International Journal of Computer and Communication Technology

Volume 8 | Issue 2 Article 5

April 2017

DETECTION OF DATA LEAKAGE

MAMTA SINGH
Department of Computer Science and Engineering, Institute of Technology and Management, GIDA Gorakhpur, India, singhmmt2010@gmail.com

PRITI TRIPATHI
Department of Computer Science and Engineering, Institute of Technology and Management, GIDA Gorakhpur, India, prititripathi@gmail.com

RENUKA SINGH
Department of Computer Science and Engineering, Institute of Technology and Management, GIDA Gorakhpur, India, renukasingh@gmail.com

DETECTION OF DATA LEAKAGE

MAMTA SINGH[1], PRITI TRIPATHI[2], RENUKA SINGH[3]

[1,2,3]Department of Computer Science and Engineering, Institute of Technology and Management, GIDA Gorakhpur, India
E-mail: singhmmt2010@gmail.com

Abstract- Sometimes sensitive data must be handed over to supposedly trust third parties. With the extensive application of database systems, the owners of the databases have urgent requirements to protect their copyright of databases. Some of the data is leaked and found in an unauthorized place the distributor must assess the likelihood that the leaked data came from one or more users. This paper contains concept of data leakage, its causes of leakage and different techniques to protect and detect the data leakage. In the field of IT huge database is being used. This database is shared with multiple people at a time. But during this sharing of the data, there are huge chances of data vulnerability, leakage or alteration. So, to prevent these problems, detection of data leakage system has been proposed. This paper includes brief idea about data leakage detection and a methodology to detect the data leakage persons. Data leakage is the main hindrance in data distribution. Traditionally this data leakage is handled by watermarking technique which requires modification of data.

Keywords- *Data Allocation Strategies, Data Leakage, Fake Object.*

I. INTRODUCTION

Data leakage is the unauthorized transmission of data or information from within an organization to an external destination or recipient [8][12]. Data leakage is defined as the accidental or intentional distribution of private or sensitive data to an unauthorized entity. Sensitive data of companies and organization includes intellectual property, financial information, patient information, personal credit card data and other information depending upon the business and the industry. Furthermore, in many cases, sensitive data shred among various stakeholders such as employees working from outside the organizational premises, business partners and customers. This increases the risk of confidential information falling into unauthorized hands.
Furthermore, in many cases, sensitive data is shared among various stakeholders such as employees working from outside the organizational premises (e.g., on laptops), business partners and customers. This increases the risk of confidential information falling into unauthorized hands.

In the course of doing business, sometimes data must be handed over to supposedly trusted third parties for some enhancement or operations. Let's take the example; a hospital may give patient records to researcher who will devise new treatments. Similarly a company may have partnership with other companies that require sharing of customer data. Another enterprise may outsource its data processing, so data must be given to various other companies. Owner of data is termed as the distributor and the supposedly third parties are called as the agents. In this project, our goal is to identify the guilty agent when the distributor's sensitive data have been leaked by some agents. Perturbation and watermarking are techniques which can help in such situations. Perturbation is a very useful technique where the data is modified and made less sensitive before being handed to agents. For example, one can add random noise to certain attributes or one can replace exact values by ranges on the original record. However in some cases, it is not important to alter the original record. Suppose if an outsourcer is doing our payroll, he must have the exact salary and customer bank account numbers. If medical researchers treating the patients (as opposed to simply computing statistic) they may need accurate data for the patients.

Traditionally, leakage detection is handled by the watermarking. For example a unique code is embedded in each distributed copy. If that copy is later found in the hands of an unauthorized party, the leaker can be identified. Watermarks can be very useful in some cases but again, involve some modification of the original data. Furthermore, watermarks can sometimes be destroyed if the data recipient is malicious [7].

In this paper, we develop an algorithm of data allocation strategies for finding the guilty agents that improves the chances of identifying a leaker. We also consider the option of adding fake objects to the distributed set. Such object do not corresponds to real entities but appear realistic to the agents. Means that fake objects act as a type of watermarks for the entire set, without modifying any original data. If it turns out that an agent was given one or more fake objects that were leaked, then the distributor can be more confident that agent was guilty.

