A NOVEL ROBUST SCHEME OF WATERMARKING DATABASE

Jia-jin Le [1], Qin Zhu [2, 3 *]
1. School of Computer Science and Technology, Donghua University, Shanghai 200051, P. R. China
2. College of Computer Science and Technology, Nantong University, Nantong 226019, P. R. China

Ying Zhu [3]
3 College of Information Science and Technology, Donghua University, Shanghai 200051, P. R. China

Keywords: Relational database, Database watermarking, Robustness, Chaos sequence.

Abstract: A scheme for watermarking relational database is proposed in this paper. It is applied to protect the copyright of numeric data. The chaos binary sequences are generated under the control of the privacy key, and are utilized as the watermark signal and the control signal for watermark embedding. Both the privacy key and the primary key determine the watermarking position, and the watermark is embedded into the numeric data by changing the parity of their low order digits, thus avoids the syndrome phenomena caused by the usual Least Significant Bit (LSB) watermarking scheme. The embedment of the watermark meets the requirement of the synchronous dynamic updating for the database, and the detection of the watermark needs no original database. Both the theoretical analysis and the practical experiments prove that this scheme possesses fine efficiency, imperceptibility and security, and it is robust against common attacks towards the watermark.

1 INTRODUCTION

Database, which runs for decades, is probably the most valuable asset of information systems. For example, the commodity sales database stores large numbers of customer's information and business transaction details. Another example is the med-care information database that holds costly statistics for research. For the secure holes of the networks and the operation systems, together with the feature of being easily copied and redistributed data files, databases could be the aims of those pirates. Therefore, the copyright protection of database has received considerable attention of researchers. In newly born outsourced database system (Hacigumus et al. 2002), data owners delegate their database needs and functionalities to a third party that provides database services to the users. Since the providers are not fully trustable or compromised, it is also important to provide copyright protection for outsourced database (Zhu et al. 2006).

Traditionally, data security of database can be implemented via cryptography, which transforms the meaningful plain text to meaningless cipher text. However, the randomness of cipher text usually exposes the importance of the message, thus causes the new insecurity. Meanwhile, the encryption, decryption and the management of privacy keys are all associated with complicated operation, thus limit the availability of the database. Furthermore, once the cipher text is decrypted, nothing can be done for the data security of the database.

Watermarking database is a new technique for data security of database. An amount of information, which is imperceptible and un-removable except the carrier object is damaged entirely, is embedded into the database through signal processing method to protect the copyright. Although watermarking database is rare applied for real-time prevention from database piracy, it is able to convincingly prove copyright in court, thus constitutes the significant deterrent to the attackers.

Compared with the multimedia data watermarking, it is hard to find distinguishable redundant space for embedding watermarks into database due to the different peculiarities of the database from multimedia works. To protect copyright, database watermark should possess at least four features:

Le J., Zhu Q. and Zhu Y. (2007).
A NOVEL ROBUST SCHEME OF WATERMARKING DATABASE.
In *Proceedings of the Second International Conference on Software and Data Technologies - Volume ISDM/WsEHST/DC*, pages 97-102
DOI: 10.5220/0001336900970102

(1) Security. The scheme of database watermarking should be secure, the search space of privacy key should be large enough, and the management mechanism for privacy keys should be reliable. Generally, the watermarking algorithm is public, and the security of the system mainly depends on the privacy key.

(2) Imperceptibility. As for the legal users of the database, the watermark should not be perceived, and should not influence the availability of the database.

(3) Robustness. The watermark of database should tolerate common attacks towards it, such as subset extracting attack, subset modification attack, subset addition attack, and so on. The watermark should be hard to be erased or be forged, and the probability of false positive and false negative decision should be small enough.

(4) Blind detection. No original database is needed during the watermark detection, and the copyright is judged only requiring the privacy key of the watermarking algorithm.

In practice, both the imperceptibility and the robustness are the rubs of watermarking scheme, since they are usually contradictory, and need to be reasonably traded off.

