MULTIPLE WATERMARKING OF RELATIONAL DATABASES AND OWNERSHIP CLAIM

Manoj Kumar, O.P.Verma
Delhi Technological University, Delhi, India

***Abstract*—** Database Watermarking methods are used for security and copyright protection of relational databases. Many techniques have been popular for watermarking multimedia digital assets like images, audio, video, text etc. Methods used for these objects are usually not applicable with numerical database, because to insert a watermark into a data, small error is created in data, called mark. An error in relational data is usually not acceptable, so a different approach need to be develop to create a mark into the numerical database. Many different approaches have been discussed in previous researches for relational database watermarking. This paper explores enhancing security with multiple watermarking on a given database. Additionally it suggest a solution to the problem of copyright claim when more than one entities claim ownership of same database. In case both claimant have inserted their own watermarks, it is possible to extract both watermarks and decide which watermark was inserted first. The scheme can be applied to any existing database watermarking algorithms but a flexible algorithm is suggested for observing effects of multiple watermarking.

Keywords—database copyright, digital watermarking, watermark embedding.

I. INTRODUCTION

The technology of data mining have matured enough. Use of the technology of data mining can mine a lot of potential, worthful, and new knowledge from unregulated data materials. Applying the new mined knowledge can improve the work, raise the efficiency and can take someone ahead of their competitors. Recent business models are being modeled based on result of data mining. Big giants in the area of online marketing continuously improve their business based of result of data mining on their and competitors' past sales data. Advertizing a product to user on internet/social media is also based on data mining. As data mining is extensively used to stay ahead in business, more and more research institutions are buying the databases from various sources for analysis purpose. For example the consumer's personal data like name, birthday, phone number, address are available in any online store. Online store is also storing data related to products purchased and searched. Means customer's consumption ability, likes, interest, need, choice of brand and many more details can be stored by the merchant. This data may not any meaning for general people, but it will be the most important information about "Consumer's psychology", "Consumer behavior" for the marketing research institutions and organizations. Sometimes, the enterprises itself would like to sell their data warehouses to these institutions to do their research if there are no concern related to customer's personal data. The market of databases is flourishing because this kind of demand and supply market is developed. But the database is easier to be copy and abuse, and the internet is so popular that the information propagates more rapidly. The information passes through the internet without any monitoring and could be destroyed or altered. The consumer of the information are usually not having any idea about the validity of the information received. If someone puts fake data in the information intentionally, the researcher who analyzes the data would make a misleading conclusion and it might have great effects on related research. So the public authentication which has public trust comes with the tide of fashion. Using the public authentication protects the digital information on the internet, proving the copyright and integrity of digital information is most important. So digital watermarking is developed. The concept of digital watermarking comes from Information Hiding which is popular technique for multimedia data like images, video and audio. If people argued the copyright of protected information, we can extract the embedded watermarks to prove the copyright. Digital watermarking technique mainly applies to copyright protected and integrity of information content authenticated. In the copyright protection, the watermarks must have robustness. Having robustness means if the data is altered maliciously then also it is possible to extract the hidden watermarks with some possible noise. In the integrity of information content authenticated, the watermarks have to make sure whether the data is attacked. In the past, digital watermarking technique is widely used on image process. At present, it used on databases because the markets of databases is on rise. Rich body of literature on watermarking multimedia data are available [1]. Most of these techniques were initially developed for still images but later extended to watermark video and audio sources. There are many technical challenges if we think about watermarking relational databases due to the differences in the characteristics of relational database and multimedia data. Multimedia objects consist of a large number of bits, with considerable redundancy. Thus, the watermark has a large cover in which it can be embedded. A database relation consists of tuples, each of which represents a separate object. The watermark needs to be spread over these separate objects. Other challange is that the relative spatial/temporal positioning of various pieces of a multimedia object typically does not change whereas. collection of tuples of a relation constitute a set [2]. There is no implied ordering of these tuples in database. Portions of a multimedia object cannot be dropped or replaced arbitrarily without causing perceptual changes in the object. However, the pirate of a relation can simply drop some tuples or substitute them with tuples from other relations. Because of these differences, techniques developed for multimedia data cannot be directly used for watermarking relational databases. For instance, pixels in a neighborhood in a real image are usually highly correlated and this assumption forms the basis of many techniques such as predictive coding for deciding watermark locations. Several techniques first apply a transform (e.g. discrete Fourier, discrete