II. EXISTING SYSTEM

Leakage detection is handled by watermarking, e.g., a unique code is embedded in each distributed copy. If that copy is later discovered in the hands of an

unauthorized party, the leaker can be identified. Watermarks were initially used in images, video and audio data whose digital representation includes considerable redundancy [1]. Watermarking aims to identify a data owner and, hence, is subject to attacks where a pirate claims ownership of the data or weakens a merchant's claims.

A. Watermarking Methodology

Nowadays, the digital assets such as software, images, video, audio and text are pirated which is a strong concern for owners of these assets. The protection schemes for such assets are based upon the insertion of digital watermarks into them. In this process, a particular object or record from the data is selected for the purpose of watermarking satisfying criteria. The criterion is these marks should have insignificant impact on the usefulness of data[10]. The procedure of watermarking introduces small errors into the object being watermarked. These intentional errors are called marks and all the marks together constitute the watermark. These marks are chosen in such a way that it has least impact on the data and placed such that a malicious user cannot destroy them. In traditional technique, leakage detection is handled by watermarking which is method of implanting a unique code on each of the distributed copy. When this copy is later discovered in the hands of an unauthorized party, the leaker can be identified.

Drawbacks of watermarking

However, there are two major disadvantages of the above algorithm:

1. It involves some modification of data i.e. making the data less sensitive by altering attributes of the data. This alteration of data is called perturbation. However in some cases, it is important not to alter the original distributed data. For example, if an agent is doing the payroll, he must have the exact salary. We cannot modify the salary in this case.
2. The second problem is that these watermarks can be sometimes destroyed if the recipient is malicious.

III. PROPOSED SYSTEM

Our project is based on FRIL tool which is a java based tool. It distinct data entries that refer to the same entity in two or more input files. The process is important for both data cleaning and integration. It detects the data leakage between the given files as the output.

A. FRIL technology

FRIL is tool for comparative record linkage. The tool extends traditional record linkage tools with a richer set of parameters. Users may systematically and iteratively explore the optimal combination of parameter values to enhance linking performance and accuracy. Results of linking birth defects monitoring program and birth certificate data using FRIL show 99% precision and 95% recall rates when compared to results.

The goal of record linkage is to find syntactically distinct data entries that refer to the same entity in two or more input files. The process is important for both data cleaning and integration in birth defects surveillance and research.

B. FRIL Architecture

The figure shows the workflow of the FRIL architecture. The user specifies the initial input files. Each run involves the user specifying the search method, the distance function in the attribute comparison module and the decision model.

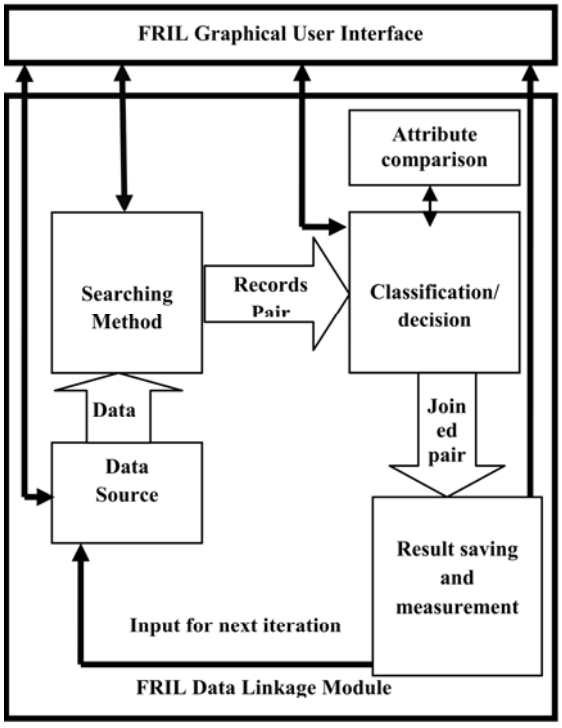

Figure1. FRIL Architecture

FRIL Modes

In this project FRIL provides two modes that is

1. Linkage of the data files
2. Detection of the data

Configuration of the Data Files

The input files which will be linked and detection will be performed may have the following types of data files that are-

CSV File

Comma-separated values (CSV) file stores tabular data (numbers and text) in plain-text form. Plain text means that the file is a sequence of characters, with

no data that has to be interpreted instead, as binary numbers. A CSV file consists of any number of records, separated by line breaks of some kind; each record consists of fields, separated by some other character or string, most commonly a literal comma or tab.

