ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

A Robust and Reversible Watermarking Technique for Relational Data

Tharun Kumar Vulkundakar
M.Tech Student
Department of CSE
Sphoorty Engineering.

Srikanth Lakumarapu
Assistant Professor,
Department of CSE
Sphoorty Engineering.

Deepthi Janagama
HoD
Department of CSE
Sphoorty Engineering.

ABSTRACT

Advancement in information technology is playing an increasing role in the use of information systems comprising relational databases. These databases are used effectively in collaborative environments for information extraction; consequently, they are vulnerable to security threats concerning ownership rights and data tampering. 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 of which, the data quality gets compromised. Reversible watermarking is employed to ensure data quality along-with data recovery. However, such techniques are usually not robust against malicious attacks and do not provide any mechanism to selectively watermark a particular attribute by taking into account its role in knowledge discovery. Therefore, reversible watermarking is required that ensures; (i) watermark encoding and decoding by accounting for the role of all the features in knowledge discovery; and, (ii) original data recovery in the presence of active malicious attacks. In this paper, a robust and semi-blind reversible watermarking (RRW) technique for numerical relational data has been proposed that addresses the above objectives. Experimental studies prove the effectiveness of RRW against malicious attacks and show that the proposed technique outperforms existing ones.

INTRODUCTION

In the digital world of today, data is excessively being generated due to the increasing use of the Internet and cloud computing. Data is stored in different digital formats such as images, audio, video, natural language texts and relational data. Relational data in particular is shared extensively by the owners with research communities and in virtual data storage locations in the cloud. The purpose is to work in a collaborative environment and make data openly available so that it is useful for knowledge extraction and decision making. Take the case of Walmart—a large multinational retail corporation that has made its sales database available openly over the Internet so that it may be used for the purposes of identifying market trends through data mining. However these openly available datasets make attractive targets for attacks. For example there are documented attack incidents where data containing personal information related to customers using certain Walmart video services was stolen. According to a survey related to the security of outsourced customer data, it is reported that 46 percent of organizations do not consider security and privacy issues while sharing their confidential data. Therefore, 64 percent organizations have to face data loss repeatedly. Similarly, data breaches in the health care and medical domain are increasing alarmingly. Therefore it is imperative, that in shared environments such as that of the cloud, security threats that arise from un-trusted parties and relational databases need to be addressed along with the enforcement of ownership rights on behalf of their owners.

Watermarking techniques have historically been used to ensure security in terms of ownership protection and tamper proofing for a wide variety of data formats. This includes images, audio, video, natural language processing software, relational databases, and more. Reversible watermarking techniques can ensure data recovery along with ownership protection. Fingerprinting, data hashing, serial codes are some other techniques used for ownership protection. Fingerprints also called transactional watermarks, are used to monitor and identify digital ownership by watermarking all the copies of contents with different watermarks for different recipients. Primarily this type of digital watermarking tries to identify the source of data leakage by tracing a guilty agent. In hashing, digital contents can be saved by performing a one-way hash function whereby the data contents do not change. If the hash of the original and tampered data is the same, data authenticity can be verified but ownership cannot be proved easily. Serial or classification codes are used for filtering of inappropriate contents over the Internet and are mainly applicable to images, audio and video. Watermarking has the property that it can provide ownership protection over the digital content by marking the data with a watermark unique to the owner. The embedded watermark can subsequently be used for proving and claiming ownership.

Digital watermarking of multimedia content is more commonly known. Particularly image watermarking—a derivative of Steganography is an age-old practice allowing covert transmission of messages from one party to another by exploiting redundancy in common image formats. However the basic process of multimedia watermarking is very different from that used to watermark relational databases because of a fundamental difference in the properties of the data. Multimedia data is highly correlated and continuous whereas relational data is independent and discrete. With the advent of modern copyright protection and information hiding techniques, database watermarking can be used to enforce ownership rights of relational data. However a major drawback of these techniques is that they modify the data to a very large extent which often results in the loss of data quality. There is a strong need to preserve the data quality in watermarked data so that it is of sufficiently high quality and fit for use in decision making as well as in planning processes in different application domains. Data quality can be defined as the appropriateness of data for its intended applications.

EXISTING SYSTEM

The first reversible watermarking scheme for relational databases was proposed by Zhang and Yang. In this technique, histogram expansion is used for reversible watermarking of relational database.

Histogram expansion technique is used to reversibly watermark the selected nonzero initial digits of errors. This technique is keeps track of overhead information to authenticate data quality. However, this technique is not robust against heavy attacks.

Difference expansion watermarking techniques (DEW), exploit methods of arithmetic operations on numeric features and perform transformations. The watermark information is normally embedded in the LSB of features of relational databases to minimize distortions.

