© 2017 IJSRSET | Volume 3 | Issue 7 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology
National Level Technical Conference on Advanced Computing Technologies- n'CACT'17, Department of Computer Science & Engineering, Ammini College of Engineering, Kannampariyaram, Mankara, Palakkad, Kerala, India
In association with
International Journal of Scientific Research in Science and Technology

An Efficient Watermarking Technique Using Genetic Algorithm for Relational Data

Sreesha K S[1], Jincy Easow[2], Prof. Jisha P Abraham[3]
[1,2,3] Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, India

ABSTRACT

Watermarking is advocated to enforce ownership rights over shared relational data and for providing a means for tackling data tampering. When ownership rights are enforced using watermarking, the underlying data undergoes certain modifications; as a result the data quality gets compromised. Reversible watermarking is employed to ensure data quality along with data recovery. Reversible watermarking tries to overcome the problem of data quality degradation by allowing recovery of original data along with the embedded watermark information. An optimal watermark value is created through the Genetic Algorithm (GA) and inserted into the selected feature of the relational database. It mainly comprises a data preprocessing phase, watermark encoding phase, attacker channel, watermark decoding phase and data recovery phase. In data preprocessing phase, secret parameters are defined and strategies are used to analyze and rank features to watermark. An optimum watermark string is created in this phase by employing GA an optimization scheme that ensures reversibility without data quality loss.
Keywords: Genetic Algorithm(GA)

I. INTRODUCTION

A watermark is a visible overlay of copyright information usually in the form of text or an image logo added to photos or other digital documents. A watermark protects digital intellectual property, such as photos and artwork, from unauthorized use. It identifies the rightful owner of the work, which discourages other people from using it as their own. While it's hard to prevent people from doing so, having a well-placed watermark containing a copyright symbol, name and URL of the owner can go a long way to discourage this popular,

- Data Preprocessing
- Watermark Encoding
- Watermark Decoding
- Data Recovery

Watermarking technique provides the protection to the database by embedding information. Invisible watermarking, the watermark information such as text or a logo which identifies the owner of the media, that is visible in the picture or video. Here 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 watermarking technique which are irreversible may causes modification of data undergoes through it at the certain extent. This method tries to overcome the problem of data quality degradation by allowing recovery of original data along with the embedded watermark information. Watermarking scheme modify original database as a modulation of the watermark information, and causes impossible to prevent permanent distortion to the original database, and affects to meet the integrity requirement of many application.

II. LITERATURE SURVEY

Several watermarking techniques are proposed which includes watermarking for images, databases, audio and

video. The watermarking is primarily developed for the images, the research in audio is started later. There are less watermarking techniques are proposed for audio compared to the images/video. Embedding the data in audio is difficult compared to the images because the Human auditory system is more sensitive than the Human visual System. In last ten years there is a lot of advancement in watermarking few of them are discussed here. This chapter reviews the literature of information hiding in sequences. Scientific publications included into the literature survey have been chosen in order to build a sufficient background that would help out in identifying and solving the research problems. During the last decade watermarking schemes have been applied widely. These schemes are sophisticated very much in terms of robustness and imperceptibility. Robustness and imperceptibility are important requirements of watermarking, while they conflict each other.

Non-blind watermarking schemes are theoretically interesting, but not so useful in practical use, since it requires double storage capacity and communication bandwidth for watermark detection. Of course, non-blind schemes may be useful for copyright verification mechanism in a copyright dispute. On the other hand, blind watermarking schemes can detect and extract watermarks without use of the unwatermarked. Therefore it requires only a half storage capacity and half bandwidth compared with the non-blind watermarking scheme. The relational data defers from multimedia data in many respects: (i) Few Redundant Data: Multimedia objects consists of large number of bits providing large cover to hide watermark, whereas the database object is a collection of independent objects, called tuples. The watermark has to be embedded into these tuples, (ii) Out-of-Order Relational Data: The relative spatial/temporal positions of different parts or components in multimedia objects do not change, whereas there is no ordering among the tuples in database relations as the collection of tuples is considered as set, (iii) Frequent Updating: Any portion of multimedia objects is not dropped or replaced normally, whereas tuples may be inserted, deleted, or updated during normal database operations, (iv) There are many psychophysical phenomena based on human visual system and human auditory system which can be exploited for mark embedding. However, one cannot exploit such phenomena in case of relational databases.

III. PROPOSED SYSTEM

Advancement in information technology is playing an increasing role in the use of information systems comprising relational databases. A robust and semi-blind reversible watermarking technique for numerical relational data as well as the string data has been proposed that addresses the above objectives. This project proposes one such reversible watermarking technique that keeps the data useful for knowledge discovery. GA - an optimization algorithm is employed in this proposed project to achieve an optimal solution that is feasible for the problem at hand and does not violate the defined constraints. The proposed system consist of four stages: data preprocessing, watermark encoding, watermark decoding and data recovery. The watermark preprocessing phase computes different parameters for calculation of an optimal watermark. These parameters are used for watermark encoding and decoding. The main focus of watermark encoding phase is to embed watermark information in such a way that it does not affect the data quality.

