Hindawi
Scientific Programming
Volume 2022, Article ID 2847590, 10 pages
https://doi.org/10.1155/2022/2847590

Research Article

Copyright Protection of Literary Works Based on Data Mining Algorithms

Liying Che

Law School of Case Western Reserve University, Cleveland, OH 44106, USA

Correspondence should be addressed to Liying Che; lxc516@case.edu

Received 14 December 2021; Revised 28 December 2021; Accepted 5 January 2022; Published 28 January 2022

Academic Editor: Tongguang Ni

In order to improve the copyright protection effect of literary works and improve the healthy dissemination of digitized literary works, this paper combines data mining technology to conduct research on the copyright protection of literary works and constructs a literary copyright protection system. In digital literary works, watermarking algorithms can be used to watermark the characteristics of literary works to obtain digital literary works that have been watermarked. After that, this paper can combine data mining algorithms to perform text feature recognition and feature classification and improve the copyright protection effect of literary works. The experimental research results verify that the effect of the copyright protection system of literary works based on data mining algorithms is very good.

1. Introduction

The rapid development of computer storage technology and network technology has brought massive amounts of information to people. This information usually takes images, videos, audios, animations [1], and texts as the main manifestations, among which texts have the widest range of dissemination and the highest frequency of use. The massive dissemination of information brings convenience to people's work and life, but it also has shortcomings, such as many copyright disputes and illegal copying problems, which urgently needs author identification methods that can resolve copyright disputes. Through research, it is found that texts written by different authors or authors have greater style differences, and different texts written by the same author have the same writing techniques, usual sentence structure, vocabulary, etc. [2]. The author recognition method first extracts and counts the features of a large number of texts written by different authors and trains the classifier. Then, for the controversial text, it uses effective feature extraction methods to obtain statistical vectors and input them into the trained classifier. Finally, it outputs specific classification categories or specific authors. The method of text author recognition can assist in resolving copyright disputes of disputed works (especially disputed works of well-known authors), combating piracy, and maintaining integrity. The key part of the text author recognition method is training and building a classifier [3].

Classification is a typical machine learning method with teachers, and it is also an important research topic in the field of data mining. The classification function or classifier is obtained by continuously learning training data. When classification is needed, the test data can use the obtained function or classifier to output a given category. How to choose a suitable classification model in the application is an important issue. Text classification technology can be widely used in fields such as natural language processing and understanding, information management, data evaluation, and information filtering. The more common text classification methods include support vector machine, K-nearest neighbor method, Bayesian classification, neural network, and decision tree classification. Support vector machine is mainly used in pattern recognition and other fields. It is a pattern recognition method based on statistical learning theory. Its characteristic is that it can maximize the geometric edge area and minimize the empirical error at the same time. According to the situation of the known samples, the nearest neighbor algorithm can determine whether the

new sample and the known sample are in the same category. The nearest neighbor algorithm has many developments and improvements, but the general idea is to store all or part of the training samples first and then calculate the distance between the test sample and the training sample through the similar function and finally determine the type of the test sample. The nearest neighbor algorithm can quickly achieve classification, especially in the field of statistical-based pattern recognition. The principle of the neural network is to simulate the structure of the human brain and treat the sample as a connected input/output unit. The training sample learns by adjusting the unit value.

Based on this, this paper combines data mining technology to conduct research on the copyright protection of literary works, constructs a literary copyright protection system, and improves the copyright protection effect of modern digital literary works.

2. Related Work

Literature [4] proposed Triangle Similarity Quadruple (TSQ) and Tetrahedral Volume Ratio (TVR). The TSQ algorithm constructs the Macro Embedding Primitive (MEP) and selects the ratio of the side length of the triangle or the ratio of the base to the height in the MEP as the watermark embedding primitive: the TvR algorithm selects the four sides after constructing the tetrahedral sequence. The volume ratio between the volumes is used as the watermark embedding primitive. Literature [5] calculates the distance from each vertex of the model to the center of the vertex field and the distance from the center of the model and embeds the watermark by modifying the ratio between the two. This algorithm is a non-blind watermarking algorithm, which can resist similar transformation, noise, simplification, and their joint attacks. However, the transparency of the watermark is insufficient.

Literature [6] proposed two digital watermarking algorithms based on local distance: Vertex Flood Algorithm (VFA) and Triangle Flood Algorithm (TFA). The VFA algorithm divides the vertex set according to the distance from the vertex of the model to the center of the selected triangle and embeds the watermark by modifying the distance from the vertex in each set to the center of the selected triangle; the TFA algorithm continuously selects the triangle and connects the adjacent triangles of the triangle, sorting into a triangle traversal sequence according to the distance from the non-shared vertex to the shared edge, and then modifying the height of each triangle in the traversal sequence to achieve the purpose of embedding the watermark. Literature [7] embeds the watermark by modifying the distance from the model vertex to the center of the model. As a global geometric feature, this distance can well reflect the shape of the 3D model and can maintain sufficient stability without changing the visual effect of the model. Therefore, the algorithm has better robustness against noise and simplification attacks; literature [8] improves the transparency of the watermark by controlling the intensity of local watermark embedding, and uses a weighting method to improve the simplification and reduction of the watermark during watermark extraction. Robustness of noise attacks: literature [9] embeds both robust and fragile watermarks in the 3D model by modifying this distance and uses the method of adding weights to improve the robustness of the algorithm when extracting the watermark. Literature [10] proposed a multiple digital watermarking algorithm. This algorithm uses the distance from the vertex to the center of the model to embed the watermark and at the same time introduces the affine invariant range and embeds the second watermark by modifying the vertex order of the triangle face. The complementary advantages of the two watermarks increase the types of algorithms against attacks. Literature [11] focuses on improving the transparency of watermarking. Literature [12] improves the method of controlling the embedding strength of local watermarks. Literature [13] uses the K-means clustering method to select a specific set of vertices according to the curvature of the vertices and uses genetic algorithms to embed the watermark.