In this paper, we propose a scheme for relational database watermarking, which is applied to protect the copyright of numeric data. The chaos binary sequences are generated under the control of the privacy key, and are utilized as the watermark signal and the control signal for watermark embedding. The watermark is embedded into the numeric data by changing the parity of their low order digits, thus avoids the syndrome phenomena caused by the usual Least Significant Bit (LSB) watermarking scheme. The embedding watermark meets the requirement of the synchronous dynamic updating for the database, and the detection of the watermark needs no original database.

The rest of the paper is organized as follows: Section 2 surveys the related research of watermarking database. Section 3 circumstantiates the algorithms of the proposed scheme for watermarking database based on chaos-sequence, including generating watermark signal, embedding and detecting watermark. Section 4 analyses the proposed scheme in respects of security, robustness, imperceptibility, and overhead. Section 5 provides an experimental evaluation. Section 6 concludes this paper with summaries and suggestions for future work.

2 RELATED WORK

Agrawal and Kiernan are the pioneers in the field of watermarking database (2002), and they firstly proposed the scheme of embedding watermark into relational database and implemented it. This scheme assumes that numeric attributes can tolerate modifications of some LSB. Tuples are firstly selected for watermark embedding. Then certain bits of some attributes of the selected tuples are modified to embed watermark bits. Therefore, this scheme is also referred as LSB scheme, and it is the most frequently used scheme in watermarking database. However, due to the different precision and word length between different databases, the same algorithm operated in different databases may modify different bits of the same data, thus may result in syndrome phenomena. This becomes an unavoidable drawback of the LSB scheme.

Niu Xiamu et al. precisely defined the constraints of the availability of the database based on the LSB scheme, indexed the tuples which can be marked and divided them into groups, thus embedded multi-bits watermark into the database (2003). This scheme is the extending of the LSB scheme, so it inherits the same drawback of the LSB scheme that may lead to syndrome phenomena.

Sion et al. embedded watermark into database by secretly sorting tuples and dividing subsets (2004). All tuples are divided into non-intersecting subsets, and a single watermark bit is embedded into tuples of a subset by modifying the distribution of tuples values. The same watermark bit is embedded repeatedly across several subsets and the majority-voting algorithm is employed to detect the embedded bits. This scheme involves expensive operation, and the capacity of the watermark is rather limited.

Francesc et al. embedded a watermark into each attribute of a multivariate continuous numerical dataset based on the theory of statistics (2006). Data quality is assured to the extent that the watermarked data nearly preserve the attribute means and the co-variance matrix from the original dataset. The scheme is claimed to be robust against random noise addition attacks. However, due to each watermark embedding based on the statistic value of completely static dataset, this scheme is not suitable for database that needs frequent updating.

3 ALGORITHM

There is redundancy of precision among the low order digits of numeric data in database. For example, 0.1 degree is enough for the precision of air temperature, but there may be two or more decimal fractions stored in the database. Another example is the forest cover type in database, in case of the measure unit being square meter, it hardly does harm to the availability of the database when adding 1 to or subtracting 1 from the unit's order of the data. If the precision of the modified data is limited within curtain extent, the carrier channel for watermark signal is provided. In fact, data owners are usually willing to obtain the ability of asserting ownership at the cost of mini distortion of the data.

3.1 Watermark Signal

Based on its properties of non-repetitive iterative operation and sensitiveness to the initial input, the chaos model is applied to generate the pseudo-randomized series, and is sue for the watermark signal and the control signal for watermark embedding after curtain transform. Even very tiny alteration of the input can have tremendous impact on the output of the chaos system, and the outputs are non-repetitive. Therefore, the chaos model possesses some property of one-way cryptographic Hash function.

Acting as a randomized series generator, the Logistic chaos equation is

$$x_{n+1} = \mu x_n(1 - x_n),\quad \mu \in [1,4],\quad n = 0,1,2,... \tag{1}$$

We set μ = 3.93 in this computation (Yen, 2001), and use the normalized bit conjunction of the privacy key and the primary key for the initial value of the chaotic series generator. The pseudo-randomized bits sequence is generated by transforming the series numbers to 0 or 1 according to whether they are below 0.5.

