Journal on Big Data
DOI:10.32604/jbd.2020.010590
Article

Research on Copyright Protection Method of Material Genome Engineering Data Based on Zero-Watermarking

Lulu Cui[2,3,*] and Yabin Xu[1,2,3]

[1]Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing, 100101, China
[2]Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing Information Science and Technology University, Beijing, 100101, China
[3]School of Computer, Beijing Information Science & Technology University, Beijing, 100101, China
[*]Corresponding Author: Lulu Cui. Email: 18239926741@163.com
Received: 01 April 2020; Accepted: 10 August 2020

Abstract: In order to effectively solve the problem of copyright protection of materials genome engineering data, this paper proposes a method for copyright protection of materials genome engineering data based on zero-watermarking technology. First, the important attribute values are selected from the materials genome engineering database; then, use the method of remainder to group the selected attribute values and extract eigenvalues; then, the eigenvalues sequence is obtained by the majority election method; finally, XOR the sequence with the actual copyright information to obtain the watermarking information and store it in the third-party authentication center. When a copyright dispute requires copyright authentication for the database to be detected. First, the zero-watermarking construction algorithm is used to obtain an eigenvalues sequence; then, this sequence is XORed with the watermarking information stored in the third-party authentication center to obtain copyright information to-be-detected. Finally, the ownership is determined by calculating the similarity between copyright information to-be-detected and copyright information that has practical significance. The experimental result shows that the zero-watermarking method proposed in this paper can effectively resist various common attacks, and can well achieve the copyright protection of material genome engineering database.

Keywords: Material genome engineering; copyright protection; zero-watermarking; majority voting

1 Introduction

Material genome engineering is a subversive frontier technology emerging in the field of international materials in recent years. Its basic idea is to integrate high-throughput computing, high-throughput experiments and material big data technology, accelerate the research and development process of materials from discovery, manufacturing to application through collaborative innovation and reduce costs. Data + Artificial Intelligence is the core of material genome engineering. Material genome engineering will be based on an unprecedented large number of data, fusion of artificial intelligence technology to achieve the design and prediction of advanced materials and processes [1].

For a long time, the field of material research and application has accumulated abundant material data resources. With the advent of the era of cloud computing, big data and artificial intelligence, people begin to integrate and share these data, which give birth to the emergence of material genome engineering database, providing important support and guarantee for the analysis and mining of material data [2–3]. However, it is found that the copyright protection of material genome engineering data is one of the main factors restricting the development of material genome engineering.

Therefore, this paper proposes a zero-watermarking copyright protection method for material genome engineering database. This method is divided into two parts: watermarking construction and watermarking detection.

The construction idea of zero-watermarking is as follows: first, important attribute values are selected and marked according to the characteristics of the material genome engineering database. Next, group the attribute values and extract the eigenvalues in each group. Then, these eigenvalues are formed a sequence of eigenvalues by using the majority voting method, and the watermarking information is obtained through the operation between this eigenvalue sequence and the binary sequence of copyright information. Finally, the watermarking information is stored in the third-party authentication center.

The detection idea of zero-watermarking is as follows: when the copyright authentication of the database to be detected is required. First, the zero-watermarking construction algorithm is used to obtain the eigenvalue sequence. Then, by calculating this sequence and the watermarking information stored in the third-party authentication center to obtain the copyright information to be detected. Finally, the ownership is determined by calculating the similarity between the copyright information to be detected and the copyright information.

The innovations of this paper are as follows:

(1) Aiming at the problem that the existing zero-watermarking method cannot accurately extract eigenvalue after the database was attacked, a method of remainder was proposed to group and sort the attribute values, which can effectively improve the accuracy of the extracted eigenvalue, and then improve the robustness of zero-watermarking algorithm.

(2) In order to solve the problem that the zero-watermarking algorithm is weak in resisting attack, this paper proposes a method to extract the eigenvalue sequence through majority voting, then combine the eigenvalue sequence with copyright information to generate watermarking information, which can reduce the complexity and improve the anti-attack ability of the zero-watermarking algorithm effectively.

