International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 12 3911 - 3915

Information Leak Detection System using Fingerprint of data

Rudrani Sinha
Student M.Tech(CSE)
Rungta College of Engineering & Technology Bhilai, Chattisgarh,India
rudranis29@gmail.com

Prof. Chaitali Choudhary
Associate professor
Rungta College of Engineering & Technology Bhilai, Chattisgarh,India
chaitali.choudhary@gmail.com

Abstract: This proposed method is used for information leak detection is to generate fingerprint of the confidential document and generate token id that is available in Database and then check it with confidential document. In this paper, we propose an improvement for this approach to offer a much faster processing time with accuracy. The core idea of our solution is to eliminate types of phrases from the fingerprinting process. Types of phrases are identified by looking at available public documents of the organization that we want to protect from information leaks and different phrases are identified with the help Databases.

Keywords: *Information Leak Detection, Fingerprint of data , information privacy.*

*****

1. Introduction

Information leaks are a major problem of computer systems. The leak of confidential data either be it accidental or intentional, may cause huge losses to the data owner. These losses may not be only financial, but also reputation loss whose cost cannot be estimated easily

. There may be 2 solutions for information leak detection. The first one is to use specific expressions, keywords or phrases for identify confidential information. For example, a leak of a MasterCard number can be detected by searching expressions of 16 digits starting with two digits in the range from 51 to 55. In this case[2], the alternative solution is to generate fingerprints of confidential information, which are of any structure, and check the generated fingerprints against fingerprints obtained from outgoing file.

We here provide a solution based on information retrieval[1] to identify only phrases containing sensitive information for fingerprinting. The main idea of our solution is to identify the popularity of phrases before fingerprinting in two ways. We first look at available public documents of the company that we want to protect from information leak. . Our solution can improve the accuracy of detection by reducing false positives caused by public and common phrases. Our work makes the following major contributions

- We propose[12] a solution to improve the performance of the traditional approach for information leak detection. Our main idea is to identify non-sensitive phrases as well as common phrases, and remove them from the fingerprinting process of confidential documents.
- To evaluate the popularity of a long combined phrase we provide a technique to split the phrase into sub-phrases and finding out the popularity of the phrase based on its divided phrases.

2. Related Work

An information retrieval based solution to improve the performance of the traditional cyclical hashing approach for information leak detection. The core idea of IRILD[1] is to identify and remove public phrases (found in public documents) and common phrases (identified by checking the number of results returned by Google when querying the phrases)from the fingerprinting process, since these types of phrases do not contain sensitive information. Specifically, IRILD achieved a much faster leak detection speed. As a result, IRILD can be utilized by systems with a large numbers of sensitive documents. Also, IRILD achieved much higher accuracy when compared with traditional cyclical hashing due to the removal of false positives related to public and common phrases.

Data leakage is the big challenge in front of the industries & different institutes. Though there are number of systems designed for the data security by using different encryption algorithms, there is a big issue of the integrity of the users of those systems. It is very hard for any system administrator to trace out the data leaker among the system users. It creates a lot many ethical issues in the working environment of the office.

The data leakage detection industry is very heterogeneous as it evolved out of ripe product lines of leading IT security vendors. A broad arsenal of enabling technologies such as firewalls, encryption, access control, identity management, machine learning content/context-based detectors and others have already been incorporated to offer protection against various facets of the data leakage threat. The competitive benefits of developing a "one-stop-shop", silver bullet data leakage detection suite is mainly in facilitating effective orchestration of the aforementioned enabling technologies to provide the highest degree of protection by ensuring an optimal fit of specific data leakage detection

technologies with the "threat landscape" they operate in. This landscape is characterized by types of leakage channels, data states, users, and IT platforms. Traditionally, 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[3] 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. E.g. A hospital may give patient records to researchers who will devise new treatments. Similarly, a company may have partnerships with other companies that require sharing customer data. Another enterprise may outsource its data processing, so data must be given to various other companies. We call the owner of the data the distributor and the supposedly trusted third parties the agents.