During watermark embedding, data gets modified according to the available bandwidth (or capacity) of the watermark information. The bandwidth of the watermark should be sufficiently large to ensure robustness but not so large that it destroys the data quality. The data owner decides the amount of data modification such that the quality is not compromised for a particular database application before-hand and therefore defines usability constraints to introduce tolerable distortion into the data. After watermarking, the data is released to the intended recipients over a communication channel that is assumed to be insecure and termed as the “attacker channel” in this research domain. The data may undergo several malicious attacks in the attacker channel. The efficiency and effectiveness is described through robustness analysis determined by its response to subset insertion, alteration and deletion attacks. The Watermark decoding phase recovers watermark information effectively for detection of the embedded watermark. Data recovery phase mainly comprises the important task of successful recovery of the original data.

A. Preprocessing Phase

In the preprocessing phase, two important tasks are accomplished:
1) Selection of a suitable feature for watermark embedding
2) Calculation of an optimal watermark with the help of an optimization technique.

Feature Analysis and Selection

For developing a decisive information model of various features of the dataset, all the features are ranked according to their importance in information extraction, subject to their mutual dependence on other features. For this purpose, mutual information (MI), is exploited, that is an important statistical measure for computation of mutual dependence of two random variables. Mutual information of every feature with all other features is calculated by

$$MI(A, B) = \Sigma a \Sigma PAB(a,b) log \frac{PAB(a,b)}{PA(a)PB(b)}$$

Where MI(A,B) measures the degree of correlation of features by measuring the marginal probability distributions as PA(a), PB(b) and the joint probability distribution PAB(a, b). Then MI of one feature with all other features is computed using the relation. The value of 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 for this particular scenario, a secret threshold can be used for selecting the feature for watermark embedding. In this context, the data owner can define a secret threshold based on MI of all the features in the database. The feature(s) having MI lower than that threshold can be selected for watermarking. The attacker will not attack the features having large MI as in that case the usability of the data will be compromised. Therefore, he will be forced to attack the feature(s) with lower MI without concrete knowledge (due to the use of secret threshold) of which features have been watermarked.

Watermark Creation Using Genetic Algorithm

For the creation of optimal watermark information, that needs to be embedded in the original data, use an evolutionary technique GA. GA is a population-based computational model, basically inspired from genetic evolution. GA evolves a potential solution to an optimization problem by searching the possible solution space. In the search of optimal solution, the GA follows an iterative mechanism to evolve a population of chromosomes. The GA preserves essential information through the application of basic genetic operations to these chromosomes that include: selection, crossover, mutation and replacement. The GA evaluates the quality of each candidate chromosome by employing a fitness function. The evolutionary mechanism of the GA continues through a number of generations, until some termination criteria is met. During watermark creation phase, we employed the following major steps of the GA for getting optimal watermark information:

1) Initial random population of binary strings called chromosomes is generated. Gene values of each chromosome represents l-bit watermark string.

2) Fitness of each chromosome is evaluated by employing a constrained optimized fitness function

3) Tournament selection mechanism is applied to get the most appropriate individuals as parent chromosomes.

4) Genetic operations of crossover and mutation are performed on parent chromosomes to create off-springs. A single point crossover operator is applied to evolve high quality individuals, inheriting parental characteristics, by exchanging information between two or more chromosomes. A uniform mutation operator is applied to bring diversity in population through small random changes in gene values of binary chromosomes. The values of crossover fraction and mutation rate are set empirically

5) Elitism strategy is applied to hire two individuals with best fitness value; as elites to the next generation without genetic changes.

6) Remaining population of the next generation is created by replacing less fit individuals of the previous generation with the most fit newly created off-springs.

7) Steps 2 to 6 are repeated until MIO and MIW reach approximately equal values for a certain number of generations.

8) Both, optimal watermark information string and best fitness value (b) is returned after the fulfillment of the termination criteria.

B. Watermark Encoding Phase

Watermark information calculation is formulated as a CO problem to meet the data quality constraint of the data owner. A GA is used to create optimal watermark information that includes:

1) Optimal chromosomal string (watermark string of length l)

2) b value, b is a parameter that is computed using GA and represents a tolerable amount of change to embed in the feature values.