Literature [14] proposed a digital watermarking algorithm based on Extended Gauss Image (EGI). The algorithm builds a set of triangle faces based on the normal vector of the triangle face and embeds the watermark by modifying the statistical feature of the mean value of the normal vector of each set. Literature [15] divides the vertices of the 3D model into 6 regions, and each region establishes an extended Gaussian image of the normal vector, which realizes the repeated embedding of watermark information in each region and optimizes the method of modifying the vertex coordinates. Literature [16] proposed a digital watermarking algorithm based on complex extended Gaussian image (Copmlex EGI), which establishes a complex weight for each partition and selects the partition with larger weight to embed the watermark, which effectively improves the robustness. Literature [17] uses the vertex neighborhood of each vertex to calculate an average vector and embeds the watermark by modifying the length of the average vector. The algorithm can handle polygonal mesh models with arbitrary topologies and has good robustness to affine transformations, but it cannot resist attacks such as mesh reconstruction and mesh simplification. Literature [18] uses the model center and principal component analysis method to transform the model into an affine invariant space and transforms the vertex coordinates into spherical coordinates and then constructs a histogram reflecting the value distribution of the radial component of the vertex according to the spherical coordinates. The histogram moderately changes the distribution of the radial component to embed the watermark. The algorithm can resist similar transformation and simplification attacks, but it cannot resist shearing attacks, and it has weak resistance to noise attacks. Literature [19] defines the distance from the vertex of the 3D model to the center of the model as the vertex norm and proposes a highly robust blind watermarking algorithm based on the statistical characteristics of the vertex norm. This algorithm establishes a histogram of all vertex norms, divides the histogram into several partitions according to the number of watermarks, and embeds the watermark by slightly changing the mean or variance of the vertex norm of each partition. This algorithm combines the stability of both

the global geometric features and statistical features of the 3D model and has achieved good robustness against various common attacks. However, the algorithm depends on the center position of the model, so it cannot resist shearing attacks. And there are also shortcomings in transparency.

3. Literary Works Watermarking Algorithm Based on Text Data Mining

By analyzing the characteristics of common BIM model format DXF files, this paper combines the existing two-dimensional vector graphics digital watermarking algorithm to propose a digital watermarking algorithm for data copyright protection based on the BIM model. This paper selects the vertex coordinates of the multiface mesh of the entity of the BIM model data to embed the watermark. In order to solve the problem that the vertex coordinates in the BIM model have more identical values and less effective carriers used to embed the watermark in practical applications, random noise is added to the original coordinate data within the error tolerance to increase the embedding capacity of the watermark. In order to enhance the ability to resist pruning attacks, the watermark information needs to be embedded as evenly as possible in the X and Y coordinates of all multiface mesh vertices of the BIM model data. In order to maintain the synchronization relationship between data and watermark and realize blind watermark detection, the idea of coordinate mapping is adopted. At the same time, the security of the watermark is improved by Logistic scrambling of the watermark image. In this algorithm, firstly, it extracts the vertex coordinates of all the multiface meshes in the data to construct a vertex set and obtains the high-level part of the coordinate data. After that, it establishes a mapping relationship with the watermark through a one-way mapping function to use the low-order part of the coordinate value as the embedding carrier of the watermark and embeds the watermark into the vertex coordinate position using the quantization modulation method. Moreover, it selects the initial value of chaotic transformation as the key for watermark extraction. When the watermark is extracted, no original data is needed, and blind detection is realized. The embedding process of the watermark is shown in Figure 1.

Logistic mapping, also known as insect mouth model, is a typical chaotic sequence in chaos theory, and its equation form is formula (1). Chaos phenomenon is a random-like process that appears in a deterministic system. The process is bounded, non-convergent, and sensitive to initial values. The use of chaotic sequences to encrypt the watermark not only is simple and easy to use, but also has no periodicity and is difficult to crack, which can improve the security of the watermark. For an image of $M \times N$ size, a one-dimensional chaotic encryption sequence w is obtained after $M \times N$ iterations.

$$L_{i+1} = \mu L_i (1 - L_i), (i = 0, 1, 2, \ldots, m \times n). \quad (1)$$

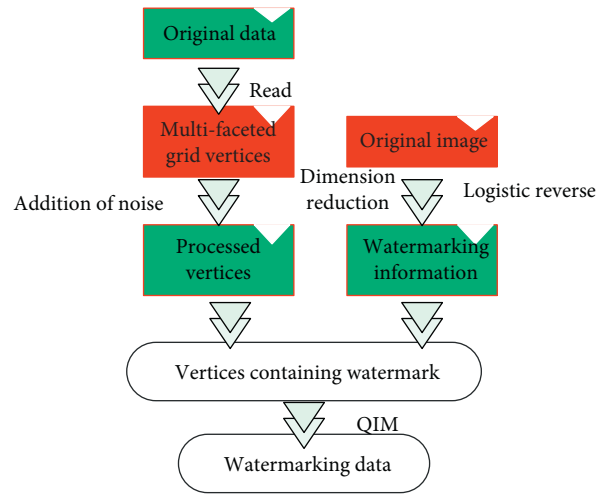

Figure 1: Flow chart of watermark embedding.

When the condition $0 < L_i < 1, 3.5699456 \ldots < \mu \leq 4$ is satisfied, the Logistic mapping works in a chaotic state. In particular, when μ is close to 4, the iteratively generated value is a pseudo-random distribution state. This paper uses the Logistic chaotic map to encrypt an image of 32×64 size and then reduces the dimensionality of the generated binary watermark image to obtain a one-dimensional sequence w with a length of $S = w_m \times w_n$. The initial value $L_0 = 0.98$ of the chaotic transformation is selected for many trials. Figure 2(a) is the original image used in the experiment, Figure 2(b) is the chaotic image after scrambling, and Figure 2(c) is the decrypted image after inverse scrambling [20].

