Relational Data based on Watermarking Technique

[1]Shubham Dahiwal, [2]Vaibhav M. Dhanve, [3]Pratik Borkar, [4]Prof. Smita Makade
[1,2,3,4]Smt. Indira Gandhi College of Engineering, Navi Mumbai, Maharashtra, India.
[1]dahiwal.shubham@gmail.com, [2]vaibhavmdhanve@gmail.com, [3]pratik12951@gmail.com,
[4]smita.sakharwade@gmail.com

Abstract— Data is excessively generated by information systems these days. Compromise of ownership, data tampering may occur in relational databases. Watermarking is a solution used to overcome this issue. Robust and Reversible Watermarking technique is a tactic applied to assure recovery as well as the quality of the data. But, these techniques usually lacks resilience in front of malevolent attacks and selectively watermarking a particular attribute is not possible. So, reversible watermarking is needed that assures; (i) encoding and decoding of watermarks by considering the role of knowledge discovery features; and, (ii) recovery of actual data while attacks are done. In this paper, a watermarking technique for numerical and string relational data has been proposed that addresses the above concerns. Experiments accords the robustness against malicious attacks and show that the proposed technique is more efficient.

Keywords - Genetic algorithm, data recovery, data quality, robustness, numerical data, non-numerical data.

1. INTRODUCTION

In this era of digital devices, vast amount of data generated because of high use of internet and cloud computing [6]. sharing of relational data to the research communities by owners, ownership need to be maintained. For that, watermarking on relational data need to be applied. 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 knowledge discovery into account.

The technique tries to solve the complication of data quality deterioration, by permitting the actual data's recovery with the embedded watermark information [11]. This paper contains one such watermarking technique that keeps the data worth for discovery of knowledge [3]. Modifications of data are considered to an extent such that the data quality, prior to embedding and post extraction, is satisfactory for the extraction of knowledge [7]. Consequently, the discovery of knowledge turns successful in decision support systems where high quality data recovery is important [3]. Achieving attack resilience in such a scenario is a defiance work as the two features of data recovery and assurance of robustness , may be potentially opposing [1]. But, we attempt for finding the most appropriate way to handle this task for both numerical and non-numerical data.

1.1 Problem Statement

The proposed watermarking technique is used to assure the data quality as well as the recovery. But, these techniques mostly fail to be resilient while conflicting with malevolent attacks and there is no provision for selectively watermarking specific attribute in selective manner by considering the role of the selected attributes in the discovery of knowledge [8].

In the existing system, there is no provision to handle the non-numerical (string) data.

1.2 Existing system

Experiments accords for the effectuality of the implemented technique, showing that the proposed technique performs better than the existing ones. Existing techniques, do not consider the non-numerical data. RRW outperforms existing state of the art reversible watermarking techniques including Difference Expansion Watermarking, Genetic Algorithm Difference Expansion Watermarking and Prediction Error Expansion Watermarking [4].

1.3 Proposed System

The main aim of this project is to maintain the ownership of Relational Data and also minimizing distortion in the watermarked content. A novel watermarking technique for numerical as well as non-numerical relational data has been presented, with respect to the above needs. This paper introduces one such watermarking technique that retains the

usefulness of data. Genetic algorithm is proposed in this paper to reach a best solution that is acceptable for the problem, without violating the defined problems [1].

Advantage of Proposed System

1. More diverse type of relational data can be worked upon viz. the non-numerical string data that is introduced in this paper.

2. Reversible watermarking techniques can guarantee retrieval of actual data as well as preserving the legitimacy of the owners [10].

II. PLANNING & FORMULATION

2.1 Architecture

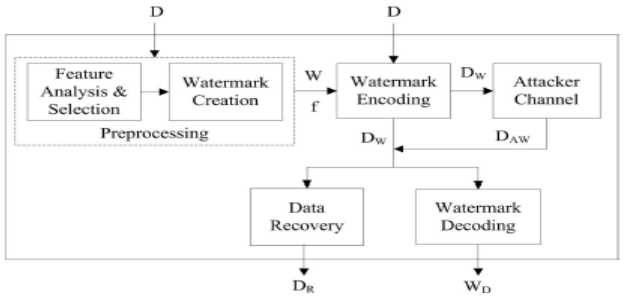

Fig 1. Main Architechture of watermarking technique.

