International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 08 | Aug-2016 www.irjet.net p-ISSN: 2395-0072

A Robust and Resilient Watermarking Approach for Numeric and Non-Numeric Relational Data

Aswany Shaji[1], Ms. Vidhya P.M.[2]

[1]MTech Cyber Security, Dept. of CSE, SNGCE, Kadayiruppu, Kerala, India.
[2]Asst. Prof. MTech Computer Science, Dept. of CSE, SNGCE, Kadayiruppu, Kerala, India.

Abstract - *Today's world is digital world. Nowadays, in every field there is enormous use of digital contents. Information handled on internet and multimedia network system is in digital form. The copying of digital content without quality loss is not so difficult .Due to this , there are more chances of copying of such digital information. So , there is great need of prohibiting such illegal copyright of digital media. Digital watermarking (DWM) is the powerful solution to this problem. Digital watermarking is nothing but the technology in which there is embedding of various information in digital content which we have to protect from illegal copying. This embedded information to protect the data is embedded as watermark. Reversible watermark technology allows the distortion-free recovery of relational databases after the embedded watermark data are detected or verified. I propose a robust technique of embedding reversible watermark in a relational database with non-numeric attributes. In this project, a robust, resilient and reversible watermarking scheme for nonnumeric data that can be used to provide proof of ownership for the owner of a relational database.*

Keywords: **Reversible Watermarking, data recovery, data quality**

1. INTRODUCTION

The advent of the Internet has resulted in many new opportunities for the creation and delivery of content in digital form. Applications include electronic advertising, real time video and audio delivery, digital repositories and libraries, and Web publishing. The recent surge in the growth of the Internet results in offering of a wide range of web-based services, such as database as a service, digital repositories and libraries, e-commerce, online decision support system etc. These applications make the digital assets, such as digital images, video, audio, database content etc, easily accessible by ordinary people around the world for sharing, purchasing, distributing, or many other purposes. As a result of this, such digital products are facing serious challenges like piracy, illegal redistribution, ownership claiming, forgery, theft etc[1],[2].

Watermarking is a technique which is used to deter data piracy and tamperproof the data during its transmission from one machine to another. There are realms of watermarking namely image watermarking, video watermarking and audio watermarking. The watermarking of the relational databases is a rare and currently developing sector. But as the need of the database applications is increasing, it's exerting more pressure on the data providers to create services that allow the users to access and search their data remotely. Now this is a threat for the data providers which in turn leads to the increasing demand of the database watermarking to detect the pirated copy and protect it from being tampered with. One thing which shouldn't be assumed is that database watermarking techniques are similar to the watermarking techniques of multimedia objects. Watermarking of the database requires techniques which differ from those of the conventional ones used for multimedia watermarking purposes. Multimedia objects cannot be dropped or replaced arbitrarily but in databases the tuple insertion, deletion and update are the main norms in database setting. Because of this, different techniques developed for multimedia data cannot be directly used for watermarking of the relational databases. Text properties and semantics can be exploited[3].

The process of embedding information into another object/signal is termed as digital watermarking. In visible watermarking, the information is visible in the picture or video. Typically, the information is text or a logo which identifies the owner of the media. The image on the right has a visible watermark. When a television broadcaster adds its logo to the corner of transmitted video, this is also a visible watermark. In invisible watermarking, information is added as digital data to audio, picture or video, but it cannot be perceived as such (although it may be possible to detect that some amount of information is hidden). The watermark may be intended for widespread use and is thus made easy to retrieve or it may be a form of Steganography, where a party

communicates a secret message embedded in the digital signal. In either case, as in visible watermarking, the objective is to attach ownership or other descriptive information to the signal in a way that is difficult to remove. It is also possible to use hidden embedded information as a means of covert communication between individual. The purpose of embedding the information depends upon application and need of user of digital media. Digital watermarking provides the solution for difficult problem of providing guarantee to organizer and consumer of digital content about their legal rights .Copyright protection for multimedia information is nothing but a golden key for multimedia industry[1],[3].