Due to the large number of coordinate repeated values in the BIM model, there are fewer effective carriers for embedding the watermark. To solve this problem, this paper adds random noise to the original coordinate data within the error tolerance to increase the embedding capacity of the watermark. The repeated coordinate values in the vertices set of the polyhedral mesh of the original data are subjected to the noise adding operation shown in formula (2) to obtain the processed vertex set V_e [21].

$$\begin{cases} x_e = x + \text{rand} \times Q \\ y_e = y + \text{rand} \times Q \end{cases}. \quad (2)$$

Here, (x_e, y_e) represents the vertex coordinates of the polyhedral mesh after adding noise, (x, y) is the vertex coordinates of the original data, rand is a random function that generates a random number within (0,1), and Q is the allowable range of error.

This algorithm embeds the watermark with the multifaceted mesh vertices of the BIM model data entity as the object. The vertices of the multifaceted mesh of the BIM model data are set V_K, denoted as $V_K = \{V_i\}, V_i = (x_i, y_i), i \in \{0, 1, 2, \ldots, K-1\}$. Among them, V_i represents the vertex of each polyhedral mesh, (x_i, y_i) is the X, Y coordinate value of the vertex, and K represents the number of vertices of the polyhedral mesh.

(a) (b) (c)

Figure 2: Original watermark.

The specific process of watermark embedding is as follows:

Step 1. The algorithm reads the BIM model data, extracts all the multiface mesh vertices in the model object entity, and constructs the multiface mesh vertex set V_K.

Step 2. The algorithm adds noise to the two coordinate values (x_i, y_i) of each vertex V_i in the set V_K and at the same time enlarges it by 10 times, which is denoted as V_i' , $V_i' = (x_i', y_i')$. Among them, V_i' represents each polyhedral mesh vertex after noise processing, and (x_i', y_i') is the two coordinate values after noise is added to the vertex.

Step 3. The algorithm selects the embedded bit x_w of the watermark according to the data accuracy requirements, and the selection method is as in formula (3). Then, the algorithm gradually modifies the vertex coordinates of the multiface mesh according to the mapping relationship between the high part of the data and the watermark bit $w(x_w)$;

$$x_w = \text{floor}\left(\text{mod}\left(\frac{x_i'}{p, S}\right)\right) + 1. \tag{3}$$

Here, floor represents rounding down, the mod function is the modulo operation and returns the remainder after dividing x_i'/p by S, p is the difference between the magnification and the most significant digit after the decimal point, and S represents the length of the watermark, and $p = 1000$ is selected in this paper.

Step 4. The algorithm uses quantization modulation technology to embed the watermark into the processed coordinate value x_i' and calculate the embedded watermark data x_i^w, where the quantization amplitude is R. There are two cases according to the value of the embedded watermark, as follows [22]:

$$\begin{aligned} x_i^w &= x_i' - \frac{R}{2}, \text{ if } w(x_w) = 0 \text{ and } \text{mod}(x_i', R) \geq \frac{R}{2}, \\ x_i^w &= x_i' + \frac{R}{2}, \text{ if } w(x_w) = 1 \text{ and } \text{mod}(x_i', R) < \frac{R}{2}. \end{aligned} \tag{4}$$

In the same way, according to the different embedded watermarks and the QIM method, the watermark is embedded in the y_i' coordinate of the vertex V_i' of the multifaceted mesh.

Step 5. The algorithm reduces the coordinate value (x_i^w, y_i^w) in V_i^w after the watermark is embedded by 10^t times, and merges the unmodified data with it to generate the watermarked BIM model data.

The extraction of watermark is the reverse process of watermark embedding (Figure 3). The specific steps to extract the watermark are as follows:

Step 1. The algorithm reads the BIM model data to be detected, extracts all the vertices of the multifaceted mesh that can be watermarked, and magnifies the vertex coordinates by 10^t times, where the selection of magnification index t is the same as the value of t when the watermark is embedded.

Step 2. According to the mapping relationship established by the one-way mapping function and the watermark, the algorithm finds the position x_w of the watermark.

Step 3. The algorithm performs QIM operation based on the quantized value R when the watermark is embedded, and extracts the value of the watermark bit $w'(x_w)$ by formula (6).

$$\begin{cases} w'(x_w) = w(x_w) - 1, \text{mod}(x_i^w, R) < \dfrac{R}{2} \\ w'(x_w) = w(x_w) + 1, \text{mod}(x_i^w, R) \geq \dfrac{R}{2} \end{cases}. \tag{5}$$

Step 4. In this algorithm, the same watermark is embedded multiple times, and the value of the watermark bit $w'(x_w)$ can be used to determine the value of the extracted watermark information w'':

$$\begin{cases} w'' = 1, w'(x_w) < 1 \\ w'' = 0, w'(x_w) \geq 1 \end{cases}. \tag{6}$$

This shows that when the value of the extracted watermark bit is less than 1, the value of the watermark information is 1; otherwise it is 0.

Step 5. The algorithm performs dimension increase processing on the obtained one-dimensional watermark information w' and inversely scrambles to obtain the watermark image W'.