The watermark pre-processing phase computes different parameters for calculation of an optimal watermark, encoding and decoding. The basis of encoding phase is that the embedded watermark information gets embedded while retaining the data quality. Data gets altered during the embedding process as per the availability of the bandwidth (or capacity) of the watermark information [9]. The watermark bandwidth must be adequately, be large enough to assure resilience but should not be magnified to an extent, that the data quality gets deteriorated [5]. 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. Numerical features can be taken under consideration from any dataset and a suitable feature is determined to embed watermark on the basis of mutual information [1]. After watermarking, the data is released to the intended recipients over a communication channel that is assumed to be insecure and termed as the [1] "attacker channel" in this research domain [1]. The data may undergo several malicious attacks in the attacker channel [1]. The efficiency and effectiveness is

described through robustness analysis determined by its response to subset insertion, alteration and deletion attacks[2]. The Watermark decoding phase recovers watermark information effectively for detection of the embedded watermark [1]. Data recovery phase mainly comprises the important task of successful recovery of the original data [1].

III. ALGORITHM DEVELOPMENT

PROPOSED SYSTEM ALGORITHMS

The presence of data quality constraint [3]. Lately done research profess Genetic Algorithm, which is one of the computational intelligence techniques [3].

GA - an optimization algorithm is employed in the robust and reversible watermarking technique (RRW) proposed in this paper to achieve an optimal solution that is feasible for the problem at hand and does not violate the defined constraints

Genetic Algorithm based on Difference Expansion watermarking [2] (GADEW)

technique is used in a proposed robust and reversible solution for relational data.

The watermark encoding algorithm starts the embedding process with the MSB (most significant bit) of the watermark [1]. For this reason, the algorithm considers one

Notations Used in the Paper

Symbol	Description	Symbol	Description
D	Original database	b	The watermark bit
D_W	Watermarked database	$min(a_w)$	The minimum value a of a feature after watermarking
R	Total number of tuples/rows/records in a table (or dataset)	r	A tuple in the database table
η_r	Detected amount of percentage change in encoding	∇	A matrix containing percent change in data values
$max(a)$	The maximum value a of a feature	ζ	The watermark encoder used for watermark encoding
A	a feature/column/attribute selected for watermarking (D)	D'_W	A watermarked database after the malicious attacks
$min(a)$	The minimum value of a feature	$max(a_w)$	The maximum value of a feature a after watermarking
l	The length of the watermark	η_{d_r}	Detected amount of change in the value of a feature after an attack on the watermark bit b
w	Watermark bits	η_{Δ_r}	The difference between the changes detected in the value of a feature during the encoding and decoding process
A_W	The watermarked feature	W_D	Decoded watermark
F	Total number of features in the database	D_W	D watermarked by the proposed scheme
dtW	detected watermark bit	D_r	Recovered Data
MI_O	Mutual information of original data	MI_W	Mutual information of watermarked data
Acc_O	Classification accuracy of original data	Acc_W	Classification accuracy of watermarked data
Sn_O	Sensitivity of original data	Sn_W	Sensitivity of watermarked data
Sp_O	Specificity of original data	Sp_W	Specificity of watermarked data
Δ_{MI}	Change in value of mutual information	a	value of feature A
β	An optimized value to watermark a feature	λ	The usability constraints defined by the data owner
ζ	The watermark decoder used for watermark decoding	RRW	proposed Robust and Reversible Watermarking technique
μ_D	Mean of the original data of RRW	σ_D	Variance of the original data of RRW
μ_{D_W}	Mean of the watermarked data of RRW	σ_{D_W}	Variance of the watermarked data of RRW
μ_d	Mean of the original data of PEEW Technique	σ_d	Variance of the original data of PEEW Technique
μ_{d_w}	Mean of the watermarked data of PEEW Technique	σ_{d_w}	Variance of the watermarked data of PEEW Technique

tuple at a time [1].