The benefit of an Agent-based Information Leak-age Detection system[4] lies in the ability to modify and add detection capabilities, modularize those capabilities, and then conditionally employ such capabilities at the discretion of a central control mechanism (in our system, the Controller Agent). The use of mobile agents as described in this paper, and in general, reduces the per-host administrative complexity as once the initial agent environment is properly installed and configured, all further necessary actions are performed by the agents themselves. Additionally, mobile agents are able to provide unique reporting capabilities that, for the purposes of our research, may benefit the analysis of information leakage and the underlying covert channels through which information has been leaked. While the information leakage detection approach detailed here is based on the work of [20], future work in this area may lead to the inclusion of techniques aimed at detecting and blocking covert channels prior to the occurrence of information leakage. Given the highly varied nature of covert channeling methods, detecting all such methods is likely a matter for which a solution can only be obtained through the liberal use of techniques rooted deeply in the field of artificial intelligence.

Data plays a pivotal role in IT systems. Especially when sensitive data has to be sent to other places through trusted agents, it is very challenging and important to detect leakage when they deliberately leak it to others. The scenario where a distributor gives sensitive data to his trusted agents and the data is intentionally leaked to others. The distributor should identify or detect this leakage and its means that is who leaked it as well. This is the problem of paper [6] Towards this we propose new data allocation strategies for improving the probability of detecting leakages accurately. The system should detect leakage correctly and the means as well as against to the leakage by other means. The proposed methods [5] do not relay on the alterations of released data. It is also possible to inject “looks genuine but fake” data in order to improve the probability of detecting leakage and tracing the party who actually leaked it.

Sharing of data should proceed by considering assumptions specified and may reduce the leakage through our efficient algorithm and by the process of asymmetric key encryption algorithm for the fake object creation and which includes our chances of detection process even when the intended persons are colluded. Our future work includes the inquiring of agent guilt models that capture leakage scenarios that are not studied in this paper[15]. For instance, what is the appropriate model for cases where agents can collude and identify fake tuples? Another open problem is the extension of our allocation strategies so that they can handle agent requests in an online fashion.

Data leakage happens every day when confidential business information such as customer data, bank details, source code or design specifications, intellectual property and trade secrets are leaked out. When these are leaked out it leaves the company insecure state and it goes outside the jurisdiction. Because it may not be certain if a leaked object came from an agent or from some other source, since certain data cannot admit watermarks. So this uncontrollable data leakage put business in a susceptible position. In spite of these difficulties, this system[19] shown that it is possible to assess the likelihood that an agent is responsible for a leak, based on the overlap of his data with the leaked data and the data of other agents. The presented model assesses the “guilt”of agents on basis of general method as well as fake elimination method. So in this system we can find particular one guilty agent. The main focus of this project is the data allocation problem. It specifies how the distributor can “intelligently” give data to agents in order to improve the chances of detecting a guilty agent. Finally, by adding fake objects to distributed set, the distributor can find the guilt agent easily.

3. Implementation study

To avoid false positives involving phrases as shown in the example of Figure 1 and also reduce the un-necessary cost of generating and checking fingerprint of the phrases from Database, we propose Fingerprint based method that is able to identify phrases and eliminate them from the fingerprinting process. In our method, we evaluate the popularity of phrases by submitting them to in database that contains large number of phrases.

Information leak detection system will also eliminate phrases that can be found in those public documents from the fingerprinting process because these phrases contain already known information.

 ISSN: 2321-8169
3911 - 3915

This method generates fingerprints for confidential documents in three steps. First we take input text file and then secondly we generate token ID for the specific text file by removing and third steps is to verify this text file as illustrated in fig. 1.

Note that while the fingerprint generation of information leak detection system is different from that of the popular approach, the information leak detection of these two approaches is still the same, i.e., fingerprints of confidential documents in the database are used to check against fingerprints of outgoing documents for information leak detection. Also, note that this method is not used for encoded, encrypted, or compressed data.