Step 6. Finally, the watermark similarity is evaluated by calculating the normalized correlation coefficient

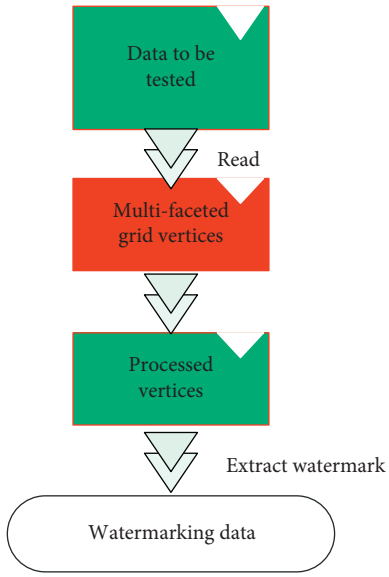

Figure 3: Flow chart of watermark detection.

between the original watermark and the extracted watermark. The calculation formula is as follows:

$$NC = \frac{\sum_{M}^{m=1}\sum_{N}^{n=1}\left[W'(m,n)\times W(m,n)\right]}{\sqrt{\sum_{M}^{m=1}\sum_{N}^{n=1}\left[W'(m,n)\right]^2}\times\sqrt{\sum_{M}^{m=1}\sum_{N}^{n=1}\left[W(m,n)\right]^2}}. \quad (7)$$

Here, NC is a measure of similarity. The greater the value, the greater the similarity. The size of the watermark image is $\times N$, $W(m,n)$ represents the original watermark information, and $W'(m,n)$ is the extracted watermark information.

The BIM model data is a digital expression of the physical function characteristics of the engineering project facility. Based on 3D digital technology, it integrates engineering data model data of various related information of construction projects. The diversity of BIM professional software has led to the diversification of data formats. The format of BIM model data is very important for the selection of hidden domains. The research and development of existing application systems are all based on geometric data models, and data exchange is mainly carried out through graphics information exchange standards such as IGES, DXF, and DWG.

DXF data model is often used for information exchange between AutoCAD and other software. It is mainly composed of graphic objects and non-graphic objects and also contains limited attribute information, which is convenient to operate. For BIM model data in DXF format, the vertices of the multifaceted mesh are an important feature position of the model data. However, the coordinates of the vertices of the multiface mesh in the BIM model data have many repeated values, and there are fewer effective carriers for embedding watermarks. In order to solve this problem, random noise is added to the frequency domain amplitude coefficient after transformation of the original coordinate data within the error tolerance range to increase the watermark embedding capacity. As shown in Figure 4, W1 is the watermark image extracted without any processing on the original data, and the image has serious noise, and W2 is the watermark extracted after the noise preprocessing, and the watermark image is clearly visible.

The algorithm proposed in this paper includes watermark embedding part and watermark extraction part. First, this paper selects the multiface mesh elements in the BIM model data as the unit and constructs a complex number sequence with all the multiface mesh vertices as characteristic points. Moreover, this paper uses the DFT transform to obtain the amplitude coefficient as the embedding carrier of the watermark, uses the QIM method to embed the watermark on the amplitude coefficient of the DFT frequency domain, and then performs IDFT transform to obtain the watermarked BIM model data. When it is attacked, the watermark is extracted, the watermark is extracted through the voting principle, and the correlation method is used to detect. At this time, the original data is not needed, and blind detection is realized. In order to enhance the ability to resist the attack of deleting entities, the watermark information is evenly embedded in the X and Y coordinate transformation coefficients of all multiface mesh vertices in the BIM model data as much as possible. In order to reduce the excessive influence on the original data, the amplitude value is enlarged. In order to maintain the synchronization relationship between data and watermark and realize blind watermark detection, the idea of coordinate mapping is adopted. According to the nature of DFT transformation, in order to avoid the large error caused by the translation attack on the data, the watermark is not embedded on the first transformation coefficient amplitude value of the set of vertices of the multiface mesh. To ensure the security of the watermark, Logistic chaotic mapping is used to scramble the original watermark image. The flowchart of the algorithm is shown in Figure 5.

First, the BIM model data in the space domain needs to be DFT-transformed to the frequency domain. The specific process of the transformation is as follows:

Step 1. $V_d = \{v_j\}, v_j = (x_j, y_j)$ represents the set of all polyhedral mesh vertices in the original BIM model data, where $j = 1, 2, \ldots, N$, v_j is the coordinates of the polyhedral mesh vertices, (x_j, y_j) is the X, Y coordinate value of the vertices, and N is the number of polyhedral mesh vertices. Using multiface mesh elements as the unit, the complex number sequence $\{a_j\}$ is generated as follows:

$$a_j = x_j + iy_j, j \in \{1, 2, 3, \ldots, N\}. \quad (8)$$

Step 2. For the N point sequence $\{a_j\}$, its DFT transformation is shown as follows:

$$A_l = \sum_{j=1}^{N} a_j\left(e^{-2\pi i/N}\right)^{jl} j, l \in \{1, 2, \ldots, N\}. \quad (9)$$

Here, A_l represents the data after DFT transformation. a_j in the formula can be a complex value. In practice, a_j is a real

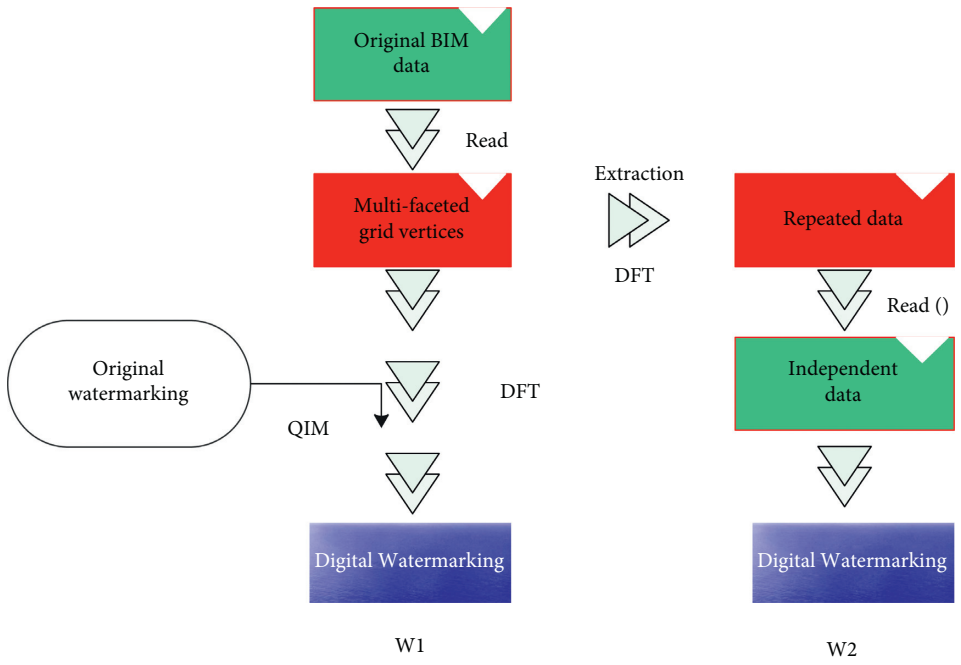

Figure 4: Preprocessing of raw data.

