Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 3459605, 10 pages
https://doi.org/10.1155/2022/3459605

Research Article

Key Information Extraction Algorithm of Different Types of Digital Archives for Cultural Operation and Management

Xiulun Ma

University of Jinan, Jinan 250022, China

Correspondence should be addressed to Xiulun Ma; shc_maxl@ujn.edu.cn

Received 4 July 2022; Accepted 11 August 2022; Published 29 August 2022

Academic Editor: Baiyuan Ding

In order to improve the effect of key information extraction from digital archives, a key information extraction algorithm for different types of digital archives is designed. Preprocess digital archive information, taking part of speech and marks as key information. Self-organizing feature mapping network is used to extract the key information features of digital archives, and the semantic similarity calculation results are obtained by combining the feature extraction results. Combine with mutual information collection, take that word with the highest mutual information value as the collection cent, traverse all keywords, and take the central word as the key information of digital archives to complete the extraction of key information. Experiments show that the recall rate of the algorithm ranges from 96% to 99%, the extraction accuracy of key information of digital archives is between 96 and 98%, and the average extraction time of key information of digital archives is 0.63 s. The practical application effect is good.

1. Introduction

Generally speaking, there are no management problems in the spontaneous stage of cultural production. Cultural production and business activities are the product of commodity economy [1, 2]. When material and cultural production develop to a certain extent, social division of labor is further clarified, and professionals and professional groups engaged in cultural production appear, both the ruling class and the ruled class try to use cultural production to serve the interests of their own class, which is the conscious stage of cultural production [3]. Only at the conscious stage of cultural production can cultural operation and management be put on the agenda. In a modern capitalist society, everything into a commodity culture products without exception has become a part of capitalists, for-profit special goods. The vast majority of cultural activities are restricted by the value of commodity production rules [4]. The tendency of commercialization of cultural production and cultural activities has become a common social phenomenon in the field of capitalist culture. The law of market economy dominates the management of cultural operations and management activities, and the quality of management is the key to the success or failure of specific cultural products in the free competition of the cultural market. Especially with the rapid development of social economy, different types of digital archives are gradually increasing, the main products of these are social culture, in order to better manage for these different types of digital files, need to study a new key to different types of digital archives information extraction algorithm, to enhance the management level of the digital scheme. Therefore, it is of great significance to study a key information extraction algorithm of different types of digital archives.

In the digital archives, information extraction is an important research topic, reference [5] proposed a digital book records mass data fast extraction algorithm. Based on the range characteristics of large-scale data attributes, the distribution samples of digital book archive data are divided into multiple subintervals to achieve data classification. By constructing a neuron model, the error terms of output are determined according to the data output of the hidden layer and output layer, and the weight of each layer of the BP neural network is adjusted. This method builds a fast extraction model based on BP neural network and realizes the fast extraction of massive archive data. Reference [6] proposed a key information extraction algorithm based on TextRank and cluster filtering. First, the key information is

extracted and vectorized for Word2Vec. Then, TextRank is improved by constructing a graph model integrating word eigenvalues and edge weights, and the stable graphs obtained by iterative convergence are merged and clustered to form clusters. Then, a cluster quality evaluation formula was designed for cluster filtering, and TextRank was applied to form the final clustering. Finally, annotate the information type of the cluster. For testing the text, by comparing the key information vector distance cluster heart vector and the words information types, combines information type and key information to get the key of the text information. Reference [7] proposed a hidden Markov model based on an improved extraction algorithm of key information extraction. The web document is converted into D0M tree and preprocessed, and the information item to be extracted is mapped to state and the observation item to be extracted is mapped to vocabulary. The improved hidden Markov model is used to extract key information of the text. Reference [8] proposed a key information extraction algorithm based on word vector and location information. Vector representation model by word learning vector of each word in the target document said, will the reflect of the latent semantic relations between the word and the word vector combined with location feature fusion to the PageRank score model, choose a few top words or phrases as the key target document information, in order to complete the digital archives of key information extraction. Reference [9] proposed a key information extraction algorithm of unstructured text in the knowledge database. Six yuan group was used to optimize the hidden Markov model, probability model, and smooth processing of incomplete training samples. Initialization and termination operations were carried out for the sequences of observation values released at different times to obtain the optimal state sequence. After decoding the observation sequence, the positive sequence and reverse sequence were obtained by comparing them to filter out the states without decoding ambiguity and complete ambiguity elimination. According to the maximum probability state sequence, the text key information to be extracted is defined and the key information is extracted.

However, the above-mentioned key information extraction algorithm is suitable for different types of digital files, and the effect is not ideal because the boundary of key information extraction is uncertain. Therefore, this paper designs a new key information extraction algorithm for different types of digital archives. Firstly, the algorithm divides the main categories of key information, takes parts of speech and marks as features, and introduces the self-organizing feature mapping neural network to traverse the center of word set, thus realizing the extraction of key signals quickly and accurately. The effectiveness of the algorithm is verified by experiments.

2. Materials and Methods

2.1. Digital Archive Processing

2.1.1. The Text Participle. In the process of cultural management, there are many types of digital archives. Before extracting the key information of different types of digital archives, it is necessary to preprocess the key information of digital archives. The preprocessing process includes word segmentation and marking. Word segmentation refers to classifying the words in the text and setting the marks according to the categories, which lays the foundation for the key information extraction of digital archives in the future [10].