3.1 System design:

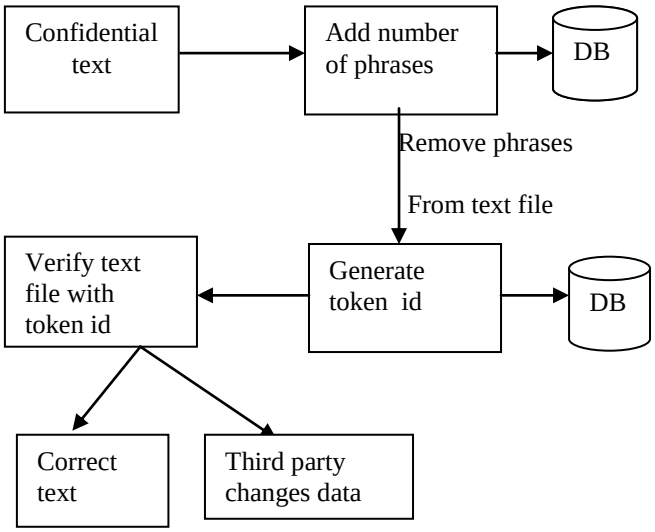

Figure 1: System design for Information leak Detection System

This method generates fingerprints for confidential documents in three steps. First we take input text file and then secondly we generate token ID for the specific text file by removing and third steps is to verify this text file

If the result is true then text is correct and if the result is false then we conclude that someone tampered with this text.

3.1 Experimental study:

We implement this proposed method in Netbeans IDE 7.2.1 using Java Script pages and for storing phrases and tokens in Databases we use My SQL by using Xamp server v3.2.1. The experiment can be done in three steps:
First we take input text file as shown in fig-2 and then secondly we generate token ID for the specific text file by removing and third steps is to verify this text file.

If the result is true then text is correct and if the result is false then we conclude that someone tampered with this text. The overview of the Information leak detection system is shown in figure 2.