3.2 Watermark Carrier

In our scheme, the watermark is embedded into the numeric data by changing the parity of their low order digits, thus gets the carrier channel for the watermark signal. What the scheme modifying is the low order digits rather than the least significant bits, so it avoids the syndrome phenomena of the LSB scheme caused by the different types of databases or operation systems.

Definition 1: Suppose that the scheme of a database relation is R (P, A_1 , ... A_j , ... A_v), where P is the primary key attribute, A_j ($1 \le j \le v$) is the attribute of R, r_i ($1 \le i \le n$) is the tuple in R, $r_i.A_{ji}$ is the value of attribute A_j of tuple r_i.

Definition 2: A candidate contribute A_j ($1 \le j \le v$) in R is that which is a numeric attribute, and of which there is redundancy of precision among the low order digits.

For the sake of simple description of the problem, we assume that all v attributes in R are integer numeric candidate contributes in this paper.

Definition 3: A candidate digit candidate contribute A_j ($1 \le j \le v$) is that which is the decimal low order digit of A_j, and in which there is redundancy of precision. The number of candidate digits in candidate contribute A_j is denoted as ξ_j, and the candidate digits sort in descending order as follows: $d_{\xi j}$, $d_{\xi j\ -\ 1}$, ... , d_1 .

According to the above definitions, the available carrier channel for watermark in database relation R(P, A_1 , ... A_j , ... A_v) is the candidate digits d_{kj} ($1 \le k_j \le \xi_j$) of candidate contributes A_j ($1 \le j \le v$).

3.3 Watermark Embedding

Our scheme determines the watermarking position according to both the privacy key and the primary key in a tuple, i.e. a certain candidate digit in a certain candidate attribute, meanwhile, determines whether the watermark to be embedded according to the admissible error. To a determined candidate digit to be marked, the parity of it is modified according to the corresponding bit of both the watermark signal and the embedding control signal.

Definition 4: The admissible error of a candidate attribute A_j ($1 \le j \le v$) is that the error which is within the extent of available distortion of A_j , and it is denoted in percentage terms as δ_j .

Definition 5: Suppose that $r_i.A_j$ ($1 \le i \le n, 1 \le j \le v$) is changed to $r_i.A_j{}'$ after embedding watermark, the relative error is denoted in percentage terms as

$$\sigma_{ij} = \frac{|r_i.A_j{}' - r_i.A_j|}{|r_i.A_j|} \times 100\% \ . \tag{2}$$

Rule 1: The necessary condition for selecting $r_i.A_j$($1 \le i \le n, 1 \le j \le v$) as watermark carrier is that there is candidate digit in $r_i.A_j$, and $\sigma_{ij} \le \delta_j$.

Rule 2: Suppose that the length of watermark is p, the total number of database relation R is n, and the interval number between two adjacent marked tuples is γ, then the necessary condition for repetitively embedding watermark into R is that $p < n / \gamma$.

Algorithm 1: Watermark embedding

(1) Input two privacy keys *key*1, *key*2; normalize them to be the initial value of chaos equation (1);

(2) Generate two chaotic binary sequence $c1[p]$, $c2[p]$ according to the algorithm described in section 3.1;

(3) Input the value of gap γ ;

(4) Input array ξ [v], which contains numbers of each candidate digit in each candidate attribute;

