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

Recognition of Transformed Data Leaks

Prarthana Ranganathan[1], Sweta Barua[2], Swarupa Paul[3]

B.Tech in Computer Science & Engineering, SRM University, Chennai, India

---

Abstract - In this paper we are trying to deal with data leaks from organizational point of view. We are trying to prevent the sensitive information of an organization from getting leaked due to various attacks such as side channel analysis attack, inadvertent attack, malicious attacks, to name a few. We use Lucene search framework algorithm and Levenshtein distance algorithm to achieve the same. Using these algorithms, we can intercept the sensitive information before it goes out of the organization or falls into the wrong hands.

***Key Words*: Lucene,Levenshtein,Data leak detection, inadvertent attacks, mail server**

1. INTRODUCTION

The most common sources of data loss are through either malicious attacks, such as virus or trojan horse, or side channel analysis attacks, which occur through electronic radiation, photonic emission and many more[1]. In an organization, the most standard way of data leak occurs because of inadvertent attacks. Inadvertent attack basically refers to accidental data loss by legitimate user of an organization. It mainly happens due to human errors such as not using encryption or forwarding mails without care [2]. In some cases, sensitive information are duplicated and made vulnerable [4].

One way of preventing the leakage is by the using an intrusion detection system [6] and [7]. But this method just prevents the attacker from accessing information from outside. Sometimes, sensitive information of an organization can be transformed before their release such that they are not recognized by detection system of the organization [3] and [5]. Thus a method must be implemented which can be used to find all the possible combination of the information and prevents the data from getting leaked. So we implement Lucene search framework and Levenshtein distance algorithm.

Lucene is a simple search library. It is open source and scalable. This high-performance library is used to index and search virtually any kind of text. It provides the core functionalities which are required by any search application. Apart from these basic operations, a search application can also provide administration user interface and help administrators of the application to control the level of search based on the user profiles.

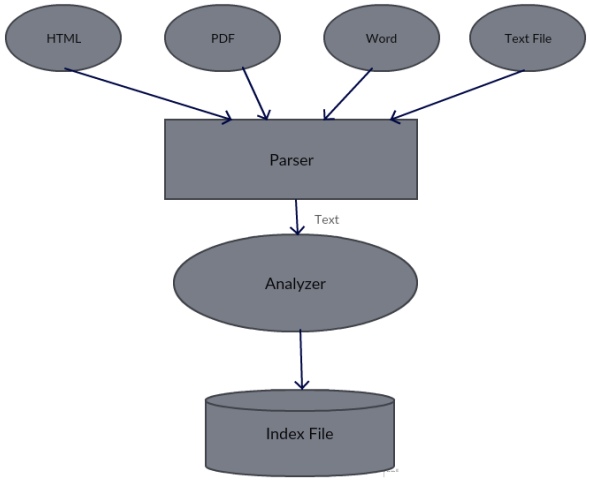

Fig -1: Lucene Analyzer

Indexing process is one of the important functionality provided by Lucene. Fig 2 shows the indexing process and use of classes. IndexWriter is the most important component of the process.

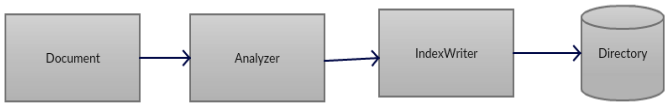

Fig – 2: Indexing

2. SYSTEM DESIGN

Normally, in an organization, employees register themselves in mail server with their name, authorized job position and their authorized e-mail domain. They use this to transfer files and mostly without any restriction of sensitive content checking. There is no content checking and domain filtering on their transformed sensitive data. Sensitive content is outsourced from one organization to another. Here outsourcing mechanism of transferred data is offending over the protocol. Thus we use the combination of Lucene search framework and Levenshtein distance algorithm.