The difference between the word segmentation process of the reverse maximum matching algorithm and the forward maximum matching algorithm is that the scanning of the reverse maximum matching algorithm starts from the end of the string. Each unsuccessful match removes the preceding word until the match is successful. Then the basic idea of the bidirectional maximum matching algorithm is: When segmenting different types of digital archives information, firstly, a forward word-for-word maximum matching algorithm is applied to the character string to be processed, then a reverse word-for-word maximum matching algorithm is applied, and the output result is used to complete the word segmentation processing. Assuming bidirectional maximum matching word segmentation for $S = (C_1, C_2, \ldots, C_i)$, the algorithm process can be described as follows:

(1) First take out the first word C_1 in S, and search in the dictionary to see if there are any words with C_1 as the prefix. If there are, save them as word marks [11].

(2) Take a word C_2 from S and match it with the dictionary to determine whether there is a word with C_2 as the prefix.

(3) If it does not exist, split C_1 from string S, ending with a word split.

(4) If there is, to determine whether C_1C_2 into words, calculate the number n headed by C_1C_2 words.

(5) If $n = 0$, the participle ends once [12].

(6) If n is not 0, then take a word C_i from S and match it with the dictionary to determine whether there is a word prefixed with $C_1, C_2, \ldots, C_i$.

(7) If yes, go to Step 6.

(8) If it does not exist, split $C_1, C_2, \ldots, C_{i-1}$ from string S, ending with a word split.

(9) Continue word segmentation from string C_i of S, repeat the above steps until the end of string S forward segmentation.

(10) Take out the last word C_n in S and match it in the dictionary to find whether there is a word with suffix C_1. If so, save it as a word mark [13].

(11) Then take out a word C_{n-1} from S and match it with the dictionary to judge whether there is a word with suffix C_1C_2.

(12) If it does not exist, it splits C_n from string S, ending with a word split.

(13) If there is, then judge whether $C_{n-1}C_n$ is a word and count the number of words starting with $C_{n-1}C_n$, expressed by n.

(14) If $n = 0$, then the participle ends.

(15) If n is not 0, take out a word C_i from S and match it with the dictionary to determine whether there is a word with $C_i, \ldots, C_{n-1}C_n$ as the suffix.

(16) If yes, go to Step (15).

(17) If it does not exist, $C_i, \ldots, C_{n-1}C_n$ will be cut out from string S and a word segmentation will end.

(18) Continue word segmentation from word C_i of string S, and repeat the above steps until the end of reverse segmentation of string S, so as to remove the stop word. The specific implementation process is shown in Figure 1.

2.1.2. The Part of Speech Tagging. Part of speech is a grammatical attribute of vocabulary, which generally indicates the type of a word in the corpus. Part-of-speech tagging refers to the process and method of tagging the part of speech of each word. Some words contain multiple parts of speech, with different parts of speech and completely different ways of expression [14, 15]. However, in general, when a word contains one or more parts of speech, the frequency of its commonly used parts of speech is far greater than that of other parts of speech, so the accuracy of POS tagging can be ensured on the whole, and the POS tagging method can be applied to most application scenarios [16]. Conditional Random Field Algorithm (CRF) was proposed by Lafferty et al. in 2001. It is an undirected graph model combining the characteristics of the maximum entropy model and hidden Markov model. In recent years, good results have been achieved in sequence tagging tasks such as word segmentation, part-of-speech tagging, and named entity recognition [17]. One of the simplest conditional random fields is the chain structure, in this special conditional random field, the chain structure is composed of several character marks. In CRF models with only one order chain, the fully connected subgraph covers the set of the current marker and one marker before it, as well as the maximum connected graph of any subset of the observation sequence. The chained conditional random field is shown in Figure 2, and the set of vertices can be regarded as the maximum connected subgraph.

In the sequence labeling task, random variable $X = \{x_1, x_2, \ldots, x_n\}$ represents the observable sequence, random variable $Y = \{y_1, y_2, \ldots, y_n\}$ represents the corresponding marker sequence of the observed sequence [18], and the chained conditional probability distribution of the random variable Y is:

$$p(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x) + \sum_{i,k} \lambda_k' f_k'(y_i, x) \right). \tag{1}$$

In the above formula, $f_k(y_{i-1}, y_i, x)$ is the state feature function for edge and capture mark transfer features. $\sum_{i,k} \lambda_k' f_k'(y_i, x)$ is the non-negative factor for each node. $f_k'(y_i, x)$ is the state feature function that captures the current marked feature for the edge. λ_k and λ_k' are learning model parameters [19], said the weight of characteristic function. $Z(x)$ is a normalizing factor dependent only on the observation sequence. The specific calculation formula is as follows:

$$Z(x) = \exp\left(\sum_i \sum_k \lambda_k f_k(y_{i-1}, y_i, x) \right). \tag{2}$$

Conditional random field reasoning refers to finding a marker sequence $Y = \{y_1, y_2, \ldots, y_n\}$ corresponding to the most probable one given an observation sequence $X = \{x_1, x_2, \ldots, x_n\}$. In the distribution function of conditional random fields, the normalized factor is completely independent of the marker sequence [20]. Therefore, given the model parameters, the most likely marker sequence can be expressed as:

$$\begin{aligned} Y^* &= \operatorname{argmax}_y p(y|x) \\ &= \operatorname{argmax}_y \exp\left(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x) + \sum_{i,k}^{k'} \lambda_k' f_k'(y_i, x) \right). \end{aligned} \tag{3}$$

When the current sequence position is i and the current label is y, the algorithm can be used to obtain the unnormalized probability value of the optimal label sequence to the current position. Its recursive form is:

$$\theta(i, y) = \max y' \left\{ \theta\left(i - 1, y' * l \sum_k \lambda_k f_k(x, y, y', i) \right) \right\}. \tag{4}$$

2.2. Key Information Feature Extraction of Digital Archives. Self-organizing feature mapping neural network was proposed by a professor of neural network expert self-organizing feature mapping network of University of Helsinki, Finland in 1981 [21]. This network simulates the function of self-organizing feature mapping of the brain nervous system. It is a kind of competitive learning network, which can carry out self-organizing learning without supervision in learning [22]. This paper uses this method to extract the key information features of different types of digital archives. This can improve the accuracy and efficiency of extracting key information from archives.

