ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)

https://doi.org/10.17559/TV-20230927000965
Original scientific paper

Tamper-Resistant Corpus Retrieval Using Perceptual Hashing

Die HU, Weili HU*

Abstract: Conventional corpus retrieval tools are susceptible to malicious attacks, leading to the tampering of corpus resources. To solve this problem, in order to improve the security of corpus retrieval, a tamper resistance retrieval method of corpus based on perceptual Hash computer algorithm is proposed. First, the four dimensional chaotic map is used to encrypt the corpus resources to achieve Tamper resistance processing of the original corpus resources. Then, the robust features of the corpus resources are extracted, and after decomposition and dimensionality reduction, the feature sequence is transformed into a perceptual hash sequence, which facilitates matching the retrieval keywords with the hash sequence in the corpus during retrieval, further avoiding unauthorized modifications. Finally, the perceptual hash sequence is input into the lightweight neural network for training, and a combination of coarse and fine granularity is used to match the perceptual hash corpus in the corpus with the retrieval hash sequence input by the user, obtaining the retrieval results. The experimental results show that the retrieval accuracy of this method is higher than 98%, the tampering rate of the corpus is 0, and the retrieval performance and practical application value are greatly improved.

Keywords: chaotic mapping; corpus; corpus retrieval; perceived hash algorithm; tamper resistance

1 INTRODUCTION

Corpus is a collection of large amounts of language materials collected, organized, and stored from reality, with high value and importance. Due to the large scale and usually public resources of corpora, ensuring the integrity and reliability of their data is crucial for ensuring the credibility and traceability of research.

For linguistics, text book research, information technology and other fields, the corpus collects a large amount of corpus information, which can assist traditional language and Natural language processing research from data statistics or case analysis, it is also possible to conduct targeted corpus collection and updating work according to the research needs of the field, or annotate the collected corpus information in different ways, facilitating the smooth development of related research.

With the development and optimization of corpus construction and key technologies, the corpus contains increasingly rich corpus materials, which not only effectively promotes the development of subject education and teaching in professional fields, but also provides resource support for scientific knowledge popularization, scientific research practice, industrial production research and development, etc. [1]. However, the extensive collection of corpus resources in the corpus has greatly increased the difficulty of resource retrieval in the corpus. The research on corpus retrieval tools has received a lot of attention and achieved certain research results in the early stages of their application.

However, due to the technical design of the corpus itself, it is prone to malicious attacks and tampering. How to efficiently and accurately retrieve resources from corpora and ensure that resource retrieval is not tampered with has become a focus of current research. Many domestic and foreign corpora have already designed built-in retrieval tools in the early stages of construction, but the retrieval efficiency of built-in retrieval tools is relatively low and the retrieval ability is relatively limited, which not only cannot meet the specific needs of searchers, but also cannot guarantee the authenticity of the retrieval corpus, which is not conducive to the effective application of corpus resources [2]. In addition, with the continuous development of technology, the abuse of digital technology and the threat of information tampering are increasing. Especially in the era of the Internet, the dissemination of information has become extremely easy, leading to issues with the authenticity and credibility of information, and also posing challenges to the integrity of corpora. Through research on anti tampering retrieval methods for corpus, it is possible to effectively resist information tampering behavior and ensure the security and credibility of corpus data.

At present, some achievements have been made in the research of retrieval Tamper resistance methods, and the performance of retrieval methods has been improved in different aspects. Among them, reference [3] focuses on the metadata of cloud logs, chain of custody, and cloud files, utilizing blockchain technology to ensure data integrity on cloud virtual machines and achieve tamper proof access among stakeholders. However, this method cannot be transferred to corpus tamper prevention work. Reference [4] has designed a special information retrieval system for Formulaic language in the corpus based on the search engine Elastic search. The system expands the corpus according to the common Language change and modifier change laws of Formulaic language. In the query process, we can realize the corpus search by capturing the changes of derivatives or other Formulaic language. However, in practical applications, it has been found that although language has certain regularity, the changes in idioms and idioms are often very specific and individual. These changes include local changes in language, changes in formal modifiers, and changes in derivative words. There are also differences between different idioms and idioms. Therefore, fully matching all possible changes is a very difficult task. This leads to a low retrieval accuracy of this method. Reference [5] utilizes web online tools to provide corpus retrieval functionality for essay/correction systems, and combines it with the CALL (Computer Aided Language Learning) exercise system to create more effective retrieval based on learners' misuse. However, learners may have inaccurate grammar, spelling, or vocabulary when submitting their essays. These errors can lead to noise or incorrect data in the learner's corpus, thereby reducing the ability of the method to perform accurate matching. Reference [6] proposed to combine the segmented fuzzy C-means clustering algorithm with the Inverted index algorithm for document content retrieval, and achieve the matching of retrieval resources and retrieval requirements with Hamming distance as the standard, although Hamming distance is a commonly used measure of the difference between two binary strings. However, in document retrieval, Hamming distance only considers differences in document content, without considering differences in document structure and semantics. This means that even if two documents are semantically identical but structurally slightly different, their Hamming distance may still be significant. Therefore, using only Hamming distance to match retrieval resources and requirements may result in ineffective identification of tampered or modified documents. Reference [7] improved the fuzzy rough subset theory using region mining algorithms to enhance the correct decision-making ability of fuzzy rough subsets for uncertain data. Remove redundant attributes from the database to improve the efficiency of data resource retrieval and the accuracy of retrieval result sorting. In this study, the region mining algorithm is a data mining algorithm used to discover hidden patterns and association rules in datasets. Although it has certain data analysis capabilities, it may not be well applied in tamper proof scenarios. Anti tampering mainly involves verifying and protecting the integrity and authenticity of data, and region mining algorithms may not provide effective solutions to these problems.

