ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

Enhanced and Resilient Watermarking Technique for Non-numerical Relational data

[1] Kshitija G. Patil, [2] Shalu Chopra
[1] M.E Student, Department of I.T, V.E.S Institute of Technology, Mumbai, India,
[2] Head of Department, Department of I.T, V.E.S Institute of Technology, Mumbai, India,

Abstract: **- Nowadays there is very much need of data to be available online for the objective of research studies, current market analysis, predictions to be made for decision making and various other factors. In particular Relational database is shared by owners for the above purposes over the internet that is to data storage locations in a cloud which is being used in the collaborative environment. Consequently, arises a need for a technique to protect the available digital data against security threats like ownership rights and tampering of data. For which Robust and Reversible Watermarking Techniques works towards ensuring ownership rights over shared Numeric relational data and protects against tampering of data without compromising much with data quality having high usability of data and ensures the recovery of original data. The relational database shared by organizations cannot be only limited to have numeric data and they have non-numeric data as well. Therefore this paper proposes and explains in detail the Semi-blind Robust and Reversible watermarking technique for non-numeric data that is Textual data.**

Keywords**: Genetic Algorithm, Mutual Information, Non-Numerical, Robust and Reversible Watermarking.**

I. INTRODUCTION

From a long time, digital watermarking techniques are being used for ownership protection of images, audio, video [1], [2], and natural language processing software's. With increasing need of sharing of databases across the Internet, the same requirement has evolved for relational databases [3]. A large number of techniques available for multimedia watermarking which proved to be effective but these cannot be applied directly to database. As the properties of data are different basic process of watermarking multimedia data differs with watermarking of relational databases. As the relational data is very much independent and discrete but multimedia data is correlated and continuous.

Data owners allow their data to be accessed and used remotely; therefore, may become a victim of data theft. Although, watermarking technology helps them to prove their ownership through identifying data piracy, yet introduces permanent modifications into the data which are irreversible and the watermarked data is different from the original content. Consequently, data analysis and decision making on distorted version of data is not acceptable. There is very much need to preserve the quality of data in watermarked data so that it is of high quality and having high usability with fit for use in decision making as well as in planning processes in different application domains. Reversible watermarking of relational databases is a relatively emerging area. Reversible watermarking techniques overcomes the drawback of degraded data quality by ensuring recovery of original data along with the embedded watermark information. Many reversible watermarking techniques are already available in literature. But these techniques do not allow to selection of features based on their importance in Knowledge discovery and these techniques are either applicable to numeric or non-numeric data. Overcoming the drawbacks of previous techniques came up the Semi-Blind Robust and Reversible Watermarking technique for Numeric Relational data [4] and this paper will focus on extension of RRW to be implemented for Non-numeric data.

II. LITERATURE SURVEY

The First Reversible Watermarking Technique was introduced by Zhang for Relational databases in [5]. This technique focuses on use of Histogram expansion to implement reversible watermarking for relational database. It involves distributing the error between two evenly distributed variables and selected some initial nonzero digits of error to form histogram. This technique involves reversibly watermarking the selected non zero initial digits of errors. It keeps track of overhead information to ensure data quality. But this technique is not effective against heavy attacks which targets large volume of data that is more number of tuples in other words its not robust technique. A Robust and Reversible watermarking Technique came up which was Difference Expansion watermarking technique based on Genetic algorithm (GADEW) [6]. It overcomes the drawbacks of previously introduced Watermarking solutions of DEW (Difference Expansion Watermarking)[7] which involved embedding of

All Rights Reserved © 2018 IJERCSE

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

watermark information in LSB of features of relational databases to reduce distortions, whereas GADEW focuses on increasing watermark capacity and reducing false positive rate to minimize distortions in the data. In this technique basically GA is used for increasing capacity of watermark hence reducing distortions. Overcoming all the drawbacks of already introduced reversible watermarking techniques like DEW, GADEW and PEEW, proposed in [7], [6], [8] respectively, Saman Iftikhar, M. Kamran and Zahid Anwar [4] proposes a semi-blind reversible and robust watermarking technique specifically for numerical relational data. Basically in all previous watermarking techniques attributes or features are selected for watermarking without taking in consideration their role in discovery of knowledge from raw data. And were embedding the watermark in partitions of data to have reduced distortion, which resulted in low quality data being recovered as original data from watermarked data and lacked robustness. Quality of Data is preserved in RRW by taking into consideration relevance of the feature in knowledge discovery while selecting a feature for watermarking that is RRW takes into account mutual information measure for finding out relative importance of features. And in RRW all or large fraction of tuples of selected feature can be watermarked leading to have high robustness against malicious attacks. So basically RRW is configurable allowing the data owner to choose a fraction or large volume of tuples for watermarking but having watermark for all tuples is not required.