The structure of self-organizing feature mapping neural network is shown in Figure 3.

We set the number of neurons in the input layer to be n, and the number of neurons in the competition layer to be $M = m^2$. The input layer and the competition layer form a two-dimensional planar array. The two layers are connected, and sometimes neurons in the competing layer are also connected by edge inhibition [23]. There are two kinds of connection weights in the network, one is the connection weights of neurons responding to external inputs, and the other is the connection weights between neurons, whose size controls the size of interactions between neurons [24, 25].

The connections of neurons at the competitive layer of each input neuron in the self-organizing feature mapping

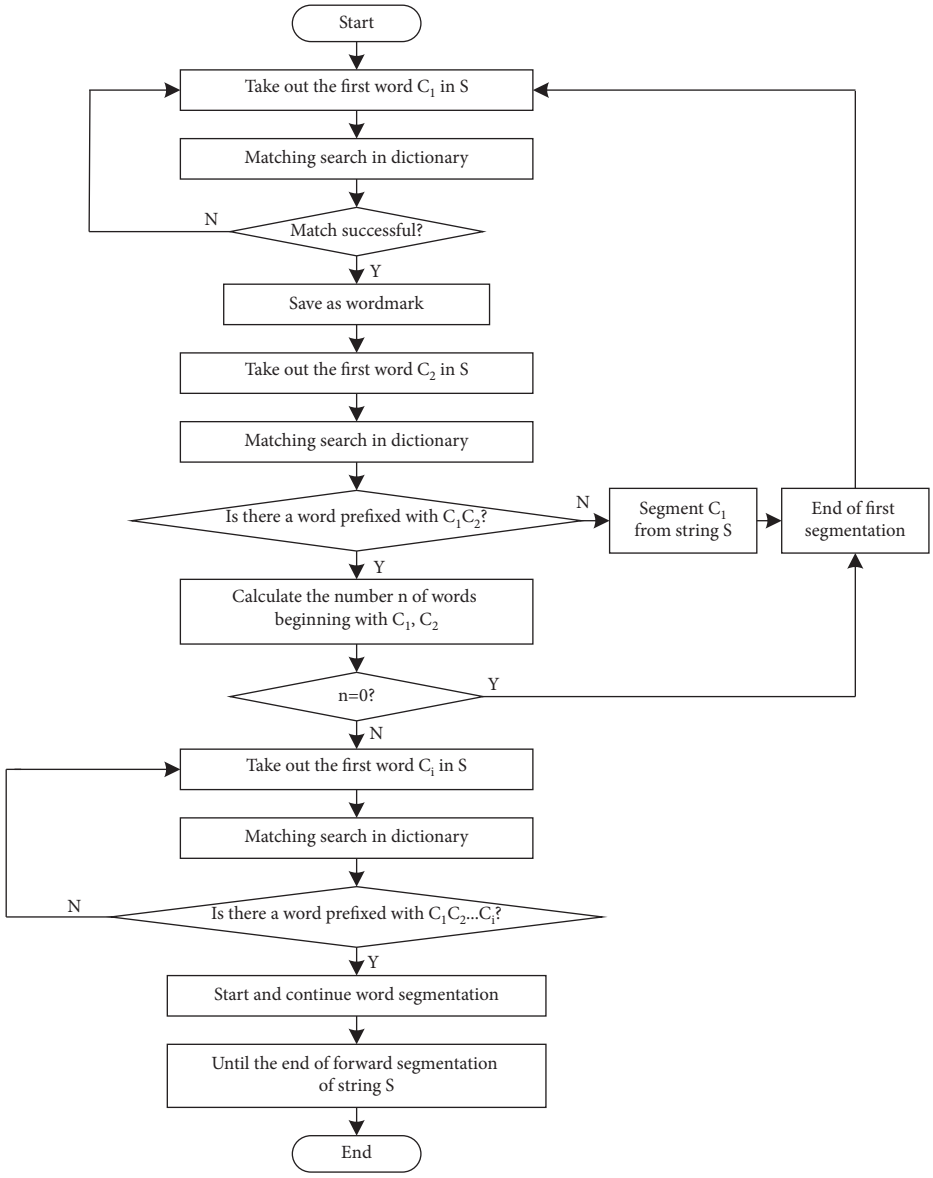

Figure 1: Text word segmentation and the process of removing stop words.

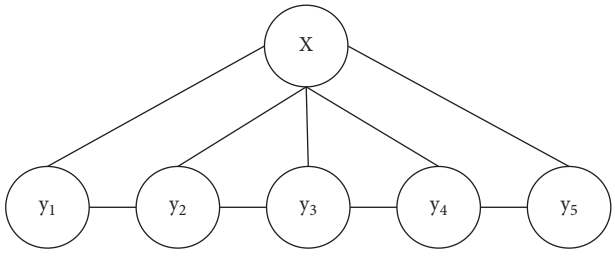

Figure 2: Chained conditional random fields.

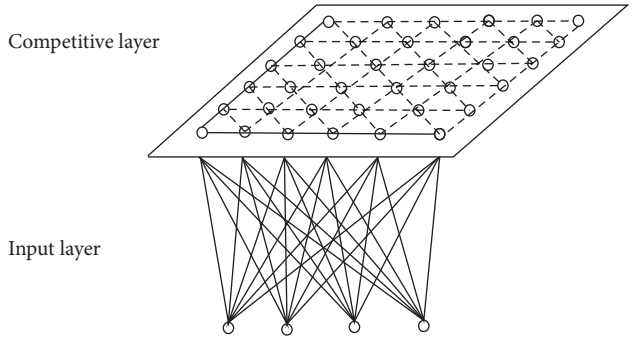

Figure 3: Structure of self-organizing feature mapping neural network.

network structure shown in Figure 3 are extracted, as shown in Figure 4.

