Preventing the Disclosure of Data Leaks in Mail Transactions

[1]Aishwarya M B, [2]Saileela R, [3]Ankitha p, [4]Aparnashree R, [5]Sharan Lional Pais

[1]Engineering, [2]Engineering, [3]Engineering, [4]Engineering, [5]Assistant Professor
Department of Information Science and Engineering
ALVA'S INSTITUTE OF ENGINEERING AND TECHNOLOGY

***Abstract*: The sensitive data leaks on computer systems poses a serious problems to organizational security. Old reports shows that the lack of proper encryption on files and communications due to human mistakes is one of the main causes of data loss. Organization need tools to identify the exposure of sensitive data by screening the content in storage and transmission to detect sensitive information have been stored or transmitted in the clear. However, detecting the sensitive information is challenging due to data transformation in the content. Transformations result is highly dramatically leak patterns. We are using some alignment techniques for detecting complicated data–leak patterns. Our algorithm is designed for preventing long sensitive data patterns. This prevention is paired between comparable sampling algorithm, that allows one to compare the between two separately sampled sequences. Our systems achieves best detection accuracy in recognizing transformed leaks. We implementing a sound change of our algorithms in graphics processing unit that results a high analyses is throughput. Our intention is to bring the high multi-threading scalability of the data leak detection method required through a commune.**

***Keywords*: Sensitive data, organizational security, data–leak patterns, parallelized version**

I. INTRODUCTION

Data leakage is a term that have been used in the information security field to give the information about unwanted disclosures of information. Unlike the traditional security threats from the outside, data leakage is mainly caused by the people who work in their organization though who may leak data to the outside unauthorized entities. So, traditional security measures i.e. firewalls, intrusion detection systems, anti-virus are no longer valid due to lack of understanding of data semantics.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 will going to happen from time to time, period to period and brings a serious damage continuously. To address the problem of data leakage detection (DLD), plenty of research work has been done with the use of hash fingerprinting, n-gram, statistical methods and so on. With them rapid development of Internet, many new communication technologies e.g., Device-to-Device technology have emerged. As a result, the volume of data grows dramatically and the forms of data becomes much complicated. This brings new challenge to DLD. Therefore, new DLD method is required with better tolerance of data transformation and higher efficiency to deal with the large amounts of unstructured data in long patterns.

II. 1.1 EXISTING SYSTEM:

Volume of data grows dramatically and the forms of data becomes much complicated. This brings new challenge to DLD. Therefore, new DLD method is required with better tolerance of data transformation and higher efficiency to deal with the large amounts of unstructured data in long patterns.

III. DISADVANTAGES OF EXISTING SYSTEM:

The Data transformation is one of the real application scenarios and very complicated. As data transformation is lightly common in real situation, a large number of sensitive data may be leaked once the detection can't tolerate the transformed data. Hence, how to better tolerate the data transformation is of vital importance for DLD.

IV. PROPOSED SYSTEM:

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

ADVANTAGE PROPOSED SYSTEM

Advantages of Proposed System: AGW is proposed to handle extensively transformed data. We use the adaptive context weight mechanism to better preserve the information of key terms and their context,which can tolerate a large degree of data transformation. We also propose a low-complexity algorithm with a weight reward and penalty mechanism, to deal with the large amounts of data in big data scenario. Thus, AGW can improve the detection accuracy towards transformed data and has a low computational complexity.

Technique: Term Frequency Technique
Term frequency–inverse document frequency it is a numeric measure that is used to score the importance of a word in a document based on how they did it appear in that document and more collection of documents. If a word appears again and again in a document, then it should be more important and we should give that word a highest score. But if a word appears in too many other documents, it's probably not a same identifier, therefore so we should assign a lower score to that word. The math formula for this measure:

tfidf(t,d,D)=tf(t,d)×idf(t,D)tfidf(t,d,D)=tf(t,d)×idf(t,D)Where t

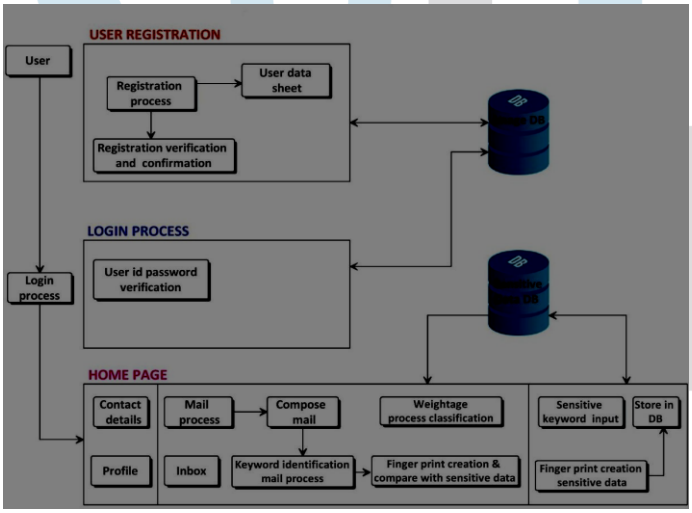
Denotes the terms; d denotes each document; D denotes then collection of documents.

V. IN THIS DOCUMENTATION, WE ARE TAKING THIS FORMULA FOR USING

four small documents to illustrate.
Library(tm)
Library (proxy)
Library (dplyr)
Doc<_c ("The rose is red."" The rose is bright today.")
("The rose in the flowers is bright", "We can see the shining rose, the bright flowers.")

