ROBUST DATABASE WATERMARKING TECHNIQUE OVER NUMERICAL DATA

HOSSEIN MORADIAN SARDROUDI

A dissertation submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

DECEMBER 2010

To my beloved parents, thank you for always being there for me, supporting me and encouraging me to be the best that I can be.

ACKNOWLEDGMENT

This research would not have been possible without the support of many people. I am heartily thankful to my supervisor, Assoc. Prof. Dr. Subariah Ibrahim, who was abundantly helpful and offered invaluable assistance, support and guidance. Without her continued support and interest, this thesis would not have been the same as presented here. My sincere appreciation also extends to all my friends who have provided assistance at various occasions. Their views and tips are useful indeed. Also I wish to express my love and gratitude to my beloved families; for their supports and endless love, through the duration of my studies. Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the research.

ABSTRACT

Proving ownership rights on outsourced databases is a vital issue. This research proposed an improved algorithm for relational database watermarking to minimize data variation as well as robust watermarking. For minimizing data variation two minimizing methods are proposed, the first method avoids embedding watermark bits in attributes with small values, the second method flips one of the unmarked bits in marked attribute in order to obtain the nearest number to the original attribute value. For embedding watermark bits, this research transforms the watermark image into a two dimensional matrix elements and embeds a gray (black and white) image in a database table. Furthermore in extracting process, a recovering phase is added. Recovering phase finds the missing extracted watermark elements and increases the robustness, so that the watermark can be detected even in a small subset of a watermarked database. In extracting process, majority voting method is employed to retrieve the correct watermark. For obtaining accurate results this study use~~s~~ three various images and three separate database tables with different sizes. By applying these methods, variation of data after applying watermark is extremely less and also an upper Correction Factor for watermark can be obtained. Detecting the watermark neither requires the original database nor the watermark. The results have illustrated that the proposed method are convincing compared to other related methods.

ABSTRAK

Membuktikan hak pemilikan pada pangkalan data sumber luaran adalah masalah penting. Penyelidikan ini mencadangkan penambahbaikan algoritma untuk peneraan air pangkalan data hubungan untuk meminimumkan perbezaan data serta menghasilkan tera air yang teguh. Untuk meminimumkan perbezaan data, dua kaedah meminimumkan dicadangkan, kaedah pertama mengelakkan pembenaman bit tera air pada atribut ber nilai kecil, kaedah kedua menukarkan salah satu bit yang tidak bertanda dalam attribute bertanda bagi mendapat nilai terdekat dengan nilai attribute asal. Untuk membenam bit tera air, kajian ini menukar imej tera air kepada elemen-elemen matriks dua dimensi dan membenam imej kelabu (hitam dan putih) ke dalam jadual pangkalan data. Selanjutnya pada proses ekstraksi, fasa pemulihan ditambah. Phasa pemulihan mencari elemen-elemen yang hilang dalam ter air yang diekstrak dan meningkatkan keteguhan tera air supaya era air boleh dikesan walaupun dalam subset kecil dari pangkalan data yang telah ditera air. Dalam proses ekstraksi, kaedah suara majoriti digunakan untuk memperolehi tera air yang betul. Untuk mendapatkan keputusan yang tepat kajian ini menggunakan tiga imej pelbagai dan tiga jadual pangkalan data dengan ukuran yang berbeza. Dengan menggunakan kaedah ini, perbezaan data menjadi kecil setelah tera air dibenam dan nilai Factor Pembetulan yang lebih tinggi bagi ter air diperolehi. Pengesanan tera air tidak memerlukan pangkalan data asal mahupun tera air. Hasil kajian menggambarkan bahawa kaedah yang dicadangkan meyakinkan berbanding dengan kaedah lain yang berkaitan.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

The speed of computer technologies and the growth of Internet have made duplication and distribution of digital information simpler. Copyright protection of intellectual properties has, therefore, become an important issue. One way for copyright protection is *watermarking.* It means securely embedding of some specific information about the copyright holder (company logos, image, text, ownership descriptions, etc.) into the digital object to be protected (Petitcolas, Anderson and Kuhn, 1999). Since several years ago, as an adequate technique to perform copyright protection, digital watermarking method has been widely used in image, audio, video and other areas, and gains productive issues (Chaokun *et al.*, 2008).

