An Intelligent Email Filtering System to Prevent Data Leakage

Dipali Patil, Rajnish Singh, Ashish Kumar Rai, Tanmoy Mani, Pintu Singh

***Abstract*— Data leakage is defined as the accidental or unintentional distribution of private or sensitive data to an unauthorized entity. Most of the organizations face the problem of data leakage. Security practitioners have always had to deal with data leakage issues that arise from various ways like email and other internet channels. Hence, there is a need to filter e-mails. This can be done by using an intelligent system which will filter email for organization's confidential data. Concept used in Email classification is we check mail based on the message bodies, black lists consisting of hash of organization's secured data are computed and stored. Then computed hash of email contents is computed with the black list data and depending on the match the email is either blocked or forwarded.**

***Index Terms*— Email classification, Data leakage, secured Data Hash, Black list, Waitlist, white list, Email continuity, word threshold value.**

I. INTRODUCTION

A data leak is a security incident in which confidential data is stolen or used by an individual unauthorized to do so. In the course of business communication, mostly confidential data must be distributed to supposedly trusted third parties .The owner of the data is the distributor and the supposedly trusted third parties are the agents. Despite of many security rules and regulation in use today employees of organization are involved in risky behaviors that put organization's data and personal data at risk. Emailing system is a most used tool for leaking sensitive data. Considering these data leakage threats through email, an organization blocks the email of employees or restricts the access of email hosting sites by using firewall. Email filtering system is an intelligent system which allows employees of an organization to compose and forward email by filtering email content to leakage of organization's sensitive data. Any type data can be leaked through emails like video,audio,text,zip folders, images and files .In our proposed system owner of the origination is the administrator and trusted agents employees of firm. The aim of the system is to prevent the leakage of organization's sensitive data through email.

Manuscript received February 08, 2015.

Dipali Patil is a prof. in computer dept of sinhgad institute of technology of Savitaribai phule pune University,Pune,Iindia

Rajnish Singh is a final year student in computer enginnering in sinhgad institute of technology of Savitraribai phule pune university ,Pune,India

Ashish Kumar Rai is a final yaer student in computer engineering in sinhgad institute of technology of Savitari bai phule pune university,Pune ,India

Tanmoy Mani is a final yaer student in computer engineering in sinhgad institute of technology of Savitari bai phule pune university,Pune ,India

Pintu Singh is a final yaer student in computer engineering in sinhgad institute of technology of Savitari bai phule pune university,Pune ,India

II. RELATED WORK

Currently most of the organizations are using firewall to prevent data leakage through email by blocking email hosting sites like Gmail, rediffmail, yahoo etc. By this approach employees cannot send emails within the organization. The drawback of this approach is that it cannot operate on individual files and email, as firewall only works on packets. Because to this drawback firewall cannot block emails by filtering their contents. Some previous approach uses the method of watermarking the data to identify the guilty agent. So the drawback of this system was watermark can be easily removed and it can't be applies to all types of data. A system previously developed uses the concept of fake objects.

Fake objects are added to the databases which appear exactly as the original data. The algorithm used for classifying the emails is the K-nearest algorithm. The problem with fake objects concept is that it is only suitable for the fixed data. Most of the organization used Microsoft outlook express to handle the emails. Outlook provides a heterogeneous environment but has no provision for email filtering. Firewall approach is combined with outlook to prevent data leakage. The major drawback of using outlook was node to node internet connectivity is required. Considering all these limitations there was an urgent need to develop a system which can cover these drawbacks. Our proposed system fulfills that need.

III. PROPOSED SYSTEM

Due to many drawbacks in existing system it is not possible to prevent data leakage without affecting the email continuity. Another drawback of existing system is that they are not cost effective. So our proposed system will cover all these limitations and prevent data leakage but at same time it will provide email continuity which is very part of organization's work. It can used by large as well as small organizations because it very cost effective. The diagram shown below shows the architecture of proposed system.