However, when using the above methods for corpus resource retrieval, it was found that although the above methods can improve the resource security of the corpus to a certain extent, the corpus information of the corpus has always been exposed in the network during retrieval work, and attackers can maliciously tamper with the corpus content, causing unpredictable impact on the research and application of the corpus resources.

Based on the above analysis, this study designed a tamper-resistant corpus retrieval method using perceptual hashing. The design idea is as follows:

① Using chaotic mapping algorithm to encrypt the original corpus in the corpus, thereby reducing the probability of tampering during the process of uploading the corpus to a public server.

② Robust feature extraction and processing of encrypted corpus. By using wavelet analysis and Laplacian mapping, robust features of the corpus are obtained. After dimensionality reduction, a hash corpus sequence is generated according to perceptual hash algorithm rules.

③ The fingerprint features of the perceptual hash corpus sequence are used as similarity matching objects, and after being trained and processed by a lightweight neural network, the retrieval results are output in descending order of similarity.

2 RESEARCH ON TAMPER RESISTANCE RETRIEVAL METHODS OF CORPUS

2.1 Tamper Resistance Processing of Corpus

The majority of corpus resources stored in a corpus are text resources, which are labeled as different categories during collection. When users use a corpus for corpus resource retrieval, the search keywords they use are matched with the annotation labels to obtain the retrieval results. However, in the above process, the corpus data is highly susceptible to malicious tampering, and the tampered corpus resources are difficult to distinguish through simple sequence correlations, leading to corpus application problems. The Tamper resistance processing of corpus is mainly to encrypt the original corpus when the corpus is uploaded to the physical server of the corpus using encryption algorithms, so as to ensure the security of the corpus stored in the open server of the corpus. The resources in the corpus are directly stored in the form of digital Character encoding, which makes malicious tampering difficult. This study converts all document type resources in the corpus into images to improve the construction efficiency of subsequent perceptual hash algorithms. In this study, the chaotic mapping principle is used to encrypt the original corpus image to prevent Tamper resistance.

Chaotic mapping is a highly unpredictable and sensitive initial value dependent dynamic system. Using chaotic mapping to encrypt corpus resources can enhance data security and prevent unauthorized access and tampering. At the same time, the sequences generated by chaotic mapping have nonlinear characteristics, which makes it more difficult for attackers to infer and recover the original information through mathematical methods after encryption of corpus resources.

In the process of corpus image encryption, chaotic mapping can insert scrambling sequences between different pixel points, ensuring that the probability of real corpus being tampered with is reduced when the corpus is maliciously tampered with [8]. Fig. 1 shows the flowchart of chaotic corpus encryption processing.

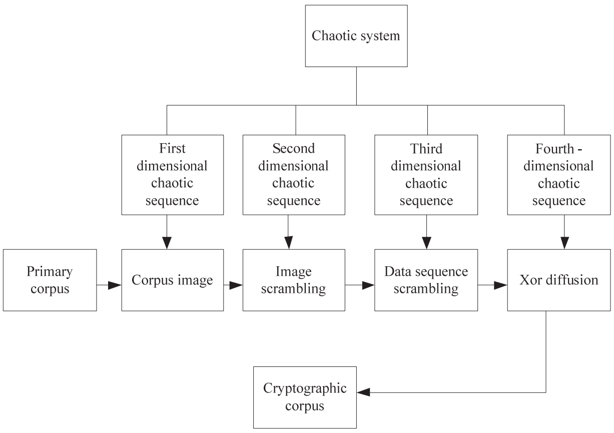
Chaotic system
First dimensional chaotic sequence
Second dimensional chaotic sequence
Third dimensional chaotic sequence
Fourth - dimensional chaotic sequence
Primary corpus
Corpus image
Image scrambling
Data sequence scrambling
Xor diffusion
Cryptographic corpus

Figure 1 Block diagram of chaotic encryption for corpus images

The principle formula of Henon mapping is as follows:

$$\begin{cases} u_{n+1} = 1 - \alpha u_n + v_n \\ v_{n+1} = \beta u_n \end{cases} \tag{1}$$

In the formula, α and β are the control parameters of the Henon map, respectively; u, v is a chaotic sequence variable, respectively.

Set the control parameter values of Henon mapping according to Initial condition such as the size of the corpus image and the number of pixels. Among them:

① The size of the corpus image: The size of the corpus image can affect the selection of control parameters for Henon mapping. Larger images may require more complex and diverse control parameters to ensure that the generated chaotic sequence has sufficient randomness and diffusion.

② Number of pixels: The number of pixels in the corpus image can also serve as a reference for setting Henon mapping control parameters. A higher number of pixels may require a larger parameter range or more complex parameter settings, thereby increasing the complexity and security of chaotic sequences.

③ Computational efficiency: Computational efficiency is an important factor in determining the parameter values of Henon mapping control. Overly complex parameters may lead to excessive computational burden, affecting the real-time and practicality of the system. Therefore, it is necessary to balance safety and computational efficiency when setting parameters.

Due to the fact that corpus resources contain attributes of different dimensions and quantities when uploaded, in order to fully cover all possible attributes, two positive Lyapunov exponents are introduced to form a hyperchaotic system [9]. The system introduces two new state variables *z* and *e*, and couples the two variables into the above mapping formula to obtain the following mathematical model:

$$\begin{cases} u_{n+1} = \alpha\left(u_n - v_n\right) \\ v_{n+1} = u_n z_n - \beta u_n + e_n \\ z_{n+1} = {u_n}^2 - \beta z_n \\ e_{n+1} = \varepsilon\left(u_n + v_n\right) \end{cases} \tag{2}$$

In the equation, ε is the control parameter, and its value is determined based on the corpus image.