2 Research Status at Home and Abroad

The traditional digital copyright protection technology mainly includes cryptography and digital watermarking [4]. Cryptography uses cryptographic principles to encrypt files. Authorized users can only use files after obtaining the decrypted key. Obviously, this method is not conducive to open and share of data. Digital watermarking is an information hiding technology. It can be used to confirm the content creator, buyer, transmit secret information or judge whether the carrier has been tampered with.

Currently digital watermarking methods used for copyright protection are divided into two categories. One is embedded watermarking methods, such as literature [5–10]. This method embeds watermarking information by modifying the unimportant bits of important attribute values in the database. That is, it embeds watermarking information by modifying data items, which makes a contradiction between the digital watermarking technology and the availability of the database.

The other is zero-watermarking method, which solves the contradiction between watermarking robustness and data usage worth in embedded watermarking method. The zero-watermarking method does not modify any host information in the database, but constructs the watermarking by extracting the characteristics of important attributes in the database, and deposits the constructed watermarking in a trusted third party [11].

For the relational database zero-watermarking research, in literature [12] constructed zero-watermarking information by extracting the highest bit. Literature [13] added chaotic sequences to filter data items, and then to chaotically scrambled the watermarking information. Literature [14] generates zero-watermarking information by selecting more important binary bits of data as the carrier and combining them with copyright information of practical significance. Literature [15] groups attributes according to the length of binary to be embedded in copyright information, but this method does not take full advantage of copyright information and grouping, leading to the low robustness of the watermarking.

To sum up, although the methods proposed or adopted in the above literatures can realize the protection of data copyright, the following two problems still exist in the copyright protection of material genome engineering data: (1) Material genetic engineering database is different from ordinary database, it has special structure form and data presentation way. At present, no researchers have carried out copyright protection research on it, let alone put forward targeted watermarking research and design methods; (2) Through the analysis of material genome engineering data, it is found that the material genome engineering data has higher data accuracy and lower redundancy. Direct or indirect modification of material genome engineering data may cause great deviation to experimental results. Therefore, the existing embedded watermarking algorithm cannot be applied in the material genome engineering database. Some zero-watermarking algorithms are generally not robust enough to resist all kinds of attacks.

3 Selection and Marking of Attribute Values

3.1 Attribute Value Selection

For the material genome engineering database, the more important data is not only the important characteristics of the material genome engineering database, but also its main value, which is also the most worth of protection. Therefore, when extracting the characteristics of the material genome engineering data to construct zero-watermarking, the important attribute values in the material genome engineering database should be selected.

Through the analysis of material genome engineering data, the data in chemical composition list is selected as the important attribute value. The data in the chemical composition list is made up of some tuples, each of which is made up of some attributes. Data users who want to use genome engineering data to analyze materials must do so on a tuple basis, that is, they need to know all the attributes in the tuple and the value of each attribute. Otherwise, the data is of no value to the user.

Based on this, when selecting the attribute values in the chemical composition list, we select one attribute value from each tuple for eigenvalue extraction. In order to improve the robustness of watermarking algorithm, it is required to extract as many attribute values as possible and distribute them in the chemical composition table. Through the statistics of the number of attributes in the chemical composition list, we selected the ‘Co’, ‘Cr’, ‘W’, ‘Al’, ‘Ti’, ‘Ta’, ‘Ni’, ‘Mo’, ‘Hf’, ‘B’, ‘C’, ‘Zr’ attribute of the chemical composition list for the selection of attribute values, a total of 8132 attribute values were selected. The distribution of the number of extracted attribute values in each attribute is shown in Fig. 1. It can be seen from the attribute distribution diagram that the number of selected attributes is basically evenly distributed in each selected attributes. Therefore, the selection of attribute values can be basically evenly distributed in the chemical composition table.