(5) Input array δ [v], which contains admissible errors of each candidate attribute;

```
(6) for $i$ = 1 to $n$
  (6.1) if LGS (NRM ($key$1o $P_i$)) mod $\gamma = 0$, then
    (6.1.1) $k$ = NXT ( LGS (NRM (k$ey$1o $P_i$))) mod $p$ +1;
    (6.1.2) $j$ = NXT ( LGS (NRM (k$ey$1o $P_i$))) mod $v$ +1;
    (6.1.3) $d$ = NXT ( LGS (NRM (k$ey$1o $P_i$))) mod $\xi$ ($j$) +1;
    (6.1.4) if $r_i.A_j * \delta ( j ) \geq 10^{d-1}$ ,then
      (6.1.4.1) if ($c1$ ($k$), $c2$ ($k$)) = (0, 1) AND ($value$ ($d$) mod 2 ) = 1, then $value$ ($d$) = $value$ ($d$) +1;
      (6.1.4.2) if ($c1$ ($k$), $c2$ ($k$)) = (0, 0) AND ($value$ ($d$) mod 2 ) = 1, then $value$ ($d$) = $value$ ($d$) -1;
      (6.1.4.3) if ($c1$ ($k$), $c2$ ($k$)) = (1, 1) AND ($value$ ($d$) mod 2 ) = 0, then $value$ ($d$) = $value$ ($d$) +1;
      (6.1.4.4) if ($c1$ ($k$), $c2$ ($k$)) = (1, 0) AND ($value$ ($d$) mod 2 ) = 0, then $value$ ($d$) = $value$ ($d$) -1;
```

In the embedding algorithm, function LGS ( ) generates chaotic randomized series based on the Logistic chaos equation, and processes the series numbers as integers by omitting the decimal points. Function NXT (LGS ( )) takes next number of the chaotic series. (key1o P_i) denotes the bit catenation operation of the privacy key and the primary key. Function NRM( ) is a normalization operation.

3.4 Watermark Detection

The process of watermark detection is similar to the embedding algorithm. First, the chaos binary sequences are generated under the control of the privacy key. Then, to each tuple, the detection position is determined by both the privacy key and the primary key, and the values of the candidate digits are read and then connected to form a bit sequence according to their parity. Since the same bit in the chaos sequence may be repeatedly embedded, the algorithm of majority voting is operated to get a binary sequence finally (Sion et al. 2004).

A normalized correlation detector is deployed to compute the similarity between the detected binary sequence and the initial watermark sequence.

Definition 6: Suppose that the watermark sequence is $c[p]$, and the detected sequence is $c'[p]$, then the normalized correlation of the two sequence is denoted as

$$Z_{nc}(c,c') = \frac{\sum_{i=1}^{p} c(i)\cdot c'(i)}{\sqrt{\sum_{i=1}^{p} c^2(i)\cdot \sum_{i=1}^{p} c'^2(i)}} \,. \qquad (3)$$

If the value of the normalized correlation is above the threshold that is set in advance based on statistical analysis, it indicates that the database is watermarked by the relative chaos sequence.

4 ANALYSIS

The probability of that k odd (or even) numbers to be detected in n candidate digits satisfies the binomial distribution:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, k = 0,1,2,...,n \ \cdot \qquad (4)$$

Before the database is watermarked, the probability of a digit being odd (or even) is 0.5, and the probability distribution of watermark match rate (k/n) reaches its maximum in 0.5. The substance of watermarking database is changing the probability distribution of watermark match rate into small probability extent of the binomial distribution. The small probability proves the ownership of the database for the owner; meanwhile, it indicates the difficulty of attack to the attackers.

In respect of security: (1) Both the embedding and detecting of watermark are under the control of the privacy key, on which the security of system mainly depends. The key is a pseudo-randomized binary sequence, and the search space of it is large enough. (2) For erasing a watermark in a tuple, the attacker who knows no the privacy key must correctly guess whether the tuple is marked, which attribute and which low order digit are marked, and the value marked together. Therefore, the opportunity of attacking watermark successfully is very little.

In respect of robustness: (1) The chaos sequence is utilized as the watermark, so the distribution of the watermark has stochastic behaviour. (2) The watermark embedding only depends on the privacy key and the primary key, so it has the ability to resist any tuple-rearrangement attack. (3) Repeatedly embedding the watermark into the database by over-sampling the chaos sequence, thus improves the robustness of the watermark against resist attacks.