The algorithm for watermark decoding and data recovery algorithm is proposed [1]. In the watermarking creating period, we follow the steps of the GA, shown below, for obtaining the information of optimal watermark [1]:

1) Randomly, the first population of binary strings is formed. Gene values of each chromosome represents l-bit watermark string [1].
2) Evaluation of each chromosome is done to obtain fitness by appointing a fitness function, which is optimized and constrained [1].
3) Tournament selection mechanism is applied to get the most appropriate individuals as parent chromosomes.

4) Crossover and mutation are done on peer chromosomes to generate off-springs. A crossover is done so that high quality entities, deriving their peer characteristics, by exchanging information between more than or equal to two chromosomes [1]. Mutation is performed so as to diversify the population by performing changes in values of the genes of binary chromosomes [1]. The rate of mutation and crossover fraction values are set factually [5].

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

6) Residual population of the new generation is generated by replacing less fit entities of the older generation with the more-fit freshly generated childs.

7) Above steps are performed again and again until MIO and MIW approximates to an equal value for noticeable numer of generations.

8) Both, best amongst fitness function values and the best watermark information string, is returned after the fulfilment of the termination criteria.

3.1 Watermark Encoding Phase

Watermark information calculation is formulated as a CO problem to meet the data quality constraint of the data owner. We use Genetic Algorithm to obtain most favourable watermark information that contains: (1) Optimal chromosomal string (watermark string of length l); and (2) b value [1]. b is a parameter that is formulated using GA and represents a forbearing amount of change to embed in the feature values [1]. Once the optimum value of b for each candidate feature A is found, it is saved for use during watermark encoding and decoding. A bit string watermark and an optimum value (b) is used to handle the data, with satisfaction of usability constraints [1]. The value b is added into every tuple of the selected feature A when a given bit is 0; else, the value of b is decremented from the feature value [3].

Algorithm 1. Watermark Encoding

Algorithm 1. Watermark Encoding

Input: D, w, β
Output: D_W, ∇
for $w = 1$ to l **do**
//loop will iterate for all watermark bits w from 1 to length l of the watermark
for $r = 1$ to R **do**
//loop will iterate for all tuples of the data
if $b_{r,w} == 0$ **then**
// the case when the watermark bit is 0
changes are calculated by using Equation

$$\eta_r = D_r * \zeta.$$

data is watermarked by using Equation

$$D_{W_r} = D_r + \beta.$$

insert η_r *into* ∇
end if

if $b_{r,w} == 1$ **then**
// the case when the watermark bit is 1
changes are calculated by using Equation

$$\eta_r = D_r * \zeta.$$

data is watermarked by using Equation

$$D_{W_r} = D_r - \beta.$$

insert η_r *into* ∇
end if
end for
end for
return D_W, ∇

Algorithm For Non Numeric (String)Data:

Input: Non Numeric (String) data, First 3 bit of watermark bits(starting from MSB).

Output: Encoded Non-Numeric(String)Data.

Step 1: Loop Will Iterate through watermark bits.
Step 2: Store 3 watermark Bits in temp variable.
Step 3: Loop will iterate through all Non-numerical(string) data.
Step 4: Store all data(String) into array, characterwise.
Step 5: Consider one character at a time.
Step 6: Convert considered character into ASCII form & add '2' into ASCII value.
Step 7: Store above value into another Temp2 variable.
Step 8: Convert 3 Stored watermark bits(Stored in Temp) into Decimal(Integer) Format & store it into Temp3 Variable.
Step 9: Add Temp3 value and Temp2 value & store it in Temp4 variable.
Step 10: if((Temp4 >= 65 && Temp <= 90) || (Temp>=97 && Temp4<=122))
{ Convert Temp4's value(ASCII) into Non-Numeric(String) equivalent.
} else
{ No further encryption possible and will be encoded as it is.
}

3.2 Watermark Decoding Phase

In the watermark decoding process, the first step is to locate the features which have been marked. In Encoding phase the optimization process through GA is not needed

[2][3]. 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 decoding is performed on a single at a time by watermark decoder [1]. In the decoding phase, hdr is calculated and constitute the change found in the watermarked data. The value of hdr , hr and hDr is calculated using the values of tuple r and so, it could be different for every r [1]. The parameter hDr is computed by calculating the difference between the original data change amount hr and the watermark detected change amount hdr.