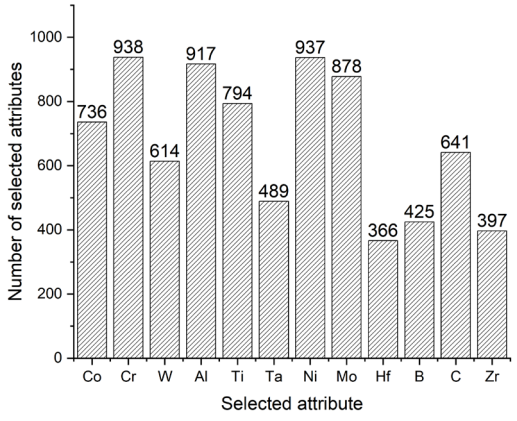

Figure 1: Statistical figure of distribution of attribute values

3.2 Attribute Value Markers

During watermarking detection, it is necessary to find the attribute values used in feature extraction. In order to correctly find the same attribute values after the data is attacked, it is necessary to mark the attribute values, that is, assign a tag similar to the ID number to the attribute values.

Because the position of tuples and attributes in the material genome database can be changed arbitrarily, but it does not affect the use of the material genome database. However, the primary key in each tuple is not allowed to change, if you change the tuple's primary key value, the database becomes unavailable. Therefore, we chose to hash each property value with a primary key value, property name, and property value to uniquely identify each property value.

Common hash algorithms include MD4, MD5, SHA-1 and others. MD5 is a one-way encryption algorithm that converts the input information into a 128-bit fixed-length hash value, which is used to verify the integrity of the data transmission process. Once the data is tampered with, the calculated MD5 value must be different. Therefore, we select the Index = MD5(K, R.P, Zm, Ai) function as the marker function of attribute value. Where, K represents the user'key, R.P represents the primary key of the tuple, Zm represents the name of the property field and Ai represents the value of the property.

In order to clearly show the selection and marking of attribute values and the construction process of zero-watermarking in the material genome engineering database, a small amount of data in the material genome engineering database was selected and the copyright information (W') was assumed to be "110". For example, for the selected user key of "cll", tuple primary key of "195720", property field name of "Cr", property value of "15". Hash function calculation result Index = MD5(cll, 195720, Cr, 15) = "3ba1a44e6df3113b226bd89c67903d14". In turn, according to different attribute values, hash calculation is conducted to finally get different tag values. The result of attribute value selection and marking is shown in Tab. 1.

Table 1: Selection of attribute values and marking results

N	K	R.P	Zm	Ai	Index
1	cll	195720	Cr	15	3ba1a44e6df3113b226bd89c67903d14
2	cll	195719	W	4.0	3ee152568d94551a434c3fec52620f3d
3	cll	195718	Al	4.5	6465fe837bcbd6ac76bb1cbc7b025708
4	cll	195717	Ti	2.5	6d216b02b9dd5ff9f492d5da22070584
5	cll	195716	Ta	2.0	5abaaf94d60552b0ff5da48e378a2874
6	cll	195715	Ni	68.7	06599e3382265aa5273999c4b39011e5
7	cll	195714	Mo	2.0	202349659cca12fcec12e1ade2bb8ecf
8	cll	195712	B	0.01	c5a45c321d27e71dee6cb2ccd96781c0
9	cll	195711	C	0.05	d701d6dc9e65e650e8e64593872004d7
10	cll	195710	Zr	0.15	bd208d7bb1dc29e8ae6390cd51861315

As you can see from Tab. 1, the MD5 algorithm converts different attribute values into a fixed-length hash value, thereby achieving a unique token for the attribute value.

4 Construction and Detection of Zero-Watermarking

4.1 Construction of Zero-Watermarking

The zero-watermarking construction method of material genome engineering database proposed by us is mainly divided into three steps: (1) Group the selected attribute values; (2) Extraction of eigenvalues; (3) The generation of watermarking information.