Agrawal and Kiernan, developed a database watermarking technique for protecting relational framework (Agrawal and Kiernan, 2002). The digital watermarking methodology on relational data is gradually expanding, and several relational data watermarking algorithms provide required copyright protection on relational data. Relational data watermarking technology can provide the necessary copyright protection on the basis of ensuring relational data availability (Chaokun *et al.,* 2008).

Watermarking in relational frameworks is a very young technology that has just begun its maturity cycle towards full deployment in industry-level applications. The main motivation for relational database watermarking is to protect relational databases mainly those published online (e.g., parametric specifications, surveys, and life sciences data), from tampering and pirated replicas. A watermark can be regarded as some sort of data that is embedded into elemental data for tamper detection, localization, ownership proof, and/or traitor tracing reasons. Database watermarking techniques complement the database protection (Yingjiu, 2008). The basic idea is change some of attributes value to another value, if such change is tolerable in certain applications; or change physically the order of tuples, if changing the order of tuples is acceptable in related application and database. Database watermarking technology enables us to hide an imperceptible, robust, and secure data in relational database.

For implementing copyright of digital products like relational databases, database watermarking method is a powerful way. It embeds copyright authentication data which is called mark or watermark into original data to keep safe copyright of digital product data. Relational database watermark is generally invisible or imperceptible. It combines the watermark with original data and hides it in original data without destroying it, and could live through some operations which do not ruin the important value or commercial value of original data (Chen *et al.,* 2008)

As shown and described in Figure 1.1, relational database watermarking comprises of two major basic processes. First watermark insertion and then watermark detection. For inserting (embedding) watermark information, in most methods a private secret key is used to embed watermark information into an original relational database in order to create the watermarked database, for sharing, publication or distribution database. By giving proper private secret key and watermark information, a watermark detection process can be applied to any mistrustful database so as to find out whether or not a legal watermark can be detected. A doubtful database can be any watermarked database or innocent database, or a mixture of them because of several database attacks. Copyright protection, ownership proof, traitor tracing, tamper detection and localization, are main purposes of database watermarking (Yingjiu, 2008).

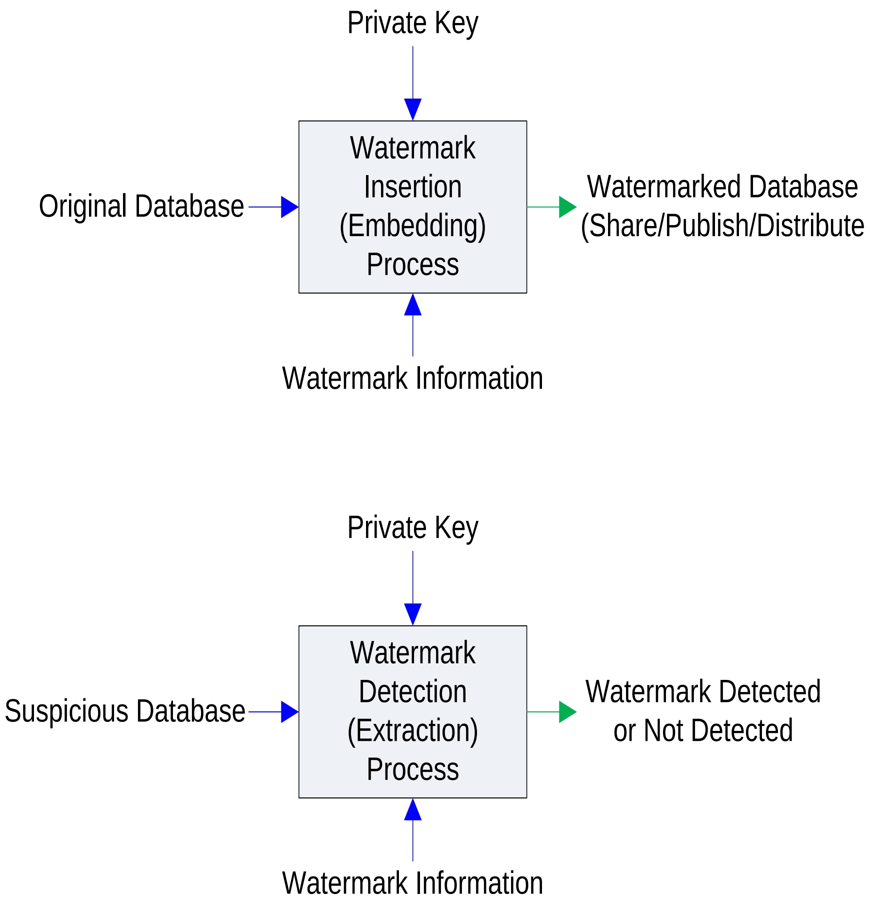
Private Key
Original Database
Watermark Insertion (Embedding) Process
Watermarked Database (Share/Publish/Distribute
Watermark Information
Private Key
Suspicious Database
Watermark Detection (Extraction) Process
Watermark Detected or Not Detected
Watermark Information