Set the input mode of the network as $P_k = (p_1^k, p_2^k, \ldots, p_n^k), k = 1, 2, \ldots, q$ and the neuron vector of the competition layer as $A_j = (a_{j1}, a_{j2}, \ldots, a_{jm})$, $j = 1, 2, \ldots, m$. Where P_k is a continuous value and A_j is a numerical quantity. The connection vector between neuron j of the competition layer and neuron of the input layer is $W_j = (w_{j1}, w_{j2}, \ldots, w_{jm}), j = 1, 2, \ldots, M$.

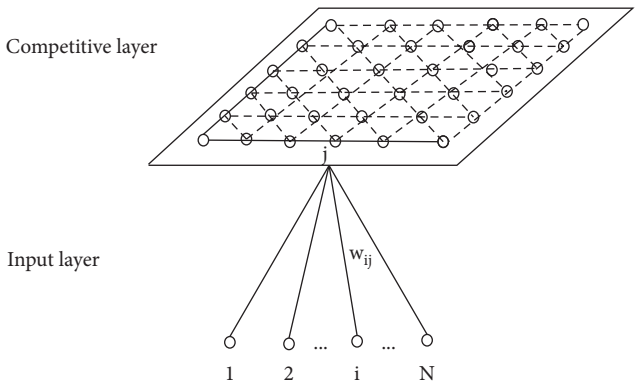

Figure 4: Self-organizing feature mapping network structure after extracting competitive layer neurons.

The self-organizing learning process of the self-organizing feature mapping network can also be described as: for each input of the network, only part of the weight is adjusted to make the weight vector closer to or more deviated from the input vector. This adjustment process is competitive learning. With continuous learning, ownership vectors are separated from each other in vector space, forming a class of patterns representing input space, respectively, which is the clustering function of automatic feature recognition in a self-organizing feature mapping network. The learning and working rules of the network are as follows:

(1) Initialization

Assign the network connection weight $\{w_{ij}\}$ to the random value $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, M$ in the interval [0, 1]. The initial value of learning rate $\eta(t)$, $\eta(t), 0 < \eta(t) < 1$ was determined. Determine the initial value $N_g(0)$ of neighborhood $N_g(t)$. Neighborhood $N_g(t)$ is essentially a region centered on the winning neuron g and contains several neurons. This area is generally uniformly symmetrical, most typically a square or circular area. The value of $N_g(t)$ represents the number of neurons in the neighborhood during the t-th learning. Determine the total number of studies T.

(2) One of the q learning modes P_k, P_k is provided to the input layer of the network and normalized. The specific calculation formula is as follows:

$$\overline{P}_k = \frac{P_k}{P_k} = \frac{\left(p_1^k, p_2^k, \ldots, p_n^k\right)}{\left[\left(p_1^k\right)^2 + \left(p_2^k\right)^2 + \cdots + \left(p_n^k\right)^2\right]^{1/2}}. \quad (5)$$

(3) Normalize the connection weight vector $W_j = (w_{j1}, w_{j2}, \ldots, w_{jN})$ and calculate the Euclidean distance between $\overline{W}_j$ and $\overline{P}_k$. The calculation formula of $\overline{W}_j$ is as follows:

$$\overline{W}_j = \frac{W_k}{W_k} = \frac{\left(w_1^k, w_2^k, \ldots, w_n^k\right)}{\left[\left(w_1^k\right)^2 + \left(w_2^k\right)^2 + \cdots + \left(w_n^k\right)^2\right]^{1/2}}. \quad (6)$$

The Euclidean distance between $\overline{W}_j$ and $\overline{P}_k$ can be calculated by the following formula:

$$d_j = \left[\sum_{i=1}^{N}\left(\overline{P}_i^k - \overline{W}_i^k\right)^2\right], \quad j = 1, 2, \ldots, M. \quad (7)$$

(4) Find the minimum distance d_g and determine the winning neuron g.

$$d_g = \min\left[d_j\right]. \quad (8)$$

(5) Adjust the connection weights, and modify the connection weights between all neurons in neighborhood $N_g(t)$ of the competition layer and neurons of the input layer. The specific formula is as follows:

$$\overline{w}_{ji}(t+1) = \overline{w}_{ji}(t) + \eta(t), \left[\overline{P}_i^k - \overline{w}_{ji}(t)\right]. \quad (9)$$

In the above formula, $\eta(t)$ is the learning rate at moment t.

(6) Select another learning mode to provide to the input layer of the network and return to step (3) until all q learning modes are provided to the network.

(7) Updated learning rate $\eta(t)$ and neighborhood $N_g(t)$.

$$\eta(t) = \eta(0)\left(1 - \frac{1}{T}\right). \quad (10)$$

In the above formula, $\eta(0)$ is the initial learning rate, t is the number of learning, and T is the total number of learning.

Assume that the coordinate value of a certain neuron g in the competition layer in the two-dimensional array is (x_g, y_g), then the range of neighborhood is point $(x_g + N_g(t), y_g + N_g(t))$ and point $(x_g - N_g(t), y_g - N_g(t))$ as the square in the upper right corner and the lower left corner, and the modified formula is as follows:

$$N_g(t) = \mathrm{INT}\left[N_g(0)\left(1 - \frac{1}{T}\right)\right]. \quad (11)$$

In the above formula, INT $(\cdot)$ is the integral function.

