A Detection Mechanism for Data Leakage

Prof. Umarani Chellpandy
Sam Soloman
MCA
INFORMATION SECURITY MANAGEMENT SERVICES
Jain (Deemed-to-be-University) University Bangalore, India

Abstract: Data leakage has become one of the growing internal risk in InfoSec and among many enterprises and personnel. Multiple ways have been created and deployed to inscribe the issues of data leakage prevention (DLP). Nevertheless, huge amounts of unorganized data must be analyzed and secured in our ever growing and evolving data. Because the size of data is growing exponentially and the structure of data have become more complicated and denser to manage, which has become a new threat for DLP and brings the question how these data can be processed and used efficiently. As a solution we introduce a method called Flexible weighted Graph model (FGM) which helps by aligning it to the elements of the weighted graphs. By adapting this approach, we can solve the problem in three steps. First, the weighted graphs can evaluate the sensitivity of the approved data based on the information passed. The second step is, the improved label propagation which can increase the scalability of the raw data. The third step is a low-complexity measuring algorithm is designed to analyze the furthest sensitiveness.

I. INTRODUCTION

Leakage of Data (or data loss) can be described as intentional or nonintentional loss of data which is caused by people or a process within or outside an organization. As we know many of the security threats happens from outside attackers, but in case of data leakage, which is mainly caused by an insider due to negligence or a disgruntled employee leaking information to an outsider.

Thus, traditional security measures (i.e. firewalls, intrusion detection systems, anti-virus) are no longer valid due to lack of understanding of data semantics. However, data leakage incidents happen from time to time, bringing serious damage continuously. To avoid the risk of data losses, plenty of researches have been done with the use of hash fingerprinting, n-gram, statistical methods and so on. With the breakthrough advancements of internet technology, many new communication technologies such as (e.g., Device-to-Device technology) have emerged. As a result, the size of data has grown exponentially, which has made handling of data more tedious due to the emerging varieties of data. This brings new challenges to data leakage detection. Therefore, the new DLD method is equipped with better tolerance for data diversity and higher efficiency to deal with the large amounts of unorganized data in a long term.

In this method, a model named Flexible weighted Graph model (FGM) is deployed to detect transformed data. In this model, all the documents are represented by graphs. The sensitive context weights in the form of node weights and edge weights are defined in the graph to improve the detection accuracy towards the transformed data. The context weight can quantify the sensitivity of the keywords adaptively based on the context around the keywords. The proposed solution aims to detect large amounts of newly generated, extensively transformed data accurately and efficiently. The main contributions are as follows.

- To better tolerate the long-transformed data, we define an adaptive context weight mechanism to measure the sensitivity of the keyword based on its text. The complex documents are further represented by weighted context graphs, containing both key terms and contextual information.
- To make up for the limitation of the template and increase the scalability, we also take the data semantics into consideration. An improved label propagation algorithm (LPA) is used to tag the same label on the highly relevant terms of the tested

graphs. We also merge the context graphs of each file into a general one - the template graph, to preserve more correlation information between key terms and enhance the overall sensitive context.

- To deal with the large amounts of data, we propose an algorithm with low-complexity. With a weight reward and penalty mechanism, the algorithm quantifies the sensitivity of the tested documents by one walk on their graphs, which allows the detection to be implemented in real time.

2.Scope of the Product

The proposed of system of DLD (Data Leakage Detection) helps to identify data leakage in an organization which is caused by human errors, with help of adaptive graph mechanism and tokenization. It can prevent majority of data leakage and prevent loss of data, which can provide data security and flexibility for an organization. This system helps to warn users and as well as the protect sensitive information which is shared among users with help of tokenization and adaptive weighted graph model to check the weightage of sensitive keywords in a message and providing fast results.

3.Existing System

To define the problems of data loss detection (DLD), plenty of research work has been done with the help of hash fingerprinting, n-gram, statistical methods and so on. With the breakthrough development which occurred in Internet, many new communication technologies (e.g., Device-to-Device technology) have emerged. As a result, data has grown dramatically, and the data varieties have become much more complicated. This brings new challenge to data leakage detection. Therefore, new DLD method is equipped with better tolerance of data transformation and higher efficiency to deal with the large amounts of unorganized data in long patterns.