Figure 1.1 Basic Database Watermarking Processes

1.2 Background of the Problem

The piracy of digital assets like software, images, video, audio and text or data has long been a bother for owners of these properties. Inserting digital watermarks into these assets is one major way for protecting these data ownership. Most of the watermarking algorithms add a few errors into the object being watermarked. These deliberate errors are called *marks*. The marks must have an insignificant contact on the usefulness of the data and should manage in such a way that a malicious attacker cannot demolish marks without making the data useless. All these marks together make the *watermark*. It should be noticed that watermarking action does not prevent copying of digital objects, but it can prevent unofficial copying by presenting methods for verifying the original ownership of a digital objects (Agrawal, Haas and Kiernan, 2003).

The watermarking techniques can be fragile, robust, or semi-fragile. Fragile watermarks do not survive transformations to the original host signal and their purpose is tamper detection and localization of the original signal. Robust watermarking scheme provides a mark that can only be removed when the original content is destroyed as well. Typically, many of the applications for copyright protection, ownership proof, or traitor tracing involve relatively high quality original content and the imperceptibility criterion is critical for such applications. Semi-fragile watermarking techniques differentiate between common signal transformations and deliberate attack. It is robust to common signal transformations attack but is fragile to deliberate attack (Yingjiu, 2008).

In watermarking scheme, there is some difference between database watermarking with other multimedia watermarking. In relational database watermarking it has specific restrictions on embedding watermarks for relational databases with regular structured data and well defined semantics, even minimal modification of that may impact the data use. The first restriction is, the change of attributes value will render the whole database useless. Secondly, some methods are not blind detection and too fragile to be applicable to the data that is constantly updated (Xiangrong, Xingming and Minggang, 2007).

Watermarking methods are greatly focused in the area of image, audio, video, text, natural language and software, but less studied on relational databases, and studying in this area can help to improve the methods and techniques for the purpose of protecting data.

In relational database area, watermarking issues provides copyright protection of database data by embedding or hiding the digital information called mark or watermark. The embedded data can later be detected or extracted from the watermarked object (database) for identifying the copyright owner. To achieve this objective, various robust watermarking methods have been rapidly developed in the current decade. Proof-of-ownership watermarks are secret marks embedded into database using **a** secret private key without which a detector cannot determine its presence. These marks are designed for use as proof of rightful ownership, perhaps in a court of law. Attacks on copyright marking systems are quite various and different, especially since the set of “un-sensible” or "acceptable" alterations to data is difficult to identify (Craver, Wu and Liu, 2001).