(8) Let $t = t + 1$, return to step (2), until $t = T$.

2.3. Key Information Extraction Algorithm of Digital Archives. Key in the process of information extraction, in the digital archives to effectively extract the digital archives of key information, cannot individually understand the individual words of digital archives, and words or similar to each other in the digital archives correlation words combined into a block, a comprehensive understanding of the whole text content and the exact meaning of each word. Therefore, the semantic similarity between words is used as the clustering distance. All the semanemes of a word will form a hierarchical structure similar to a tree according to their upper and lower positional relations, which is traversed through the tree. Finally, the distance between words can be used to judge the similarity of word meaning. The formula for calculating word distance is as follows:

$$\mathrm{Sim}(p_1, p_2) = \frac{\alpha}{\alpha + \mathrm{dist}(p_1, p_2)}. \quad (12)$$

In the above formula, p_1 and p_2 represent two semesters, which are variable parameters. $\mathrm{dist}(p_1, p_2)$ represents the length of the path between two sememes of a word. The semantic origin of describing concepts is divided into four parts: The first basic semantic origin, the symbolic semantic origin, the relational semantic origin, and other independent semantic origin. The overall similarity between concepts is calculated by the following formula:

$$\mathrm{Sim}(s_1, s_2) = \sum_{i=1}^{n} y_i \prod_{j=1}^{i} \mathrm{Sim}_j(p_1, p_2). \quad (13)$$

In the above formula, s_1 and s_2 represent two concepts, and y_i represents the result of feature extraction. If there are two words w_1 and w_2 in the set, among which word w_1 has n concept descriptions and word w_2 has m concept descriptions, the maximum similarity between concepts w_1 and w_2 can be used as the semantic similarity of the two words, and the calculation formula is as follows:

$$\mathrm{Sim}(w_1, w_2) = \max_{i=1,\ldots,n, j=1,\ldots,m} \mathrm{Sim}(s_{1i}, s_{2j}). \quad (14)$$

The process of key information extraction algorithm of digital archives is as follows:

Preprocessing: Word segmentation for digital archival text, stop word overconsideration.

Step 1: Calculate all candidate words and semantic similarities between w_i and w_j in digital archival text $\mathrm{Sim}(w_i, w_j)$.

(1) TF-IDF value is calculated, and word $W = \{W_1, W_2, \ldots, W_N\}$ with word frequency greater than the threshold t is selected as the candidate key information. The calculation formula of TF-IDF value is as follows:

$$\frac{TF}{IDF_i} = \frac{tf_i \times \log(N/n_i)}{\sqrt{\sum_j (tf_i \times \log(N/n_j))^2}}. \quad (15)$$

In the above formula, tf_i is the number of occurrences of the word in the current digitized archival text, N is the total number of digitized archival text, and n_i is the number of digitized archives containing the word w_i in the database.

(2) During initialization, each word $\{W_i\}$ in the candidate word has a cluster Z_i, a total of n clusters, and all of them are set with unaccessed markers.

(3) Among all non-visited word clusters, select the cluster pair (C_l, C_k) with the largest similarity, that is, the closest distance, by calculating the maximum value of $\mathrm{Sim}(w_i, w_j)$. If $\mathrm{Sim}(C_l, C_k)$ is less than the given threshold, turn to (6); otherwise, merged clusters C_l and C_k are new clusters $C_0 = C_l \cup C_k$. Set to current cluster C, C to no access flag, C_l and C_k to access flag.

(4) Calculate the semantic similarity among all unaccessed word clusters, and transfer to (4).

(5) After clustering, the first k words with better quality are selected from each cluster Z_i as the final key information, so as to obtain the candidate word set $W = \{C_1, C_2, \ldots, C_m\}$.

Step 2: Treat each word in the text as a set C_i, a total of N sets (N is the number of words in the text).

Step 3: Select the two sets C_i and C_j with the greatest similarity from the N sets, and combine the two sets into a new set C.

Step 4: Select the center point of the current set: calculate the mutual information sum of the words in the current set and other words outside the set, and select the word with the largest mutual information value as the center point of the current set. If the calculated mutual information value between words is large, it indicates that they are also relatively large, on the contrary, it indicates that they are relatively small. The mutual information between w_i and w_j, that is, the public information between w_i and w_j, is calculated as follows:

$$I(w_i, w_j) = \log \frac{p(w_i|w_j)}{p(w_i)} = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}. \quad (16)$$

In the above formula, $p(w_i, w_j)$ is the common frequency of w_i and w_j, $p(w_i)$ is the separate frequency of w_i, and $p(w_j)$ is the separate frequency of w_j. According to the above formula, when $I(w_i, w_j) > 0$, the greater the value, the more public information between w_i and w_j and the stronger the correlation; when $I(w_i, w_j) = 0$, there is less public information between w_i and w_j and the correlation is weak; when $I(w_i, w_j) < 0$, there is no correlation between w_i and w_j.

Step 5: Among other words outside the set, select the word with the highest similarity with the center point of the set. If the similarity value is greater than the threshold, add it to the current set C; calculate the mutual information between the central point of the current set and the words outside the set, and add the word with the largest mutual information value to the current set C.

Step 6: Turn to step 4 to update the current collection center point until all words are accessed. If the mutual information value between the central point of the set and other words outside the set is less than 0, perform step 3 for the remaining unreachable words until all the words are accessed and divided.

Step 7: In the final cluster set, select its first K central words as the key information of the text. The key information extraction algorithm flow of different types of digital archives is shown in Figure 5.

