International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 07 | July -2017 www.irjet.net p-ISSN: 2395-0072

Genetic Algorithm Based Reversible Watermarking Approach for Numeric and Non-Numeric Relational Data

Gayatri R. Ghogare[1], Prof. Aparna Junnarkar [2]

[1]Student, ME, Department of Computer Engineering, P.E.S Modern College of Engineering, Shivajinagar, Pune-05, Maharashtra, India.
[2]Assistant Professor, Department of Computer Engineering, P.E.S Modern College of Engineering, Shivajinagar, Pune-05, Maharashtra, India.

Abstract - *With the course of time, immense growth in the use of Internet has led to the generation and consumption of huge amount of relational data. Relational databases are made remotely available for the users within collaborative environments for knowledge gain and decision making purposes. But, eventually, this has led to the relational data thefts and data degradation threats. Reversible watermarking has proved a best candidate solution for overcoming such tedious threats. Reversible watermarking schemes are used for imposing ownership rights and tackling data tampering. But, these schemes are not effective against protecting relational databases from active malicious attacks such as insertion, alteration and deletion attacks. Also, relational data is modified to a very large extent resulting in data quality degradation. In this paper, a genetic algorithm(GA) based reversible watermarking scheme for numeric and non-numeric(text) relational data is proposed. This scheme ensures: i) recovery of original relational data along with the embedded watermark; i) data quality; and iii) attack resilience against various malicious attacks. Experimental studies prove the effectiveness and performance of the scheme for both numeric as well as non-numeric relational database.*

Keywords: **Data Quality, Data Recovery, Feature Extraction, Genetic Algorithm, Numerical Data, Non-numerical Data, Reversible Watermarking, Robustness, Relational Databases.**

1. INTRODUCTION

Digitization has encouraged various organizations in making their relational databases openly accessible over the Internet. But with this, organizations have to face various relational database related threats such as data loss, data degradation, malicious attacks on databases.,etc. Digital watermarking for relational databases has proved to be a prominent solution for imposing copyright protection, tamper detection and preserving database integrity. However, the core process involved in watermarking relational databases is far different than that of the multimedia watermarking due to difference in the type of data, structure of data storage, etc.

Watermarking of relational databases needs to be performed in such a way that the data quality will not get compromised and it will be later useful in decision making as well as in planning process. Reversible watermarking is a relatively new and emerging area for maintaining integrity of the relational databases. The data format of the relational data is different than that of audio, video, images and natural languages. The techniques used in watermarking relational databases needs to consider following constraints: i) a database consists of tuples/records, which is where the watermark needs to be embedded; ii) the ordering of the records in a database relation; and iii) data operations such as insertion, deletion and alteration that occurs normally in the database. Additionally, a proper embedding mechanism to embed some information along with the type of data stored in the database should be taken into consideration to achieve better results.

Irreversible watermarking helps in protecting ownership rights, but in these schemes the embedding process alters the data to a large extent. Reversible watermarking consists of same process as that of irreversible watermarking i.e., encoding and decoding process with third additional data recovery process. In genetic algorithm(GA) based reversible watermarking approach for numeric and non-numeric relational data, the concept of Mutual Information(MI) is used which statistically measures the correlation between the features(i.e. dependency of the features on each other in terms of information stored in the relational databases). This concept makes it easier in embedding the secret parameter values and manipulating the database information so that the suspicious attackers are unable to attack the databases. The importance of the features in watermark embedding process is not taken into account in previous works. The GA based reversible watermarking method considers the importance of database features and performs the embedding mechanism accordingly. This technique gives a robust data recovery option which is reversible in nature and attack resilient. Also, it paves a way for efficient decision making and knowledge discovery using relational databases within shared environments.