value, that is, the imaginary part is 0. At this time, the formula can be expanded to

$$A_l = \sum_{j=1}^{N} a_j\left(\cos\left(2\pi l\frac{j}{N}\right) - i\times\ \sin\left(\pi l\frac{j}{N}\right)\right), l \in \{1,2,\ldots,N\}. \quad (10)$$

The sequence coefficient has two values, the amplitude coefficient $|A_l|$ and the phase coefficient $\angle A_l$, as shown in formula (11). The set of amplitude coefficients is denoted as $\{|A_l|\}$, and the set of phase coefficients is $\{\angle A_l\}$.

$$\begin{cases} |A_l| = \sum_{j=1}^{N} a_j \cos\ 2\ \pi l\frac{j}{N}, l \in \{1,2,\ldots,N\} \\ \\ \angle A_l = \sum_{j=1}^{N} -a_j \sin \pi l\frac{j}{N}, l \in \{1,2,\ldots,N\} \end{cases}. \quad (11)$$

The specific steps of the watermark generation and embedding algorithm are as follows:

(1) The generation of watermark information. The algorithm reads an image with a size of $M \times M\ (M \geq 2)$ pixels as the original watermark image. In order to improve the security of the watermark, the original watermark is scrambled by Logistic mapping, and the dimensionality of the scrambled binary matrix is reduced to obtain a one-dimensional binary sequence W, where the sequence expression formula is $W = \{w_m = 0, 1 | m = 0, 1, \ldots, P-1\}$, and P represents the length of the watermark.

(2) The algorithm reads the BIM data, the amplitude coefficient $\{|A_i|\}$ obtained by DFT transformation of $\{a_j\}$ is expanded by 10^7, and the noise is added.

(3) The algorithm uses the QIM method to embed the watermark into the amplified amplitude coefficient and obtain the embedded watermark amplitude coefficient $|A_l|$ through the following equation:

$$\begin{cases} |A_l'| = |A_l| - \frac{R}{2}, w_m = 0\ \text{Andmod}(A_l, R) \geq \frac{R}{2} \\ \\ |A_l'| = |A_l|, w_m = 0\ \text{And mod}(A_l, R) < \frac{R}{2} \\ \\ |A_l| = |A_l| + \frac{R}{2}, w_m = 1\text{And mod}(A_l, R) < \frac{R}{2} \\ \\ |A_l'| = |A_l|, w_m = 1\ \text{And mod}(A_l, R) \geq \frac{R}{2} \end{cases}. \quad (12)$$

(4) The algorithm scales the obtained $|A_l|$ to restore it to the original data size, and the reduction factor is equal to the enlargement factor.

(5) The algorithm combines the obtained embedded watermark amplitude value with the unmodified phase coefficient to generate a new coefficient $\{A_l\}$, and then IDFT transforms it to obtain the complex number sequence $\{a_l\}$ after embedding the watermark.

(6) The algorithm modifies the vertices of the multi-faceted mesh according to $\{a_l'\}$ and obtains the set of multifaceted vertices V_d', $V_d' = \{v_j' = (x_j' + y_j')\}$, $j \in \{1,2,\ldots,N\}$, after the watermark is embedded, so as to obtain the BIM data after the watermark is embedded.

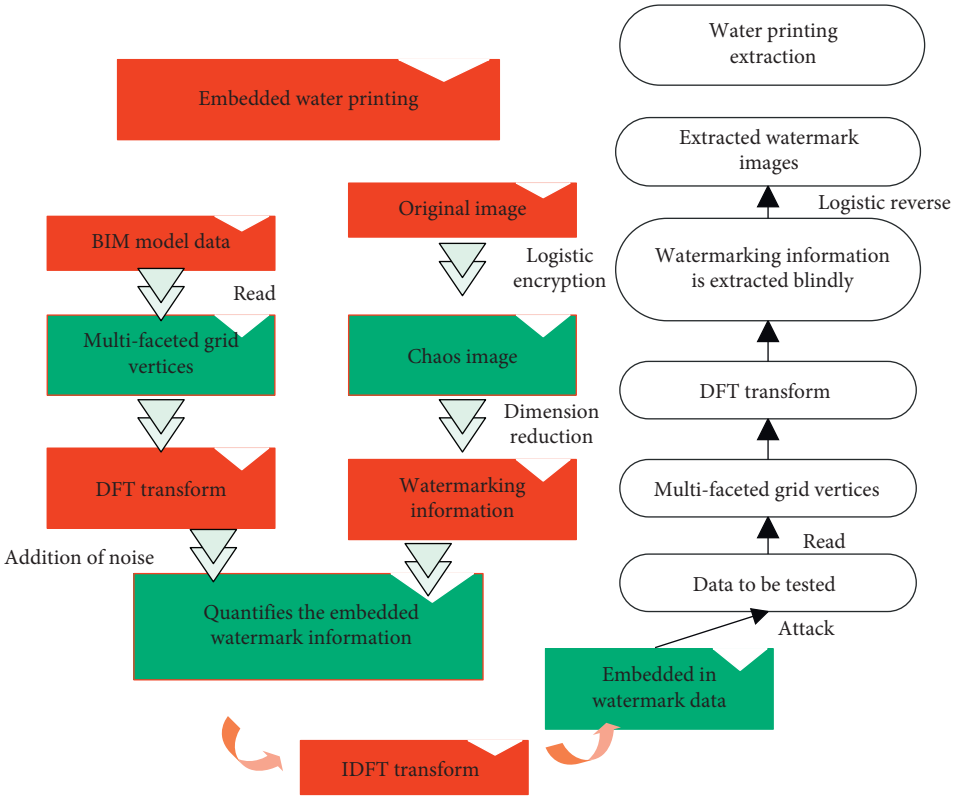

Figure 5: Watermark embedding and extraction framework.