2.2 Technique: Pre-processing Technique

Pre-processing is one of the data mining technique that involves the transforming raw data into an understandable format. Real-world data is often not completed, inconsistented, or lacking in certain in there behaviors or trends, and it contain many errors. Data preprocessing is a known method of resolving such issues. it prepares raw data for future processing and also it is used database-driven applications such as customer relationship management and rule-based applications (like neural networks).
Data goes through a series of steps during pre-processing:
1. Cleaning: Data is cleaned from the processes such as we are going to fill the missing values.
2. **Integration**: Data with different representations are put together and conflicts within the data are resolved.
3. **Transformation:** Data is normalized, aggregated and generalized.
4. **Reduction:** This step aims to present a reduced representation of the data in a data warehouse.
5. **Discretization:** Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals



CONCLUSIONS AND FUTURE WORK

Presented a content inspection technique for detecting leaks of sensitive information in the content of files or network traffic. Our detection approach is based on aligning two sampled sequences for similarity comparison. Our experimental results suggest that our alignment method is useful for detecting multiple common data leak scenarios. The parallel versions of our prototype provide substantial speedup and indicate high scalability of our design. Future work is planned to explore data-movementtracking approaches for data leak prevention on a host.

REFERENCES

[1] X. Shu, J. Zhang, D. Yao, and W.-C. Feng, "Rapid and parallel content screening for detecting transformed data exposure," in Proc. 3rd Int. Workshop Secure. Privacy Big Data (Big Security), Apr. /May 2015, pp. 191–196.
[2] S. Kumar, B. Chandrasekaran, J. Turner, and G. Varghese, "Curing regular expressions matching algorithms from insomnia, amnesia, and calculi," in Proc. 3rd ACM/IEEE Sump. Archit. Newt. Common. Syst. (ANCS), 2007, pp. 155–164.
[3] R. Hoyle, S. Patel, D. White. Dawson, P. Whalen, and A. Kapadia, "Attire: Conveying Information exposure through avatar apparel," in Proc. Conf. Compute. Supported Cooperate. Work Companion (CSCW), 2013, pp. 19–22.

[4] J. Croft and M. Caesar, "Towards practical avoidance of information leakage in Enterprise networks," in Proc. 6th USENIX Conf. Hot Topics Secure. (Hot Sec), 2011, p.7.
[5] V. P. Emeril's, V. Pappas, G. Portokalidis, and A. D. Keromytis, "iLeak: A lightweight system for detecting inadvertent information leaks," in Proc. 6th Eur. Conf. Compute. Newt. Defense, Oct. 2010, pp. 21–28.
[6] L. De Carli, R. Sommer, and S. Jha, "Beyond pattern matching: A concurrency model for stateful deep packet inspection," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., 2014, pp. 1378–1390
[7] S. Jha, L. Kruger, and V. Shmatikov, "Towards practical privacy for genomic computation," in Proceedings of the IEEE Symposium on Security and Privacy, 2008. [50] V. O. Polyanovsky, M. A. Roytberg, and V. G. Tumanyan, "Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences," Algorithms Mol Biol, vol. 6, no. 1, p. 25, 2011.
[8] A. Z. Broder, M. Charikar, A. M. Frieze, and M. Mitzenmacher, "Minwise independent permutations," J. Comput. Syst. Sci., vol. 60, no. 3, pp. 630–659, 2000.
[9] P. K. Agarwal and R. Sharathkumar, "Streaming algorithms for extent problems in high dimensions," in Proceedings of the 21st Annual ACMSIAM Symposium on Discrete Algorithms, ser. SODA '10, 2010, pp. 1481–1489.
[10] D. Feldman, M. Monemizadeh, C. Sohler, and D. P. Woodruff, "Coresets and sketches for high dimensional subspace approximation problems," in Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms, ser. SODA '10, 2010, pp. 630–649.
[11] M. O. Rabin, "Fingerprinting by random polynomials," Center for Research in Computing Technology, Harvard University, Tech. Rep., 1981, tR-15-81.
[12] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," J. Mol. Biol., vol. 147, no. 1, pp. 195–197, March 1981.
[13] C. Kalyan and K. Chandrasekaran, "Information leak detection in financial e-mails using mail pattern analysis under partial information," in Proceedings of the 7th WSEAS International Conference on Applied Informatics and Communications, ser. AIC'07, vol. 7, 2007, pp. 104– 109.
[14] C. Wuest and E. Florio, "Firefox and malware: When browsers attack," Symantec Corporation, Tech. Rep., October 2009.
[15] W. Liu, B. Schmidt, G. Voss, A. Schroder, and W. Muller-Wittig, "Biosequence database scanning on a GPU," in Proceedings of the 20th International Parallel and Distributed Processing Symposium, 2006.
[16] S. Manavski and G. Valle, "CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment," BMC bioinformatics, vol. 9, no. 2, p. 10, 2008.
[17] Y. Liu, D. Maskell, and B. Schmidt, "CUDASW++: optimizing SmithWaterman sequence database searches for CUDA-enabled graphics processing units," BMC Research Notes, vol. 2, no. 1, p. 73, 2009.
[18] M. A. Jamshed, J. Lee, S. Moon, I. Yun, D. Kim, S. Lee, Y. Yi, and K. Park, "Kargus: a highly-scalable software-based intrusion detection system," in ACM Conference on Computer and Communications Security, 2012, pp. 317–328.
[19] K. Lee, H. Lin, and W. Feng, "Performance characterization of dataintensive kernels on AMD fusion architectures," Computer Science - Research and Development, pp. 1–10, 2012.