2. Literature Review

Reversible watermarking has proved to be the most prominent technique for protecting relational databases in

present era. Reversible watermarking has paved a way in increased use of relational databases in knowledge discovery and decision making purposes. Several techniques are introduced and implemented for securing relational databases and maintaining database consistency. Some of the popularly known approaches used in ownership protection are fingerprinting, data hashing and serial codes. Also, encryption is the oldest method used in achieving data security. Fingerprinting is one of the widely used aspect of encryption approaches. Fingerprinting i.e. transactional watermarks are used to keep track and locate digital ownership by embedding all the copies of contents each with different individual watermarks for different recipients. One of the main disadvantage fingerprinting in encryption holds is that once it is stolen it can be easily used by attacker. This results in a privacy issue in using bio-metric schemes. Also, in encryption, data are transformed altogether and cannot be used in any decision making and planning operations. Whereas, watermarking approaches encode the data in such a manner that it remains helpful for recipients in latter use[14].

The first and foremost approach used in irreversible watermarking technique lacked the capability to retrieve the original relational data along with the preservation of it's data quality[20]. The histogram expansion technique describes the first reversible watermarking method used for watermarking the relational databases. In this approach, the selective nonzero digits of errors are used to form histograms and are then reversibly watermarked. Here, the original data can be fully reconstructed using the untampered watermarked relational datasets. In this technique, with the use of validation of datasets only the data owner has the capability to recover the original database state. But, this scheme is not suitable against malicious attacks that are subjected to a large number of tuples [15].

In Difference Expansion Watermarking(DEW) technique, a concept of difference expansion is applied on integers and reversibility is achieved in watermarking of relational databases. Here, arithmetic operations are carried out on numeric features and transformations are performed for manipulating data within database. The important advantages of this scheme are reversed to huge condition of original content, legitimate owner recognition, resistance against subordinate attacks, and storing of the original dataset at a protected secondary storage is not at all required. The key generation here is based on hash of the secret key and primary key[12], [13]. The advanced version of DEW technique is based on support vector regression(SVR) prediction for watermarking relational databases. This scheme aims at providing ownership proof and safeguard the database from been tampered. However, this scheme is prone to alteration attacks which makes the original database recovery difficult and nearly impossible [5].

Genetic Algorithm based on Difference Expansion Watermarking(GADEW) technique provides a better solution over the drawbacks of DEW technique. In DEW approach, only two features of the selected record are taken into consideration depending on the amount of distortion tolerated which in turn results in low amount of watermark capacity and higher amount of distortion rate. In GADEW technique, genetic algorithm along with difference expansion is implemented for minimizing the distortion rate in the data. But, with this, the watermark capacity is decreased with the increase in watermarked tuples[2]. A Prediction-Error Expansion Watermarking (PEEW) technique incorporates a predictor in place of difference operator to select candidate pixels or features for embedding watermark information. PEEW technique has been extended to handle the situation where the multiple owners claim the ownership of watermarked data. PEEW scheme is able to embed a large watermark into the image with relatively low distortion by implementing a new adaptive embedding and pixel selection strategies. This results in better quality of restored image after watermark extraction [16], [17], [18].

Web-based databases are watermarked using a robust, resilient and reversible watermarking method. In this method, an algorithm is proposed which uses public watermarks for watermarking based on parameterized tuple partitioning and white spaces. This technique results in high distortion rate within the tuples in database[6]. A robust and reversible watermarking technique using genetic algorithm is favorable method to secure data from getting manipulated and maintaining the data integrity. In this technique, Mutual Information(MI) concept provides an appropriate mechanism for embedding watermark information. Here, a robust data recovery with high data quality is achieved along with protection of data from active malicious attacks(insertion, alteration and deletion attacks)[1].

These schemes only considered the numeric features for watermarking relational data. This provides a way for attacker to attack the relational database using certain mechanisms for non-numeric datasets. A robust technique of embedding watermarks in a relational database with text attributes is implemented where non-numeric features are considered for watermarking. Here, an attribute is selected using a pseudorandom generator and the watermark is embedded into it. In this technique, minor changes are embedded such that the usability of database is not degraded and copyright protection is also achieved[4]. This method is robust against different malicious attacks such as subset deletion, addition and modification attacks.