As we know the implementation of RRW for numerical relational data, this paper will focus on implementation of RRW for non-numeric data using the same methodology with bit different methodology for textual data as no any database shared by data owners or organization can be with only numeric data it will be with non-numeric data as well there arises the need to have RRW implementation methodology for non-numeric data which this paper focuses on.

III. PROPOSED RRW TECHNIQUE FOR NON-NUMERIC DATA

A. Overview of RRW Technique

Basically, RRW technique works into four phases [4].

(1) Preprocessing Phase- This is basically preprocessing phase of watermark in which watermark to be embedded in Data is generated and the appropriate or most feasible feature to be watermarked is selected;

(2) Encoding Phase – Using the above generated watermark the selected feature is being encoded without affecting the quality of data;

(3) Decoding Phase of Watermark- In this phase, the watermark that is embedded or used for encoding is found out;

(4) Data recovery – It works towards the important task of all recovery of original data.

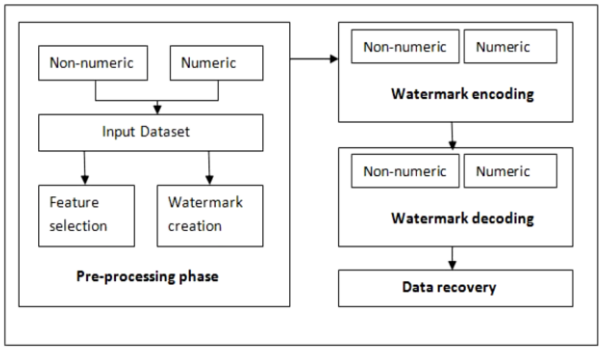

Fig 1- RRW Overview

B. Preprocessing Phase

Preprocessing phase basically focuses on implementation of optimization methodologies for selection of most appropriate features from the relational database to be watermarked based on their relative importance in knowledge extraction from raw data and computation of optimized watermark string using optimization technique of Genetic Algorithm.

a. Finding out appropriate features for watermarking- Based on the Mutual dependency of one feature on another all the features from the dataset are ranked to develop a model which will help to take decision which feature is more suitable for watermarking. So to compute mutual dependency the statistical measure of Mutual Information(MI) is computed for features having Numeric data values. MI is responsible for calculation of mutual dependence between two random variables.

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

MI basically works to compute Marginal and Joint Probability distributions for finding out the mutual dependency on one another and having the features ranked accordingly as per MI values. And for Non-Numeric Features in place of MI that is marginal and joint probability, Entropy is calculated which serves the same purpose as MI to find the most suitable feature for watermarking from the databases for non-numeric features.

Lets take an example to understand finding out of MI value for Non-numeric feature- Considering one attribute "cp" and its 4 tuples from a Health database.

Cp
type_angina
Asympt
non_anginal
atyp_angina

All Rights Reserved © 2018 IJERCSE

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

Step1:- Count how many times a tuple value repeats itself in a particular feature column like typ_angina is coming 1 time and etc and total number of rows in our case is 5. Similarly for other tuple values its as follows in below table I

Table I

Tuple values	typ_ angina	Asympt	Non-anginal	Atyp_angina
No. of occurrences	1	2	1	1
No. of occurrences/ Total No. of tuples	1/5	2/5	1/5	1/5
Probabilities	0.2	0.4	0.2	0.2

Step 2:- Find the subsets with different combinations of tuple values as (pr1,pr2) for a attribute
Subset1- (typ_angina, Asympt)(typ_angin,non_anginal)
(typ_angina, atyp_angina).
Subset 2- (Asympt, non_anginal)(Asympt, atyp_angina).
Subset 3- (non_anginal ,atyp_angina)
Step 3:- Follow this algo for all the subset combinations above.

```
If probability of pr1== probability of pr2
{
MI = pr1 * ((pr1 * pr2) / pr1);
}
Else {
MI = pr1 * pr2;
}
MI (cp) = MI (cp) + MI;
End;
```

Following the above steps for each combinations in subsets comparing the values in Table I we get the values for each of the combinations in Subsets are as follows-
Subset1- (0.2*0.4) + (0.2*(0.2*0.2)/0.2=0.040) +
(0.2*(0.2*0.2)/0.2=0.040)
Subset2- (0.080=0.4*0.2) + (0.080=0.4*0.2)
Subset3- (0.2*(0.2*0.2)/0.2=0.040)
So adding all the values we got above for each of them we get MI value for the non-numeric feature "cp" as 0.080+0.040+0.040+0.080+0.080+0.040 = 0.36
Therefore MI (cp)= 0.36
In this way MI values for each non-numeric feature can be found out.