Algorithm 2. Watermark Decoding

Algorithm 2. Watermark Decoding

Input: D_W or D'_W, ∇, l
Output: W_D
  for $r = 1$ to R **do**
    //loop will iterate for all tuples of the data
    for $b = l$ to 1 **do**
      //loop will iterate for all watermark bits b from 1 to length l of the watermark
      $\eta_{d_r} \Leftarrow D'_{W(r)} * \zeta$
      $\eta_{\Delta_r} \Leftarrow \eta_{d_r} - \eta_r$
      if $\eta_{\Delta_r} \leq 0$ **then**
        detected watermark bit (dtW) is 1
      else if $\eta_{\Delta_r} > 0$ and $\eta_{\Delta_r} \leq 1$ **then**
        detected watermark bit (dtW) is 0
      end if
    end for
  end for
  $W_D \Leftarrow mode(dtW(1, 2, ..., l))$
  return W_D

3.3 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 [1]. The following Algorithm 3 presents the data recovery mechanism [1].

Algorithm 3. Data Recovery

Algorithm 3. Data Recovery

Input: D_W or D'_W, b
Output: D_r
  for $r = 1$ to R **do**
    //loop will iterate for all tuples of the data
    for $b = l$ to 1 **do**
      //loop will iterate for all watermark bits b from 1 to length l of the watermark
      if $dtW(r, b) == 1$ **then**
        // 0 or 1 watermark bit is detected from every tuple r
        data is recovered by using Equation

$$D_r = D'_{W_r} + \beta,$$

      else
        data is recovered by using Equation

$$D_r = D'_{W_r} - \beta.$$

      end if
    end for
  end for
  return D_r

IV. CONCLUSION

Reversible watermarking techniques 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 technique for watermarking numerical as well as non-numerical (string) data of relational data is presented. The main contribution of this work is that it allows recovery of a large portion of the both numerical & non-numerical data even after being subjected to malicious attacks. It is also evaluated through attack analysis where the watermark is detected with maximum decoding accuracy in different scenarios.

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

REFERENCES

[1] *RRW- A Robust and Reversible Watermarking Technique for Relational Data*., Saman Iftikhar, M. Kamran, and Zahid Anwar. April 2015.

[2] *Genetic algorithm and difference expansion based reversible watermarking for relational databases*., Khurram Jawad, Asifullah Khan., 2013.

[3] *Reversible Fragile Database Watermarking Technology using Difference Expansion Based on SVR Prediction*., Jung-Nan Chang, Hsien-Chu Wu., 2012.

[4] *A blind reversible method for watermarking relational databases based on a time-stamping protocol*., Mahmoud E. Farfoura, Shi-Jinn Horng, Jui-Lin Lai, Ray-Shine Run, Rong-Jian Chen, Muhammad Khurram Khan., 2012.

[5] *A robust watermarking approach for large databases*., Erik Sonnleitner., 2012.

[6] *A Method for Trust Management in Cloud Computing: Data Coloring by Cloud Watermarking*., Yu-Chao Liu, Yu-Tao Ma, Hai-Su Zhang, De-Yi Li, Gui-Sheng Chen., August 2011..

[7] *Efficient Reversible Watermarking Based on Adaptive Prediction-Error Expansion and Pixel Selection*., Xiaolong Li, Bin Yang, and Tieyong Zeng., Dec 2011.

[8] *A New Approach for Relational Database Watermarking Using Image*., Hossein Moradian, Sardroudi Subariah Ibrahim., 2010.

[9] *Expansion Embedding Techniques for Reversible Watermarking* Diljith M. Thodi and Jeffrey J. Rodríguez, Senior Member, IEEE., 2007.

[10] I. Cox, M. Miller, J. Bloom, and M. Miller, *Digital Watermarking. Burlington*, MA, USA: Morgan Kaufmann, 2001.

[11] M. Kamran and M. Farooq, "*An information-preserving watermarking scheme for right protection of EMR systems*," IEEE Trans. Knowl. Data Eng., vol. 24, no. 11, pp. 1950–1962, Nov. 2012.