1.3 Statement of the Problem

One of the most important issues of digital watermarking method is copyright protection and ownership identification for digital objects. Till now, a lot of digital watermarking methods, for copyright protection of multimedia data (image, video, audio), have been proposed to keep away from their misusage. Implementation of these watermarking schemes require more focus on robustness, trustworthiness and imperceptibility (Craver, Wu and Liu, 2001)

Because of the difference in data properties, while the basic processes in relational database watermarking are totally comparable to those in watermarking multimedia issue, the approaches developed for multimedia watermarking cannot be directly used for databases. In general, database relations vary from multimedia data in significant ways and therefore need a different class of information-hiding techniques. Unlike multimedia data whose elements are vastly correlated, database relations composed of separate objects or tuples. The tuples can be inserted, deleted, or modified usually in either legal updates or malicious attacks. No existing watermarking methods for multimedia data are designed to support such tuples operations (Yingjiu, 2008)

Altering the value of attributes in database is the most important point about relational database watermarking and should be noted. According to majority of watermarking methods, a watermark process, alters the value of item being watermarked, it inserts a mark in the database such that, first, the insertion of the mark does not destroy the significance of the data and it should be still useful for the intended motivation; and second, it complicate an attacker to erase or modify the mark beyond detection without ruining the value of data. If the data value to be watermarked cannot be altered without disturbing its value then a watermark cannot be inserted. The vital issue is not to avoid modification, but to limit modification, to at acceptable range with regard to the intended use of the data. Thus, an important first step in adding a watermark, by modifying data, is to identify changes that are acceptable by application. In fact, the level of such change depends on the application for which the watermarked data is to be used. Obviously, the nature of value or utilization of the data becomes thus the principle to the watermarking process. For instance in some applications, the value may be in ensuring the same computation, whereas for natural language text it may be in conveying the equivalent meaning. For example synonym replacement is acceptable. Equivalently, for a group of numbers, the utilization of the data may be placed in real values, in the relative values of the numbers, or in the distribution (Sion, 2007).

Rights protection for some data is more critical issues where it is vital, sensitive, valuable, and about to be outsourced. Data mining application is an example, where data is sold in pieces to parties specialized in mining it, for example, sales patterns database, oil drilling data, or financial data. Airline reservation and scheduling portals are other scenarios in which data is made available for direct, interactive use. According to the nature of most of the data, it is very hard to associate rights of the originator over it. To solve this issue, watermarking techniques can be used (Sion, 2007).

Another critical issue related to database watermarking is attack to watermark. A trade-off exists between the expected level of marking resilience and resistance to various attacks, and the ability to keep data quality in the result, according to the original data. It is simple to find that, if the encoded watermark is to be very "strong" one can easily alter the whole database data aggressively, but at the same time almost certainly also destroy its real value. As data quality requirements become increasingly restrictive, any applied watermark is unavoidably more weak and open to attack (Sion, 2007).

A robust approach that can minimize and measure the data variation should be proposed. Therefore, the following questions are answered in this study:

i) How to embed watermark bits into database with minimum data variation?

ii) Which policy will perform better so as to get robust database watermarking?

iii) How to measure data variation and analyze results in comparison with other methods?

1.4 Research Aim

A digital watermark is a technique that embeds a mark in a digital object by making small changes in the digital object. The three key methods of digital watermark are:

i) The method of prepare watermark data.
ii) The watermark embedding algorithm.
iii) The watermark extraction algorithm.

This research explores a new algorithm for information hiding as a rights assessment and data ownership tool in a relational database context. The aim of the research is to embed mark data in relational database, considering the most desirable requirements of invisibility and robustness. The basic idea is to ensure that some bits in some of the attributes of some of the tuples contain specific values. The tuples, attributes within a tuple, and specific bit values in some attributes are all algorithmically determined under the control of a private key known only to the owner of the database. This bit pattern constitutes the watermark. Only if one has access to the private key can detect the watermark with high probability (Agrawal, Haas and Kiernan, 2003).

1.5 Objectives of Research

This research objectives can listed as below:

i) To study existing database watermarking methods.

ii) To enhance a method for embedding marks into database with minimum data modification.

iii) To propose a robust watermarking method against intentional and unintentional attacks, and evaluate the proposed method by comparing the results with previous methods.