The method of generating watermarking information through grouping and XOR operation with copyright information can better resist various attacks compared with the existing. The principle is as follows: the attribute values were divided into groups with the same length L as copyright information, and an eigenvalue was taken from each group to form an eigenvalue sequence with the length of L. After that, the watermarking information was obtained by XOR operation with the binary sequence of copyright information. Repeat this step to generate multiple sets of watermarking information and store the results in a third-party authentication center. In case of copyright disputes, multiple sets of copyright information to be detected are obtained through XOR operation between the extracted multiple sets of eigenvalue sequences and multiple sets of watermarking information of a third party, determine the copyright information to be detected through the majority election method. Therefore, the anti-attack ability of watermarking algorithm can be improved effectively.

The process of grouping attribute values is as follows: M attribute values were selected from the chemical composition table of the material genome engineering database, and the marker values of these attribute values were calculated (Index). The attribute values are divided into L groups by using the result of Index mod the length L of the copyright letter. The attribute values in each group are arranged in ascending order according to the result of the mod M/L by Index.

The extraction process of eigenvalues is as follows: Calculate the number and mean of attributes in each group, find the group with the least number of attributes, and assume that the number of attributes of this group is A. Each group takes the first A attribute value and extracts the eigenvalue according to the size relation between attribute value and mean value. The eigenvalue L group with length A is obtained.

Algorithm: Watermarking information generation

Input: material genome database R, copyright information with practical significance W

Output: watermarking information W2;

1：The important attributes in the table of chemical composition list in the material genome database were selected, and a total of M attribute values were selected;

2：Hash marker the selected M attribute values, Index=MD5(K, R.P, Zm, Ai);

3：W’ = str to bit (W);

4：L = length(W’);

5: Group the attribute values into L groups according to the length of L, j=index%L+1(1<=j<=L);

6: Count the length of the number of attributes in each group and find the group with the least number of attributes. The number of attributes in this group is denoted as A;

7: For (i = 1; I <= L; i++)

8: The attribute values in each group j are arranged in ascending order according to the result of the attribute value mod M/L

9： If (A==0):

10： Each group adds a 0 at the end, and adds 1 to the value of A

11: end if;

12: The first A attributes were taken from each group, and the characteristics of these attributes were extracted

13: Eigenvalue extraction rules:

$$M_i^* = \begin{cases} 0, M_i < \text{AveNums} \\ 1, M_i \geq \text{AveNums} \end{cases}$$

14: end for;

15: A total of L groups, each group has A sequence of (0,1);

16: for (i = 1; I <= L; i++)

17: The (0,1) sequence of each group is elected by majority and L_i is obtained

18: end for;

19: Combine all L_i to get an eigenvalue sequence W1 of length L;

20: W2 = W1 ⊕ W’// XOR operation;

21: The watermarking information W2 is stored in ZWMC.

The watermarking information generation algorithm is listed in the above Figure.

The results of grouping and eigenvalue extraction are shown in Tab. 2; the difference between the attribute value and the mean is represented by D; the eigenvalues of each group is represented by G_E; take the first A eigenvalues for each group is represented by T_E.

Table 2: Grouping results and extraction of eigenvalues

G	Index	Ai	D	G_E	T_E
1	3ba1a44e6df3113b226bd89c67903d14	15.0	>=0	1	1
	5abaaf94d60552b0ff5da48e378a2874	2.0	<0	0	0
	202349659cca12fcec12e1ade2bb8ecf	2.0	<0	0	0
	d701d6dc9e65e650e8e64593872004d7	0.05	<0	0	
2	c5a45c321d27e71dee6cb2ccd96781c0	0.01	<0	0	0
	06599e3382265aa5273999c4b39011e5	68.7	>=0	1	1
	bd208d7bb1dc29e8ae6390cd51861315	0.15	<0	0	0
3	6465fe837bcbd6ac76bb1cbc7b025708	4.5	>=0	1	1
	6d216b02b9dd5ff9f492d5da22070584	2.5	<0	0	0
	3ee152568d94551a434c3fec52620f3d	4.0	>=0	1	1