Embedding watermarks in database relations is a challenging problem because there is little redundancy present in a database relation. One important property of digital watermarks is invisibility. Usually, in a watermarking scheme, a watermark is embedded by slightly modifying the cover data. To ensure invisibility, the modifications are limited to some acceptable level. This requires that the cover data can tolerate these modifications. In the context of multimedia data, this requirement is not a problem. Since multimedia data are highly correlated, there is a lot of redundant information present in multimedia data. Although compression techniques can remove some of the redundant information, currently, no compression technique is perfect enough to remove all the redundant information. So we move for watermark embedding. A watermark can be embedded as a part of the redundant information without affecting the quality of the multimedia data. Furthermore, some properties of the human vision (auditory) system can be incorporated to the watermark embedding so that the strength of the embedded watermark can be adjusted adaptively. All of these make it easy to ensure invisibility for multimedia watermarking. In contrast, database relations contain large number of independent tuples. A tuple can be added, deleted, or modified without affecting other tuples. All tuples and all attributes are equally important. There is little redundancy present in the tuples. Thus, it is a challenge to embed an invisible watermark in a database relation.

In general, the database watermarking techniques consist of two phases: watermark embedding and watermark verification. Watermark embedding phase includes a private key K (known only to the owner) which is used to embed the watermark bits into the original database to form watermarked database. The watermarked database is then made publicly available. During the embedding phase, a private key K is used to embed the watermark W into the original database. To verify the ownership of a suspicious database, the verification process is performed where the suspicious database is taken as input, and by using the key K (the same which is used during the embedding phase) the embedded watermark (if present) is extracted and compared with the original watermark information. A suspicious database can be any watermarked database or innocent database, or a mixture of them under various database attacks[2].

2. BACKGROUND

We first give formal introductions and information's regarding digital watermarking in database. Now we will see background information regarding the proposed system.

2.1 Genetic Algorithm

Genetic Algorithms are a family of computational models inspired by evolution. These algorithms encode a potential solution to a specific problem on a simple chromosome like data structure and apply recombination operators to these structures so as to preserve critical information Genetic algorithms are often viewed as function optimizers although the range of problems to which genetic algorithms have been applied is quite broad. An implementation of a genetic algorithm begins with a population of (typically random) chromosomes. One then evaluates these structures and allocates reproductive opportunities in such a way that those chromosomes which represent a better solution to the target problem are given more chances to "reproduce" than those chromosomes which are poorer solutions. The "goodness" of a solution is typically defined with respect to the current population[4].

2.2Elitism Strategy

Genetic algorithms are stochastic search methods based on the principles of natural genetic systems. They perform a multi-dimensional search in providing an optimal solution for evaluation(fitness)function of an optimization problem. GAs are empirically found to provide global near optimal solutions of various complex optimization problems in the fields of operation search, VLSI design ,pattern recognition ,image processing etc. While solving an optimization problem using GAs, each solution is coded as a string(called "chromosome")of finite length over a finite alphabet A. Each string or chromosome is considered as an individual. A collection of M (M is finite) such individuals is called

populations. GAs start with a randomly generated population of size M. In each iteration, a new population of the same size is generated from the current population using three basic operations on the individuals of the population .The operators are selection, crossover, and mutation. The new population obtained after selection, crossover and mutation is then used to generate another population. Note that the number of possible populations is always finite since A is a finite set and M is finite .Sometimes the knowledge about the best string obtained so far is preserved within the population .Such a model is called a genetic algorithm with an elitist model or EGA.

Genetic algorithm with elitism is a way to solve a general problem by probabilistic search method For both numerical and non-numerical data, for the creation of optimal watermark information, that needs to be embedded in the original data, we use an evolutionary technique; elitism strategy . Elitist selection is a selection strategy where a limited number of individuals with the best fitness values are chosen to pass to the next generation, avoiding the crossover and mutation operators. Elitism prevents the random destruction by crossover or mutation operators of individuals with good genetics. The number of elite individuals should not be too high, otherwise the population will tend to degenerate. A practical variant of the general process of constructing a new population is to allow the best organism(s) from the current generation to carry over to the next, unaltered. This strategy is known as elitist selection and guarantees that the solution quality obtained by the GA will not decrease from one generation to the next .we take the strategy of replacing the worst string of the new population with the best string of the current population. Genetic algorithms with this strategy are referred as genetic algorithms with elitism or EGA [2].