Gupta and Pieprzyks' proposed reversible watermarking technique introduces distortions as a result of the embedding process. Changes in the data are controlled by placing certain bounds on LSB. On the contrary, to limit the distortions, the data outside the limited bounds is left unwatermarked. As a result, the watermark robustness gets compromised.

Sonnleitner proposed a robust, blind, resilient and reversible, image based watermarking scheme for large scale databases. The bit string of an image is used as a watermark where one bit from the bit string is embedded in all tuples of a single partition and the same process is repeated for the rest of the partitions. This technique demonstrates a remarkable decrease in watermark detection rate during various types of heavy attacks, and the database tuples get highly distorted.

PROPOSED SYSTEM

This paper proposed a robust and semi-blind reversible watermarking (RRW) technique for numerical relational data.
RRW 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.

In the watermark encoding phase, the watermark information is embedded in the selected feature(s). Two parameters, b the optimized value from the GA and hr a change matrix are used in the watermark encoding and decoding phases.

Finally, the watermarked data for intended recipients is generated. The attacker channel comprises subset alteration, subset deletion and subset insertion attacks generated by the adversary. These malicious attacks modify the original data and try to degrade its quality.

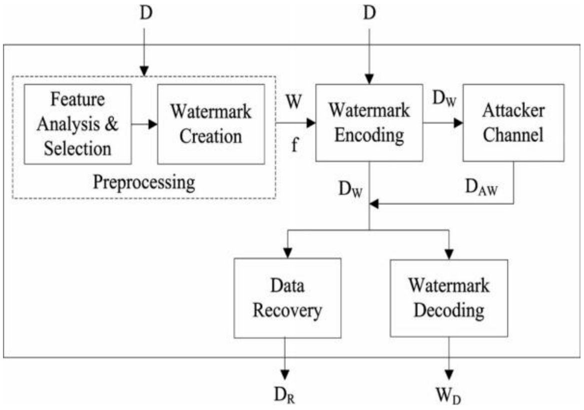
In the watermark decoding phase the embedded watermark is decoded from the suspicious data. In order to achieve this preprocessing step is performed again, and decoding strategies (feature selection on the basis of MI, b the optimized value from the GA and hr the change matrix) are used to recover the watermark.

Semi-blind nature of RRW is used mainly for data reversibility in case of heavy attacks (attacks that may target large number of tuples). Original data is recovered in data recovery phase, through post processing steps for error correction and recovery.

Advantage

- RRW is robust
- The data quality remains intact after watermarking.

SYSTEM ARCHITECTURE:



IMPLEMENTATION

Modules

- Watermark Preprocessing
- Watermark Encoding
- Watermark decoding
- Data Recovery

Module Description:

Watermark Preprocessing

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

Mutual information is calculated for every feature and the lowest feature is selected for watermarking.

Mutual information of every feature with all other features is calculated by using

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

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).

Watermark Creation through Genetic Algorithm (GA)

In this step the genetic algorithm is used to create the watermark bits. The optimal fitness value obtained through GA is basically the change—β—to be embedded in the original data that needs to be watermarked. The purpose of getting an optimum value of b is to justify the amount of change that a feature value can withhold without compromising the data quality.

Watermark Encoding

In this module the optimized value of β is embedded in the particular selected features.

A GA is used to create optimal watermark information that includes: (1) Optimal chromosomal string (watermark string of length *l*); and (2) β value.

β 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 β is used to manipulate the data provided it satisfies the usability constraints λ. The value β 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.

Watermark decoding

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. This module uses a watermark decoder ζ, 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.

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 ηr.

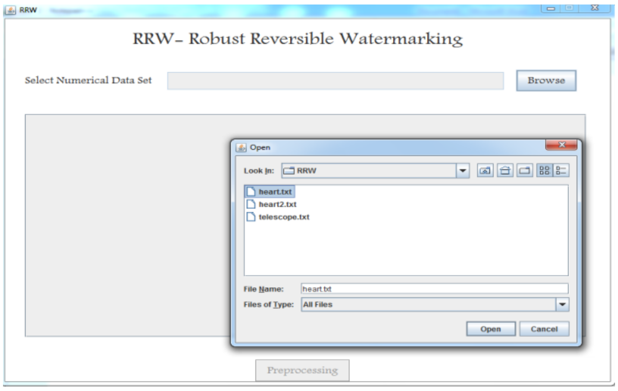
Step 2. The bits are then decoded according to the percentage change values of watermarked data. If $\eta\Delta r \leq 0$, the detected watermark bit will be 1. If $\eta\Delta r > 0$ and $\eta\Delta r \leq 1$, the detected watermark bit will be 0.

The final watermark information is retrieved through a majority voting scheme using
$W_D \Leftarrow mode(dtW(1,2,\ldots,l))$.