As can be seen from Tab. 2, the group with the least number of attributes has three attribute values, so the first three attribute values of each group are taken for the extraction of eigenvalues. Therefore, a total of 3 eigenvalues with length 3 are obtained, which are G1 = [100], G2 = [010] and G3 = [101]. A majority of the eigenvalues in G1, G2 and G3 were elected to obtain an eigenvalue sequence "001" of length 3, which was then XOR operated with W' "110" to obtain watermarking information "111" and stored in the third-party authentication center ZWMC.

4.2 Detection of Zero-Watermarking

The detection process of zero-watermarking is the inverse process of zero-watermarking construction. The detection of zero-watermarking in material genome engineering database is mainly divided into four steps: (1) Group the selected attribute values; (2) Extraction of eigenvalues; (3) The generation of watermarking information to be detected; (4) Similarity calculation and judgment of copyright ownership.

Since the process of grouping the selected attribute values and the extraction of eigenvalues are the same as the watermarking construction process, only the specific process of generating the watermarking information and calculating the correlation is given below.

The generation process of watermarking information to be detected is as follows: just like the watermarking generation process, a binary sequence of length L is obtained through the majority voting method and this binary sequence is XOR operation with the watermarking information in ZWMC to obtain the copyright information W" to be detected.

The correlation calculation process is as follows: the similarity was measured by Eq. (1) normalization correlation coefficient NC.

$$NC = \frac{1}{H}\sum_{i=1}^{H} b_k \tag{1}$$

where, bk = XNOR (W', W''), W' is the binary sequence of copyright information, W'' is the binary sequence of copyright information to be detected and H is the length of watermarking information.

5 Experiment

At present, the open material genome engineering database mainly includes material genome engineering database and material science data sharing network. In this paper, the chemical composition list of nickel-based super alloy in the materials science data sharing network was selected for the construction and detection of zero-watermarking, with a total of 11352 tuples and 112668 attribute values.

The zero-watermarking algorithm proposed by us and the zero-watermarking algorithm adopted in literature [15] were compared on the same data set material genome engineering database for subset selection, subset increase and subset change attacks, and the normalized correlation coefficient NC was calculated. During the experiment, we selected the copyright information with practical significance as "Beijing Information Science and Technology University".

5.1 Subset Selection Attack Experiment

The subset selection attack means that the data stealer does not use all the data in the relational database, but only some data of attributes or tuples, so that the watermarking information cannot be detected. When carrying out the subset selection attack experiment, we selected different proportions of data and calculated the NC value, as shown in Fig. 2:

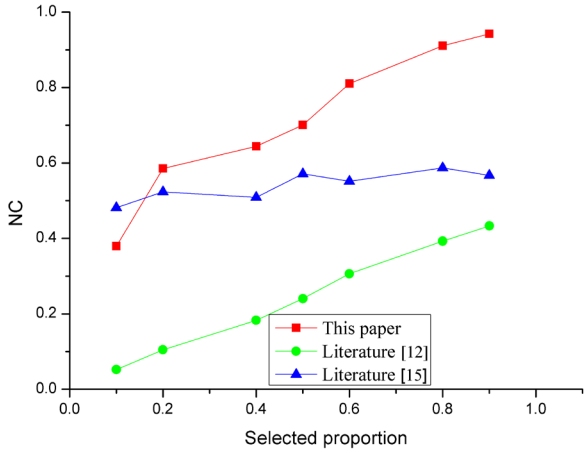

Figure 2: Subset selection attack

As can be seen from Fig. 2, with the increase of selection ratio, the NC value is on the rise. The NC values of this algorithm and Literature [15] are significantly higher than Literature [12]. When the selection ratio reaches above 60%, the NC value of this algorithm is significantly higher than literature [15]. The reasons are as follows: first of all, with the increase of the selection ratio, the data volume keeps increasing, the number of selectable eigenvalues of each group also increases, and the number of generated eigenvalues will increase, at which time the advantages of majority voting will be more obvious. In addition, due to the sorting of the attribute values in each group, the proportion of correctly finding the attribute values used to extract the eigenvalues in the watermarking detection process is also increasing. At this time, the amount of watermarking information correctly extracted will increase as the selection ratio increases. Therefore, compared with literature [12] and literature [15], the NC value of this algorithm is higher.