EGA is a general purpose algorithm and one can safely use it without doing a great deal of research on computing every time a new combinatorial optimization problem appears.

2.2 Random Number Generation Algorithms

A degree of randomness is built into the fabric of reality. It is impossible to say for certain what a baby's personality will be, how the temperature will fluctuate next week, or which way dice will land on their next roll. A planet in which everything could be predicted would be bland, and much of the excitement of life would be lost. Because randomness is so inherent in everyday life, many researchers have tried to either harvest or simulate its effect inside the digital realm. There are two main types of random number generators. The first type attempts to capture random events in the real world to create its sequences. It is referred to as a true random number generator, because in normal circumstances it is impossible for anyone to predict the next number in the sequence. The second camp believes that algorithms with unpredictable outputs (assuming no one knows the initial conditions) are sufficient to meet the requirements for randomness. The generators produced through algorithmic techniques are called pseudo-random generators, because in reality each value is determined based off the system's state, and is not truly random [8].

3. PROPOSED SYSTEM

The proposed system discusses RRW for reversible watermarking of relational databases that improves data recovery ratio. It is a robust technique of embedding reversible watermark in a relational database with numeric and non-numeric attributes. A robust,resilient and reversible watermarking scheme for numeric and non-numeric data that can be used to provide proof of ownership for the owner of a relational database. The main architecture of RRW is presented in Fig. 1 RRW includes the following four major phases: (1) watermark preprocessing; (2) watermark encoding; (3) watermark decoding; and (4) data recovery.

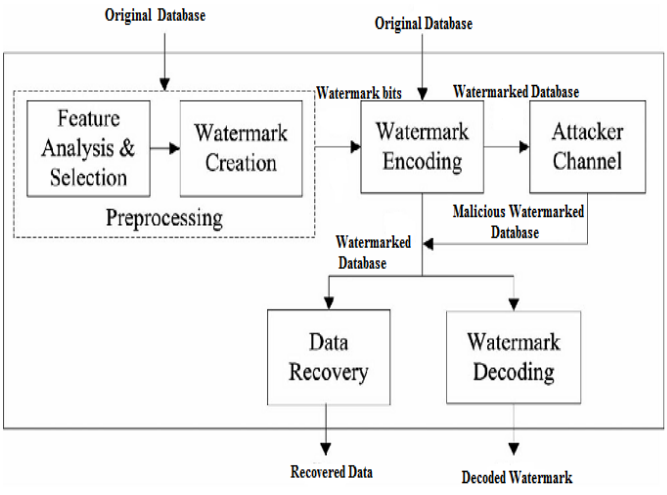

Fig. 1 Architecture of RRW

3.1 Watermark Preprocessing Phase

The watermark preprocessing phase computes different parameters for calculation of an optimal watermark. These parameters are used for watermark encoding and decoding.

In the preprocessing phase, two important tasks are accomplished :(1) selection of a suitable feature for watermark embedding; (2) calculation of an optimal watermark.

3.1.1 Feature Analysis and Selection

For developing a decisive information model of various features of the dataset, compute mutual information of features. Mutual information of every feature with all other features is calculated by using Equation (1).

$$MI(A,B) = \sum_{a}\sum_{b} P_{AB}(a,b)\log\frac{P_{AB}(a,b)}{P_A(a)P_B(b)}. \qquad (1)$$

The value of mutual information (MI)of each feature is then used to rank the features. The attacker can try and predict the feature with the lowest MI in an attempt to guess which feature has been watermarked. To deceive the attacker, a secret threshold can be used for selecting the feature for watermark embedding [1].