Then as Prediction of feature with lowest MI will be easier for an attacker in an attempt to identify watermarked feature, so for this purpose a secret threshold is set based on the MI of all features in database. So the features having MI lower than threshold will be chosen for watermarking.

b. Watermark Creation using Genetic Algorithm

Optimal Watermark to be used for subsequent phases of encoding and decoding is generated using Genetic Algorithm [9].
Optimized Watermark string is generated based on the implementation of following steps of Optimization technique of Genetic Algorithm
1) Initialization- Generate Random population of Binary Strings called Chromosomes. Gene values of each chromosome represent l-bit watermark string.
2) Evaluation of fitness- Calculate Fitness of each chromosome by employing a constrained optimized fitness function.
3) Selection of the fittest chromosome as a parent chromosomes- Tournament selection mechanism is applied for this.
4) Crossover and Mutation- Genetic operations of crossover and mutation are performed on parent chromosomes to create offsprings. 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) Implementation of elitism- Elitism strategy is applied to hire two individuals with best fitness value; as elites to the next generation without genetic changes. 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. Digital numbers of each non-numeric attribute we are embedding the watermark information.
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 MI of original data and MI of watermarked data 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 fulfilment of the termination criteria.

All Rights Reserved © 2018 IJERCSE

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 3, March 2018

C. Encoding Phase

Data owner can select any number of features for watermark embedding based upon secret threshold, MI and entropy of the feature(s). Encoding for non-numeric data is done as: In this watermark generated in preprocessing step and which is different every time is used to manipulate data. Each character of selected non numeric feature is converted into its ASCII equivalent. Then generated watermark is converted into its integer equivalent. Additions of these two values are performed. Then this addition is converted into equivalent character.

D. Decoding Phase

In this phase of Watermark Decoding first step is taken in finding out the attributes that are watermarked. So in this decoding of watermark string is done. Decoding is done bit by bit on watermarked database based on the change matrix which has percentage of change in data values and based on the knowledge of length of watermark. A Watermark string is generated from every tuple. Consequently final watermark string is regained with the help of a majority voting scheme. And basically this phase is with similar way of implementations for numerical[4] and non-numerical.
So output of this phase is the original watermark string which was used for encoding of original data to make it watermarked data.

E. Data Recovery Phase

For non-numeric data, Detected watermark is used to manipulate data. Each character of selected non numeric feature is converted into its ASCII equivalent. Then detected watermark is converted into its integer equivalent. Subtractions of these two values are performed. Then this subtraction is converted into equivalent character.

III. CONCLUSION

Irreversible watermarking techniques had led to data quality getting compromised. Reversible watermarking is a solution to this problem because they are able to recover original data from watermarked data and ensure data quality to some extent. Then except one technique that is semiblind robust and Reversible watermarking technique other techniques do not consider watermark encoding and decoding by accounting the role of the features in knowledge discovery. Proposed technique is extension of this technique. This paper basically focused in detail how RRW can be applied for non-numeric relational data as the shared relational data from data owners cannot be with only numeric values.

REFERENCES

[1] S. Subramanya and B. K. Yi, "Digital rights management," IEEE Potentials, vol. 25, no. 2, pp. 31–34, Mar.-Apr. 2006.

[2] J. T. Brassil, S. Low, and N. F. Maxemchuk, "Copyright protection for the electronic distribution of text documents," Proc. IEEE, vol. 87, no. 7, pp. 1181–1196, Jul. 1999.

[3] Y.-C. Liu, Y.-T. Ma, H.-S. Zhang, D.-Y. Li, and G.-S. Chen, "A method for trust management in cloud computing: Data coloring by cloud watermarking," Int. J. Autom. Comput., vol. 8, no. 3, pp. 280–285, 2011.

[4] SamanIftikhar, M. Kamran, and Zahid Anwar, ―RRW—A Robust and Reversible Watermarking Technique forRelational Data,‖ In IEEE Trans. on knowledge and Data Engineering, VOL. 27, NO. 4, APRIL 2015.

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

[6] 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

[7] G. Gupta and J. Pieprzyk, "Reversible and blind database watermarkingusing difference expansion," in Proc. 1st Int. Conf. Forensic Appl. Tech. Telecommun., Inf., Multimedia Workshop, 2008, p. 24.

[8] 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.

[9] M. Mitchell, An introduction to genetic algorithms. Cambridge, MA, USA: MIT Press, 1996

All Rights Reserved © 2018 IJERCSE