cosine, Mellin-Fourier, Wavelet) to the image, insert the watermark in the transformed space, and then invert the transform. The noise introduced by the watermarking signal is thus spread over the whole image. A direct application of these techniques to a relation will introduce errors in all of the attribute values, which might not be acceptable. Furthermore, such a watermark might not survive even minor updates to the relation. Watermarking techniques for text exploit the special properties of formatted text. Watermarks are often introduced by altering the spacing between words and lines of text. Some techniques rely on rephrasing some sentences in the text. While these techniques might be useful to watermark relational databases containing CLOBs (character large binary objects), their applicability to relations consisting of simple data types is suspect.

II. RELATED WORK

Idea of relational database watermarking was introduces by Rajesh Agrawal and Jerry Kiernan in 2002. [3]. They proposed an idea which ensures that some bit positions of some of the selected database attributes of some of the selected tuples contain some specific values known as watermarks. A secret private key known only to the owner of the data is used for selection of the tuples which actually contain the watermark bit in some of the attributes. Bit positions in an attribute, and specific bit values are all determined algorithmically using cryptographic functions [4]. Cryptographic encryption function with a private secret key is applied on primary key attribute of tuple to determine eligibility of tuple to have watermark in its attributes. The bit pattern embedded in attributes of selected tuples constitute the watermark. Embedded watermark bits can be extracted with high probability using the same secret private key. Original unmarked database and watermark are not required to detect the watermark. The watermark can be detected even in a small subset of a watermarked relation as long as the sample contains some of the marks.

In 2004, Sion et al. [5] proposed a method which targets selected tuples to embed the watermark bits into partition statistics. Statistics were changed according to the distortion tolerance (usability constraints). Distortion tolerance was responsible to keep a check on the values of the attributes so that the change made to attribute values do not exceed a limit. It was further enhanced an optimization techniques[6].

The watermarking method in [7] embeds random digits (between 0 to 9) at LSB positions of the candidate attributes for some algorithmically chosen tuples. Two sectret keys are used here. During watermark embedding phase, the database tuples are securely partitioned into groups using a cryptographic hash function and only the first m (which is equal to the length of the watermark) groups are considered. The decision whether to mark i^{th} ($1 \leq i \leq m$) group depends on the i^{th} bit of the owner's watermark, whereas the selection of the tuples in a group is based on a secret key (which is different from that used during partitioning) as well as the information at second LSB positions of the numeric candidate attributes. Finally, for the selected tuples random numbers (between 0 and 9) are embedded at LSB positions in the attribute values of those tuples. Observe that although the owner has a watermark of length m, it is not actually embedded. Rather, it is used to identify some valid groups to embed the random values which acts as embedded watermark information. The detection phase determines the presence of mark in a group if the maximum occurrence frequency for a value between 0 and 9 for that group exceeds a threshold.

Among the most recent works, Gupta et al. [8] proposed a reversible watermarking scheme. This is a modified version of scheme proposed by Agrawal and Kiernan [3]. In this scheme, at the time of watermark detection, the original unwatermarked database can also be recovered along with the watermark as ownership proof. The watermark extraction process first extracts a bit *OldBit* from the integer part of the attribute value before replacing it by the watermark bit and then inserts it in the fraction portion of the attribute value. Thus, the watermark bit can be recovered during detection and the attribute can be restored to its unmarked value by replacing the watermark bit with the original bit *OldBit* extracted from the fraction part. They also propose another algorithm to defeat any attempt of additive or secondary attack which relies on the obvious fact that the database relation must be watermarked by the actual owner before Mallory.

Genetic algorithm (GA) and pattern search (PS) were used to insert watermark in statistics of relational database by minimizing or maximizing the hiding function, while keeping distortion tolerance intact. GA was used in watermark signal processes to embed watermark in data statistics. However, their focus was to make watermarking signal more correlated to the original database and thus make watermark detection easy. The shortcoming of aforementioned watermarking techniques is that they are not able to recover the original cover work exactly from the watermarked data. This problem was solved by the introduction of reversible watermarking techniques in the domain of relational databases. Difference expansion based watermarking (DEW) technique was used to achieve reversibility in context of relational databases [9]. DEW is able to restore the original database exactly. Additionally, it also allows adding distortion into the database using distortion tolerance of the attribute. It also encourages the owner to distribute the trial version of the database, which can only be reverted by those users who have purchased the key. Similarly, Gupta et al. [8] solved the problem of secondary watermarking attack by using reversible watermarking.

Gupta and Pieprzyk proposed a zero-bit watermarking method, Farfoura *et al.* [10], Franco et. al [11] suggest watermarking the fractional part of one numerical attribute by means of prediction-error expansion modulation proposed by Alattar in [12][13]. Although this method is said to be robust against some common database manipulations like insertion of tuples, deletion of tuples, a rounding integer operation may destroy the watermark. More generally, difference expansion modulation has not been designed for being robust to attributes' values modifications (this is similar to the one used for images). In order to overcome the above issues, they proposed to exploit the robust lossless watermarking modulation originally proposed for images by De Vleeschouwer *et al.* [14] and integrate it within a common database watermarking scheme. As we see, this one manipulates circular histograms of data and is less or not at all sensitive to the rounding integer operation or dependent on the existence of attributes with fractional parts. Moreover, this method does not depend on the storing structure of the database, thus making it robust to tuple reordering in a relation.

III. PROPOSED WORK

Relational database watermarking techniques are having better space for watermarking as single watermark is usually inserted only in a small fraction of database. Only selected tuples are used to embed watermark information. A single watermark usually populates around 5-20 percent of tuples only leaving bigger space for additional watermarks. Additional watermark can add additional level of security. It is proposed that primary watermark can have sufficiently large percentage of watermarked tuples, followed by additional watermarkings each one using very small fraction of tuples for hiding their watermark. It is also the case that a watermarked database sold to one party is watermarked again using similar or different algorithm and key and sold again to another party. Here original/first watermark is not removed. This type of multiple watermarking on samee database leads to conflict on copyright claims. But if is very simple to detect the order in which multiple watermarks are inserted into the database.

Consider case of double watermarking of a given database DB. Watermark W1 is inserted first followed by W2. There are two possible ways of watermarking: (i) two different keys but same algorithm is used for both W1 and W2. (ii) Two different keys and algorithm for W1 and W2. Let $\alpha 1$ and $\alpha 2$ ($0< \alpha 1 <1$ and $0 < \alpha 2 <1$) are fractions of tuples affected by two watermarkings respectively. Probability that a tuple holds information from watermark W1 is $\alpha 1$, and probability that a tuple holds information from watermark W2 is $\alpha 2$.

probability that a tuple which holds watermark information from W1 is again selected to hold watermark from W2. Probability that a tuple which is watermarked for W1 is again selected for W2 is $\alpha 1.\alpha 2$. Table 1 shows number of such tuples for different values of $\alpha 1$ and $\alpha 2$.

Table 1: Effect of double watermarking on database with 200000 tuples.

α1	**α2**	**Number of tuples marked for W1**	**Number of tuples marked for W2**	**Number of tuples watermarked twice**
0.05	0.05	10000	10000	25
0.05	0.10	10000	20000	1000
0.10	0.10	20000	20000	2000
0.10	0.15	20000	30000	3000
0.15	0.15	30000	30000	4500
0.15	0.20	30000	40000	6000
0.20	0.20	40000	40000	8000

It is observed that for large database and watermarking schemes with $\alpha > 0.05$, sufficient number of tuples are watermarked by both algorithms. Thus tuples which are watermarked again loose their first watermark as is is overwritten by second watermark. Let S1 is set of tuples watermarked by W1, S2 is set of tuples watermarked by W2 and S12 is set of tuples watermarked by both W1 and W2. Now we concentrate on this small set S12 of tuples which were first watermarked by W1 and then again watermarked by W2. If majority of tuples from S12 are containging watermark from W2, we say order of watermarking was W1 followed by W2. Otherwise the order was W2 followed by W1. In absence of any modification attack on database, all tuples in set S12 must have watermark information from W2 if W2 is done after W1. Otherwise all tuples in set S12 will have watermark from W1 only. Thus copyright claim disputes in these situations can be easily resolved just by computing the set S12 and checking which watermark information exists on set S12.

But there are cases where copyright disputes can not be settled if relational database is having multiple watermarks from various claimants. This happens when each watermarking is done on different disjoint sets of attributes. Watermark impossed by one claimant in set of attributes SA1 is not disturbed as onother claiant has embedded watermark in different set of attributes SA2. As sets SA1 and SA2 are disjoint, they do not disturb each other. When watermark is extracted, both watermarks are retrieved successfully and it is impossible to detect which watermark was inserted first. But this situation is having one advantage also. If single owner of database insertes multiple watermarks using disjoint sets of attributes, watermarking becomes very robust. Even if some watermarks are distorted due to various attacks, some other watermarks can be successfully retrieved and verified for ownership.

One such algorithm for multiple watermarking is propose here which will have enhanced robustness against various attacks. Proposed algorithm is extention of algorithm discussed in previous section. Consider the case of double watermarking W1 and W2. Let SA1={A1, A2,...,Ak} is set of attributes to be used for W1, and SA2=(Ak+1,Ak+2,...,Am} be set of attributes used for W2. It is clear here that sets SA1 and SA2 are disjoint. Selection of attributes in SA1 and SA2 can be made random ensuring they are disjoint. Let us assume for simplicity, size of each set is 2^n . Following algorithm will insert first watermark W1 in set of attributes SA1 using sectret key K1.

1. for each tuple $r \in R$
1.1 E_{PK} = Encrypt (PK, K1): encrypt primary key PK using secret key K1.
1.2 If (E_{PK} mod α1 =0)
1.2.1 Select n bits $b_1..b_n$ from E_{PK}. It is used to select attribute from set of attributes SA1 for watermark insertion.
1.2.2 Select another three bits (b_1,b_2,b_3) from E_{PK} to choose one of the bit from the above selected attribute.
1.2.3 Transform the above selected bit to '1'.

Similar algorithm can be used to insert second watermark W2 in set of attributes SA2 using secret key K2.

1. for each tuple $r \in R$
1.1 E_{PK} = Encrypt (PK, K2): encrypt primary key PK using secret key K2.
1.2 If (E_{PK} mod α2 =0)
1.2.1 Select n bits $b_1..b_n$ from E_{PK}. It is used to select attribute from set of attributes SA1 for watermark insertion.
1.2.2 Select another three bits (b_1,b_2,b_3) from E_{PK} to choose one of the bit from the above selected attribute.
1.2.3 Transform the above selected bit to '0'.

Minimum one attributes are needed here in each set SA1 and SA2, but when we are having 2^n attributes in each set SA1 and SA2, randomness of bit distribution among 2^n attributes makes watermark more robust.

Another algorithm proposed in previous section can also be modified to insert multiple watermarks. This algorithm insert watermark in selected pair of tuples by aligning selected bits of two attritutes. Following is algorithm to insert first watermark W1 in set of attributes SA1 using secret key K1.

1. Select A_x and A_y, two attributes from set SA1 for hiding watermark W1.
2. For each tuple $r \in R$
2.1 E_{PK} = Encrypt (P_{K1},K1)
***2.2* If (E_{PK} mod α1 =0)**
***2.2.1* Select 3 bit from E_{PK} to decide value *i*, and select another 3 bits from E_{PK} to decide value of *j*.**
2.2.2 Make j^{th} bit of attribute A_y identical to i^{th} attribute of A_x.

Following is algorithm to insert second watermark W2 in set of attributes SA2 using secret key K2.

1. Select A_x and A_y, two attributes from set SA2 for hiding watermark W2.
2. For each tuple $r \in R$
2.1 E_{PK} = Encrypt (P_{K1},K2)
2.2 If (E_{PK} mod α1 =0)
2.2.1 Select 3 bit from E_{PK} to decide value *i*, and select another 3 bits from E_{PK} to decide value of *j*.
2.2.2 Make j^{th} bit of attribute A_y identical to i^{th} attribute of A_x.

IV. RESULTS AND DISCUSSIONS

Watermarks are embedded in selected tuples and it is not always embedded in one selected attribute. If we take n=1, means we fix the attribute for hiding watermark bit. when using n attributes, watermarking bits are distributed randomly among n attributes. Experiment have been done for n=1 and n=2 (four attributes) on a database having 200000 tuples and different densities

Table 2: Double watermarking on database with 200000 tuples, with n=1 and different combinations of α values.