Data Recovery

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
$D_r = D'_{Wr} + \beta$
$D_r = D'_{Wr} - \beta$

SCREEN SHOTS



RRW

RRW– Robust Reversible Watermarking

Select Numerical Data Set

Open

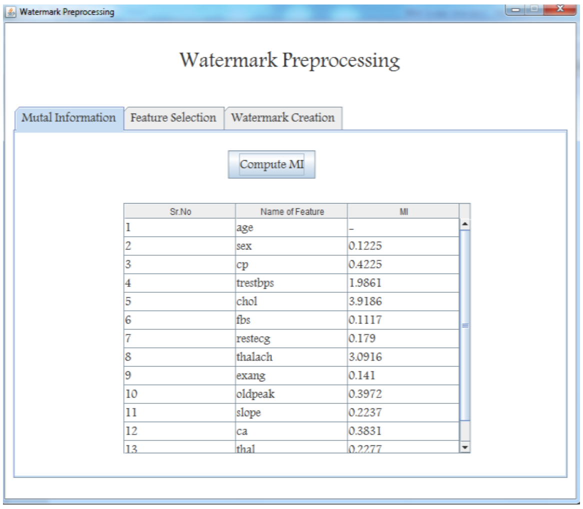
Look In: RRW

heart.txt
heart2.txt
telescope.txt

File Name: heart.txt
Files of Type: All Files

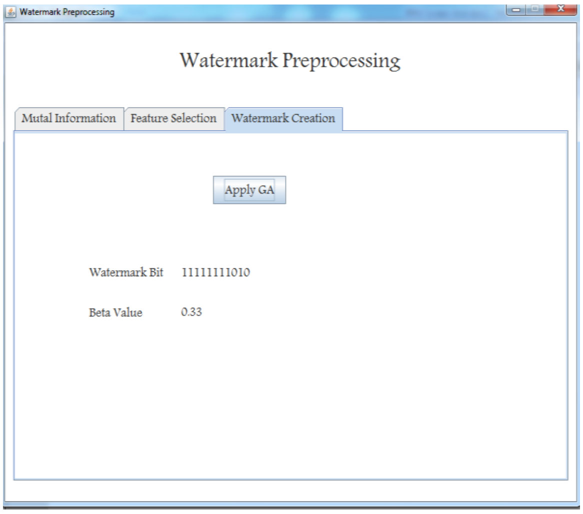
Mutual information:

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal



Sr.No	Name of Feature	MI
1	age	-
2	sex	0.1225
3	cp	0.4225
4	trestbps	1.9861
5	chol	3.9186
6	fbs	0.1117
7	restecg	0.179
8	thalach	3.0916
9	exang	0.141
10	oldpeak	0.3972
11	slope	0.2237
12	ca	0.3831
13	thal	0.2277

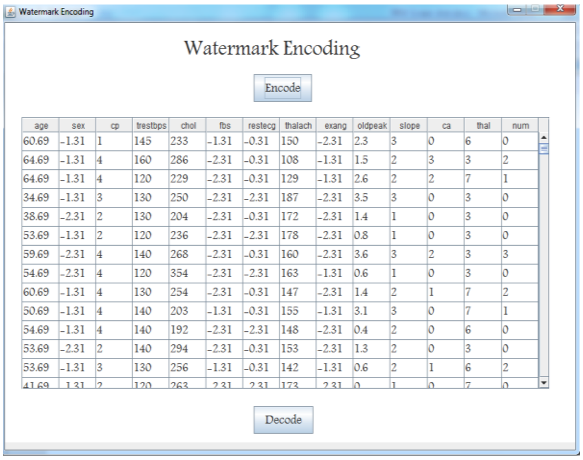
Watermark Creation:



Watermark Bit 11111111010

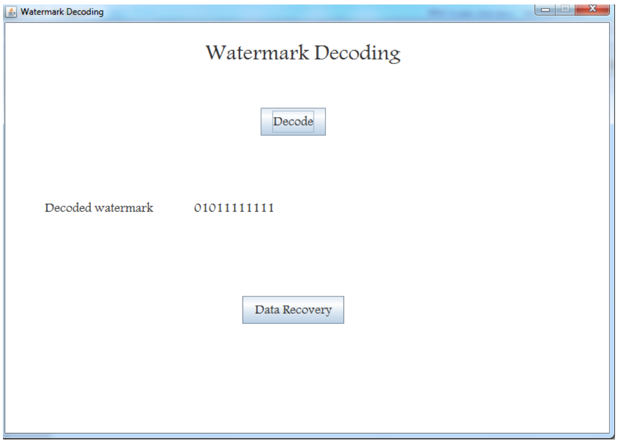
Beta Value 0.33

Watermark Encoding:



age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
60.69	-1.31	1	145	233	-1.31	-0.31	150	-2.31	2.3	3	0	6	0
64.69	-1.31	4	160	286	-2.31	-0.31	108	-1.31	1.5	2	3	3	2
64.69	-1.31	4	120	229	-2.31	-0.31	129	-1.31	2.6	2	2	7	1
34.69	-1.31	3	130	250	-2.31	-2.31	187	-2.31	3.5	3	0	3	0
38.69	-2.31	2	130	204	-2.31	-0.31	172	-2.31	1.4	1	0	3	0
53.69	-1.31	2	120	236	-2.31	-2.31	178	-2.31	0.8	1	0	3	0
59.69	-2.31	4	140	268	-2.31	-0.31	160	-2.31	3.6	3	2	3	3
54.69	-2.31	4	120	354	-2.31	-2.31	163	-1.31	0.6	1	0	3	0
60.69	-1.31	4	130	254	-2.31	-0.31	147	-2.31	1.4	2	1	7	2
50.69	-1.31	4	140	203	-1.31	-0.31	155	-1.31	3.1	3	0	7	1
54.69	-1.31	4	140	192	-2.31	-2.31	148	-2.31	0.4	2	0	6	0
53.69	-2.31	2	140	294	-2.31	-0.31	153	-2.31	1.3	2	0	3	0
53.69	-1.31	3	130	256	-1.31	-0.31	142	-1.31	0.6	2	1	6	2
41.69	-1.31	2	120	263	-2.31	-2.31	173	-2.31	0	1	0	7	0

Watermark Decoding:



Decoded watermark 01011111111

CONCLUSION

Irreversible watermarking techniques make changes in the data to such an extent that data quality gets compromised. 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. However, these techniques are not robust against malicious attacks—particularly those techniques that target some selected tuples for watermarking. In this paper, a novel robust and reversible technique for watermarking numerical data of relational databases is presented. The main contribution of this work is that it allows recovery of a large portion of the data even after being subjected to malicious attacks. RRW is also evaluated through attack analysis where the watermark is detected with maximum decoding accuracy in different scenarios. A number of experiments have been conducted with different number of tuples attacked. The results of the experimental study show that, even if an intruder deletes, adds or alters up to 50 percent of tuples, RRW is able to recover both the embedded watermark and the original data. RRW is compared with recently proposed state-of-the-art techniques such as DEW, GADEW and PEEW to demonstrate that RRW outperforms all of them on different performance merits. One of our future concerns is to watermark shared databases in distributed environments where different members share their data in various proportions. We also plan to extend RRW for non-numeric data stores.

REFERENCES

1. Y. Zhang, B. Yang, and X-M Niu, “Reversible watermarking for relational database authentication,” J. Comput., vol. 17, no. 2, pp. 59–66, 2006

2. G. Gupta and J. Pieprzyk, “Reversible and blind database watermarking using difference expansion,” in Proc. 1st Int. Conf. Forensic Appl. Tech. Telecommun., Inf., Multimedia Workshop, 2008, p. 24

3. 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

4. M. E. Farfoura and S.-J. Horng, “A novel blind reversible method for watermarking relational databases,” in Proc. IEEE Int. Symp. Parallel Distrib. Process. Appl., 2010, pp. 563–569.

5. D. M. Thodi and J. J. Rodriguez, “Expansion embedding techniques for reversible watermarking,” IEEE Trans. Image Process., vol. 16, no. 3, pp. 721–730, Feb. 2007

6. 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, no. 12, pp. 3524– 3533, Dec. 2011.

7. E. Sonnleitner, “A robust watermarking approach for large databases,” in Proc. IEEE First AESS Eur. Conf. Satellite Telecommun., 2012, pp. 1–6.

Author Details

Mr. V. Tharun Kumar received the Bachelor’s of Technology in Computer Science and Engineering from Aurora’s Scientific Technological and Research Academy, JNTU-H. His interest subjects are Database, Data mining, and Operating System.

Mr.L.Srikanth, received the Master of Technology degree in Computer Science & Engineering from the Holy Mary Institute of Technology & Science-JNTUH, He is currently working as Assistant Professor in the Department of CSE with Sphoorthy Engineering College, Hyderabad. His interest subjects are Information Security, Computer Programming, Java, DBMS, Operating Systems, Computer Organization and etc.

Mrs,Deepthi Janagama has done Master of Science in Texas A&M University Commerce, Texas, USA and received the Bachelor Of Technology degree from Kamala Institute Of Technology And Science. She is currently working as Associate Professor and Head of the Department of CSE with Sphoorthy Engineering College, Nadergul. Her interested subjects are Database management system, Data minning and Data Warehousing, Digital logic design, computer programming.