The Journal of Systems and Software 86 (2013) 2742–2753

Contents lists available at ScienceDirect

The Journal of Systems and Software

journal homepage: www.elsevier.com/locate/jss

Genetic algorithm and difference expansion based reversible watermarking for relational databases

Khurram Jawad, Asifullah Khan*

Pattern Recognition Lab, Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Nilore, 45650 Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 25 April 2012
Received in revised form 26 May 2013
Accepted 6 June 2013
Available online 28 June 2013

Keywords:
Relational database
Reversible watermarking
Difference expansion
Genetic algorithm
Robust watermarking
Distortion
Watermark capacity
Watermark attack

ABSTRACT

In this paper, we present a new robust and reversible watermarking approach for the protection of relational databases. Our approach is based on the idea of difference expansion and utilizes genetic algorithm (GA) to improve watermark capacity and reduce distortion. The proposed approach is reversible and therefore, distortion introduced after watermark insertion can be fully restored. Using GA, different attributes are explored to meet the optimal criteria rather than selecting less effective attributes for watermark insertion. Checking only the distortion tolerance of two attributes for a selected tuple may not be useful for watermark capacity and distortion therefore, distortion tolerance of different attributes are explored. Distortion caused by difference expansion can help an attacker to predict watermarked attribute. Thus, we have incorporated tuple and attribute-wise distortion in the fitness function of GA, making it tough for an attacker to predict watermarked attribute. From experimental analysis, it is concluded that the proposed technique provides improved capacity and reduced distortion compared to existing approaches. Problem of false positives and change in attribute order at detection side is also resolved. Additionally, the proposed technique is resilient against a wide range of attacks such as addition, deletion, sorting, bit flipping, tuple-wise-multifaceted, attribute-wise-multifaceted, and additive attacks.

1. Introduction

Easy access to internet has boosted the growth of business and research. Nowadays, sharing information online is an important activity for business and research, which also involves buying/selling of databases. For example, sharing of data related to weather, stock market, power consumption, consumer behavior, medical, scientific, etc. is frequently performed. Consequently, there is a great need for providing security of databases to discourage illegal copying and distribution in today's internet based application environment (Chamlawi et al., 2010). In this context, proof of ownership and tamper-proof-transportation of the databases are the most challenging issues these days (Sion et al., 2004).

Watermarking provides solutions for many of the problems encountered in the distribution of different multimedia objects such as, image, text, and audio (Cox et al., 2002). Similarly, watermarking is effective in protection of relational databases. However, a common problem with relational database watermarking is the available bandwidth for watermark insertion, which is very limited compared to other multimedia objects. The data in relational databases can tolerate very small distortion. Increasing the amount of distortion may cause the values to become meaningless.

Rakesh and Jerry (2002) did initial efforts in the domain of database watermarking to hide watermark bits in the least significant bit (LSB) of its target values. Their technique is based on numeric set watermarking. It was assumed that the relational database can allow small amount of distortion in a database. However, inserting watermark bit into LSB may result in loss of data as it can be easily manipulated by the attacker.

In another study, Sion et al. (2004) targeted the collection of selected tuples to hide watermark bits into partition statistics. Statistics were changed according to distortion tolerance (usability constraints). Distortion tolerance was responsible to keep check on the values of the attributes so that the change did not exceed a limit. Shehab et al. (2008) further enhanced the above method by using optimization techniques.

Genetic algorithm (GA) and pattern search (PS) were used to insert watermark in statistics of relational database by minimizing or maximizing the hiding function, while keeping distortion tolerance intact. Further, Mailing et al. (2008) employed GA in watermark signal processes to embed watermark in data statistics. However, their focus was to make watermarking signal more

* Corresponding author. Tel.: +92 512207380; fax: +92 512208070.
E-mail addresses: khurramjawad@pieas.edu.pk (K. Jawad), asif@pieas.edu.pk, khan.asifullah@gmail.com (A. Khan).

0164-1212/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jss.2013.06.023

correlated to the original database and thus make watermark detection easy. The shortcoming of aforementioned watermarking techniques is that they are not able to recover the original cover work exactly from the watermarked data. This problem was solved by the introduction of reversible watermarking techniques in the domain of relational databases.

Difference expansion based watermarking (DEW) technique was used by Gupta and Pieprzyk (2008) to achieve reversibility in context of relational databases. DEW is able to restore the original database exactly. Additionally, it also allows adding distortion into the database using distortion tolerance of the attribute. It also encourages the owner to distribute the trial version of the database, which can only be reverted by those users who have purchased the key. Similarly, Gupta et al. (2009) solved the problem of secondary watermarking attack by using reversible watermarking.

Previous watermarking techniques (Chang and Lin, 2008; Gupta et al., 2009; Jun, 2003; Wu et al., 2009) do not explore the combination of different attributes for suitability, rather these techniques use predefined criteria for the selection of attributes. However, insertion of watermark into the database depends upon the distortion tolerance of the selected attribute. If changes are within the distortion limit then, the watermark can be embedded. Otherwise, the attribute is left unwatermarked. This is the essence of our current methodology that we are able to explore the combination of different attributes for suitability of watermark bit insertion.