3. System Architecture

The genetic algorithm(GA) based reversible watermarking approach for numeric and non-numeric relational data focuses on recovering the relational data completely along with successful watermark detection. Also,

the goal is to maintain the data consistency in presence of various active malicious attacks such as insertion, alteration and deletion. These goals can be achieved through appropriate selection of secret parameter value and watermark string which can be utilized for encoding and decoding the watermark easily. Appropriate conditions are applied to improve the performance of Genetic Algorithm(GA) and decrease the computation cost. The system architecture provides the process for watermarking the relational database using numeric and non-numeric attributes of database aiming to recover the watermarked data successfully. The architecture in "Figure 1" comprises of following four phases: 1)Watermark pre-processing phase; 2)Watermark Encoding phase; 3)Watermark Decoding phase; and 4)Data Recovery.

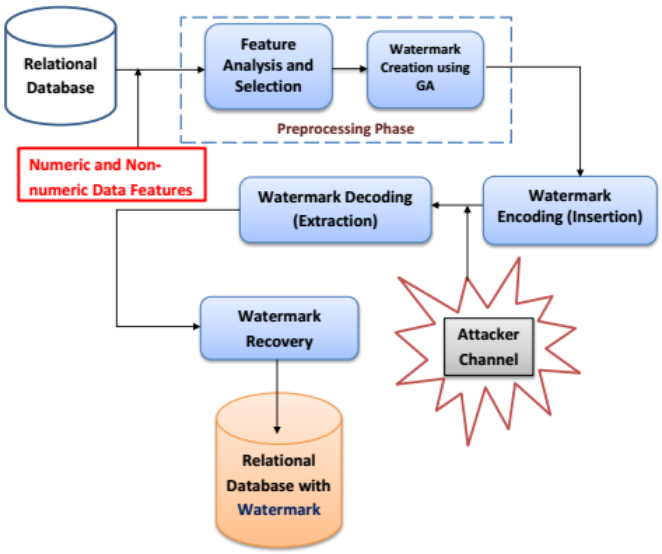

Figure 1: Architecture Diagram of Genetic Algorithm(GA) based Reversible Watermarking Technique.

The processing of the phases takes place as follows:

1) Watermark Pre-processing Phase:

In pre-processing phase, different parameters for defining most favorable watermark value are calculated. A secret threshold is defined using which user can analyze and rank the features according to their importance. This phase consists of following two tasks:

- Feature Analysis and Selection: For nonnumeric(textual) attributes, we cannot watermark these attributes in string format. For this purpose, initially the textual data is converted to numeric values and weight of each is calculated. This weight is further used for employing the Mutual information(MI) amongst the features. A suitable non-numeric data feature is selected for watermark embedding. Here, all the features are ranked according to their importance and mutual dependence. The MI concept is used for defining the correlation between one feature with the other. The probability distribution amongst the features is calculated and a secret threshold range is defined within which the watermark mechanism should be incorporated. Here the feature whose MI value is less than the threshold is selected for watermarking. This provides a way of protecting data from attacker as the proper knowledge of embedding process is not known to the attacker.
- Watermark Creation: Here, an optimization algorithm is used for finding optimal watermark value. Genetic Algorithm(GA), an optimization technique is used for calculating an appropriate watermark string and the fitness parameter i.e., Beta value. This beta value provides the amount of change that is tolerable in the selected features while watermark embedding. For finding an optimal solution, an iterative mechanism is followed by GA that evolves a population of chromosomes. The basic genetic operations to these chromosomes in GA are: selection, crossover, mutation and replacement. Here, the fitness value is calculated based upon generation of binary string, calculating parent chromosome by applying the tournament selection, offspring creation using crossover and mutation operation and lastly by applying the elitism strategy two individuals are obtained. This results in finally in generation of optimal watermark string and the best fitness value beta. The MI values for original data and watermarked data are compared for number of generations. The mean and variance provide the results for preserving the data quality of the obtained watermarked relational data.