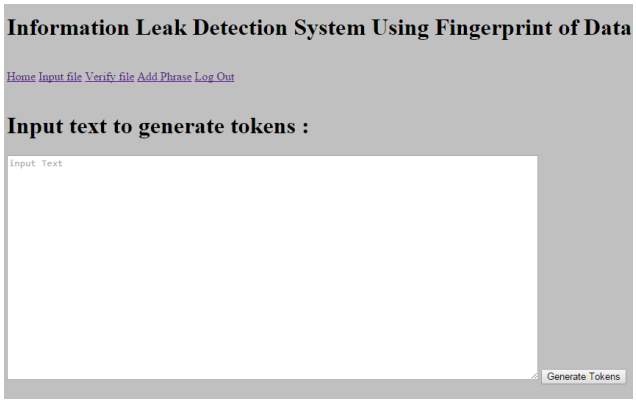

Fig-2 overview of generating tokens of input file

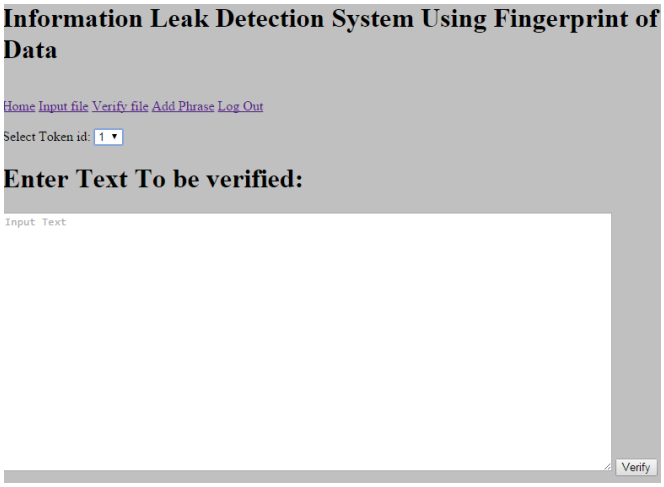

Figure-3:overview of ILD file verification

4. Conclusion

In this paper, we introduced information leak detection system to improve the performance of the traditional cyclical hashing approach. The core idea of Information retrieval is to identify and remove public phrases (found in public documents) and common phrases from the fingerprinting process, since these types of phrases do not contain sensitive information. Furthermore, we conducted extensive experimental evaluation of our solution, and proved that it significantly outperformed the cyclical hashing approach. As a result, Information retrieval can be utilized by systems with a large numbers of sensitive documents. Also, Information retrieval achieved much higher accuracy when compared with traditional cyclical hashing due to the removal of false positives related to public and common phrases.

5. Future Enhancement

This process is only used for only text file. and if the text file contains some Unicode characters in such cases this process can't give correct result.

6. Acknowledgement

I would like to express my special thanks of gratitude to my supervisor Prof. Chaitali choudhary, Associate Professor (CSE) who gave me the golden opportunity to do this wonderful project on the topic "**Information Leak Detection System using Fingerprint of data**", which also helped me in doing a lot of Research and i came to know about so many new things.

I am thankful to Mr. Toran Verma, M.Tech. Coordinator(CSE), Prof. K. J. Satao, HOD (IT) and Prof. Ajay Kushwaha, HOD(MCA) for giving thoughtful suggestions during my work.

I owe the greatest debt and special respectful thanks to Mr. Santosh Rungta Sir, Chairman, Dr. Sourabh Rungta, Director(Tech.), Mr. Sonal Rungta, Director(Finance), Dr. S. M. Prasanna Kumar, Director, Rungta College of Engineering and Technology, Bhilai and Mr. S. B. Burje, Vice Principal, Rungta College of Engineering and Technology, Bhilai for their inspiration and constant encouragement that enabled me to present my work in this form.

I would like to take this opportunity to express my thanks towards everyone in the computer Laboratory of the Rungta College of Engineering and Technology, Bhilai.

References

[1] Eleni Gessiou , Quang Hieu Vu , Sotiris Ioannidis Institute of Computer Science, FORTH, Greece Etisalat BT Innovation Center, Khalifa University, UAE by "IRILD: an Information Retrieval based method for Information Leak Detection" IEEE 2011

[2] Sandip A. Kale, Prof. S.V.Kulkarni Department Of CSE, MIT College of Engg, Aurangabad, Dr.B.A.M.University, Aurangabad (M.S), India "Data Leakage Detection" IJARCCE Vol. 1, Issue 9, November 2012

[3] Priyanka Barge,Pratibha Dhawale,Namrata Kolashetti Ass. Prof., Department of Computer Engineering, india "A Novel Data Leakage Detection"IJMER Vol.3, Issue.1, Jan-Feb. 2013 pp-538-540

[4] Hamed Okhravi, Stephen Bishop, Shahram Rahimi and Yung-Chuan Lee "A MA-based System for Information Leakage Detection in Distributed Systems"

[5] Janga Ajay Kumar and K. Rajani Devi Nalanda Institute of Engineering and Technology, A.P, India. "AN EFFICIENT AND ROBUST MODEL FOR DATA LEAKAGE DETECTION SYSTEM" Volume 3, No. 6, June 2012 Journal of Global Research in Computer Science

[6] M. Raja Kumar, G.Samuel Vara Prasad Raju, Bathula Yedukondalu Andhra University College of Engineering, AP " An Efficient Network disk Encryption and Data Leakage Detection" IJAIR 2012.