Evolutionary techniques like GA have been successfully used for optimization in the area of pattern recognition (Afridi et al., 2011; Hayat and Khan, 2011; Khan and Mirza, 2007; Naveed and Khan, 2011; Tahir et al., 2011). Our aim is thus to use GA for improving overall performance of the watermarking system, by making an optimal/near optimal tradeoff between the basic properties of a watermarking system (Arsalan et al., 2011; Khan, 2006; Khan et al., 2006, 2008). We have used GA to improve the watermark capacity and reduce distortion of the database, while keeping distortion tolerance fixed.

DEW technique can introduce distortion in a cover work, which can make watermarked attribute more visible to the attacker. One can therefore, reduce distortion by using small value for distortion tolerance, however, this may result in limited amount of watermark embedding into the cover work. In order to keep the watermarked attribute hidden from the attacker, we thus select those attributes that are hidden in the neighborhood and cause minimum distortion in the cover work.

Another problem with DEW is the substantial false positive rate on the detection side. Due to the false positives, exact extraction of the watermark and restoration of the original data is not possible at detection side. However, we have observed that it is easy to recover the original watermark and original data exactly using a semi blind technique for watermark recovery. Therefore, the problem of false positive is resolved using side information. Similarly, changing order of attributes does not affect the usability of a relation in a database. However, reshuffling attributes at detection side may affect the process of watermark detection (Halder et al., 2010). This problem can also be resolved by passing the order of attributes to the detection module.

Existing DEW approach is mostly unable to increase watermark capacity of the relation without increasing distortion tolerance of the attributes. Attacker can thus use distortion to predict marked attributes, which may affect successful detection of the watermark. Consequently, false positives and changing order of attributes cannot be tackled at the detection side. In order to solve the above problems for database watermarking, we thus propose a GA and difference expansion based watermarking (GADEW) technique. Mean and standard deviation of the watermarked relation are used to evaluate distortion in the attributes. False positives are eliminated and problem of shuffling attributes at detection side is resolved.

Table 1
Abbreviations used in this paper.

TV	Target value
CV	Changed value
GADEW	Genetic algorithm and difference expansion based watermarking
DEW	Difference expansion based watermarking
MAC	Message authentication code
CrC	Capacity related cost
TwD	Tuple-wise distortion
AwD	Attribute-wise distortion
TC	Total cost
AV	Average value
R-dataset	Random-dataset
FCT-dataset	Forest cover type dataset
Sk	Secret key
Pk	Primary key
Std	Standard deviation
OrD	Original database
M_GADEW	Attribute wise mean of watermarked database using GADEW
M_OrD	Attribute wise mean of OrD
M_DEW	Attribute wise mean of watermarked database using DEW
S_GADEW	Attribute wise Std of watermarked database using GADEW
S_OrD	Attribute wise Std of original database
S_DEW	Attribute wise Std of watermarked database using DEW

Farfoura et al.'s prediction error expansion (PEE) is an interesting technique that uses single attribute of a relation to insert watermark, which makes it distinctive among other reversible watermarking techniques (Farfoura et al., 2012). They used only fractional portion of the numeric attribute for inserting watermark in relational databases. However, an attacker may find it easy to attack fraction portion of the numeric attribute without affecting the usability of the data. Therefore, we have used integer portion of the numeric attribute for watermarking purpose.

Farfoura et al. have used a randomly generated data set for their experiments. While we have used the Forest Cover Type dataset (FCT-dataset) that contains numeric attributes, and is used by number of database watermarking techniques (Rakesh and Jerry, 2002; Rakesh et al., 2003). PEE approach provides no check for distortion tolerance of an attribute, as a result, values can exceed distortion limit of the attribute and can affect its usability. Every selected tuple will be watermarked, if distortion tolerance of the attribute is not checked. Therefore, detection algorithm extracts watermark bit from each selected tuple. As a result, addition attack can cause high rate of false positives.

Proposed GADEW technique is able to increase watermarking capacity of the relational database at a fixed distortion tolerance. Distortion tolerance enforces limits on each attribute so that the value may not lose its meaning during watermark insertion. Distortion introduced due to watermark insertion is reduced to minimum by introducing tuple-wise and attribute-wise distortion measures. GADEW is a reversible watermarking technique that recovers both watermark and cover work exactly as it was before watermark insertion. Additionally, it is robust against different attacks including, addition, deletion, sorting, bit flipping, tuple-wise-multifaceted, attribute-wise-multifaceted, and additive attacks. Random selection of attributes also makes it tough for the attacker to predict watermark. Problem of the false positive detection is resolved and even addition attack does not result in false positive detection. The different notations used in the rest of the paper are listed in Table 1.

The rest of the paper is organized as follows: in Section 2, we provide details about the DEW technique. In Section 3, we focus on use of GA to improve the watermark capacity, invisibility, and to reduce distortion. Mapping of chromosome and fitness function of GA are also explained. Section 4 presents results and discussion. Finally, conclusions are made in Section 5.