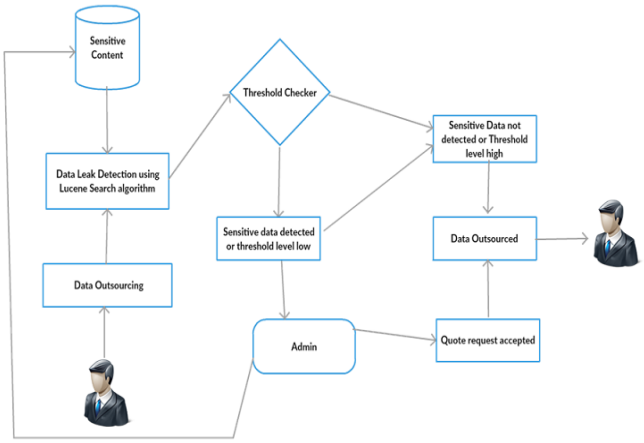

Fig – 3: Architecture Diagram

2.1 Module Description

The above fig 1 shows the architecture diagram of our proposed method. The modules are divided into three major parts :

2.1.1 Data Leak Detection(DLD) Framework Construction

In this module mail server data owner generates a Sensitive data and stored in the cloud and create the directory for lucene search framework and other data leakage detectors.

Data owner's cloud contains much sensitive information about their authorized customer's details, information technology source, and database and server details. The sensitive information is maintained by DLD. Using this DLD referenced directory perform data leak detection mechanism. The DLD consist of lucene search engine framework, levenshtein distance algorithm and our own shuffled checking algorithm. The DLD directly configured with cloud and can refer every data transformation outsourcing from authorized user transformation.

2.1.2 Data Outsourcing with DLD monitor

DLD will check all the outsourcing data before it transfer to the other organization. They are then checked with sensitive data. All the sensitive data are maintained in index file. Using this index file, DLD identifies the sensitive information concurrently with domain filtering and threshold assigning based on their email domain. DLD monitors every line of the outsourced data with the sensitive file. It will not allow any sensitive data will leak to any of the other organization. In proxy mail server the every occurrence of transformed contents are filter by users email domain. All their details are retrieved using their email. Then the threshold assigned for them based on their authorized job position and the transferred content has been tested by lucene framework search engine, levenshtein distance checking and shuffling algorithm.

2.1.3 Recognition of Sensitive Information and Request Processing

Once the DLD framework checks the outsourced content, if any data leak is identified means DLD will detect the Sensitive data. Here DLD will check not only the Sensitive data and also it will check some access condition. Every data owner maintain common access condition every file. For example, all the contents are encrypted before they outsourced. If DLD identified any Sensitive information outsourcing means they will detect the sensitive content in between of the file outsourcing. For the purpose of false alert, we maintain threshold of every domain and users position. If the Sensitive content percentage of transferred file exceeds the threshold percentage which trigger alert mail to Admin of the proxy mail server. Alert mail consists of entire details about the users even what are the Sensitive contents are pings from the transferred content by the DLD framework. After the filtration of mail in Mail server the user can claim or quote the request to admin with the quote reasons details. Finally the Mail server admin can review the quote request mails from the quote users whether he can allow or deny the mails.

3. CONCLUSIONS AND FUTURE WORKS

Hence, by using the Lucene search algorithm and Levenshtein distance algorithm, we minimize the data loss from organizational point of view. This technique can be used in steganography and military purpose where the transmission of secret information is concerned.

ACKNOWLEDGEMENTS

We would like to thank Ms.T.Malathi for her valuable guidance and cordial support throughout our research work.

REFERENCES

[1] Zhiwei Wang, Lingyu Zhou, "*Leakage-Resilient Key-Aggregate Cryptosystem with Auxiliary Input*" IEEE 978-1-5090-2279-3/16/

[2] "*Fast detection of transformed data leaks*" Xiaokui Shu, Jing Zhang, Danfeng (Daphne) yao and Wu-Chun Feng. (2016)

[3] R .S. Boyer and J. S. Moore, "*A fast string searching algorithm*," Commun. ACM, vol. 20, no. 10. pp 762-772, Oct. 1977

[4] A. Z. Broder, "*Identifying and filtering near-duplicate documents*," in Proc. 11th Anny. Symp. Combinat. Pattern Matching, 2000, pp. 1-10.

[5] X. Shu, J. Zhang, D. Yao, and W.-C. Feng, "*Rapid screening of transformed data leaks with efficient algorithms and parallel computing*," in Proc. 5th ACM Conf. Data Appl. Secur. Privacy (CODASPY), San Antonio, TX, USA, Mar. 2015, pp. 147–149.

[6] V. Paxson, "*Bro: A system for detecting network intruders in real-time*," in Proc. 7th Conf, USENIX Secur. Symp. (SSYM), vol. 7. 1998, p. 3.

[7] M. A. Jamshed et al., "*Kargus: A highly-scalable software-based intrusion detection system*," in Proc. ACM Conf. Comput. Commun. Secur., 2012, oo. 317-328.