In respect of imperceptibility: (1) Modifying the low order digits rather than the least significant bits, thus avoids the syndrome phenomena caused by the different types of databases or operation systems. (2) Due to the well-balanced characteristic of chaos sequence, which is also utilized as the watermarking control signal, the change of the mean and variance of the data is limited by modifying the parity via adding 1 to or subtract 1 from the low order digit.

In respect of overhead, the position of the watermark embedded only depends on the privacy key and the primary key of current tuple rather than any other tuples, and needs no more overhead involving other tuples, thus meets the synchronous requirement of the dynamic updating for the database.

5 EXPERIMENTS

Our experiments were performed using the Forest Cover Type data set (Agrawal and Kiernan, 2002), provided by the University of California, Irvin. The data set records the forest cover type of American land for 30 x 30 meter cells. It has 581,012 observation points, each with 54 numeric attributes. Due to many sparse data in the data set, we extracted the first 10 attributes as candidates for watermarking. The data set was transferred to the database of MS SQL Server 2000, and was added an extra attribute named ID to serve as the primary key.

The experiments set that the length of the chaos sequence was 256, and the admissible error of the data set was 3%. We embedded watermarks to the database for γ = 10, 100, 1000 respectively, and evaluated the data error caused by watermarking, the robustness of the watermark against common attacks, and the responsibility of the watermark detector.

Experiment 1: Effect on mean and variance. Table 1 shows the change of the mean and variance of values marked attributes. It can be seen that the mean of values is unchanged, and the variance of it changes very little. Compared with the experimental results of the LSB watermarking scheme (Agrawal and Kiernan, 2002), out algorithm possesses better performance of limited error.

Table 1: Change in mean and variance introduced by watermark.

Attribute	Original mean	Change in mean	Original variance	Change in variance		
				γ=10	γ=100	γ=1000
Col001	2959	0	78391.45	0	0	0
Col002	155	0	12524.68	-0.05	0.03	0.02
Col003	14	0	56.07	0	0	0
Col004	269	0	45177.23	-0.27	0.08	-0.04
Col005	46	0	3398.33	0.21	-0.13	0.07
Col006	2350	0	2431275.75	-8.41	3.65	1.29
Col007	212	0	716.63	0.11	0.04	0.01
Col008	223	0	390.80	0	0.05	-0.02
Col009	142	0	1464.94	-0.03	0.02	0
Col010	1980	0	1753492.95	-7.43	-3.36	2.31

Experiment 2: Subset modifying attack. We modified the tuples in the marked database within the extent of candidate digits in various ratios, and then detected the database. The results of the normalized correlation detected are shown in Figure 1. It can be seen that even if the ratio of the marked tuple was 0.001 (i.e. γ = 1000), and modified 90% tuples in the database, the normalized correlation reached 0.84. Therefore, the scheme is very robust against subset modifying attack towards the watermark.

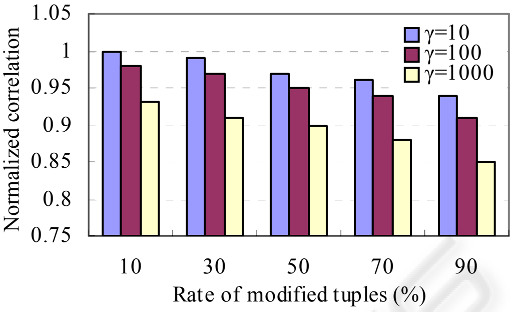

Figure 1: Watermark detection under attacks of tuples modification.

Experiment 3: Subset extracting attack. We extracted the tuples in the marked database in various ratios then detected the database. The results of the normalized correlation detected are shown in Figure 2. It can be seen that even if the ratio of marked tuple was 0.001 (i.e. γ = 1000), and extracted 90% tuples in the database, the normalized correlation reached 0.86. Therefore, the scheme is very robust against subset extracting attack towards the watermark.