The essence of watermark extraction is the reverse process of watermark embedding. When the data owner finds suspicious BIM model data, the algorithm extracts the watermark according to the following steps:

(1) The algorithm reads the vertices of the multifaceted mesh of the BIM data to be tested, forms a set V_d', and generates a complex number sequence $\{a_j'\}$ according to formula (8).

(2) The algorithm performs DFT transformation on $\{a_j'\}$ to obtain the amplitude coefficient of the coefficient $\{|A_j|\}$.

(3) The algorithm uses the parameters consistent with the embedding process and uses the QIM method to extract the value of suspicious $\{w_m'\}$. The extraction process is as follows:

$$\begin{cases} w_m' = w_m - 1, \mathrm{mod}\left(\left|A_l'\right|, R\right) < \dfrac{R}{2} \\ w_m' = w_m + 1, \mathrm{mod}\left(\left|A_l'\right|, R\right) \geq \dfrac{R}{2} \end{cases}. \tag{13}$$

(4) For the extracted one-dimensional watermark $W' = \{w_m' = 0, 1 | m = 0, 1, \ldots, P-1\}$, the algorithm performs dimensional increase processing and Logistic inverse scrambling to extract the watermark image.

(5) The algorithm uses equation (14) to calculate the normalized correlation coefficient between the extracted watermark image and the original watermark image to measure the robustness. The larger the value of NC, the more similar the two and the better the robustness.

$$NC = \frac{\sum_{m=1}^{M}\sum_{m=1}^{M} XNOR\left(W\left(m1, m2\right), W'\left(m1, m2\right)\right)}{M \times M}. \tag{14}$$

Here, $M \times M$ is the size of the watermark image, XNOR is the exclusive OR operation, $W(m1, m2)$ is the original watermark information, and $W'(m1, m2)$ is the extracted watermark information. Among them, the closer NC is to 1, the more robust the algorithm is.

4. Literary Works Protection Based on Data Mining Algorithm

In digitized literary works, we can use watermarking algorithm to watermark the characteristics of literary works to obtain digital literary works that have been watermarked. After that, we can combine data mining algorithms to perform text feature recognition and feature classification to improve the copyright protection effect of literary works.

Author recognition method mainly includes two modules: training module and classification module. The functions of the training module mainly include the process of preprocessing the original corpus, extracting key features of the text, and training to obtain the classifier. The function of the dispute text classification module is to preprocess the

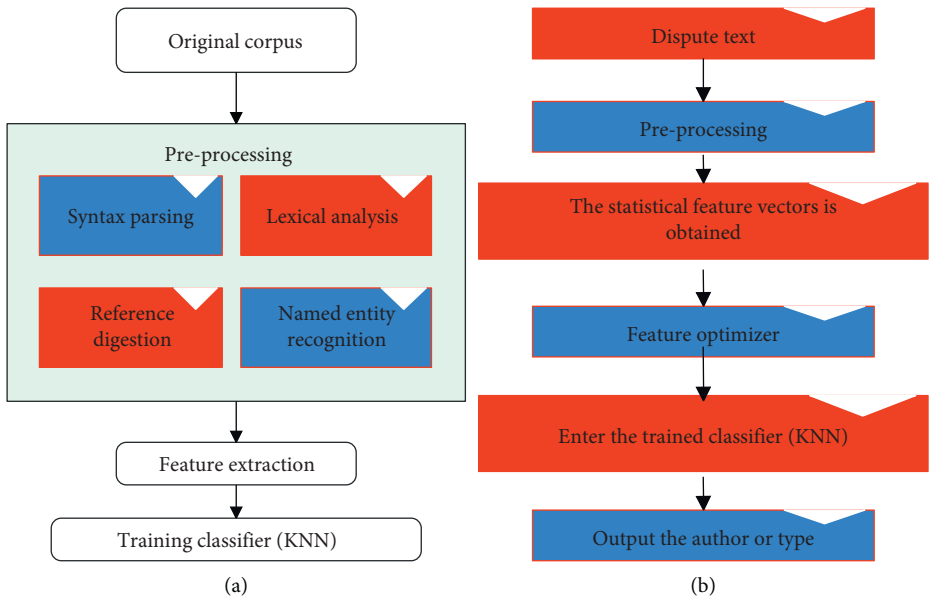

Figure 6: Model function module structure. (a) Training module diagram. (b) Classification module diagram.