Use the initial set value as the Henon mapping encryption key to process the chaotic sequence of the corpus image. The newly generated chaotic sequence is arranged in descending order according to the number of internal pixel values, and an encrypted mapping relationship corresponding to the original corpus resources is obtained. The chaotic sequences of the corpus images added with the disorder coefficient are reassembled, and the transformed sequence forms are uploaded to the public server of the corpus to complete Tamper resistance encryption. After Tamper resistance encryption of the original corpus in the corpus, the perceptual hash sequence of the corpus is constructed.

2.2 Constructing Corpus Perceived Hash Sequences

Although the possibility of tampering with the corpus is reduced after chaotic mapping encryption processing, once the encryption algorithm is cracked or some of the corpus is scrambled, it can still lead to errors and confusion in the corpus content [10]. Therefore, this article uses a perceptual hash algorithm to construct a perceptual hash sequence for all corpora in the corpus, in order to match the retrieval keywords with the hash sequence in the corpus during retrieval and prevent unauthorized modifications.

The perceptual hash algorithm has high computational efficiency and can quickly convert high-dimensional corpus features into low-dimensional binary representations. The ability to compress this dimension can reduce the computational and storage load of subsequent processing, improve the efficiency and effectiveness of retrieval. Meanwhile, the perceptual hash algorithm can map similar features onto similar hash codes by calculating the similarity between different features. This enables perceptual hash sequences to be used to quantify similarity, allowing for hash matching based retrieval operations.

The original data of the corpus resource has the characteristics of robust invariance, which is completely corresponding to the content of the corpus itself, and has particularity and uniqueness. The corpus resources themselves are textual data, not image data. However, by converting text into image representations, corpus resources can be visualized as images. Use Bag-of-words model or text vectorization technology (such as Word2Vec, BERT, etc.) to convert text to vector representation, and then use VAE image generation technology to convert vector representation to image. In this way, corresponding images can be generated from the text data of the corpus. However, it should be noted that the results of image generation are based on model training and learning, and are not directly extracted from the original corpus resources.

In two-dimensional space, the display of corpus resources is transformed into an area composed of black and white pixels, and the black area and the unconnected areas that form the font form are extracted, which are the local features of the corpus resources [11]. Starting from the black area formed by the connection, and based on the definition of the connection by morphology, assuming $X_i \subseteq CO$ is the large range of black pixel connections in the corpus image, the image connection index of X_i is based on the following equation:

$$CU\left[\left(X_1 \oplus B\right) \cap \Delta CO\right] \geq CU\left[\left(X_2 \oplus B\right) \cap \Delta CO\right] \geq \ldots \tag{3}$$

In the formula, *CU* is a subset of corpora of the same type; *CO* represents all corpora in the corpus to be searched; *B* represents an operator that performs erosion and dilation operations on corpus images.

Determine the connection of black pixels through floating transformation. When the following conditions are met, it is determined that there is a connection [12]:

$$\left[\left(X_a \cap X_b\right)_k \oplus B\right]_k \cap \left[X_b\right]_k \neq \phi \tag{4}$$

$$\left[A_a \oplus B\right]_k \cap \left[A_a\right]_k \neq \phi \tag{5}$$

$$\left[B'\right]_k \cap X_k \neq \phi \tag{6}$$

In the formula, A_a is the connection endpoint of the black pixel; *k* is the serial number of the black pixel connection part; B' is the image of the corrosion and expansion operator *B*; ϕ represents an empty set; *a*, *b* is the serial number encoding of different black pixels.

The above three formulas correspond to different black pixel connection types. After determining the connections of various black pixels in the corpus resources in two-dimensional space, the black pixel regions are extracted and subjected to Gaussian filtering and median processing.

Before constructing the perceptual hash sequence of corpus images, robust feature compression can be directly performed on corpus images due to their small size. The low-density parity Check digit is used to compress the robust features of the corpus image into the sequence, forming the robust compression sequence of the corpus content, and improving the feature compactness of the corpus image [13].

The method of information entropy is used to count the number of feature pixels contained in the robust feature regions of corpus images. The maximum and minimum values of the compressed robust image feature pixels are used as the upper and lower limits, and are evenly divided into several intervals. If the maximum number of feature pixels that can be placed in each interval is $n_i, i = 1, 2, \cdots, bins$, and $bins$ is the number of intervals divided, then the number of possible positions where pixels fall into each interval p_i is the total number of pixels in the region [14].

$$p_i = \frac{n_i}{total} \tag{7}$$

In the formula, *total* represents the total number of pixels in the region. Calculate the information entropy of regional pixels according to the following equation:

$$H = -\sum_{i=1}^{bins} p_i \times \log_2 p_i \tag{8}$$

After calculating the corpus pixels in the robust feature image based on information entropy, the wavelet decomposition algorithm shown in Fig. 2 is used to decompose the corpus image into high and low dimensions.