1.6 Research Scope

This research explores watermarking solutions in the context of relational data in which one or more of the attributes are of a numeric type. This research method marks only numeric attributes and assumes that the marked attributes can tolerate changes in some of the values; also three images are used as watermark for embedding into database. As there is no standard dataset for evaluate and compare database watermarking techniques, this research will use a random generated proper dataset for algorithm evaluation testing and comparing, also *Forest Cover Type* dataset (Blackard, Dean and Anderson, 1998) which used in quite a few publications, will be used as another dataset. To implement the proposed method, *Oracle11g Database* is selected for method database and *MATLAB Version 7.9 (R2009b)* and *Oracle Form Builder10g* will be used as programming language. The effectiveness of the research in robustness will be evaluated by some of subset attacks methods in order to comparison with previous works.

1.7 Significance of the Research

The digital products may very easily be copied by the illegal techniques; therefore the copyright owner's rights and digital assets cannot be efficiently saved. So the research of digital objects copyright protection has the essential functional significance. Relational database watermarking is the method to protect the copyright of the relational database, and called as the last line of defense for the database copyright protection

This research develops a database watermarking algorithm that minimizes the modification of database data and making the mark more robust in comparison with other methods. The proposed algorithm can be used for proof the ownership right of the data in relational database. In this way, if a third party approaches the database it is not expected to get the mark easily since all information is embedded by means of this technique. Another usage of the algorithm is that it can be used in authentication purposes.

1.8 Organization of the Research

This research consists of six chapters:

- Chapter 1: presents introduction, problem background, problem statement, aim, objectives, scopes and significant of this research.
- Chapter 2: reviews the literature in object watermarking methods and focused on database watermarking. Also some basic algorithms of database watermarking are presented in this chapter.

- Chapter 3: discusses on the methodology used in this research.
- Chapter 4: is designing of the proposed method.
- Chapter 5: discusses the experimental result.
- Chapter 6: is the conclusion and suggestion for future work.

REFERENCES

Agrawal, R., Haas, P.J. and Kiernan, J. (2003) 'Watermarking relational data: framework, algorithms and analysis', *VLDB*, vol. 3.

Agrawal, R. and Kiernan, J. (2002) 'Watermarking Relational Databases', Proceedings of the 28th VLDB Conference, Hong Kong.

Al-Haj, A. and Odeh, A. (2008) 'Robust and Blind Watermarking of Relational Database Systems', *Journal of Computer Science 4*, pp. 1024-1029.

Blackard, J.A., Dean, D.J. and Anderson, C.W. (1998) *Forest CoverType*, 28 August, [Online], Available: HYPERLINK "http://kdd.ics.uci.edu/databases/covertype/covertype.html" http://kdd.ics.uci.edu/databases/covertype/covertype.html [Jun 2010].

Chaokun, W., Jianmin, W., Ming, Z. and Guisheng, C. (2008) 'ATBaM: An Arnold Transform Based Method onWatermarking Relational Data', International Conference on Multimedia and Ubiquitous Engineering.

Chen, X., Chen, P., Yanshan, H. and Longjie, L. (2008) 'A Self-resilience Digital Image Watermark Based on Relational Database', International Symposium on Knowledge Acquisition and Modeling.

Cox, I.J., Miller, M.L. and Bloom, J.A. (2000) 'Watermarking Applications and their properties', Int. Conf. On Information Technology’2000, Las Vegas.

Craver, S.A., Wu, M. and Liu, B. (2001) 'WHAT CAN WE REASONABLY EXPECT FROM WATERMARKS', New York.

Cui, H., Cui, X. and Meng, M. (2008) 'A Public Key Cryptography Based Algorithm for Watermarking Relational Databases'.

Dong, X., Li, X., Yu, G. and Zheng, L. (2009) 'An Algorithm Resistive to Invertibility Attack in Watermarking Relational Databases'.

Dustdar, S. and Schreiner, W. (2005) 'A survey on web services composition', *Int. J. Web and Grid Services*, vol. 1, no. 1.

Guo, H., Li, Y., Liu, A. and Jajodia, S. (2006) 'A Fragile Watermarking Scheme for Detecting Malicious Modifications of Database Relations'.

