NNGT Int. J. of Information Systems , Vol. 1, July 2014

Data Leakage/Loss Prevention Systems (DLP)

Radwan Tahboub
College of Information Technology and Computer Eng.
Palestine Polytechnic University
Hebron, Palestine
radwant@ppu.edu

Yousef Saleh
Deanship Of Graduate Studies & Scientific Research
Palestine Polytechnic University
Hebron, Palestine
yousefsaleh@hotmail.com

***Abstract*- Sensitive and confidential data are a requisite for most companies, so protection for this data takes great attention by company's top management, administrators and IT managers. Data leakage causes negative impact on companies. The traditional security approaches, such as firewalls, can't protect data from leakage. Data leakage/loss prevention (DLP) systems are solutions that protect sensitive data from being in non-trusted hands. This paper is an attempt to survey and study DLP systems that will be conducted as well as a comparison with other security and data protection approaches.**

Keywords- Data Leakage Prevention (DLP), Confidential Data, Firewall, Unauthorized user, Cloud computing.

I. INTRODUCTION

Data in each company is one of the most important assets; therefore the protection of this data must take the first priority. Although the companies have security measurements and technical parries such as firewalls, still the data leakage occurs.
The data leakage happens when sensitive data is revealed to unauthorized parties whether it's intentionally or not. The data leaked may cause serious threats to a company. The loss of confidential or sensitive data can severely impact a company's reputation, customers and employee confidence, competitive advantage and in some cases lead to the closure of the company, or political crises such as WikiLeaks' leaks [31].

Data leakage problem must be solved using the Data Leakage/Loss Prevention System (DLP). DLP solutions help identifying, monitoring, protecting and reducing the risks of sensitive-data leakage. It is used to detect and prevent unauthorized user from getting sensitive data, and even to protect confidential data that can be accidentally shared [32].
In this paper we will first talk about the existing security approaches used in Data protection and in the second section we will talk about data leakage prevention systems, finally we will compare between them.

II. EXISTING TECHNOLOGIES USED IN DATA PROTECTION

There are different technologies used to protect data, most of them concentrate on protecting data from outside while the DLP system concentrate on data protection from inside.

A. IDS / IPS

Intrusion Detection Systems (IDS) is a device or software application that monitors networks or system activities for malicious activities.

Intrusion Prevention Systems (IPS) monitors networks, system activities for malicious activities it mainly identifies malicious activities, log information, attempt to block or stop activities and report activities.

IPS systems are expansion of intrusion detection systems because they both monitor network traffic, system activities for malicious activity, IPS are able to prevent or block intrusions that are detected. This can perform actions such as indicating an alarm, leaving the malicious packets.

IDS/IPS Components: IDS/IPS can be divided into two categories: network based intrusion/prevention systems (NIDS/NIPS) and host based systems (HIDS/HIPS) [6].

NIDS/NIPS checks packets on the network and looks at the data in an attempt to recognize an attack. (HIDS/HIPS) monitor traffic on one specific system, HIDS/HIPS excel at detecting/preventing unauthorized access and activity. HIDS/HIPS look at the state of a system and verify that all behaviour appear as expected. Both NIDS/NIPS and HIDS/HIPS can be used to scan for attack, track hackers' movements, and alert an administrator to ongoing attacks.

Most IDS/IPS consists of more than one application or hardware device. IDS/IPS are composed of the following parts [7].

- Network Sensors: Detect and send data to the systems.
- Central Monitoring System: Processes and analyzes data sent from sensors.
- Report Analysis: offers information about how to counteract a specific event.
- Database: store the IP and information about the attacker.
- Response Box: inputs information from the previous components and forms an appropriate response.

IDS/IPS Approach: IDS/IPS techniques can be divided into two approaches:

A signature-based or pattern matching: IDS/IPS depends on a database of known attacks. These known attacks are loaded into the system as signatures [7].

The biggest disadvantage of signature-based systems is that it can trigger only on signatures that have been loaded. A

DOI : 01.IJIS.2014.1.2

new attack may go undetected. Snort is a god example of signature-based IDS/IPS .as shown in Figure 1 [30].

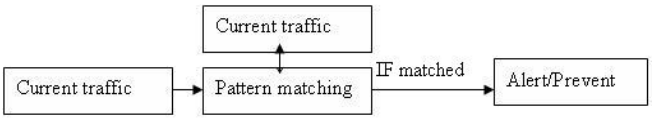

Figure 1. Signature-Based.

Anomaly-Detection systems: require the administrator to make use of profile of authorizes activates or place the IDS/IPS into learning mode so that it can learn what constitutes normal activity [8]. As shown in Figure 2.

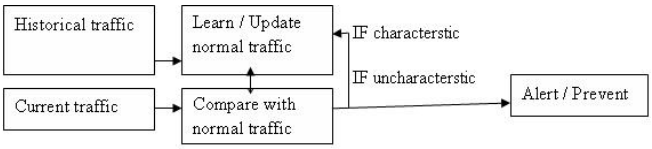

Figure 2. Anomaly -Based.