4. Proposed System

This system is a novel model named Flexible weighted Graph model (FGM) is suggested to detect transformed data. In this model, all the documents are represented by graphs. The sensitive context weights in the form of node weights and edge weights are defined in the graph to improve the detection accuracy towards the transformed data. The context weight can quantify the sensitivity of the keywords adaptively based on the context around the keywords.
The proposed solution aims to detect large amounts of newly generated, extensively transformed data accurately and efficiently.

To better tolerate the long-transformed data, we define an adaptive context weight mechanism to evaluate the sensitivity of the keyword based on the data. The complex documents are further represented by weighted context graphs, containing both key terms and contextual information.

To make up for the limitation of the template and increase the scalability, we also take the data semantics into consideration. An improved label propagation algorithm (LPA) is used to tag the same label on the highly relevant terms of the tested graphs. We also merge the context graphs of each file into a general one - the template graph, to preserve more correlation information between key terms and enhance the overall sensitive context.

To deal with the large amounts of data, we propose an algorithm with low-complexity. With a weight reward and penalty mechanism, the algorithm quantifies the sensitivity of the tested documents by one walk on their graphs, which allows the detection to be implemented in real time.

5. System overview

In the proposed system for Fast detection of Data leakage. Which helps to prevent of data leakage which occurs in organization which is mainly occurs through human error, this software will the deployed in the local server of an organization to monitor users. The process of this software works by creating a database with sensitive keywords and tokenization. When a user sends a message to another user in the company, the message will the send through the DLD system, if it contains sensitive data the admin will send a notification to the user about the sensitive data. When the message is sent it will go through the preprocess stage where the message will be tokenized and analyzed with the sensitive keywords stored in the database.

During this process the message will be checked and analyzed to find out how much sensitive information exist in the message, based on the weightage (small, medium, high) the system will decide to let the send or block the message. If the weightage is small the message will be passed but if the message is medium or high it will be blocked.

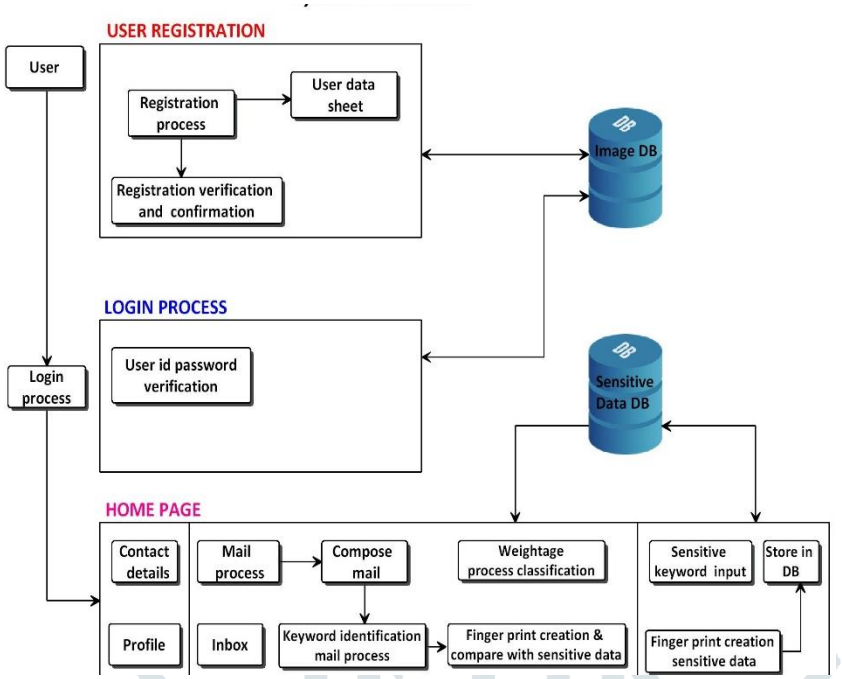

Figure:1 Process of Fast Detection of Data Leakage [5]

6. Project Module Description

- Adding sensitive keywords and weight age.

  Admin going to add sensitive keywords with weightage, those sensitive keywords is converted into fingerprints hash code by using MD5 algorithm.

- Intranet Mail System.

  With the help of this system user can able to send a mail to other users in the system and he can able attach the files to the mails.

- Detecting Sensitive Data in the text file

  This system is used to verify whether the attached files has sensitive message or not. If the attached file contains the sensitive message and user by mistake send to others which will create the data leakage problem. To overcome this from the attached files keywords are extracted and compared with the sensitive keywords.