Gupta, G. and Pieprzyk, J. (2008) 'Reversible And Blind Database Watermarking Using Difference Expansion', ICST, Adelaide.

Huang, K., Yue, M., Chen, P. and He, Y. (2009) 'A Cluster-Based Watermarking Technique for Relational Database', First International Workshop on Database Technology and Applications.

Hu, Z., Cao, Z. and Sun, J. (2009) 'An Image Based Algorithm for Watermarking Relational Databases', International Conference on Measuring Technology and Mechatronics Automation.

Jamasebi, R., Johnson, N.L., Kaffashi, F., Redline, S. and Loparo, K.A. (2008) 'A Watermarking Algorithm for Polysomnography Data', 30th Annual International IEEE EMBS Conference, Vancouver, British Columbia, 4.

Jinghai, R. and Xiaomeng, S. (2005) 'A Survey of Automated Web Service Composition Methods'.

Koz, A. (2002) *Digital Watermarkong Based On Human Visual System*.

Kutter, M. and Petitcolas, F.A.P. (1999) 'A Fair Benchmark For Image Watermarking Systems', Electronic Imaging ’99. Security and Watermarking of Multimedia Contents, Sans Jose, CA.

Li, Y., Guo, H. and Jajodia, S. (2004) 'Tamper detection and localization for categorical data using fragile watermarks', Washington, DC.

Li, Y., Swarup, V. and Jajodia, S. (2003) 'Constructing a Virtual Primary Key for Fingerprinting Relational Data', Washington, DC.

Li, Y., Swarup, V. and Jajodia, S. (2005) 'Fingerprinting Relational Databases: Schemes and Specialties'.

Meng, M., Cui, X. and Cui, H. (2008) 'The Approach for Optimization in Watermark Signal of Relational Databases by using Genetic Algorithms', International Conference on Computer Science and Information Technology.

Nikola, M. and Miroslaw, M. (2004) 'Current Solutions for Web Service Composition'.

NIST: National Institute of Standards and Technology ( ), [Online], Available: HYPERLINK "file:///D:\\Files\\Dissertation\\Watermarking\\Database%20Watermarking\\www.nist.gov" www.nist.gov .

Petitcolas, F.A.P., Anderson, R.J. and Kuhn, M.G. (1999) 'Information Hiding|A Survey', IEEE, 1062-1078.

SEOG-CHAN, O., DONGWON, L. and SOUNDAR, R.T.K. (2006) 'A Comparative Illustration of AI Planning-based Web Services Composition', *ACM SIGecom Exchanges*, vol. 5, no. 5, December, pp. 1-10.

Sion, R. (2004) 'Proving Ownership over Categorical Data', 12.

Sion, R. (2007) 'Rights Assessment for Relational Data'.

Sion, R., Atallah, M. and Prabhakar, S. (2003) 'Rights Protection for Relational Data', San Diego, California.

Sun, J., Cao, Z. and Hu, Z. (2008) 'Multiple Watermarking Relational Databases Using Image', International Conference on MultiMedia and Information Technology.

Wang, Z. and Bovik, A.C. (2002) 'A universal image quality index', IEEE SIGNAL PROCESSING LETTERS.

Wang, H., Cui, X. and Cao, Z. (2008) 'A Speech Based Algorithm for Watermarking Relational Databases'.

Xiangrong, X., Xingming, S. and Minggang, C. (2007) 'Second-LSB-Dependent Robust Watermarking for Relational Database', Third International Symposium on Information Assurance and Security.

Xiangwei, L., Fang, X.W. and GUO, J. (2009) 'Semi-automatic Web Service Composition optimization with Global Constraint', 2009 Second International Symposium on Electronic Commerce and Security.

Yingjiu, L. (2008) 'Database Watermarking: A Systematic View', in Atluri, V. and Warner, J. *Handbook of Database Security*, Springer US.

Zhang, Z.-H., Jin, X.-M., Wang, J.-M. and Li, D.-Y. (2004) 'Watermarking Relational Database Using Image', Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai.