation, that is, the innovation system and model focusing on value mining based on massive data. For the museum, it provides a new platform, content, and form support for the tasks of public education, cultural communication, scientific research, and collection of the

museum and provides data support for the accurate management of the museum. At the same time, it should also be the due meaning of the digital construction of the museum to increase the collection research direction and display angle of the museum itself to a certain extent and shape a new mode of museum work based on big data. For the above considerations, this study proposed a feature extraction method of museum big data text information based on a similarity mapping algorithm. Through the comparative analysis of simulation experiments, it can be seen that this

method not only solves the problems existing in traditional methods but also lays a foundation for the value mining of museum massive data. Future work will focus on the application of machine learning techniques for extracting museum big data features.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] H. Liang, X. Sun, Y. Sun, and Y. Gao, "Correction to: text feature extraction based on deep learning: a review," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, pp. 42–48, 2018.

[2] I. Safder, S.-U. Hassan, A. Visvizi, T. Noraset, R. Nawaz, and S. Tuarob, "Deep learning-based extraction of algorithmic metadata in full-text scholarly documents," *Information Processing & Management*, vol. 57, no. 6, Article ID 102269, 2020.

[3] C. Tyab, C. Jma, and B. Yma, "Privacy-Preserving Krawtchouk Moment feature extraction over encryptedimage data," *ScienceDirect. Information Sciences*, vol. 536, pp. 244–262, 2020.

[4] P. M. Prihatini, I. K. Suryawan, and I. Mandia, "Feature extraction for document text using Latent Dirichlet Allocation," *Journal of Physics Conference*, vol. 953, p. 960, 2018.

[5] D. Georg, K. Jonathan, and F. Georg, "Ad hoc information extraction for clinical data warehouses," *Methods of Information in Medicine*, vol. 57, no. 1, pp. e22–e29, 2918.

[6] L. Zhang and M. Jin, "A feature extraction method based on "feature dimensionality reduction" text complex network," *Information Science*, vol. 4, no. 10, pp. 20–25, 2019.

[7] Y. Xiaonan and H. Jiwei, "Simulation Research on adaptive feature extraction of illegal text on Internet," *Computer Simulation*, vol. 36, no. 6, pp. 434–437, 2019.

[8] M. Huifang, X. Yuying, and W. Shuang, "Short text feature extraction method integrating word co-occurrence distance and category information," *Computer engineering and Science*, vol. 40, no. 9, pp. 1689–1695, 2018.

[9] T. Xianlun, L. Wenxing, D. Yiming, and W. Ting, "Feature extraction and classification of short text based on serial-parallel convolution gate valve cyclic neural network," *Engineering Science and technology*, vol. 51, no. 4, pp. 125–132, 2019.

[10] P. B. Heidorn and Q. Wei, "Automatic metadata extraction from museum specimen labels," in *Proceedings of the 2008 Int'l Conf. on Dublin Core and Metadata Applications*, pp. 1–12, Berlin, Germany, September 2008.

[11] S. Giuseppe, G. Massimiliano, and M. Antonio M, "A CNN-based fusion method for feature extraction from sentinel data," *Remote Sensing*, vol. 10, no. 2, pp. 236–240, 2018.

[12] E. Izquierdo-Verdiguier, V. Laparra, J. Muñoz-Marí, L. Gómez-Chova, and G. Camps-Valls, "Advanced feature extraction for earth observation data processing," *Comprehensive Remote Sensing*, vol. 2, pp. 108–133, 2018.

[13] R. Caponetto, G. Salvatore, and M. G. Xibilia, "Data selection methods for Soft Sensor design based on feature extraction," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 132–137, 2020.

[14] L. Zhang and X. Lu, "Feature extraction based on support vector data description," *Neural Processing Letters*, vol. 49, no. 2, pp. 643–659, 2019.

[15] Z. Zhang and J. Li, "Feature extraction of big climate data – ScienceDirect," *Big Data Mining for Climate Change*, p. 52, Elsevier, Amsterdam, The Netherlands, 2019.