- Giving Sensitive Data Alert message to the user

  There is a sensitive keyword scoring mechanism by which score will be calculated comparing with sensitive words in database using hashing technique and if the score crosses the threshold level, it warns the user and block the message.

7. Data Flow Diagram

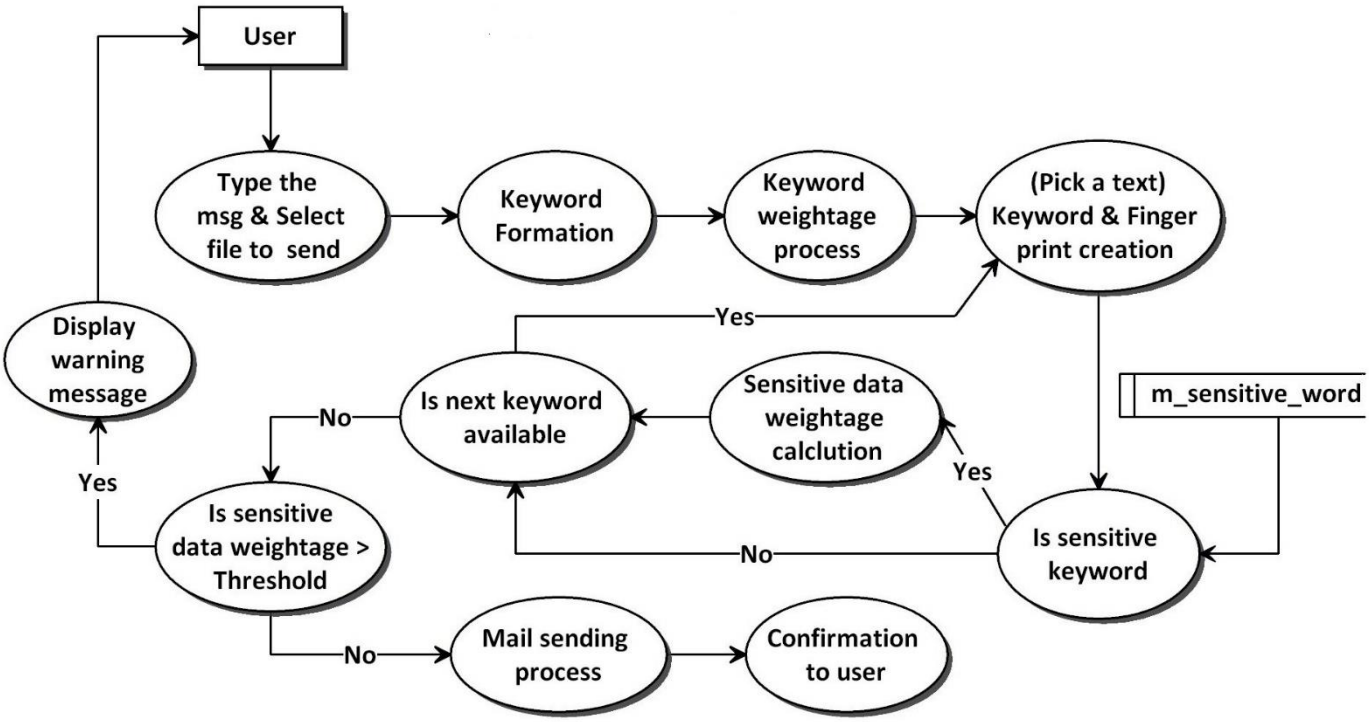

Figure: 2 Data Flow Diagram DLD [6]

8. Advantages

- This System increases the detection speed of data leakage.
- The context weight is able to quantify the sensitivity of the keywords adaptively based on the context around the keywords.
- The proposed solution aims to detect large amounts of newly generated, extensively transformed data accurately and efficiently.
- It increases the security of data which are being transferred and received.
- To deal with the large amounts of data, this algorithm is flexible and with low-complexity.
- With a weight reward and penalty mechanism, the algorithm quantifies the sensitivity of the tested documents by one walk on their graphs, which allows the detection to be implemented in real time.

9. Limitations

- One of the main disadvantages of the software is if the sensitive keyword database can be compromised by a hacker or an insider it can lead to data leakage.
- This system can be only implemented in an intranet environment as of now, because implementing in an internet environment is too vast and expensive.