EXCEL File
File saved in Microsoft excel format. FRIL works with excel file saved in office 1997-2003 file format. The configuration of excel data file require the specification of file name and sheet name.

IV. CONCLUSION

Our model is relatively simple, but we believe that it captures the essential trade-offs. The algorithms we have presented implement a variety of data distribution strategies that can improve the distributor's chances of identifying a leaker. We have shown that distributing objects judiciously can make a significant difference in identifying guilty agents, especially in cases where there is large overlap in the data that agents must receive. In spite of these difficulties, we have presented that it is possible to assess the likelihood that an agent is responsible for a leak, based on the probability that objects can be identified by other means. Data leakage is a silent type of threat. Your employee as an insider can intentionally or accidentally leak sensitive information. This sensitive information can be electronically distributed via e-mail, Web sites, FTP, instant messaging, spreadsheets, databases, and any other electronic means available – all without your knowledge. To assess the risk of distributing data two things are important, where first one is data allocation strategy that helps to distribute the tuples among customers with minimum overlap and second one is calculating guilt probability which is based on overlapping of his data set with the leaked data set.

REFERENCES

[1] N. Sandhya, G. Haricharan Sharma, K. Bhima, "Exerting Modern Techniques for Data Leakage Problems Detect," International Journal of Electronics Communication and Computer Engineering (IJECCE), Vol. 3, Issue (1) NCRTCST, ISSN 2249 –071X

[2] Sandip A. Kale C1, Prof. S. V. Kulkarni C2, "Data Leakage Detection: A Survey," IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661 Vol. 1, Issue 6 (July-Aug 2012), PP 32-35www.iosrjournals.org

[3] P. Saranya, "Online Data Leakage Detection And Analysis," IJART, Vol. 2 Issue 2, March 2012

[4] P. Papadimitriou, H. Garcia-Molina, "Data Leakage Detection," technical report, Stanford University, 2008.

[5] Shivappa M. Metagar, Sanjaykumar J. Hamilpure ,B. P. Savukar, "Water Marking Technique: An Unique Approach For Detecting The Data Leakage," Volume 2, Issue 5, Sept 2012

[6] Archana Vaidya, Prakash Lahange, Kiran More, Shefali Kachroo & Nivedita Pandey, "DATA LEAKAGE DETECTION," Vol. 3, Issue 1, pp. 315-321.

[7] Jagtap N.P., Patil S.S. And Adhiya K. P., "Implementation Of Guilt Model With Data Watcher For Data Leakage Detection System," Volume 4, Issue 1, 2012.

[8] Rohit Pol, Vishwajeet Thakur, Ruturaj Bhise, Prof. Akash Kate , " Data leakage Detection," International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 ,Vol. 2, Issue 3, May-Jun 2012, pp. 404-410.

[9] Sujana Dommala & M.SreeDevi,"Data Leakage Detection Using Fake Objects," Internati-onal Conference on Computer Science and Information Technology, ISBN: 978-93-81693-86-5, 10th June, 2012-Tirupati.

[10] R. Arul Murugan, Kavitha .E, Nivedha .M, Subashini .S, "Data Leakage Detection And Prevention Using Perturbation And Unobtrusive Analyzes," International Journal of Communications and Engineering, Volume 04– No.4, Issue: 03 March2012.

[11] Naresh Bollam, Mr. V. Malsoru, "REVIEW ON DATA LEAKAGE DETECTION," International Journal of Engineering Research and Applications (IJERA),Vol. 1, Issue 3, pp.1088-1091.

[12] Rupesh Mishra, D.K. Chitre, "Data Leakage and Detection of Guilty Agent," International Journal of Scientific & Engineering Research, Volume 3, Issue 6, June-2012.

[13] Unnati Kavali, Tejal Abhang, Mr. Vaibhav Narawade, "International Journal of Engineering Research and Applications," Vol. 2, Issue 2, Mar-Apr 2012, pp.1448-1452

[14] Rudragouda G Patil, "Development of Data leakage Detection Using Data Allocation Strategies," International Journal of Computer Applications in Engineering Sciences ,Vol. I, ISSUE II, JUNE 2011.

[15] Jayavarapu Karthik and Dr.P. Harini, "Data Leakage Detection," International Conference on Computing and Control Engineering (ICCCE 2012), 12 & 13 April, 2012.