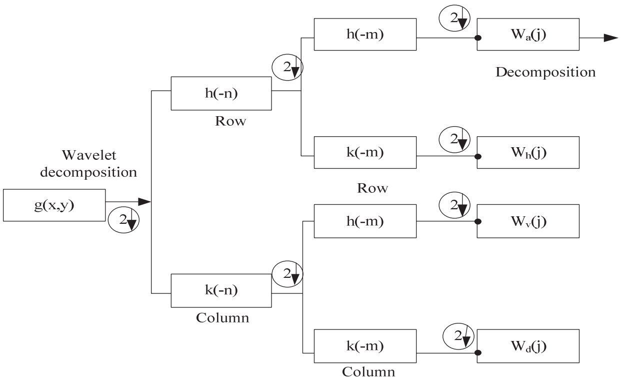

Figure 2 Wavelet decomposition of corpus images

In Fig. 2, the output $W_a(j)$, $W_h(j)$, $W_v(j)$, and $W_d(j)$ are all decomposition coefficients of wavelet decomposition with a decomposition scale of j, while p and q represent horizontal and vertical translations, respectively. The principle formula of wavelet decomposition is as follows [15]:

$$WT_s(a,b) = \frac{1}{\sqrt{a}} \int s(t) \psi^* \left(\frac{t-b}{a} \right) \mathrm{d}t \tag{9}$$

In the formula, a is the scale factor during wavelet decomposition, and its numerical value is a positive number; b is the time-shifted parameter of wavelet decomposition; $\psi(t)$ is the mother wave of wavelet changes; $s(t)$ is the original corpus image sequence that has not been decomposed by wavelet transform.

When wavelet decomposes black and white pixels in corpus images, they are decomposed into four components based on pixel connectivity and pixel distribution frequency, and subjected to downsampling decomposition processing. After downsampling and decomposition processing, the high-resolution reconstruction of the four components is performed through the inverse wavelet transform function, namely [16]:

$$fm_j^t(u,v) = \frac{\left\{ IWC_j \left[W_h{}^t(j), W_v{}^t(j), W_d{}^t(j) \right] \right\}^2}{r} \tag{10}$$

In the formula, $fm_j^t(u,v)$ is the robust feature map after the j-level decomposition and reconstruction of the corpus image, (u,v) is the coordinates of the pixels in the decomposed and reconstructed feature map, IWC_j is the reconstruction function, and r is the constraint factor.

After wavelet decomposition of the pixel features of the corpus, the information features of the low-frequency part will be discrete or lost. Therefore, this paper uses Laplacian mapping to construct the Adjacency matrix of the low-frequency part of the corpus image to obtain low-frequency robust features [17, 18].

$$\begin{aligned} & \min \sum_{i,j}^{\mathrm{n}} \left\| q_i - q_j \right\|^2 L_{ij} \\ & = min \sum_{i}^{\mathrm{n}} \sum_{j}^{\mathrm{n}} \left\| q_i - q_j \right\|^2 L_{ij} \\ & = min\, 2 \sum_{i=1}^{\mathrm{n}} D_{ii} q_i^T q_i - 2 \sum_{i}^{\mathrm{n}} \sum_{j}^{\mathrm{n}} q_j^T q_j L_{ij} \end{aligned} \tag{11}$$

where, L_{ij} represents the Adjacency matrix of the corpus image after wavelet decomposition; D represents the degree matrix of the Adjacency matrix in the form of Diagonal matrix. y represents the pixels with similar positions in the low-frequency image components after wavelet decomposition processing. The mapping target matrix of low-frequency corpus images is composed of multiple pixels. By Lagrange multiplication of the above formula, the non-zero eigenvalue obtained is the processing of the low-frequency part of the corpus image.

After combining the robust features of the high-frequency and low-frequency parts of the corpus, the robust feature parameter matrix γ corresponding to each pixel point is obtained. If each robust feature in the parameter matrix is greater than the previous feature, the previous feature value is recorded as 1 based on the perceptual hash principle [19]. Eq. (12) is the perceptual hash sequence formula for generating corpus images:

$$Ha_i(x)=\begin{cases}1,\gamma_w(j+1)>\gamma_w(j)\\0,\gamma_w(j+1)\le\gamma_w(j)\end{cases} \quad (12)$$

In the formula, *j* represents the robust feature vector of the corpus with the number *j* in the robust feature parameter matrix *γ* with the dimension of pixel *w*. After constructing the hash sequence of all corpora in the corpus, further processing is carried out to achieve the retrieval and use of corpus data.

2.3 Corpus Hash Search Matching

When users search for corpus resources in the corpus, they input the required corpus resource keywords and obtain the corresponding search results based on the matching between the search keywords and the search index. However, there is a probability of tampering at all stages of the above process, and the perceptual hash algorithm can perform unidirectional transformation on the corpus to avoid tampering. After processing the corpus resources in the corpus in the above process, the similarity matching principle is used to realize the perceptual hash computer algorithm for corpus Tamper resistance retrieval. The specific flow diagram is shown in Fig. 3.

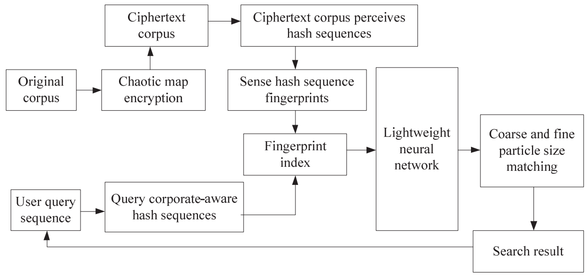
Ciphertext corpus
Ciphertext corpus perceives hash sequences
Original corpus
Chaotic map encryption
Sense hash sequence fingerprints
Lightweight neural network
Coarse and fine particle size matching
Fingerprint index
User query sequence
Query corporate-aware hash sequences
Search result

Figure 3 Schematic diagram of corpus tamper resistance retrieval process

To improve the retrieval efficiency of corpus resources, this study trains a lightweight neural network by inputting perceptual hash sequences. Lightweight neural network models typically have fewer parameters and a simpler structure, which can reduce computational and storage burden while maintaining a certain level of accuracy. Train the neural network model using a large number of annotated corpus samples using perceptual hash sequences as input. The training process mainly includes steps such as forward propagation, loss function calculation, and backpropagation. By optimizing the weights and biases of the neural network model, it can learn better corpus feature representation and similarity discrimination capabilities. After the training is completed, the trained lightweight neural network model is used to extract features from the input hash sequence, which can calculate the similarity between samples and achieve efficient retrieval of corpus resources in the corpus.