5.2 Subset Add Attack Experiment

Subset addition attack means the attacker adds some new data to the original relational database to destroy the original watermarking information. When carrying out the subset addition attack experiment, different proportions of data are added. The relationship between the NC values and the selection ratio of attribute values is shown in Fig. 3:

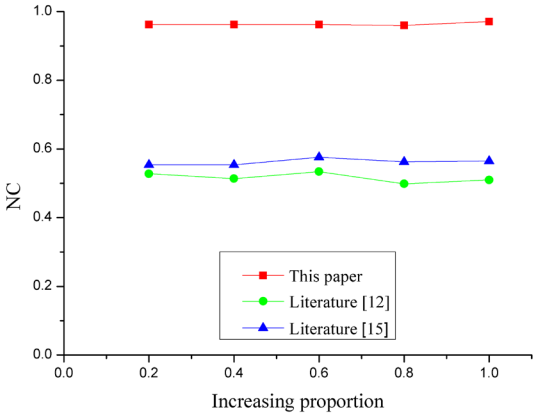

Figure 3: Subsets add attacks

As can be seen from Fig. 3, with the increase of addition ratio, the NC value of the above three watermarking algorithms does not fluctuate significantly, but the NC value of this algorithm is significantly higher than literature [12] and literature [15]. The reasons are as follows: this algorithm uses the remainder method to group attribute values and the sorts the attribute values in each group in ascending order, and only take before A attribute values for the extraction of characteristic value, so even if increase the amount of data, the attribute values used to extract the eigenvalues can still be found correctly in the construction algorithm, and with the increase of the amount of data, majority voting advantage is more obvious, the watermarking information can be correctly extracted, the NC value will be close to 1.

5.3 Subset Change Attack Experiment

The subset change attack is that the data stealer modifies some attribute values in the relational database so that no watermarking can be detected. In the subset change attack experiment, by changing different proportions of data, the relationship between the calculated NC values of the two algorithms and the selection ratio of attribute values is shown in Fig. 4:

It can be seen from Fig. 4 that the above three watermarking algorithms can resist subset change attack very well, but the effect of this algorithm is still better than the algorithm in literature [12] and literature [15]. The reasons are as follows: the subset change attack only changes the size of the attribute value, but does not change the size of the data. Therefore, this algorithm can still extract multiple groups of eigenvalues and use the majority election method to improve the accuracy of the extracted watermarking information. Therefore, the NC value of this algorithm is better than that of literature [12] and literature [15].

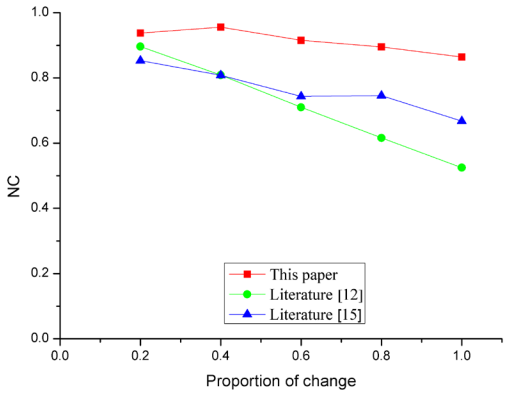

Figure 4: Subset change attack

6 Conclusion

This paper studies the copyright protection of material genome engineering data and proposes a zero-watermarking technology based copyright protection method for material genome engineering data. In the process of watermarking construction, the eigenvalues are extracted by grouping and sorting the attribute values, which effectively improves the accuracy of extracting the eigenvalues. The eigenvalue sequence is obtained through of the majority voting method and this sequence and the copyright information are XOR operation to obtain the watermarking information, which effectively improves the anti-attack ability of the watermarking algorithm. During zero-watermarking detection, the zero-watermarking construction algorithm is used to obtain the eigenvalue sequence, and then this sequence is XOR operation with the watermarking information of a trusted third party to obtain the copyright information to be detected, and the ownership of copyright is determined accordingly.