[16] M. Konnik, B. Ahmadi, and N. May, "Training AI-based feature extraction algorithms, for micro CT images, using synthesized data," *Journal of Nondestructive Evaluation*, vol. 40, no. 1, pp. 1–13, 2021.

[17] H. Yang, C. Meng, and C. Wang, "Data-driven feature extraction for analog circuit fault diagnosis using 1-D convolutional neural network," *IEEE Access*, vol. 8, no. 99, p. 1, 2020.

[18] A. Ad, B. Ts, and C. Ac, "Drilling process monitoring: a framework for data gathering and feature extraction techniques – ScienceDirect," *Procedia CIRP*, vol. 99, pp. 189–195, 2021.

[19] T. Shanableh, "Feature extraction and machine learning solutions for detecting motion vector data embedding in HEVC videos," *Multimedia Tools and Applications*, vol. 80, no. 1, pp. 1–20, 2020.

[20] M. Piernik and T. Morzy, "A study on using data clustering for feature extraction to improve the quality of Classification," *Knowledge and Information Systems*, vol. 63, no. 7, pp. 1771–1805, 2021.

[21] H. Fu, G. Sun, and J. Ren, "Fusion of PCA and segmented-PCA domain multiscale 2-D-SSA for effective spectral-spatial feature extraction and data classification in hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, no. 99, pp. 1–14, 2020.

[22] X. Ma, W.-H. Huang, E. Schnabel et al., "Data-driven I-V feature extraction for photovoltaic modules," *IEEE Journal of Photovoltaics*, vol. 9, no. 5, pp. 1405–1412, 2019.

[23] B. Li, W. Guo, and X.-L. Zhang, "A global manifold margin learning method for data feature extraction and Classification," *Engineering Applications of Artificial Intelligence*, vol. 75, pp. 94–101, 2018.

[24] D. Pantazis, P. Goodall, P. P. Conway, and A. A. West, "An automated feature extraction method with application to empirical model development from machining power data," *Mechanical Systems and Signal Processing*, vol. 124, pp. 21–35, 2019.

[25] X. U. Dongwei, Y. D. Wang, and P. Peng, "Kernel PCA for road traffic data nonlinear feature extraction," *IET Intelligent Transport Systems*, vol. 13, no. 8, pp. 1291–1298, 2019.

[26] N. Kanungsukkasem and T. Leelanupab, "Financial latent dirichlet allocation (FinLDA): feature extraction in text and data mining for financial time series prediction," *IEEE Access*, vol. 7, Article ID 71645, 2019.

[27] M. Maggipinto, A. Beghi, S. Mcloone, and G. A. Susto, "DeepVM: a Deep Learning-based approach with automatic feature extraction for 2D input data Virtual Metrology," *Journal of Process Control*, vol. 84, pp. 24–34, 2019.

[28] C. Kong, S. Zhao, and X. Weng, "Weighted summation: feature extraction of farm pigsty data for electronic nose," *IEEE Access*, vol. 7, no. 99, p. 1, 2019.

[29] J. Yogapriya, C. Saravanabhavan, R. Asokan, I. Vennila, P. Preethi, and B. Nithya, "A study of image retrieval system based on feature extraction, selection, classification and

similarity measurements," *Journal of Medical Imaging and Health Informatics*, vol. 8, no. 3, pp. 479–484, 2018.

[30] Y. Choo, J. Jang, and J. Paik, "Scene mapping-based video registration using frame similarity measurement and feature tracking," *IEIE Transactions on Smart Processing & Computing*, vol. 8, no. 6, pp. 456–464, 2019.

[31] M. Keley, A. Khademzadeh, and M. Hosseinzadeh, "Efficient mapping algorithm on mesh-based NoCs in terms of cellular learning automata," *The International Arab Journal of Information Technology*, vol. 16, no. 2, pp. 312–322, 2019.

[32] E. Yao, H. Zhang, and G. Zhang, "Robot simultaneous localization and mapping algorithm based on the vision and IMU," *Chinese Journal of Scientific Instrument*, vol. 39, no. 4, pp. 230–238, 2018.