The lightweight neural network selected in this article is based on the Bottleneck architecture, with a convolution step size of 1 and the residual value channel removed from the framework. In the first layer, the input perceptual hash sequence is convolved to increase its dimension, in the second and third layers, the sequence is processed by separating convolution kernel normalization, and in the Co representation layer, it is processed by Activation function. Finally, the processed perceptual hash sequence is convolved and dimensionally reduced, and the trained perceptual hash sequence is obtained after output.

When retrieving corpus corpus, a combination of coarse and fine granularity is used to match the perceptual hash corpus in the corpus with the retrieval hash sequence input by the user [20]. Fig. 4 shows a coarse to fine granularity matching framework for corpus aware hash sequences. By combining this framework with a lightweight neural network, retrieval results can be obtained.

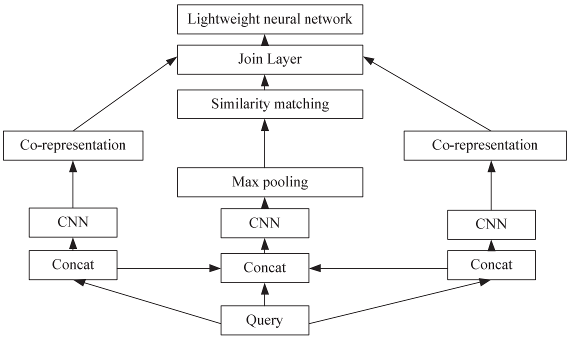

Figure 4 Coarse and fine-grained matching network for corpus

The search term input by the user is trained by a lightweight neural network, and the corresponding perceptual hash sequence is also output. By utilizing the similarity of feature fingerprints in the hash sequence, the matching degree between the retrieval target and the resources in the corpus is obtained. The definition formula for feature fingerprints is as follows:

$$R(f_i,n)=\left(2^{n-1}c_1+2^{n-2}c_2+O+2^0c_i\right)mod\,M \quad (13)$$

In the formula, *n* is the number of non repetitive robust features in the user input retrieval perception hash sequence f_i; *M* is a prime number randomly generated based on robust features; *O* is the replaceable probability value of keyword c_i.

According to the Hamming similarity principle, the similarity calculation is performed between the user input retrieval perception sequence fingerprint and the corpus perception hash sequence in the corpus. Sort the results in descending order based on similarity calculation and return them to the user, completing a corpus retrieval task [21].

3 EXPERIMENTAL VERIFICATION AND RESULT ANALYSIS

To verify the feasibility and practical value of the tamper-resistant corpus retrieval method using perceptual hashing designed above, the following experiments are designed.

3.1 Design of Experimental Plan

In order to intuitively and accurately judge the performance and practical application value of the corpus

Tamper resistance retrieval methods studied, the corpus retrieval methods mentioned in reference [4] and reference [6] are introduced as a comparison.

In order to avoid the great impact of the experiment process on the operation of the real corpus, this experiment copies the real corpus data to the experimental server, forms a mirror image and establishes an experimental LAN, physically isolating the impact of the retrieval method experiment. In this experimental server, tampering attacks are simulated manually to generate several different types of abnormal corpora. After processing all the corpus resources on the experimental server, set different retrieval goals and conduct virtual retrieval. Three corpus retrieval methods output corresponding retrieval results based on the retrieval target. Based on the quantitative indicators of retrieval method performance set in the experiment, collect and analyze experimental data, complete the experiment, and evaluate the performance of the retrieval method.

3.2 Quantitative Indicators of Retrieval Method Performance

Considering the ability to better quantify the retrieval performance of the method on the corpus, the recall and accuracy of the experimental selection method are combined to further describe the retrieval performance of the method. The higher the recall and accuracy overall, the stronger the robustness of the retrieval method and the higher its practical application value.

For retrieval methods, the ability to retrieve relevant content in the corpus reflects the comprehensive search coverage of the method. The experiment uses recall rate indicators to quantify the retrieval coverage ability of the method. The recall rate formula is defined as follows:

$$R = \frac{N_{su}}{N_{su} + N_{fu}} \tag{14}$$

In the formula, N_{fu} represents the corpus resources that meet the retrieval requirements but have not been retrieved by the method when conducting corpus retrieval according to the retrieval requirements; N_{su} represents the correct corpus resources successfully retrieved by the method.

Retrieval methods can be maliciously tampered with in practical applications, and it is important to be able to retrieve the correct corpus that meets the requirements instead of the tampered corpus.Therefore, the experiment takes the proportion of the tampered corpus in the retrieval results as a quantitative indicator of Tamper resistance prevention performance, that is, the proportion of the tampered corpus (whether it meets the retrieval requirements or not) in the retrieval results among all the corpus results retrieved according to the retrieval keywords. The calculation formula is as follows:

$$D = \frac{N_p}{N_A} \tag{15}$$

In the formula, D represents the proportion of tampered corpus in all search results; N_p represents the number of tampered corpora in all retrieved corpora; N_A represents the number of all corpora retrieved by the retrieval method based on the search keywords entered by the user.