α1	α2	Number of tuples marked for W1	Number of tuples marked for W2	Number of tuples watermarked twice
0.05	0.05	9887	9976	23
0.05	0.10	9887	20033	989
0.10	0.10	20121	20033	2015
0.10	0.15	20121	30104	3012
0.15	0.15	30016	30104	4476
0.15	0.20	30016	39875	6033
0.20	0.20	39947	39825	7988

Table 3: Double watermarking on database with 200000 tuples, with n=2 and different combinations of α values.

α1	α2	Number of tuples marked for W1	Number of tuples marked for W2	Number of tuples where same attribute is modified twice
0.05	0.05	9887	9976	8
0.05	0.10	9887	20033	247
0.10	0.10	20121	20033	504
0.10	0.15	20121	30104	761
0.15	0.15	30016	30104	1121
0.15	0.20	30016	39875	1523
0.20	0.20	39947	39825	1978

The tuples where same attribute is modified twice paly important role in deciding ownership of the database. If selected bit is '1' in these tuples, means W1 was inserted after W2, thus owner who inserted W2 is the legitimate owner of the database . On the otherhand if selected bit is '0' in these tuples, means W2 was inserted after W1, thus owner who inserted W1 is the legitimate owner of the database.

Proposed scheme can be modified in many ways keeping in mind that we must have sufficient large set of bits witch are being overwritten ahen inserting second watermark. These overwritten bits are deciding factor when a dispute over ownership is there.

V. REFERENCES

[1]. Gang Chen, "Watermarking Abstract Tree-Structured Data", Lecture Notes in Computer Science , 2005.

[2]. Khurram Jawed, Asifullah Khan, "Genetic algorithm and difference expansion based reversible watermarking for relational databases", Journal of Systems and Software, 2013.

[3]. R.Agrawal and J. Kiernan, Watermarking Relational Databases, Proc. VLDB'02 (2002), pp. 155-166.

[4]. A.A. Mohanpurkar, M.S. Joshi. "Applying watermarking for copyright protection , traitor identification and joitn ownership : A Review", 2011 World Congress on Information and Communication Technologies, (2011)

[5]. R.Sion, M. Atallah, and S. Prabhakar, Rights Protection for Relational Data, IEEE Transaction on Knowledge and Data Engineering, vol. 16, no.12, (2004), pp. 1509-1525

[6]. Shehab M,, Bretino E, and Ghofoor A, Watermarking Relational Databses using Optimization Based Technqiues, IEEE transactions on Knowledge and Data Engineering 2008,vol 20, no 1, pp 116-129

[7]. Zh-Hao Zhang, Xiao-ming-jin, Jian-min Wang and De-yi Li, "Watermarking relational database using image", in Procedings of International Conference on Machine Learning and Cybernetics, vol.3, 2006, pp. 1739-1744.

[8]. Gupta G, Pieprzyk J, Databse relation watermarking resilient against secondary watermarking attacks, Proc. 5th International Conference on Information Systems Security, pp. 222-236, (2009)

[9]. Gupta G, Pieprzyk J, Reversible and Blind Database watermarking using difference expansion, Proc. 1st International Conference on forensic applications and technqiues in Telecommunication, (2008).

[10]. Mahmoud E. Farfoura , Shi-Jinn Horng & Xian Wang,A novel blind reversible method for watermarking relational databases,Pages 87-97 ,Oct 2012.

[11]. Franco-Conteras, Javier, Gouenou Coatrieux, Fredric Cuppens, Nora Cuppens-Boulahia, and Christian Roux, "Robust Lossless Watermarking of Relational Databases Based on Circular Histogram Modulation ", IEEE Transactions on Information Forensics and Security, 2014.

[12]. Alattar AM. Reversible watermark using difference expansion of triplets. In Proceedings of IEEE International Conference on Image Processing (ICIP 2003), Spain, 2003; 501–504.

[13]. Alattar AM. Reversible watermark using the differ-ence expansion of a generalized integer transform.IEEE Transactions on Image Processing 2004;13 (8): 1147–1156.

[14]. Vleeschouwer CD, Delaigle JF, Macq B. Circular interpretation of bijective transformations in lossless watermarking for media asset management.,IEEE Transactions on Multimedia (2003); : 97–105