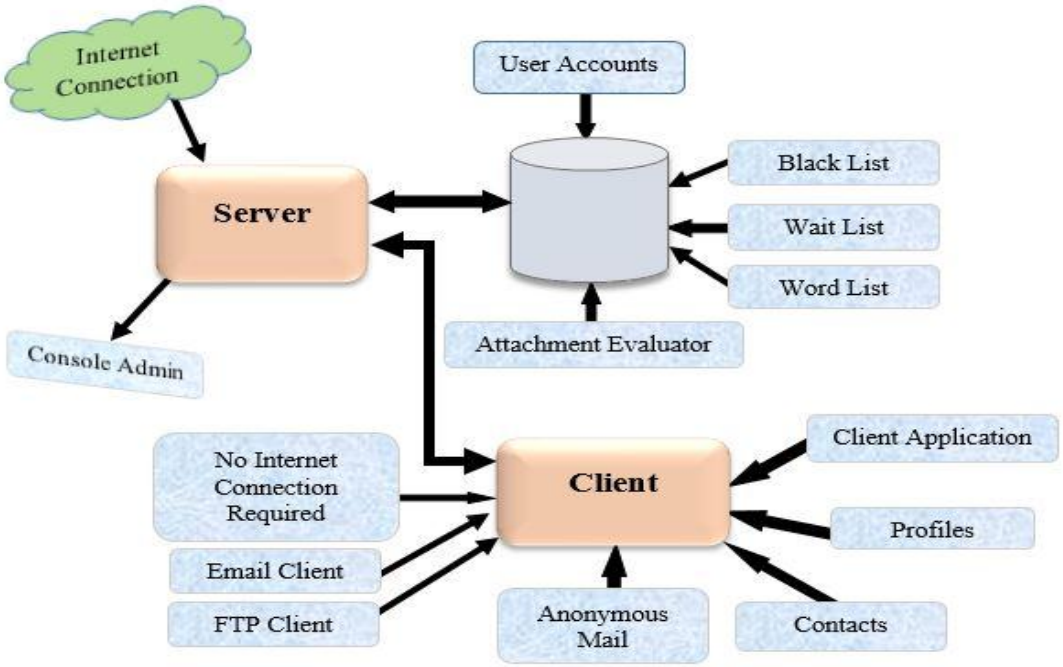

Fig. 1: Proposed system architecture [4]

The proposed system is based on client- server architecture. Employees of the organization are at the client side and administrator is at the server side. High speed internet connection is provided only at server side and clients communicate with server through servlet method. The server consists of waitlist, blacklist and word list and attachment evaluator. The black list contains hash of organization's sensitive data which is maintained by the administrator. The admin will allow or block the mails from the wait list. When employees send mail, it first comes to the server. Then server will calculate hash of the mail's data and attachment which will be compared with the black list data and depending on the match, the email will be either forwarded or blocked. If the server fails to detect whether the email contents are black listed or not, email will be sent to the wait list. The system will classify e-mail based on the message bodies, the white and black lists consisting of hash of organization's data is stored. The hash of email contents is calculated and compared with the white and black list data and depending on the match the email is either discarded or sent. The following diagram shows how and when the mail will be blocked

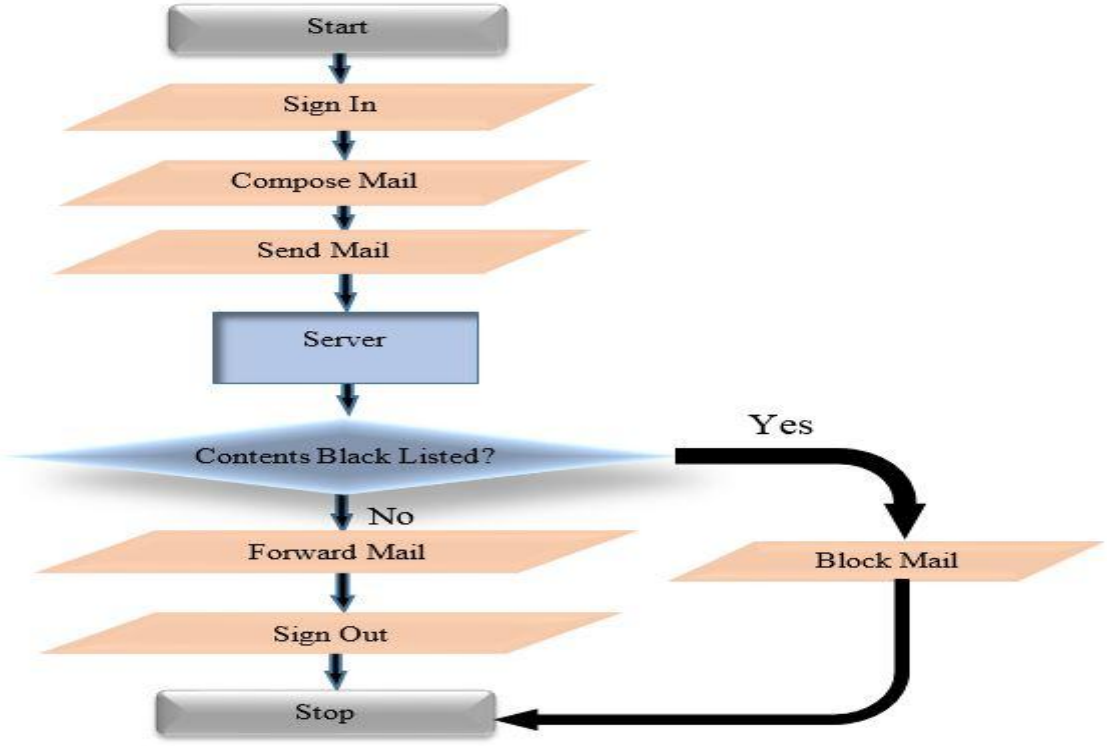

Fig. 2: Flow diagram of the system

IV. ALGORITHMS USED

A. Secure hash algorithm [sha-1]: The SHA1 is used to compute a message digest for a message or data file that is given as input. The message body and attachment will be considered to be a bit string. The length of the message is the number of bits in the message (the empty message has length 0).If the number of bits in a message is a multiple of 8, for compactness we are representing the message in hex. The SHA1 produces 160 bit message digest .The purpose of message padding is to make the total length of a padded message a multiple of 512 bit blocks. At client side the hash of the password is obtained using SHA1.At server side SHA1 is used to obtain the hash of the sensitive data and it is stored as a black listed data.