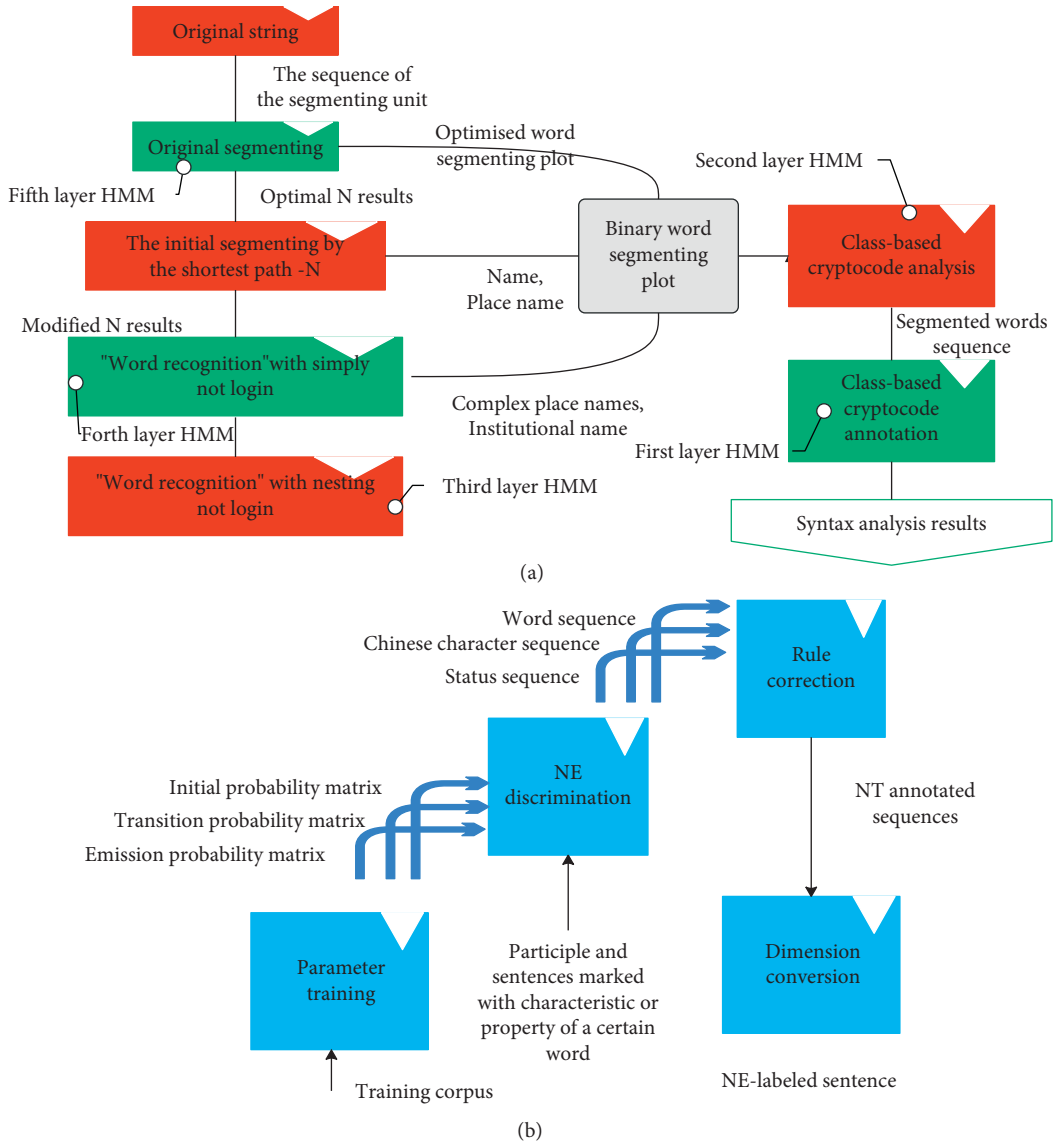

Figure 7: The watermark feature extraction function module of literary works. (a) Word segmentation system structure. (b) Structure diagram of named entity module.

Table 1: The effect of text data mining algorithm in watermarking algorithm feature recognition.

Number	Feature recognition
1	95.183
2	93.852
3	93.653
4	95.726
5	95.049
6	91.902
7	94.023
8	94.907
9	96.732
10	95.449
11	95.724
12	95.523
13	96.018
14	94.768
15	95.937
16	94.456
17	95.630
18	92.033
19	92.519
20	92.548
21	92.969
22	93.626
23	93.265
24	92.590
25	96.462
26	95.158
27	91.251
28	93.333
29	91.036
30	96.041
31	91.626
32	92.056
33	95.413
34	93.730

Table 2: Evaluation of copyright protection effect.

Number	Copyright protection
1	90.335
2	87.370
3	79.406
4	90.760
5	83.105
6	89.303
7	78.790
8	78.987
9	89.762
10	88.496
11	80.300
12	89.035
13	80.199
14	84.078
15	87.036
16	86.804
17	83.225
18	78.005
19	78.938
20	89.517
21	84.643
22	86.895
23	84.018
24	78.786
25	83.654
26	80.734
27	84.245
28	87.999
29	88.636
30	84.305
31	83.862
32	83.266
33	79.180
34	86.801

dispute text, extract the statistical feature vector from the dispute text, and then input it into the trained classifier, and finally output the author category from the classifier. The methods used in the first two stages of these two modules are exactly the same. The main function of the training module is to build a training classifier. If it is a controversial work, then extract the key statistical features from it and input it into the trained classifier, and finally judge the author's category based on the similarity value. The flowcharts of the training module and the classification module are shown in Figures 6(a) and 6(b), respectively.

The corpus must first undergo text normalization processing, and after it is expressed in a form that can be processed by the computer, the normalized text segmentation is processed. The system structure is shown in Figure 7(a). The named entity refers to the actual content of the entity expressed in the Chinese text sentence, such as unit name, person, geographic name, organization name, etc. One of the basic tasks in natural language processing technology is named entity recognition, which plays an important role in word segmentation, syntactic analysis, and automatic translation with the help of machines and other technologies. At present, the lexical analysis technology researched by the Chinese Academy of Sciences and Harbin Institute of Technology has a module for Chinese text sentence named entity recognition. The principle of this module is shown in Figure 7(b).

After combining the watermarking algorithm to obtain the above model, this paper conducts experimental verification on the model. First, the effect of the text data mining algorithm in the feature recognition of the watermarking algorithm is verified, and the results shown in Table 1 are obtained.

The above verifies that the text data mining algorithm has a very good effect in the feature recognition of the watermark algorithm. On this basis, the copyright protection effect is evaluated. This part is carried out by the expert evaluation method, and the results are shown in Table 2.