3. Results and Discussion

3.1. Experimental Scheme. In order to verify the effectiveness of the algorithm designed in this paper to extract archive information, we conducted simulation experiments. This experiment is a simulation experiment, so it is necessary to

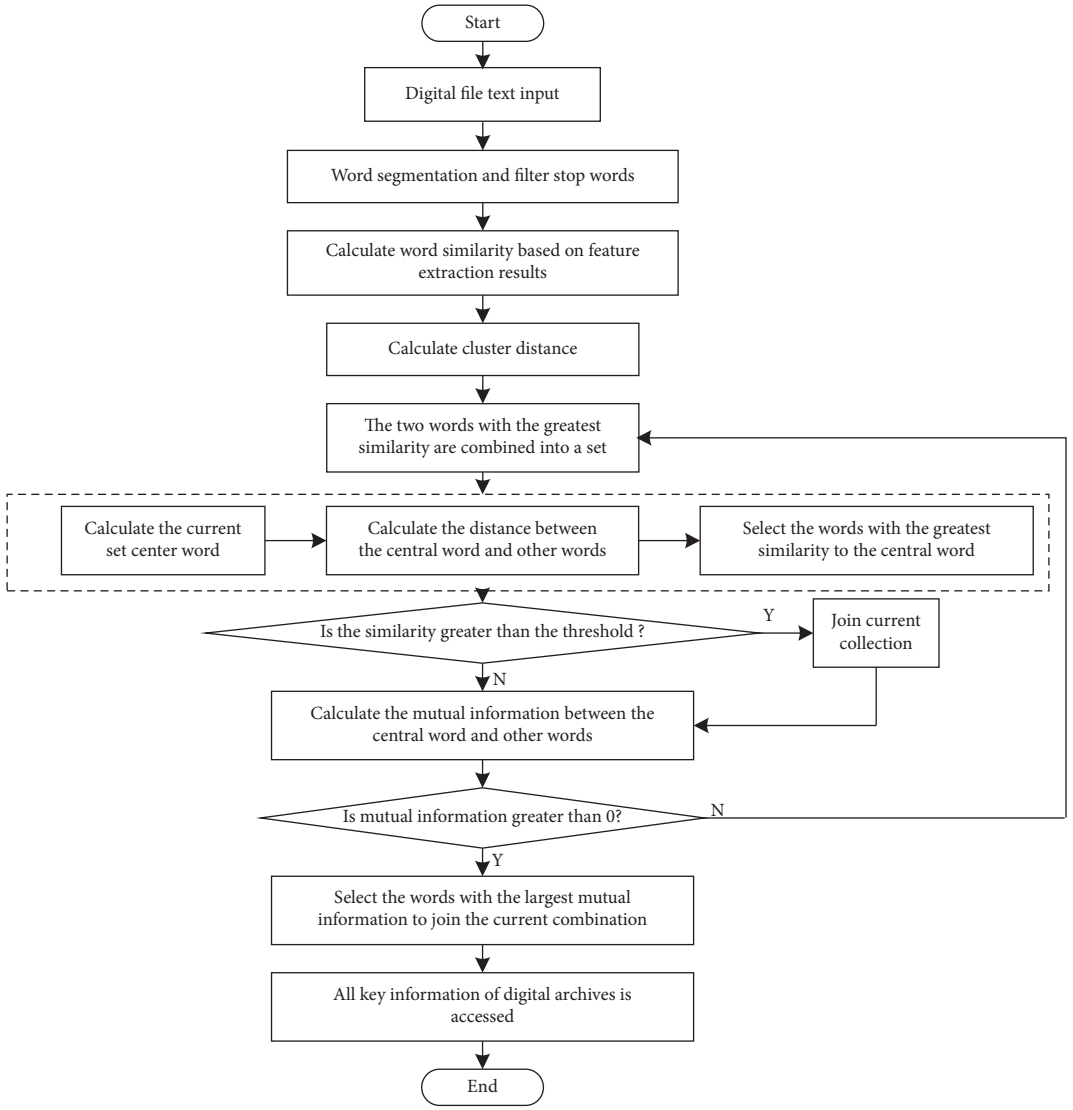

Figure 5: . Key information extraction algorithm flow of different types of digital archives.

design the experimental parameters, consider various factors, compare various types of simulation software and computers, and complete the design of environmental parameters of the simulation experiment, as shown in Table 1.

During the experiment, 500 GB digital archives were randomly selected from schools, enterprises, and relevant administrative units as data sets, and 450 GB of them were randomly selected as training sets to train this method. The remaining 50 GB were used as test sets to test the key information extraction performance of different types of digital archives. In order to ensure the objectivity of the experiment, the title and core prompt will be filtered out in the process of extracting key information. Recall rate and accuracy rate are often used as indicators of the key information extraction effect of different types of digital archives. Recall rate R and accuracy rate P adopted in this experiment are defined as follows:

$$R = \frac{j}{l} \times 100\%. \tag{17}$$

In the above formula, l represents the number of extracted key information, and j represents the actual number of key information.

$$P = \frac{L}{l} \times 100\%. \tag{18}$$

In the above formula, L represents the amount of key information accurately extracted.

The time-consuming calculation formula for extracting key information of different types of digital archives is as follows:

$$T = \sum_{i=1}^{n} t_i. \tag{19}$$

In the above formula, t_i represents the time taken for the i-th key information extraction step of digital archives.

3.2. Analysis and Discussion of Experimental Results. The recall rates of key information extraction of different types of

Table 1: Experimental environment parameters.

Experimental environment parameters	Configuration	Parameter
Hardware environment	CPU	Intel (R)Core (TM)i5-9400
	Frequency	2.90 GHz
	RAM	16.0 GB
Software environment	Operating system	Windows 10
	Analog software language	APDL
	Simulation software	Matlab 7.2

digital archives of reference [5] algorithm, reference [6] algorithm, reference [7] algorithm, and algorithm of this paper are compared. The results are shown in Figure 6.