Once the optimum value of b for each candidate feature A is found, it is saved for use during watermark encoding and decoding. A watermark (bit string) of length l and an optimum value b is used to manipulate the data provided it satisfies the usability constraints. The value b is added into every tuple of the selected feature A when a given bit is 0; otherwise, its value is subtracted from the value of the feature. It is ensured that the mutual information of a feature remains unchanged, when the watermark is inserted into the database. The watermark is inserted into every tuple for the selected feature of the dataset. The data owner can select any number of features for watermark embedding based upon a secret threshold and MI of the feature(s). After finding the optimum value of b, a parameter nr is calculated, that represents the percentage change in the watermark encoding. This parameter is calculated for a tuple r as:

$$nr = Dr * \ell$$

The parameters used in the above equation are nr, Dr, and ℓ where nr is the detected amount of percentage change in encoding, Dr is the recovered data and ℓ is the length of the watermark. Since the length of the watermark is l; nr is calculated and b is inserted l times in the database. The length ℓ of the watermark should be carefully chosen. If it is too small, it will make the watermark fragile against attacks, and if it is too large, it might compromise the data quality because the data gets altered for every bit of the watermark. In this project, the data gets altered for each watermark bit in every tuple. After a number of empirical studies, a length of 16 bits was selected. The watermark encoding algorithm starts the embedding process with the most significant bit MSB of the watermark. For this purpose the algorithm works with one tuple at a time. If the MSB of the watermark is 1, the new value of Dr, denoted by Dwr is calculated using

$$Dwr = Dr - \beta$$

The parameters used here is Dwr, Dr, and β, where Dwr is the original data to be watermarked, Dr is the recovered data and β is an optimized value, here used is 0.16. In order to embed the second MSB of the watermark, the algorithm is again employed using the same procedure, but the updated value Dr of the feature (that has now become Dwr) is used for calculating new values of nr and Dwr. If the algorithm encounters a watermark bit that is 0 then the new value of Dwr, is calculated using

$$Dwr = Dr + \beta$$

C. Watermark Decoding Phase

In the watermark decoding process, the first step is to locate the features which have been marked. The process of optimization through GA is not required during this phase. We use a watermark decoder z, which calculates the amount of change in the value of a feature that does not affect its data quality. The watermark decoder decodes the watermark by working with one bit at a time. In the decoding phase, ndr is calculated and represents the percent change detected in the watermarked data. The value of ndr, nr and nΔr is calculated using the values of tuple r and therefore might be different for every r. The parameter nΔr is computed by calculating the difference between the original data change amount nr and the watermark detected change amount ndr using

$$ndr = Dw *' \ell$$
$$n\Delta r = ndr - nr$$

The decoding phase mainly consists of two steps:
Step 1: For every candidate feature A of all the tuples in D'W, the watermark bits are detected starting from the least significant bit and moving towards the most significant bit. The bits are detected in the reverse order compared with the bits encoding order because it is easy to detect the effect of the last encoded bit of the watermark. This process is carried out using the change matrix nr.

Step 2: The bits are then decoded according to the percentage change values of watermarked data. If $n\Delta r <= 0$, the detected watermark bit will be 1. If $n\Delta r > 0$ and $n\Delta r < 1$, the detected watermark bit will be 0.
The final watermark information is retrieved through a majority voting scheme using
$wD <$ = mode(dt w(1,2,…,l))

D. Data Recovery Phase

After detecting the watermark string, some post processing steps are carried out for error correction and data recovery. The optimized value of b computed through the GA is used for regeneration of original data. The value of a numeric feature is recovered using Dr.

$$Dr = D'wr + \beta$$
$$Dr = D'wr - \beta$$

IV. CONCLUSION

Watermarking is used to enforce ownership rights over shared relational data and for providing a means for tackling data tampering. Reversible watermarking techniques are used to cater to such scenarios because they are able to recover original data from watermarked data and ensure data quality to some extent. A novel robust and reversible technique for watermarking numerical data of relational database is presented. In this project, we considered the Cleveland Heart Disease dataset, in which data includes the ID, age, type, bp, sex, cholesterol, month, right, left, etc. Firstly calculate the mutual information (MI) values between the candidate attribute and the remaining attributes. Identify the attributes that need to be watermarked, those MI values is less than the threshold value selected by the user. Then apply the genetic algorithm to generate the watermarked information that are embedded into the original selected attribute values for watermarking. Thus the encoding phase is completed. For the decoding phase, the reverse operations are performed to separate the original data and the watermarked information. These techniques are not robust against malicious attacks, particularly this techniques that target some selected tuples for watermarking. One of the future concerns of this project is to watermark the shared databases in distributed environments where different members can share their data in different proportions and also extended for non-numeric data stores.

V. REFERENCES

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

[2]. Y. Zhang, B. Yang, and X. M. Niu, "Reversible watermarking for relational database authentication", Vol.17, 2006.

[3]. Saman Iftikhar, M. Kamran, and Zahid Anwar, "A Robust and Reversible Watermarking Technique for Relational Data", Vol. 27, No. 4, April 2015.

[4]. X. Li, B. Yang, and T. Zeng, "Efficient reversible watermarking based on adaptive prediction-error expansion and pixel selection," IEEE Trans. Image Process., vol. 20, Dec. 2011.

[5]. Sonnleitner, "A robust watermarking approach for large databases," in Proc. IEEE First AESS Eur. Conf. Satellite Telecommun., 2012.

[6]. G. Gupta and J. Pieprzyk, "Database relation watermarking resilient against secondary watermarking attacks," in Information Systems and Security. New York, NY, USA: Springer, 2009.