3.1.1 Watermark Creation through Elitism Strategy and Random Number Generation

For both numerical and non-numerical data, for the creation of optimal watermark information, that needs to be embedded in the original data, we use an evolutionary technique; elitism strategy. Elitist selection is a selection strategy where a limited number of individuals with the best fitness values are chosen to pass to the next generation, avoiding the crossover and mutation operators. The number of elite individuals should not be too high, otherwise the population will tend to degenerate .A practical variant of the general process of constructing a new population is to allow the best organism(s) from the current generation to carry over to the next, unaltered. This strategy is known as elitist selection and guarantees that the solution quality obtained by the GA will not decrease from one generation to the next. The optimal fitness value obtained through elitism strategy is basically the change to be embedded in the original data that needs to be watermarked. The purpose of getting an optimum value is to justify the amount of change that a feature value can withhold without compromising the data quality. Using pre-defined digital numbers of each non-numeric attribute we embedding the watermark information [7].

Random number generators are useful for many different purposes. Aside from obvious applications like generating random numbers for the purposes of gambling or creating unpredictable results in a computer game, randomness is important for cryptography .Here we can use the random number generation algorithm for the creation of optimal watermark information. The computation time incurred in this phase for increasing data set sizes (number of tuples) where watermark bit size was kept fixed and the number of features were 14 is less compared with elitism strategy method. So we can prefer random number generation method for lesser computation time [8].

3.2 Watermark Encoding Phase

In the watermark encoding phase, the optimized value is embedded in the particular feature that is selected to be watermarked on the basis of the selection criteria. For watermark encoding algorithm we employed the following steps:

1. Input: original database, watermark bits.
2. For w=1 to l do

   //loop will iterate all watermark bits w from 1to length l of the watermark.
3. For r=1 to R do

   //loop will iterate all tuples of the data.
4. The case when the watermark bit is 0

   4.1. Detected changes are calculated during encoding process.

   4.2. Data is watermarked by adding the optimized value with the selected feature value.
5. The case when the watermark bit is 1

   5.1. Detected changes are calculated during encoding process.

   5.2. Data is watermarked by subtracting the optimized value from the selected feature value.
6. End loop.
7. End loop.
8. Return watermarked database.

3.2 Watermark Decoding Phase

In the watermark decoding process ,the first step is to locate the features which have been marked. The watermark decoder decodes the watermark by working with one bit at a time. For watermark decoding algorithm we employed the following steps:

1. Input: watermarked database or watermarked database after malicious attack, watermark bit,.

2. For r=1 to R do

//loop will iterate all tuples of the data.

3. For b=l to 1 do

//loop will iterate all watermark bits b from 1to length l of the watermark.

4. Calculate the difference between the changes detected in the value of a feature during encoding and decoding process.

5. If the detected change is less than or equal to zero ,then the detected watermark bit is 1.

6. Else if the change is greater than zero and less than or equal to one, then the detected watermark bit is 0.

7. End loop.

8. End loop.

9. Final watermark information is retrieved by taking the mode of detected watermark bit from 1 to length l of the watermark.

10. Return decoded watermark.

3.3Data Recovery Phase

After detecting the watermark string, some post processing steps are carried out for data recovery. For data recovery algorithm we employed the following steps:

1. Input: watermarked database or watermarked database after malicious attack, watermark bit.

2. For r=1 to R do

//loop will iterate all tuples of the data.

3. For b=l to 1 do

//loop will iterate all watermark bits b from 1to length l of the watermark.

4. 0 Or 1 watermark bit is detected from every tuple r then data is recovered by adding the optimized value with the tuples of the watermarked database or watermarked database after malicious attack.

5. Else data is recovered by subtracting the optimized value from the tuples of the watermarked database or watermarked database after malicious attack.

6. End loop.

7. End loop.

8. Return recovered data.

4. RESULTS AND DISCUSSION

RRW was evaluated for: (1) investigating effect on the data quality of the underlying data, (2) restoration of the original data. The data recovery, watermark detection accuracy and effect of RRW on data quality are evaluated using the case study of a heart disease medical dataset. A small set of tuples from the same dataset are also used as an example to illustrate the entire procedure step by step.

The computational time of RRW is (l*R*A) where l is the watermark length, R is the total number of tuples in the dataset and A is the feature selected for watermarking. The number of tuples are usually much larger as compared to the number of features in databases and the watermark length l; so, A <<R and l <<R. Therefore, for large databases, (R termed as “n”) the time complexity of RRW for watermark insertion and detection is O(n). For datasets involving large number of features or large number of tuples, the data owner may use a separate machine, with high computation power, for watermarking the datasets.