By analyzing the data in Figure 6, we can see that the recall rate of the algorithm in reference [5] changes in the range of 58%–85%, the recall rate of the algorithm in reference [6] changes in the range of 49%–79%, and the recall rate of the algorithm in reference [7] changes in the range of 50%–87%. Compared with the experimental comparison algorithm, the recall rate of the algorithm of this paper changes in the range of 96%–99%, which is always higher than the experimental comparison algorithm, it shows that the key information of digital archives can be extracted comprehensively by using this algorithm, and the integrity is higher.

The key information extraction accuracy of different types of digital archives of reference [5] algorithm, reference [6] algorithm, reference [7] algorithm, and algorithm of this paper are compared. The results are shown in Figure 7.

By analyzing the data in Figure 7, we can see that the extraction accuracy of key information of digital archives of reference [5] algorithm is 49%–85%, the extraction accuracy of key information of digital archives of reference [6] algorithm is 54%–80%, and the extraction accuracy of key information of digital archives of reference [7] algorithm is 56%–80%. Compared with these algorithms, the extraction accuracy of key information of digital archives of the algorithm of this paper is 96%–98%. On the whole, the key information extraction accuracy of this algorithm is relatively stable, and there is no fluctuation of too high or too low, which indicates that the reliability of this algorithm in extracting key information is high. The accuracy of information extraction is higher, which can achieve the ultimate goal of accurately extracting the key information of different digital archives.

The extraction time of key information of different types of digital archives of reference [5] algorithm, reference [6] algorithm, reference [7] algorithm, and algorithm of this paper are compared. The comparison results are shown in Table 2.

By analyzing the results in Table 2, it can be seen that the average time-consuming of digital archives key information extraction of reference [5] algorithm is 1.41 s, the average time-consuming of digital archives key information extraction of reference [6] algorithm is 1.39 s, and the average time-consuming of digital archives key information extraction of reference [7] algorithm is 1.49 s, which is the highest among the four algorithms. Compared with these algorithms, the average extraction time of key information of digital archives in this algorithm is 0.63 s, which has a shorter extraction time and higher efficiency, and can realize the rapid extraction of key information of digital archives.

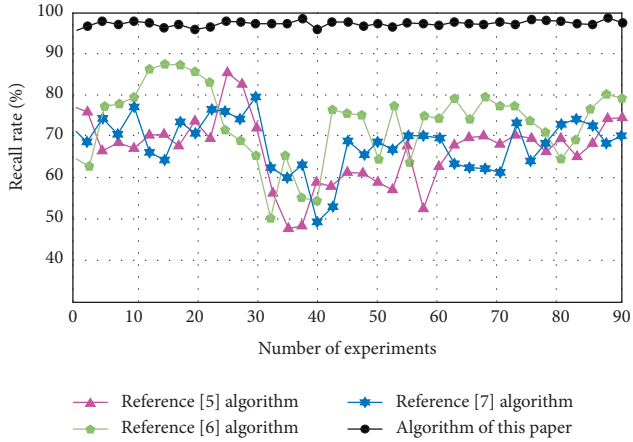

Figure 6: Comparison of recall rate.

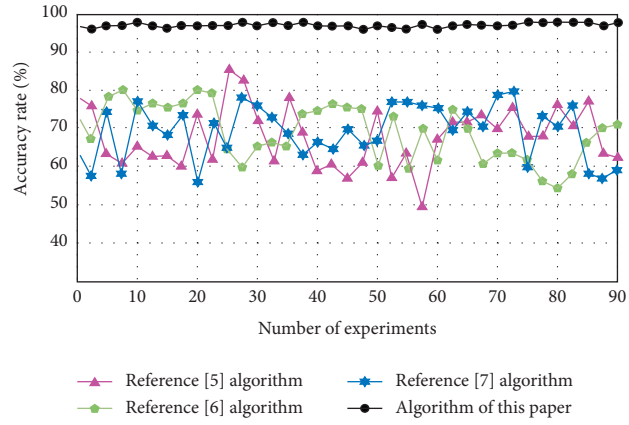

Figure 7: Comparison of accuracy.

To sum up, the recall rate of this algorithm changes in the range of 96%–99%, the accuracy of key information extraction of digital archives is 96%–98%, and the average time-consuming of key information extraction of digital archives is 0.63 s. It can achieve the goal of rapid and accurate extraction of key information of digital archives, solve a variety of problems existing in traditional methods, and can be widely used in many fields.

Table 2: Extraction time of key information of different types of digital archives.

Number of experiments	Time (s)			
	Reference [5] algorithm	Reference [6] algorithm	Reference [7] algorithm	Algorithm of this paper
10	1.25	1.33	1.47	0.47
20	1.44	1.25	1.58	0.56
30	1.23	0.96	1.56	0.58
40	1.38	1.47	1.47	0.62
50	1.45	1.58	1.52	0.85
60	1.63	1.65	1.35	0.67
70	1.45	1.58	1.47	0.81
80	1.56	1.40	1.62	0.57
90	1.29	1.31	1.41	0.51
Average value	1.41	1.39	1.49	0.63

4. Conclusions

With the continuous optimization of cultural operation and management strategies, the level of cultural operation and management has been gradually improved, and digital archives management is an important part of cultural operation and management. Therefore, extracting the key information of different types of digital archives is of great significance to the level of cultural operation and management. Therefore, this paper designs a key information extraction algorithm of different types of digital archives for cultural operation and management. The experimental results show that the recall rate of the algorithm is between 96% and 99%, the accuracy of key information extraction of digital archives is 96%–98%, and the average time-consuming of key information extraction of digital archives is 0.63 s. It can achieve the goal of rapid and accurate extraction of key information of digital archives and can be widely used in cultural operation and management, in order to improve the quality of cultural operation and management to the greatest extent, promote the further development of the cultural industry. However, the convergence of this algorithm is not tested in the process of operation. In order to avoid falling into the local optimum, it is necessary to increase the optimization of the algorithm in future research work to avoid too many iterations or high errors.