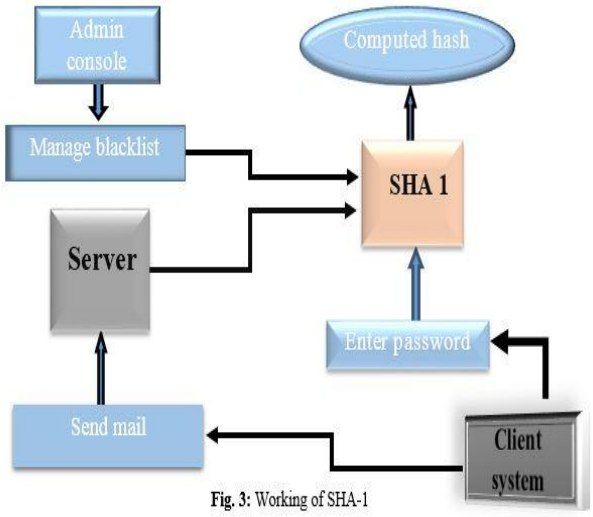

Fig. 3: Working of SHA-1

B. Term frequency: In our proposed system, we are making using of term frequency algorithm in which the data in the mail will be checked with word list. Threshold value for each word will be stored in the stored word list. If the use of any word crosses the threshold value, the mail will be blocked. Equation for Term Frequency (tf) is given as: **Wt=ct log (N/ft)** here Wt is the weight of term ft is the number of times the term in the mail, ct is number of times the term in the passage, and N is the total number of terms in mail.
Advantage: As the term frequency algorithm does not require information regarding the structure or grammar of the natural language. Therefore the algorithm may be used in many natural languages.

V. CONCLUSIONS

Many organizations handle confidential data on daily basis. The technologies that make this data easily available also increase the risk of data leakage. Some mechanisms have been implemented to prevent data leakage such as firewall mechanism which restricts access to email sites which hampers email continuity, filtering email using fake object mechanism which is only suitable for fixed database. Considering these the limitations, we have shown that it is possible to develop a system which will provide email continuity along with filtering capability for sensitive data leakage without any internet connection at client side. The algorithmic strategy used provides email filtering for any size and type of data.

ACKNOWLEDGMENT

The success of any project is never limited to an individual. Similarly, our project is also an outcome of ideas contributed by many and we would like to gratefully acknowledge them here. We express our sincere thanks to all those who have provided us with valuable guidance towards the completion of this report as a part of the syllabus of the degree course. We deeply thank our HOD Prof. Babar for his useful guidance. We also thank our Project Guide Prof. D.S. Patil without whom this project would have been a distant reality. We also thank them for giving us moral support, timely comments and discussion in all phases of the project.

REFERENCES

[1] Ankit Agarwal, Mayur Gaikwad, Department of Information Technology, University of Pune, Kapil Garg, Vahid Inamdar, " *Robust Data Leakage and Email Filtering System"* International Conference on computing Electronics and Electrical Technologies [ICCEET], 2012
[2] Panagiotis Papadimitriou and Hector Garcia- Molina, *"Data Leakage Detection"* IEEE Transactions on Knowledge and Data Engineering, Vol 23,No.1 january2011.
[3] F. Guo, J. Wang, Z. Zhang, X. Ye, and D. Li, "*An Improved Algorithm to Watermark Numeric Relational Data",* Information Security Applications, pp. 138-149, Springer, 2006.
[4] Shrihari Ahire, Vishakha Panjabi, Rahul Jagtap, Madhuri Bagul, A.S.Deokar. "*Secure Email System for SOHO (Small Office Home Office)"* Volume 4, Issue 1, January 2014 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.
[5] Savita Pundalik Teli,Santoshkumar Biradar "Effective Email Classification for Spam and Non-Spam" university of Pune,Maharastra,india, Volume 4, Issue 6, June 2014 ISSN: 2277 128X, International Journal of Advanced Research in Computer Science and Software Engineering.

Dipali Patil is a prof. in computer dept of sinhgad institute of technology of Savitaribai phule pune University,Pune,Iindia
Rajnish Singh is a final year student in computer enginnering in sinhgad institute of technology of Savitraribai phule pune university ,Pune,India
Ashish Kumar Rai is a final yaer student in computer engineering in sinhgad institute of technology of Savitari bai phule pune university,Pune ,India
Tanmoy Mani is a final yaer student in computer engineering in sinhgad institute of technology of Savitari bai phule pune university,Pune ,India
Pintu Singh is a final yaer student in computer engineering in sinhgad institute of technology of Savitari bai phule pune university,Pune ,India