Many watermarking techniques are based on different watermark information; most of these techniques are designed for numerical database. There are almost similar steps to identify attribute then tuple and then marking position for the watermark. It will be difficult for attacker to remove watermarks from different places from the database. The proposed method is able to recover both the embedded watermark and the original data. Random simulation has long been a very popular and well-studied field of mathematics. There exists a wide range of applications in biology, finance, insurance, physics and many others. So simulations of random numbers are crucial. Random number generators are useful for many different purposes. Aside from obvious applications like generating random numbers for the purposes of gambling or creating unpredictable results in a computer game, randomness is important for cryptography .Here we can use the random number generation algorithm for the creation of optimal watermark information. The computation time incurred in this phase for increasing data set sizes (number of tuples) where watermark bit size was kept fixed and the numbers of features were 14 is less compared with elitism strategy method. Fig. 2 demonstrates the time incurred in this phase for increasing dataset sizes (number of tuples) where l was kept fixed and the number of features was 14[1],[5].So we can prefer random number generation method for lesser computation time.

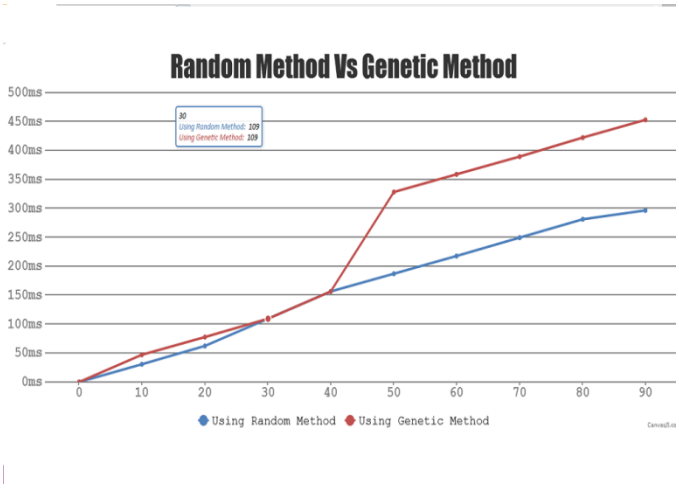

Fig.2 Computation time

5. CONCLUSION AND FUTURE WORK

The large need of networked multimedia system has created the need of copyright protection. It is very important to protect intellectual properties of digital media. Internet playing an important role of digital data transfer. Digital watermarking is the great solution of the problem of how to protect copyright. Digital watermarking is the solution for the protection of legal rights of digital content owner and customer. In this paper, we presented a new approach to watermark a non- numeric attribute in the relational database. This algorithm can be used effectively where a huge amount of relational data is transferred between owner and authenticated users.

One of our future concerns is to watermark shared databases in distributed environments where different members share their data in various proportions.

REFERENCES

[1] Saman Iftikhar, M. Kamran, Zahid Anwar ,"RRW-A Robust and Reversible Watermarking Technique for Relational Data,"IEEE Trans. on knowledge and data engineering,vol. 27, no. 4, 2015.

[2] R. Agrawal and J. Kiernan, "Watermarking relational databases," in Proc. 28th Int. Conf. Very Large Data Bases, 2002, pp. 155-166.

[3]http://arnab.org/notes/introduction-to-digital-watermarking

[4]K. Jawad and A. Khan, "Genetic algorithm and difference expansion based reversible watermarking for relational databases," J. Syst. Softw., vol. 86, no.11, pp. 2742-2753, 2013.

[5] T.Thilagam, R.Vinoth ," RRW - A Resilient Reversible Watermarking Technique for the Preclusion of Information from Cyber Punkers,"International Journal of Engineering Research-Online A Peer Reviewed International Journal, Vol.4., Issue.1., 2016.

[6] Bhupendra Nath Chaudhary, Awadesh Kumar Sharma ," A Modern Watermarking Approach for Non- Numeric Relational Database,"International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 12, December 2013.

[7] Spinoza, Ethics ,"Overview of random generation algorithms".

[8] David DiCarlo,"Random Number Generation: Types and Techniques,"Spring 2012.