The experiment used accuracy indicators to quantify the accuracy ability of corpus retrieval resources. The accuracy calculation formula is as follows:

$$P = \frac{N_{co}}{N_{co} + N_{ffu}} \tag{16}$$

In the formula, N_{co} represents the correct or unmodified corpus retrieved by the method from the required resources; N_{ffu} represents a false corpus that has been tampered with, although it meets the retrieval requirements.

In addition, for corpus retrieval methods, due to differences in application environments, the efficiency of retrieval methods is also an indicator used to quantify the performance of retrieval methods. Set different retrieval requirements based on the corpus label information in the experimental corpus. Use the method of this paper, method of reference [4], and method of reference [6] loops to complete each retrieval request. Calculate the average retrieval time required for each corpus retrieval. Choose the length of retrieval time to quantify the efficiency of the retrieval method. The shorter the retrieval time corresponding to each retrieval request, the higher the efficiency of the retrieval method in completing the corresponding retrieval work.

By comparing the retrieval performance and quantitative indicator data of method of this paper, method of reference [4], and method of reference [6] on the experimental corpus, a comparative conclusion is obtained through overall analysis, and the performance and practical application value of the corpus retrieval method studied are evaluated.

3.3 Comparative Experimental Results and Analysis

Using the experimental indicators proposed above as quantitative indicators for the performance of corpus retrieval methods, statistical analysis and processing of indicator data were conducted to obtain the recall and accuracy of the retrieval method as shown in Tab. 1.

Analyzing the data in Tab. 1, it can be seen that under the influence of constantly changing corpus retrieval objectives, the recall rate fluctuations of reference [4] and method of reference [6] are relatively small, but overall slightly lower than the method of this paper. However, the retrieval accuracy of the methods in reference [4] and [6] shows significant fluctuations, making it difficult to achieve good retrieval accuracy for each retrieval target set in the experiment. From a numerical perspective, the retrieval accuracy of method of this paper is higher than 98%, while the highest retrieval accuracy of references [4] and method of reference [6] are only 94.2% and 95.8%, respectively. In addition, there is a significant difference between the minimum and maximum retrieval accuracy of the two reference methods, resulting in poor retrieval performance. The reason for the above results is that our method uses four-dimensional chaotic mapping to encrypt corpus resources, which can effectively prevent tamperers from maliciously tampering with the original corpus resources. This can ensure the integrity and authenticity of the corpus. After decomposition and dimensionality reduction, robust features are extracted from corpus resources. These features can better distinguish different corpus resources and improve the accurate matching ability for similar corpora.

Table 1 Recall and accuracy of corpus retrieval methods / %

Number	Method of this paper		Method of reference [4]		Method of reference [6]	
	Recall rate	Accuracy rate	Recall rate	Accuracy rate	Recall rate	Accuracy rate
1	99.3	99.4	96.8	90.4	98.6	93.6
2	99.9	99.3	97.8	93.2	98	90.7
3	99.8	98.8	97.9	91.8	98.1	84.2
4	99.7	99.4	97.3	82.7	98.7	93.1
5	98.9	98.7	97.4	83.2	98.4	84.8
6	98.1	99.3	96.3	70.9	98.1	93.4
7	98.7	99.8	97.8	92.1	98	92.5
8	99.9	99.6	97.2	91.5	97.8	79.2
9	98.5	99.6	96.8	88.3	98.8	95.8
10	98.9	99.3	97.6	94.2	98.6	94.2

Fig. 5 shows the Receiver operating characteristic of the three methods.

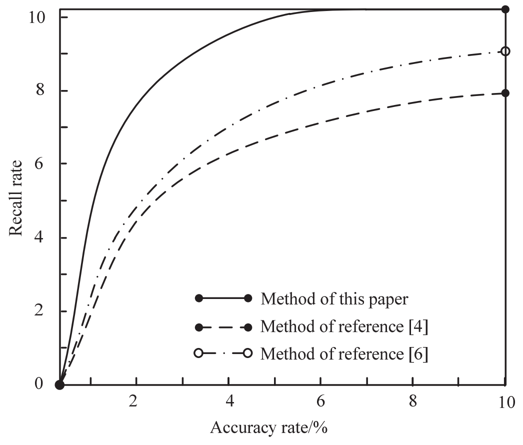

Figure 5 Change of ROC

According to the definition of the Receiver operating characteristic, the larger the area of the closed area formed below the curve, the higher the recall and accuracy of the corresponding retrieval method, and the better the balance between the two. From Fig. 5, the closed area formed by the Receiver operating characteristic of method of this paper is much larger than that in the reference [4] and method of reference [6], which indicates that method of this paper is robust for corpus retrieval. The reason for the above results is that the robust features extracted by this method from corpus resources have good anti-interference ability, which can to some extent overcome the influence of factors such as noise and image deformation. In addition, after converting the feature sequence into a perceptual hash sequence, some noise and changes in the original features can be filtered to extract more stable hash codes to represent corpus resources, which also enhances the robustness of the method proposed in this paper.

Tab. 2 shows the statistical values of the proportion of tampered results among all the results retrieved on the experimental corpus by method of this paper, method of reference [4], and method of reference [6].

Table 2 The proportion of tampered search results / %

Number	Method of this paper	Method of reference [4]	Method of reference [6]
1	0	0.14	0.01
2	0	0.11	0.13
3	0	0.11	0.12
4	0	0.31	0.11
5	0	0.13	0.14
6	0	0.15	0.08
7	0	0.22	0.11
8	0	0.18	0.09
9	0	0.31	0.15
10	0	0.12	0.12

After analyzing the data in Tab. 2, it can be seen that there are no tampered corpus resources in all search results of the method of this paper, and the proportion of tampered resources is 0. However, in the search results of references [4] and method of reference [6], a small amount of tampered resources appeared. Among them, the highest proportion of tampered corpus resources retrieved by method of reference [4] is 0.31, while the highest proportion of tampered corpus resources retrieved by method of reference [6] is 0.15%, both of which are higher than 0. The retrieval results have been tampered with and the credibility of the retrieval results is reduced. The reason for the above results is that the proposed method utilizes chaotic mapping encryption to ensure high security of the original corpus resources, making it difficult to be maliciously tampered with or modified. In the process of feature extraction and hash sequence transformation, additional recognition and filtering of tampered resources are performed by extracting robust features and using perceptual hash algorithms, further ensuring the reliability of the retrieval results.

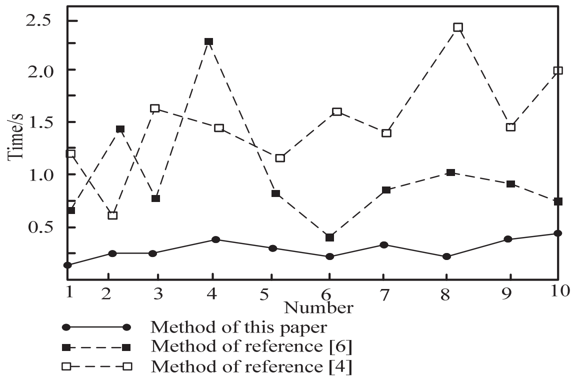

Figure 6 Comparison of retrieval efficiency

Fig. 6 shows the average retrieval time curve for each retrieval result when using three methods to perform cyclic corpus retrieval on the same retrieval target.

Analyzing the trend of line segment changes in Fig. 6, it can be seen that the overall retrieval time curve of method of this paper is located below the retrieval time curve of references [4] and method of reference [6]. Explain that the method of this paper outputs the search results as quickly as possible when searching for target requirements. And from the perspective of a single curve change, the fluctuation amplitude of the time consuming curve of the method of this paper is low, and the stability of retrieval efficiency is better. The reason for the above results is that this method uses lightweight neural networks as models for training, with fewer parameters and a simple structure, which can improve computational efficiency and reduce storage burden.

Based on the above analysis of all experimental index data, when retrieving the corpus, the Tamper resistance retrieval method based on the hash algorithm studied in this paper can retrieve the target corpus more quickly, and the corpus resources retrieved by the method are not affected by illegal tampering, and the retrieval accuracy and corpus retrieval security are greatly improved.

4 CONCLUSION

For disciplines such as linguistics and computer information language research, authentic and comprehensive corpus resources are the fundamental materials for conducting various research. Corpus carries important functions such as providing corpus resources and research concepts for research and teaching in different disciplines, achieving the sharing of corpus research results and resources. The use of corpora not only faces the challenge of limited search efficiency, but also ensures that the retrieved corpus is not maliciously tampered with to avoid adverse consequences.

This article aims to improve the efficiency and accuracy of corpus retrieval, and ensure the security of the retrieved corpus. It conducts research on a corpus tamper proof retrieval method based on perceptual hashing. Firstly, a four-dimensional chaotic map is used to encrypt the corpus resources, achieving tamper proof processing of the original corpus resources. Then, the robust features of the corpus resources are extracted and decomposed to reduce dimensionality, transforming the feature sequence into a perceptual hash sequence. During retrieval, match the search hash sequence input by the user with the perceptual hash sequence in the corpus to further avoid inaccurate retrieval results caused by tampering. Finally, the perceptual hash sequence is input into the lightweight neural network for training, and coarse to fine granularity matching is used to match the perceptual hash corpus in the corpus with the retrieval hash sequence input by the user, obtaining the final retrieval result.

This method utilizes the unidirectional nature of perceptual hashing algorithms to reduce the possibility of corpus tampering. Output correct retrieval results through similarity matching of hash sequences. Through comparative experiments with other methods, the retrieval accuracy of method of this paper has reached more than 98%, and the retrieval efficiency has been improved. The retrieved corpus is not tampered with, and Tamper resistance security is stronger.

In future research, we will consider introducing multimodal data such as images, text, audio, etc. in the process of corpus tamper proof retrieval, in order to integrate different types of information and improve the overall system performance.

Acknowledgement

2021 Annual Project of National Education Science "Fourteenth Five Year Plan" (Grant No. ELA210435); National Education Examination Research Plan 2021 Project (No. GJK2021026).

5 REFERENCES

[1] Karlsen, P. H. (2021). Educational Roles in Corpus-Based Education: From Shift to Diversification. *Nordic Journal of Language Teaching and Learning*, *9*(1), 1-12. https://doi.org/10.46364/njltl.v9i1.833

[2] Al-Surmi, M. (2022). TV shows, authenticity, and language learning:A corpus-based case study. *Register Studies*, *4*(1), 30-54. https://doi.org/10.1075/rs.19016.als

[3] Pranitha, S., Digambar, P., & Radha, V. (2022). Blockchain-based tamper-proof and transparent investigation model for cloud VMs. *The Journal of Supercomputing*, *78*(16), 17891-17919. https://doi.org/10.1007/s11227-022-04567-4

[4] Callum,H.,Maxim, F., Alison, W., & Irena, S. (2021). Leaving no stone unturned: flexible retrieval of idiomatic expressions from a large text corpus. *Machine Learning and Knowledge Extraction*, *3*(1), 263-283. https://doi.org/10.3390/make3010013

[5] Liu, S. & Liu, P. (2013). University, Yoshiyori Urano. A study of composition/correction system with corpus retrieval function. *International journal of distance education technologies: An official publication of the Information Resources Management Association: IJDET*, *11*(3), 58-78. https://doi.org/10.1109/itime.2012.6291287

[6] Gunjan,C., Anil, A., & Gaurav, D. (2023). An approach for document retrieval using cluster-based inverted indexing. *Journal of Information Science*, *49*(3), 726-739. https://doi.org/10.1177/01655515211018401

[7] Liu, L., Feng, Y., Gao, S., & Shu, J. (2022). Link quality estimation based on over-sampling and weighted random forest. *Computer Science and Information Systems*, *19*(1), 25-45. https://doi.org/10.2298/CSIS201218041L

[8] Ding, K. M., Chen, S. P., Yu, J. M., Liu, Y. N., &Zhu, J. (2022). A new subject-sensitive hashing algorithm based on multires-RCF for block chains of HRRS images. *Algorithms*, *15*(6), 213-213. https://doi.org/10.3390/a15060213

[9] Huang, Z. Q., Tang, Z. J., Zhang, X. Q., Ruan, L. L., & Zhang, X. P. (2023). Perceptual image hashing with locality preserving projection for copy detection. *IEEE Transactions on Dependable and Secure Computing*, *20*(1), 463-477. https://doi.org/10.1109/tdsc.2021.3136163

[10] Wang, X. Y. & Zeng, X. H. (2023). Deep consistency-preserving hash auto-encoders for neuroimage cross-modal retrieval. *Scientific reports*, *13*(1), 2316-2316. https://doi.org/10.1038/s41598-023-29320-6

[11] Huang, Y. B., Chen, T. F., Zhang, Q. Y., Zhang, Y., & Yan, S. H. (2022). Encrypted speech perceptual hashing authentication algorithm based on improved 2D-Henon encryption and harmonic product spectrum. *Multimedia Tools and Applications*, *81*(18), 25829-25852. https://doi.org/10.1007/s11042-022-12746-x

[12] Xie, Y. Z., Wang, Y. T., Wei, R. K., Liu, Y., Zhou, K., & Fan, L. S. (2023). A hash centroid construction method with Swin transformer for multi-label image retrieval. *Neural Computing and Applications*, *35*(15), 10891-10907. https://doi.org/10.1007/s00521-023-08273-x

[13]Birouk, W., Lahoulou, A., Melit, A., & Bouridane, A. (2023). Robust perceptual fingerprint image hashing: a comparative study. *International Journal of Biometrics*, *15*(1), 59-77. https://doi.org/10.1504/ijbm.2023.10051692

[14] Qin, C., Liu, E. L., Feng, G. R., &Zhang, X. P. (2021). Perceptual image hashing for content authentication based on convolutional neural network with multiple constraints. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(11), 4523-4537. https://doi.org/10.1109/TCSVT.2020.3047142

[15] Ding, K. M., Chen, S. P., Zeng, Y., Wang, Y. Y., & Yan, X. Y. (2023). Transformer-based subject-sensitive hashing for

integrity authentication of high-resolution remote sensing (HRRS) images. *Applied Sciences*, *13*(3), 1815-1815. https://doi.org/10.3390/app13031815

[16] Kim, K., Lee, J., Lim, H., Oh, S. W., & Han, Y. (2022). Deep RNN-Based Network Traffic Classification Scheme in Edge Computing System. *Computer Science and Information Systems*, *19*(1), 165-184. https://doi.org/10.2298/CSIS200424038K

[17] Chen, X. Y., Wan, M. H., Zheng, H., Xu, C., Sun, C. L., & Fan, Z. Z. (2022). A New Bilinear Supervised Neighborhood Discrete Discriminant Hashing. *Mathematics*, *10*(12), 2110-2110. https://doi.org/10.3390/math10122110

[18] Fang, Z. Q., Lin, X. F., Lin, Y. H., Gao, J. M., Gong, L., Lin, R. J., Pan, G. Y., Wu, J. Y., Lin, W. J., Chen, X. D., & Yi, G. B. (2022). Self-erasable dynamic surface patterns via controllable elastic modulus boosting multi-encoded and tamper-proof information storage. *Nano Research*, *16*(1), 634-644. https://doi.org/10.1007/s12274-022-4958-2

[19] Ruizan, M. & Roman, S. (2022). Research and Research in Practice" Teaching within Higher Education: Graduates, Employers, and Higher Education Face-to-Face to the Competitive Job Market Requirements. *Journal of Service, Innovation and Sustainable Development*, *3*(2), 39-50. https://doi.org/10.33168/SISD.2022.0203

[20] Sana, M., Suzan, K., Ola, A. Q., & Serene, D. (2022). Leadership Behavioral Integrity and Trust on the Employees' Organizational Trust: Examination of the Syrian Private Health Sector. *Journal of Service, Innovation and Sustainable Development*, *3*(1), 67-82. https://doi.org/10.33168/SISD.2022.0106

[21] Long, G. L., Shi, L., Xin, G., Gao, S., Zhang, W., & Xu, J. (2023). Machine-Vision-based online self-optimizing control system for line marking machines. Studies in Informatics and Control, 32(2), 93-104. https://doi.org/10.24846/v32i2y202309

Contact information:

Die HU
School of Foreign Languages,
Southwest Medical University,
No. 1, Section 1, Xianglin Road, Longmatan District,
Luzhou City, Sichuan Province, Luzhou, 646000, China
E-mail: hd2023@swmu.edu.cn

Weili HU
(Corresponding author)
School of Humanities and Management,
Southwest Medical University,Institute of Education, Xiamen University
No. 1, Section 1, Xianglin Road, Longmatan District,
Luzhou City, Sichuan Province, Luzhou, 646000, China
E-mail: 102139@swmu.edu.cn