Anomaly detention is good at monitoring behaviors that is greatly different from normal activities.

B. Anti-Malware

Malware is software designed to damage computer operations, collect sensitive data, and gain unauthorized access to computer system. Type of malwares is virus, worms, Trojans, spyware, backdoors, and root kit [10].
Malware has two categories:

- Malware which modifies the recourses such as memory, BIOS Code, PCI devices expansions EEPROMS [11].
- Malware that does not modify any of those resources, but only the resources which are dynamic by nature, like e.g. data sections, such as by modifying some function pointers in some kernel data structures, so that the attackers code gets executed instead of the original system or application.

Types of malware:

- Virus: Replicates by attaching its program instructions to an ordinary host program or document, so that the virus instructions are executed when the host program is executed. Such as file virus, boot sector virus, micro virus, e-mail virus [11].
- Network worm: Self-propagating program that spreads over a network, usually the Internet. Unlike viruses, may not depend on other programs or victim actions (such as opening an infected email attachment or clicking on a web link for a malware Web site) for replication, dissemination, or execution. Such as swarm worm [27].
- Trojan Horse (or, simply, Trojan): A destructive program that masquerades as a benign program. Stealth ware such as spyware, root kits, key loggers, trapdoors, and certain adware represents a subset of Trojans that is intentionally designed to be hard-to detect or undetectable Trojan horse software installs itself on the victim's computer when the victim opens an email attachment or computer file containing the Trojan, or clicks on a Web link that directs the victim's browser to a Web site from which the Trojan is automatically downloaded. Such as Backdoor Trojan [24].
- Spyware (non-Trojan): Non-Trojan stealth-ware that has the same objectives and performs the same types of actions as spyware Trojans. A number of bots have spyware capabilities, and are referred to as spy-bots such as Adware.
- Embedded Malicious Code: Any type of malicious logic embedded in a valid executable program by its developer, integrator, distributor, or installer. Most frequently a logic bomb, time bomb, or Trojan.

Anti-malware is any software that provide a protect computers and systems from malware, viruses, spyware and other harmful program.
Anti-malware software works in real time environment very effectively but it only looks for threats from outside, by scanning and signature validation it ensures that malware infection will be removed.

C. Firewalls

Firewalls are devices or software that permits or denies network transmissions based on a set of rules (access rule) and is used to protect networks from unauthorized access while permitting legal communication to pass [27].

Firewall is software or hardware that helps in keeping network secure, its objective is to control the incoming and outgoing traffic of networks by analyzing the data packet and determining whether it should be allowed through or not [27].

D. Vender Patches

Are the updates the systems, applications and install new hotfixes that recommended by vendors [7].

D. SIEM

Security Information Event Management (SIEM) is a tool used on enterprise data network to centralize the storage of logs which was generated by the software running on the network, as well as gathering information, analyzing the information and also presenting the information from network and security devices [33].

III. DEEP PACKET INSPECTION (DPI) METHOD

All the previous techniques are using Deep Packet Inspection (DPI) method, DPI looks at the packet, finds anomalies in the traffic and alerts the administrator or prevents the traffic .DPI classify passing traffic based on rules, these rules include information about layer 3 and layer 4 content of the packet as well as the information that describes the content of the packet's payload [12].

The following steps describe how Deep Packet Inspection Architecture works [28]:

DOI : 01.IJIS.2014.1.2

- Pattern Definition Language Interpreter uses signatures that can be written to detect and prevent against known and unknown protocols.
- The Deep Packet Inspection reassembles TCP packets arriving out-of-order.
- Deep Packet Inspection engine preprocessing involves normalization of the packet's payload. For example, a HTTP request may be URL encoded and thus the request is URL decoded in order to perform correct pattern matching on the payload.

Deep Packet Inspection engine postprocessors perform actions which may either simply pass the packet without modification, or could drop a packet or could even reset a TCP connection [34]. As shown in Figure 3.

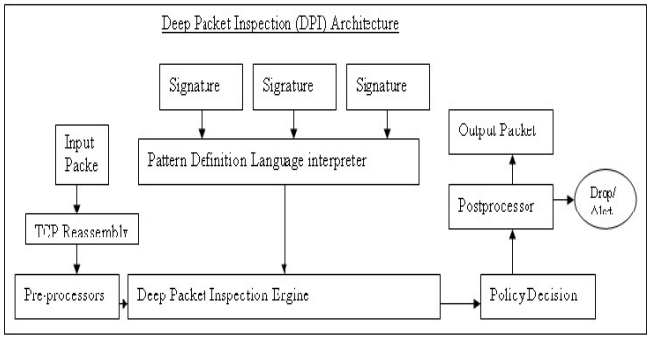

Figure 3. Deep Packet Inspection Architecture.