The above research has verified that the copyright protection effect of literary works based on data mining algorithms is very good.

5. Conclusion

While the digitization of literary works brings a new production and lifestyle to people, its own characteristics have brought a copyright crisis to itself. When digital products

exist in digital form, they can be easily edited, modified, and stored through computers or other digital equipment. At the same time, it can also carry out low-cost and lossless copying and transmission through various forms of storage media, computer networks, or other data transmission methods. The advantages of these original digital literary works make it very easy to illegally occupy, copy, edit, and disseminate unauthorized products that infringe on the owner's copyright. This paper combines data mining technology to study the copyright protection of literary works, constructs a literary copyright protection system, and improves the copyright protection effect of modern digital literary works. The experimental research results verify that the effect of the copyright protection system of literary works based on data mining algorithms is very good.

Data Availability

The labeled dataset used to support the findings of this study is available from the author upon request.

Conflicts of Interest

The author declares no conflicts of interest.

Acknowledgments

This study was sponsored by Law School of Case Western Reserve University.

References

[1] U. F. Ugwu, "Reconciling the right to learn with copyright protection in the digital age: limitations of contemporary copyright treaties," *Law and Development Review*, vol. 12, no. 1, pp. 41–77, 2019.

[2] C. Anfray, B. Arnold, M. Martin et al., "Reflection paper on copyright, patient-reported outcome instruments and their translations," *Health and Quality of Life Outcomes*, vol. 16, no. 1, pp. 224–226, 2018.

[3] E. J. Tao, "A picture2019s worth: the future of copyright protection of user-generated images on social media," *Indiana Journal of Global Legal Studies*, vol. 24, no. 2, pp. 617–636, 2017.

[4] P. Devarapalli, "Machine learning to machine owning: redefining the copyright ownership from the perspective of Australian, US, UK and EU law," *European Intellectual Property Review*, vol. 40, no. 11, pp. 722–728, 2018.

[5] H. B. Essel, R. B. Lamptey, and K. O. Asiamah, "Awareness of law students of kwame nkrumah university of science and technology (KNUST) on copyright law: emphasis on photocopying and fair use," *All Nations University Journal of Applied Thought*, vol. 6, no. 2, pp. 71–87, 2019.

[6] B. Bodó, D. Gervais, and J. P. Quintais, "Blockchain and smart contracts: the missing link in copyright licensing?" *International Journal of Law and Info Technology*, vol. 26, no. 4, pp. 311–336, 2018.

[7] T. He, "The sentimental fools and the fictitious authors: rethinking the copyright issues of AI-generated contents in China," *Asia Pacific Law Review*, vol. 27, no. 2, pp. 218–238, 2019.

[8] C. S. Myers, "Plagiarism and copyright: best practices for classroom education," *College & Undergraduate Libraries*, vol. 25, no. 1, pp. 91–99, 2018.

[9] J. H. Rooksby and C. S. Hayter, "Copyrights in higher education: motivating a research agenda," *The Journal of Technology Transfer*, vol. 44, no. 1, pp. 250–263, 2019.

[10] N. H. Sharfina, H. Paserangi, F. P. Rasyid, and M. I. K. Fuady, "Copyright issues on the prank video on the youtube," in *Proceedings of the International Conference on Environmental and Energy Policy (ICEEP 2021)*, pp. 90–97, Sanya, China, October 2021.

[11] K. Nekit, H. Ulianova, and D. Kolodi, "Website as an object of legal protection by Ukrainian legislation," *Amazonia Investiga*, vol. 8, no. 21, pp. 222–230, 2019.

[12] V. Lunyachek and N. Ruban, "Managing intellectual property rights protection in the system of comprehensive seconday education," *Public Policy and Administration*, vol. 17, no. 1, pp. 114–125, 2018.

[13] S. Schroff, "An alternative universe? Authors as copyright owners- the case of the Japanese manga Industry," *Creative Industries Journal*, vol. 12, no. 1, pp. 125–150, 2019.

[14] M. Finck and V. Moscon, "Copyright law on blockchains: between new forms of rights administration and digital rights management 2.0," *IIC—International Review of Intellectual Property and Competition Law*, vol. 50, no. 1, pp. 77–108, 2019.

[15] N. K. S. Dharmawan, "Protecting traditional Balinese weaving trough copyright law: is it appropriate?" *Diponegoro Law Review*, vol. 2, no. 1, pp. 57–84, 2017.

[16] J. P. McSherry, "The labor of literature: democracy and literary culture in modern Chile by jane D griffin," *Journal of Global South Studies*, vol. 35, no. 2, pp. 448–450, 2018.

[17] G. Carugno, "How to protect traditional folk music? some reflections upon traditional knowledge and copyright law," *International Journal for the Semiotics of Law—Revue internationale de Sémiotique juridique*, vol. 31, no. 2, pp. 261–274, 2018.

[18] R. Matulionyte, "Empowering authors via fairer copyright contract law," *The University of New South Wales Law Journal*, vol. 42, no. 2, pp. 681–718, 2019.

[19] E. Hudson, "The pastiche exception in copyright law: a case of mashed-up drafting?" *Intellectual Property Quarterly*, vol. 2017, no. 4, pp. 346–368, 2017.

[20] M. Sag, "The new legal landscape for text mining and machine learning," *Journal of the Copyright Society of the U.S.A*, vol. 66, no. 2, pp. 291–367, 2019.

[21] S. Geiregat, "Digital exhaustion of copyright after CJEU judgment in ranks and vasiļevičs," *Computer Law & Security Report*, vol. 33, no. 4, pp. 521–540, 2017.

[22] W. Slauter, "Introduction: copying and copyright, publishing practice and the law," *Victorian Periodicals Review*, vol. 51, no. 4, pp. 583–596, 2018.