[7] Sridhar Gade, Kiran Kumar Munde, Krishnaiah.R.V "Data Allocation Strategies for Leakage Detection" IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661, ISBN: 2278-8727 Volume 5, Issue 2 (Sep-Oct. 2012), PP 30-35Xiaokui Shu and Danfeng (Daphne) Yao Department of Computer Science Virginia Tech Blacksburg VA, USA "Data Leak Detection As a Service"

[8] Bijayalaxmi Purohit, Pawan Prakash Singh, Department of Computer Science Engineering, Suresh Gyan Vihar University, Jaipur "Data leakage analysis on cloud computing" IJERA ISSN: 2248-9622 Vol. 3, Issue 3, May-Jun 2013, pp.1311-1316

[9] Rekha Jadhav,G.H.Raisoni Institute of Engg. And Technology "Data Leakage Detection" International Journal of Computer Science & Communication Networks,Vol 3(1), 37-45

[10] MAMTA SINGH, PRITI TRIPATHI & RENUKA SINGH Department of Computer Science and Engineering, Institute of Technology and Management, GIDA, Gorakhpur, India

[11] Sandip A. Kale C, Prof.S.V. Kulkarni "Data Leakage Detection: A Survey" IOSRJCE ISSN : 2278-0661 Volume 1, Issue 6 (July-Aug 2012), PP 32-35

[12] Anusha.Koneru, G.Siva Nageswara Rao, J.Venkata Rao "Data Leakage Detection Using Encrypted Fake Objects" International Journal of P2P Network Trends and Technology- Volume3Issue2- 2013

[13] Ms.B.Kohila, Mrs.K.Sashi Department of Computer Science,SNR Sons College, Coimbatore-641006, India "DATA LEAKAGE DETECTION USING K-ANONYMITY ALGORITHM" International Journal of Computer Science and Management Research Vol 1 Issue 5 December 2012 ISSN 2278-733X

[14] Supriya Singh United College of Engineering & Research, Gautam Buddh Technical University, Allahabad, Uttar Pradesh, India" Data Leakage Detection Using RSA Algorithm" IJAIM Volume 2, Issue 5, May 2013 ISSN 2319 – 4847

[15] Nikhil Chaware, Prachi Bapat, Rituja Kad, Archana Jadhav, S.M.Sangve Department of Computer, ZES's DCOER,Pune "Data Leakage Detection" IJSET Volume No.1, Issue No.6, pg : 272-273 1 Dec. 2012

[16] Mr. Ajinkya S. Yadav, Mr. Ravindra P. Bachate, Prof. Shadab A. Pattekari "Detection of Data Leakage Using Unobtrusive Techniques"IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 8, Issue 4 (Jan. - Feb. 2013), PP 79-84

[17] Bhagwan D. Thorat ,P. R. Devae BVDUCOE Pune, Maharashtra – India "Identifying Guilty Agent for Data Leakage Detection System" Volume 2, Issue 3, March 2014,International Journal of Advance Research in Computer Science and Management Studies

[18] Yung-Chuan Lee Stephen Bishop ,Hamed Okhravi z Shahram Rahimi "Information Leakage Detection in Distributed Systems using SoftwareAgents" 2009

[19] B. Sruthi Patil, Mrs. M. L. Prasanthi, VCE, Hyderabad India "Modern Approaches for Detecting Data leakage Problems" International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 2 Issue 2 Feb 2013 Page No. 395-399

[20] Narendra Babu.Pamula, M.Siva Naga Prasad, K.Deepthi Computer Science And Enginnering Department, "Preventing Data Leakage in Distributive Strategies by Steganography Technique" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (2) , 2013, 220 – 223

[21] "Preventing Information Leaks in Email" Vitor R. Carvalho William W. Cohen Language Technologies Institute Carnegie Mellon University

[22] Jaymala Chavan ,Priyanka Desai Thakur College of Engg. & Technology "Relational Data Leakage Detection using Fake Object and Allocation Strategies" International Journal of Computer Applications (0975 – 8887) Volume 80 – No 16, October 2013

[23] Hugh Wimberly, Lorie M. Liebrock Department of Computer Science & Engineering "Using Fingerprint Authentication to Reduce System Security: An Empirical Study" 2011 IEEE Symposium on Security and Privacy

[24] Naresh Bollam Mr.V.Malsoru M.tech(C.S.E) Jits,karimnagar "REVIEW ON DATA LEAKAGE DETECTION" International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 Vol. 1, Issue 3, pp.1088-1091