2) Watermark Encoding:

This phase aims at embedding the watermark in the Database such that the usability of the data is not affected even after the insertion of the watermark information. For embedding the watermark, the best fitness value and the optimal watermark string obtained are used. A secret threshold defined by the owner gives the right to owner to decide the number of features to be watermarked in the database. The watermark information should be embedded in such a way that data should be protected against any sort of attack and the data quality should be maintained at the time of data recovery. An attacker channel is assumed consisting of insertion, deletion and alteration attacks.

3) Watermark Decoding:

In watermark decoding phase, the embedded data is extracted which is assumed to be received from untrusted data channel. For this purpose, initially the data features which are marked are located and then extraction process is carried from LSB towards MSB. The original data and recovered data with watermark

is compared to note the data quality and recovery percentage results.

4) Data Recovery:

This is the final phase where the original data along with the watermark is recovered. With some pre-processing the decoded watermark bits is converted to achieve watermark information. The optimized value of beta and string generated using GA is used for reconstruction of original data. The results are compared and tested for effectiveness of the recovery process.

4. Experimental Setup

4.1 Software and Hardware Requirements:

Experiments are conducted on Intel Core i3 processor with CPU of 2.5GHz and 4GB RAM with 64 bit operating system. The front end used is JAVA jdk 1.8 with Netbeans IDE 8.0.2 and the system is implemented on Windows 7 operating system. Various data-sets are available from UCI Machine Learning repository which includes, Cleveland Heart Disease data-set, MAGIC Gamma Telescope data-set and PAMAP2 Physical Activity Monitoring data-set. Here, experiment is carried on Heart Disease Medical data-set. The processing is performed on both numeric and non-numeric data features within data-set. The implemented method is analyzed for the following constraints,viz., i)effect on the data quality; ii)robustness against malicious attacks; and iii)data recovery and iv) performance of GA. A part of data-set is used to illustrate the entire process step by step.

4.2 Performance Parameters

The implemented watermarking technique is expected to provide better results than the existing approaches. Here, following Keep Performance Indicator(KPI) are used to compare the performance of the system:

1) Data Distortion Ratio

The watermark should be embedded by using optimum bandwidth. Here, the technique used is expected to maintain the quality of data without any modification within the database. Distortion effect on original data is negligible.

2) Resilience against Attack Ratio

The watermarking approaches should be robust and resilient to attacks. The GA based reversible watermarking technique is expected to provide results which depicts efficient recovery against attacks. Here, resilience against attacks such as alteration, insertion and deletion is analyzed and performance of the scheme is determined. Robustness is an important constraint in relational database concepts.

3) Watermark Detection

The proposed GA based reversible watermarking method is analyzed for detecting watermark with or without presence of the malicious attacks.

4) Data Recovery

As the GA based watermarking technique is reversible in nature, it is expected to recover the original data with or without attacks.

5. Results

To prove the efficiency of the implemented technique, extensive attack analysis is performed for generating results which shows the robustness of genetic algorithm based reversible watermarking technique. Here, operations are performed on the numeric and non-numeric features of the relational dataset. "Table 1", illustrates the features that are considered for watermarking. The Mutual Information (MI) of original data (MIo) and the watermarked data(MIw) is compared for verifying the data quality that is preserved. After comparison it is observed that the relational data content is preserved without any loss in data quality. Also, the data is recovered fully in presence of the intruder attacks such as insertion, deletion and alteration attacks.

Sr. no	Name of features	MI_o	MI_w	Δ_{MI}
1	Age	0.01010	0.01010	0
2	Chol	0.08674	0.08674	0
3	Trestbps	0.09029	0.09029	0
4	Thalch	0.08937	0.08937	0
5	Thal (defect)	0.02941	0.02941	0
6	Gender	3.91768	3.91768	0
7	Slope	4.80108	4.80108	0
8	Restecg	4.80108	4.80108	0

Table 1: Mutual Information of Selected Features Before and After Watermarking.