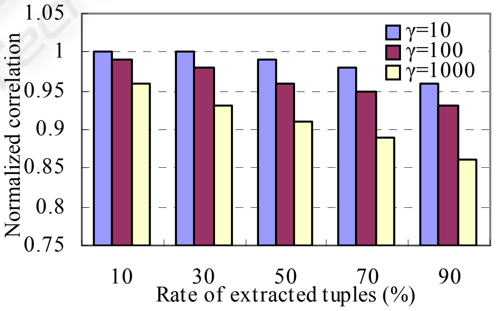

Figure 2: Watermark detection under attacks of tuples extracting.

Experiment 4: Responsibility of the watermark detector. We generated 100 chaos sequences with randomized keys, and replaced 9 sequences with what we detected in Experiment 3, labelled 10, 20, … , 90 respectively. The responses of the normalized correlation detector to each sequence are shown in Figure 3. It can be seen that the responses of the detector to the right privacy keys are obviously higher than those to other keys.

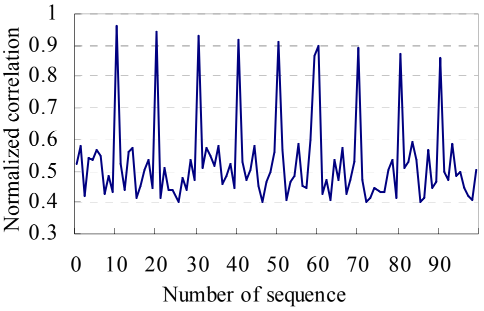

Figure 3: Detector's responses to various keys.

6 CONCLUSIONS AND FUTURE WORK

We proposed a novel scheme for watermarking database based on chaos sequence in this paper. The watermark is embedded into the numeric data by changing the parity of their low order digits, thus avoids the syndrome phenomena caused by the usual LSB watermarking scheme. The embedding watermark meets the synchronous requirement of the dynamic updating for the database, and the detection of the watermark needs no original database. Both the theoretical analysis and the practical experiment prove that this scheme possesses fine efficiency, imperceptibility and security, and it is robust against common attacks towards the watermark.

The main limitation of our scheme is that it is applicable for numeric attributes that can tolerate somewhat distortion within curtain extent. In fact, not all numeric attributes have precision redundancy among their digits, such as age, number of credit card, and so on. Besides, the scheme of watermarking non-numeric data is not concerned in this paper. Therefore, our further work includes watermarking the numeric data that cannot tolerate any distortion and watermarking non-numeric data.

ACKNOWLEDGEMENTS

This work was supported in part by grants of Production, Teaching and Research Project from the Science and Technology Committee of Shanghai, China, 2006, and by grants of the Natural Science Research Project from the Education Department of Jiangsu Province, China, 2005.

REFERENCES

Hakan Hacigumus, Bala Iyer, Sharad Mehrotra. Providing Database as a Service. In *Proc. of ICDE*, 2002

Zhu Qin, Yang Ying, Le Jia-jin, Luo Yi-shu. Watermark Based Copyright Protection of Outsourced Database. In *Proc. of 10th International Database Engineering and Application Symposium (IDEAS 2006)*, Delhi, India, 2006. IEEE Computer Society, 301-305

Rakesh Agrawal, Jerry Kiernan. Watermarking Relational Databases. In *The 28th VLDB Conference*, Hong Kong, China, 2002

Niu Xiamu, Zhao Liang, Huang Wenjun, et al. Watermarking Relational Databa se s for Ownership Protection. In *ACTA Electronica Sinica*, 2003, 31(12A): 2050～2053

Radu Sion, Mikhail Atallah, Sunil Prabhakar. Rights Protection for Relational Data. In *IEEE Transactions on Knowledge and Data Engineering*, 2004, 16(6):1509-1525

Francesc Sebe, Josep Domingo-Ferrer, Jordi Castella-Roca. Watermarking Numerical Data in the Presence of Noise. In *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2006, 14(8):495-508

J. C. Yen. Watermarks Embedded in the Permuted Image. In *The 2001 IEEE International Symposium on Circuits and Systems (ISCAS 2001)*, Sydney, Australia, 2001