Data Availability

The dataset can be accessed upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] E. Ruikyte, "Reflections on "Then & Now": arts and cultural management and the shortcomings of student-led research projects," *Exchanges The Interdisciplinary Research Journal*, vol. 8, no. 4, pp. 87–98, 2021.

[2] N. P. Gomes and W. A. Cantú, "Sociocultural trend reports as an intelligence tool of strategic cultural management," *Marketing and Smart Technologies*, vol. 191, no. 1, pp. 63–74, 2022.

[3] Y. T. Lee and N. Y. A. Gyamfi, "The Sage handbook of contemporary cross-cultural management," *Journal of International Business Studies*, vol. 53, no. 3, pp. 568–571, 2022.

[4] T. Jackson, "Book review: cases in critical cross-cultural management: an intersectional approach to culture," *International Journal of Cross Cultural Management*, vol. 21, no. 2, pp. 398-399, 2021.

[5] L. Zhang, "Simulation Research on rapid extraction of massive data from digital books and archives," *Computer Simulation*, vol. 36, no. 3, pp. 397–400, 2019.

[6] Z. B. Chen, Y. M. Li, and F. Xu, "Research on forestry text key information extraction based on textrank and cluster filtering," *Journal of Agricultural Machinery*, vol. 51, no. 5, pp. 207–214, 172.

[7] Z. Q. Liu, Y. C. Du, and S. C. Shi, "Key information extraction of web news based on Improved Hidden Markov model," *Modern Library and Information Technology*, vol. 3, no. 3, pp. 120–128, 2019.

[8] W. Fan, H. Liu, and Y. X. Zhang, "Key information extraction algorithm integrating word vector and position information," *Computer Engineering and Application*, vol. 56, no. 5, pp. 179–185, 2020.

[9] W. J. Guo and X. A. Bao, "Key information extraction model of unstructured text in knowledge database," *Computer Simulation*, vol. 38, no. 9, pp. 357–360, 2021.

[10] B. Rajyagor and R. Rakholia, "Tri-level handwritten text segmentation techniques for Gujarati language," *Indian Journal of Science and Technology*, vol. 14, no. 7, pp. 618–627, 2021.

[11] B. Geng, "Text segmentation for patent claim simplification via bidirectional long-short term memory and conditional random field," *Computational Intelligence*, vol. 38, no. 1, pp. 205–215, 2022.

[12] M. Villamizar, O. Canévet, and J. M. Odobez, "Multi-scale sequential network for semantic text segmentation and localization," *Pattern Recognition Letters*, vol. 129, no. 1, pp. 63–69, 2020.

[13] S. Gu and F. Zhang, "Applicable scene text detection based on semantic segmentation," *Journal of Physics: Conference Series*, vol. 1631, no. 1, Article ID 012080, 2020.

[14] A. Ar, "Domain adaptation for part-of-speech tagging of Indonesian text using affix information - ScienceDirect," *Procedia Computer Science*, vol. 179, no. 1, pp. 640–647, 2021.

[15] A. Chaudhary, A. Anastasopoulos, Z. Sheikh, and G. Neubig, "Reducing confusion in active learning for part-of-speech tagging," *Transactions of the Association for Computational Linguistics*, vol. 9, no. 1, pp. 1–16, 2021.

[16] A. Rkm and A. Rb, "Integration of morphological features and contextual weightage using monotonic chunk attention for part of speech tagging," *Journal of King Saud University-Computer and Information Sciences*, vol. 27, no. 1, pp. 1–10, 2021.

[17] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," *Proceedings of the IEEE*, vol. 11, 2001.

[18] Y. Liang, X. Zhao, A. J. X. Guo, and F. Zhu, "Hyperspectral image classification with deep metric learning and conditional random field," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 6, pp. 1042–1046, 2020.

[19] S. T. Cheng, C. W. Hsu, G. J. Horng, and S. Y. Chen, "Across-camera object tracking using a conditional random field model," *The Journal of Supercomputing*, vol. 77, no. 12, Article ID 14252, 2021.

[20] 裴 Pei Liang, 刘 Liu Yang, and 高 Gao Lin, "Cloud detectionof ZY-3 remote sensing images based on fully convolutional neural network and conditional random field," *Laser & Optoelectronics Progress*, vol. 56, no. 10, Article ID 102802, 2019.

[21] J. Zheng and R. Ma, "Analysis of enterprise human resources demand forecast model based on SOM neural network," *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 6596548, 10 pages, 2021.

[22] X. Chen, H. H. Wang, and B. Tian, "Visualization model of big data based on self-organizing feature map neural network and graphic theory for smart cities," *Cluster Computing*, vol. 22, no. S6, Article ID 13293, 2019.

[23] H. Zhou and K. Yu, "A novel wireless sensor network data aggregation algorithm based on self-organizing feature mapping neutral network," *Ingénierie des Systèmes d'Information*, vol. 24, no. 1, pp. 119–123, 2019.

[24] K. J. Devi, N. H. Singh, and K. Thongam, "Automatic speaker recognition from speech signals using self organizing feature map and hybrid neural network," *Microprocessors and Microsystems*, vol. 79, no. 4, Article ID 103264, 2020.

[25] D. Qiu, H. Xu, D. Luo et al., "A rainwater control optimization design approach for airports based on a self-organizing feature map neural network model," *PLoS One*, vol. 15, no. 1, Article ID e0227901, 2020.