GA Parameter	Value for Existing System	Value for Implemented System
No. of Generations	100	100
Population Size	50	50
Chromosome length	3	**5**
Selection Mechanism	Tournament Selection Tournament Size= 5	Tournament Selection Tournament Size= 5
Crossover	Type: Single Point Fraction: 0.7	Type: Single Point Fraction: 0.7
Mutation	Type: Uniform Rate: 0.1	Type: Uniform Rate: **0.05**
Elitism Count	2	2

Table 2: Genetic Algorithm Parameters

"Table 2" provides the GA parameters that are considered while performing the mechanism to obtain best

fitness value(beta) and optimal watermark string. The performance of genetic algorithm has improved with decrease in computational time as compared to existing technique. "Figure 2" provides the graphical representation depicting decrease in the computation time for GA. "Figure 3" shows comparative results for data recovery after insertion attacks. It is observed that 100% relational data is recovered for insertion attack in the implemented system. Due to use of majority voting scheme even when duplicate tuples are inserted by the attacker, the data quality is not compromised and data recovery is successful.

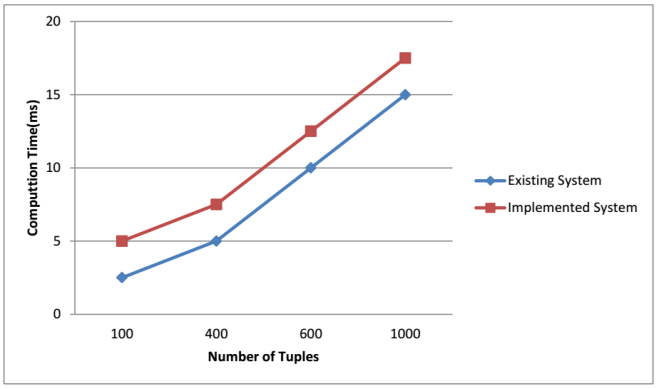

Figure 2: Computation Time for Genetic Algorithm

"Figure 4" depicts the results for data recovery after alteration attacks. For alteration attack, with decoding process the original data and unaltered tuples can be recovered. Also, some of the altered tuples can be retain if their usability remains unaffected after alteration attack. Here, it is difficult for attacker to corrupt the watermarked relational data as the secret variables are inaccessible. As compared to existing system, the implemented system gives better results in data recovery for alteration attacks. Also, "Figure 5" for data recovery after deletion attack provides details for successful data recovery even when 100% of the data is deleted. Additionally, the scheme is robust to recover the actual relational data successfully in presence of attacks.

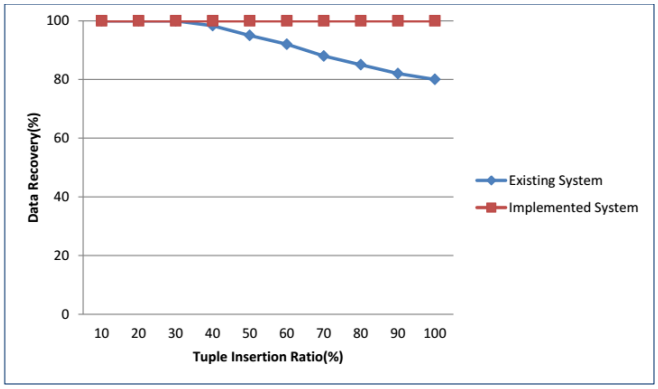

Figure 3: Data Recovery after Insertion Attacks.

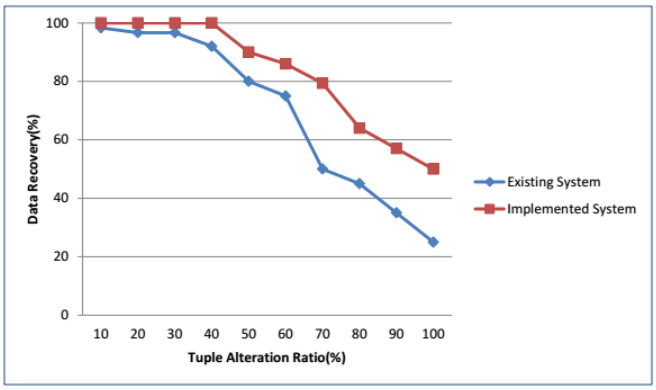

Figure 4: Data Recovery after Alteration Attacks.