Experimental results show that the proposed zero-watermarking method in this paper is more robust than the existing zero-watermarking method and can effectively protect the copyright of genome engineering data. The method proposed in this paper can also be applied to other relational databases and has strong universality and portability.

Funding Statement: This work is supported by Foundation of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research No. ICDDXN004, and Foundation of Beijing Advanced Innovation Center for Materials Genome Engineering.

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

References

[1] H. Wang, X. D. Xiang and L. T. Zhang, "Data + AI: The core of materials genomic engineering," *Science and Technology Review*, vol. 36, no. 14, pp. 15–21, 2018.

[2] H. Q. Yin, X. Jiang, R. J. Zhang, G. Q. Liu, Q. J. Zheng *et al.,* "Role of materials data in materials innovation development and thoughts on the existing problems," *Progress in Materials in China,* vol. 36, no. 6, pp. 401–405, 2017.

[3] J. N. Zuo and Y. Chen, "The analysis on the sharing mode of scientific data in the era of big data," *New Century Library*, no. 3, pp. 32–35, 2014.

[4] Y. Q. Shi, F. J. Zhang and C. Ma, "The concept, type and application of digital copyright protection technology in the field of publishing," *Technology and Publishing*, no. 3, pp. 57–59, 2012.

[5] W. Q. Zhang, Research on digital watermarking technology of relational database. Southwest Jiaotong University, 2016.

[6] X. J. Zeng, The research on digital watermarking technology for relational database. Zhejiang University of Technology, 2011.

[7] X. M. Niu, L. Zhao, W. J. Huang and H. Zhang, "Watermarking relational databases for ownership protection," *Electronic Journals,* no. S1, pp. 2050–2053, 2003.

[8] M. Kamran, S. Suhail and M. Farooq, "A robust, distortion minimizing technique for watermarking relational databases using once-for-all usability constraints," *IEEE Transactions on Knowledge & Data Engineering*, no. 12, pp. 2694–2707, 2013.

[9] Y. Y. Xue, "A digital watermarking technology by utilizing effective digit on database field," *Journal of Qinghai University (Natural Science Edition)*, vol. 32, no. 1, pp. 1–4, 2014.

[10] H. Tufail, K. Zafar and A. R. Baig, "Relational database security using digital watermarking and evolutionary techniques," *Computational Intelligence*, vol. 35, no. 4, pp. 693–716, 2019.

[11] Y. J. Meng, C. Wu, W. Zhang and X. J. Zhang, "Research on zero-watermarking registration schema for relation database," *Computer Engineering*, no. 2, pp. 133–135, 2007.

[12] Y. J. Meng and C. Wu, "Construction algorithm of zero watermark in relational database," *Journal of Lanzhou University: Natural Science Edition*, no. 6, pp. 51–55, 2007.

[13] Y. J. Meng, Y. Shi, L. Si, Y. W. Cao and J. Y. Zhang, "Zero-watermarking algorithm for database based on chaotic sequences," *Computer Engineering and Application,* no. 29, pp. 168–170, 2008.

[14] Y. W. Cao, "Copyright protection of relational database based on zero-watermarking," *Gansu Science and Technology*, vol. 26, no. 4, pp. 36–39, 2010.

[15] X. Q. Mao, "A new zero-watermarking algorithm for relational database," *Science and Technology Information*, no. 22, pp. 110–111, 2010.