10. Context Diagram

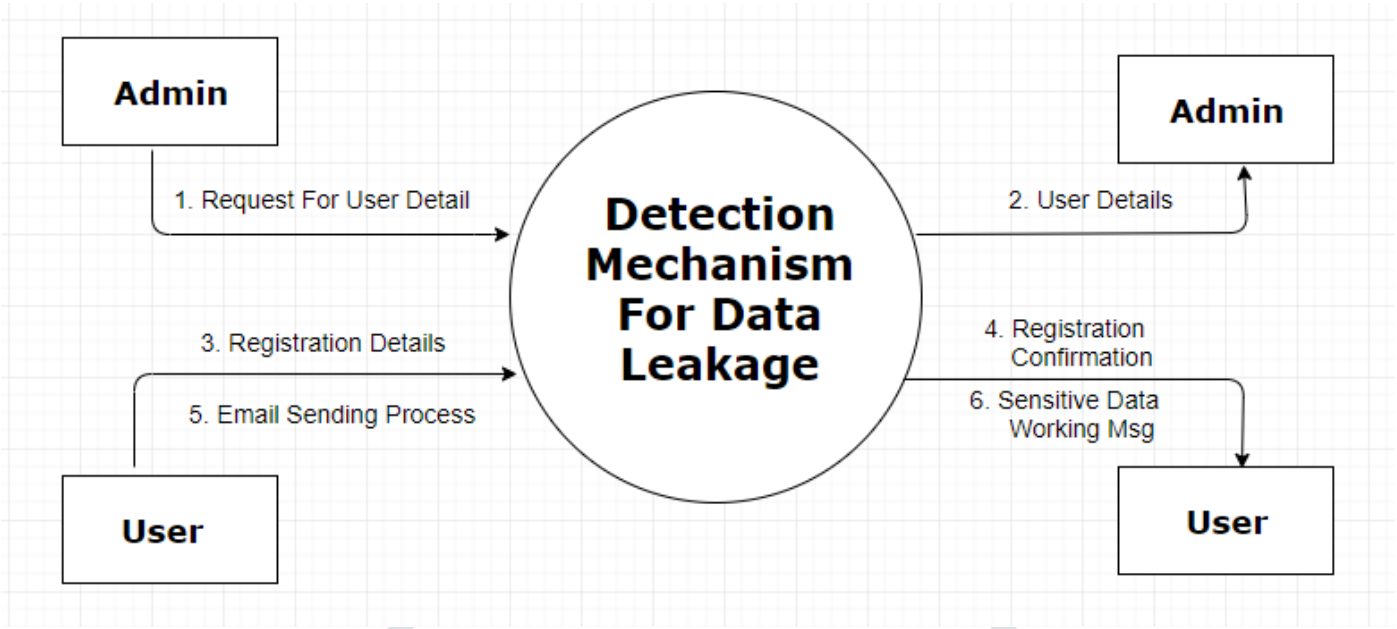

Figure: 3 Context Analysis Diagram [4]

11. Conclusion

We implement a methodology which provide a new means to avert loss of data and can prevent misusage of data in an organization, which mainly caused by carelessness. This system helps to analyze sensitive data which is shared among users in an organization, with the help of tokenization and adaptive weighted graph model. With the help of this system implemented it can help users in an organization to reduce leakage of data and protect the data which helps to improve the integrity of sensitive information which is stored in organization database.

12. Acknowledgment

This research is partially supported by Jain University Jayanagar Bangalore.

13. References

[1] https://ieeexplore.ieee.org/abstract/document/8422280

[2] https://ieeexplore.ieee.org/document/6723919

[3] https://link.springer.com/chapter/10.1007/978-3-642-22263-4_2

[4] C:\Users\samso\Pictures\Screenshots

[5] https://lh3.googleusercontent.com/mvmtMvWPfmYPa-L9tVD0TGztJIXDKw66OLy5O3lT4KpmsH4KPcIWtG4oiwvdPLs4dtkApyU=s109

[6] https://lh3.googleusercontent.com/JxUqFpf-kfAmqw8f9wNWBJ_hKBkr-Ti1xK8tyReoI-MLbbnHePe-zwj5EGR0C7bUIRxUHYg=s156

BIOGRAPHY

Prof. Umarani Chellapandy
Faculty & Guide Department of Computer Science & IT-MCA
Jain (Deemed-to-be University) Bangalore, India

Sam Soloman
Jain (Deemed-to-be-university) Jayanagar, Bangalore