Here, the experimental studies shows the effectiveness of GA based reversible watermarking technique and proves the performance accuracy in terms of data recovery, watermark detection, data quality and resilience to attacks(insertion, alteration and deletion) which is up to the mark as compared to the existing techniques.GA based reversible watermarking approach gives 100% watermark identification even when most of the tuples are attacked, but previous techniques do not provide such better results. Also, here in presence of heavy attacks data recovery is achieved completely. The performance of GA has improved significantly as compared to previous schemes.

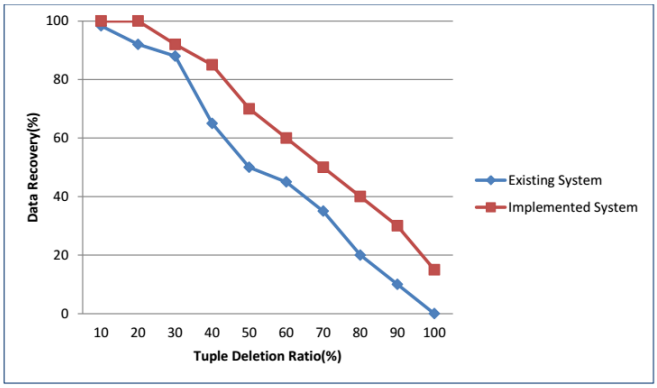

Figure 5: Data Recovery after Deletion Attacks.

6. Conclusion

Watermarking techniques for relational databases are enforced to protect the ownership of the data and prevent relational data from getting interfered from unauthorized access. The numeric and non-numeric attributes in relational database contains important sensitive information in proportion. Both attributes should be considered for watermarking of relational databases in equal proportion. The significant drawback of the irreversible watermarking is modification in relational data such that data quality is compromised to a large extent. Reversible watermarking techniques are able to overcome these drawbacks but are not effective against malicious attacks. In this paper, a

genetic algorithm based robust and reversible method for watermarking numeric and non-numeric relational databases is implemented. The Mutual Information(MI) concept used helps in obtaining the best watermark value for embedding and data recovery purposes. This approach is effective against various active malicious attacks. Experimental study shows that genetic algorithm based reversible watermarking is able to reconstruct embedded watermark and original relational data. Also, it gives good results as compared to the previously used techniques. The technique implemented gives better results in terms of performance of genetic algorithm as compared to existing schemes. A number of experiments proves that genetic algorithm based watermarking technique ensures data quality, data reversibility and is resilient to different active malicious attacks(insertion, alteration and deletion). The watermarking of shared databases in distributed environments using the robust and reversible watermarking needs to be addressed in future.

ACKNOWLEDGEMENT

With the completion of this paper, I would like to take this opportunity to express my gratitude and deep regards to thank my guide Prof.(Ms) Aparna Junnarkar, Computer Engineering Department, for her constant support and valuable guidance. I also express my sincere appreciation to the Computer Engineering Department as well as College Library. Lastly, I am thankful to my parents and well-wishers for their continuous upliftment at each and every step in fulfillment of the work.

REFERENCES

[1] Saman Iftikhar, M. Kamran and Zahid Anwar, "RRW - A Robust and Reversible Watermarking Technique for Relational Data", in Proceedings of IEEE Transactions on Knowledge and Data Engineering, val X, No : XX ,2015.

[2] K. Jawad and A. Khan, "Genetic algorithm and difference expansion based reversible watermarking for relational databases", Journal of Systems and Software, 2013.

[3] Kamran M, Suhail S, Farooq M., "A robust, distortion minimizing technique for watermarking relational databases using once-for-all usability constraints", IEEE Transactions on Knowledge and Data Engineering 2013; 25(12): 26942707.

[4] Vidhi Khanduja, Anik Khandelwal, Ankur Madharaia, Dipak Saraf, Tushar Kumar, "A Robust Watermarking Approach for Non Numeric Relational Database", in International Conference on Communication, Information Computing Technology (ICCICT), IEEE, 2012.

[5] J.-N. Chang and H.-C. Wu, "Reversible fragile database watermarking technology using difference expansion based on svr prediction", Computer, Consumer and Control (IS3C), 2012 International Symposium on IEEE, 2012, pp. 690693.

[6] E. Sonnleitner , "A robust watermarking approach for large databases", in Satellite Telecommunications (ESTEL), 2012 IEEE First AESS European Conference on. IEEE, 2012, pp. 16.

[7] M. E. Farfoura, S.-J. Horng, J.-L. Lai, R.-S. Run, R.-J. Chen, and M. K. Khan, "A blind reversible method for watermarking relational databases based on a time-stamping protocol", Expert Systems with Applications, vol. 39, no. 3, pp. 31853196, 2012.

[8] M. Kamran and M. Farooq , "An information-preserving watermarking scheme for right protection of emr systems", Knowledge and Data Engineering, IEEE Transactions on, vol. 24, no. 11, pp. 19501962, 2012.

[9] Coatrieux G, Chazard E, Beuscart R, Roux C., "Lossless watermarking of categorical attributes for verifying medical data base integrity", IEEE Annual International Conference of the Engineering in Medicine and Biology Society, EMBC, 2011, IEEE, Boston, MA, 2011;81958198.

[10] Zhang L, Gao W, Jiang N, Zhang L, Zhang Y., "Relational databases watermarking for textual and numerical data", 2011 International Conference on Mechatronic Science, Electric Engineering and Computer(MEC), IEEE, Jilin, China, 2011; 16331636.

[11] Y.-R. Wang, W.-H. Lin, and L. Yang, "An intelligent watermarking method based on particle swarm optimization", Expert Systems with Applications, vol. 38, no. 7, pp. 8024 8029, 2011.

[12] G. Gupta and J. Pieprzyk, "Database relation watermarking resilient against secondary watermarking attacks", in Information Systems Security. Springer, 2009, pp. 222236.

[13] G. Gupta and J. Pieprzyk , "Reversible and blind database watermarking using difference expansion", in Proceedings of the 1st international conference on Forensic applications and techniques in telecommunications, information, and multimedia and workshop. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2008, p. 24.

[14] S. Subramanya and B. K. Yi , "Digital rights management", Potentials, IEEE, vol. 25, no. 2, pp. 3134, 2006.

[15] Y. Zhang, B. Yang, and X.-M. Niu , "Reversible watermarking for relational database authentication", Journal of Computers, vol. 17, no. 2, pp. 5966, 2006.

[16] M. E. Farfoura and S.-J. Horng , "A novel blind reversible method for watermarking relational databases", in Parallel and Distributed Processing with Applications (ISPA), 2010 International Symposium on. IEEE, 2010, pp. 563569.

[17] D. M. Thodi and J. J. Rodriguez , "Prediction-error based reversible watermarking", in Image Processing, 2004. ICIP04. 2004 International Conference on, vol. 3. IEEE, 2004, pp. 15491552.

[18] D. M. Thodi and J. J. Rodriguez , "Reversible watermarking by prediction-error expansion", in Image Analysis and Interpretation, 2004. 6th IEEE Southwest Symposium on. IEEE, 2004, pp. 2125.

[19] A. M. Alattar , "Reversible watermark using difference expansion of triplets", in Image Processing, 2003. ICIP 2003. Proceedings. 2003. International Conference on, vol. 1. IEEE, 2003, pp. I501.

[20] R. Agrawal and J. Kiernan , "Watermarking relational databases", in Proceedings of the 28th international conference on Very Large Data Bases. VLDB Endowment, 2002, pp 155-166.