IV. Data Leakage/Loss Prevention System (DLP)

A. *Definition of DLP*

Data Leakage Prevention is a solution designed to detect potential data breach incidents in timely manner and prevent them by monitoring data while in-use (endpoint actions) or in-motion (network traffic) or at-rest (data storage). DLP is considered as a paradigm shift in information security, it addresses risks; these risks are focused around handling certain types of data with information security risk, such as personally identifiable, credit cards, financial and legal information. Exposing this data outside a company's security perimeter leads to consequences detrimental to the company. DLP solutions help to protect data from going outside company [2].

B. *Data State*

DLP solutions distinguish between three phases of data throughout their lifecycle: data-at-rest (DAR), data-in-motion (DIM) and data-in-use (DIU) [32].

Data-At-Rest are defined as all data in computer storage. To keep data-at-rest from being accessed, stolen, or altered by unauthorized people, security measures such as data encryption and access control are commonly used. A prerequisite for these security measures is content discovery, which serves to find where all the data are stored. One way to achieve this is using the content discovery features of DLP products. For example, a policy may require that customer credit card numbers be stored only on approved servers. If data are detected on an unauthorized server, they can be encrypted or removed, or a warning can be sent to the data owner.

Data-In-Use are any data with which a user is interacting. Endpoint-related systems are used to protect data-in-use and to monitor data as the user interacts with them. Usually, an agent is used to monitor the data while they are being used or transported from an endpoint device or client through different output channels to peripheral devices. The underlying idea is that if an attempt is made to send sensitive data, the potential leakage will be immediately detected and tackled (e.g., blocked) before the data can be sent. Data-in-use tools may monitor the following activities:

• Copy-paste and screen-capture operations involving sensitive data.
• Transfer of sensitive content from one place to another using portable storage device such as USB drives, CD/DVDs, smart phones, and PDAs.
• Printing or faxing sensitive content.

Data-In-Motion are data that are being sent through a network. These data may be sent inside the internal network of an organization or may cross over into an external network. DLP solutions are used to detect and inspect data which are being sent across communication channels over a network using known protocols, including email, http, instant messaging, and even unknown protocols (by simply inspecting the packets' content). If encryption or encrypted connections are permitted without the ability to decrypt the data, a DLP solution will not be able to detect leakage of encrypted sensitive data-in-motion.

C. *Data Leakage/Loss Prevention Capabilities*

Data Leakage/loss Prevention system is to deliver an unified solution to discover, monitor and protect confidential or sensitive data wherever it is stored or used, across endpoint, network, and storage systems. DLP solutions are used to address the business problem of protecting confidential data. The solution components are designed as follows and as shown in Figure 4.

- Discover where confidential or sensitive data is stored [4].
- Monitor how data is being used.
- Proactively protect data to prevent its loss.

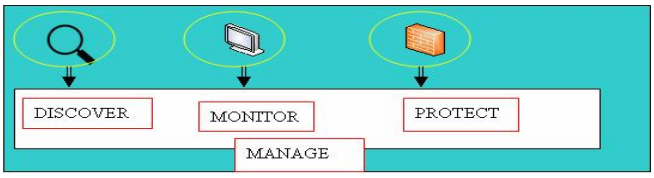

Figure 4. Loss Modes.

D. *Data Leakage/Loss Prevention Products*

DLP consists of components designed to work together to monitor and protect sensitive data wherever the data is stored and whenever it is sent outside organization [1].
DLP solutions consist of four main components as follows and as shown in Figure 5.

- Network.
- Management.
- Storage (Data Center).
- Endpoint.

Figure 5. DLP Architecture.

DOI : 01.IJIS.2014.1.2

- *Network (data in motion)*

Network coverage is referred to data in motion, coverage monitor and protects data that is being transmitted over network to Internet, examples include common business applications, including
E-mail as simple mail transport protocol (SMTP), web mail as hypertext transport protocol (HTTP), and file transfer protocol (FTP) [1].

- *Storage (discover data at rest)*

Storage coverage is referred to data at rest, scans and protects data that is stored in data repositories. Examples include public and private shares and databases [2].

- *Endpoint (data in-use, data at rest)*

Endpoint refers to devices such as laptops, desktops, or workstations. Monitors and protects data as it is moved to or off the endpoint machines. Examples include downloading data to the hard drive and copying data to removable media (USB) [3].

- *Management*

Refers to central management platform consisting of management server and databases. All users log into the management consol. This is the users interface that manages all Data Leakage/Loss Prevention policies, workflow, reporting, users, roles, system management, and security. All detection servers are managed by Management platform. Management server pushes appropriate policies, detection, and server configuration to detection servers [1].

E. DLP Methods

DLP use Deep Content Inspection (DCI) that considered the evolution of Deep Packet Inspection with the ability to look at what the actual content contains instead of focusing on individual or multiple packets. Deep Content Inspection allows services to keep track of content across multiple packets so that the signatures they may be searching for can cross packet boundaries and yet they will still be found. DCI classify passing content based on rules, these rules include information about layer 7.Figure 6 show the architecture of Deep Content Inspection (DCI)[35].

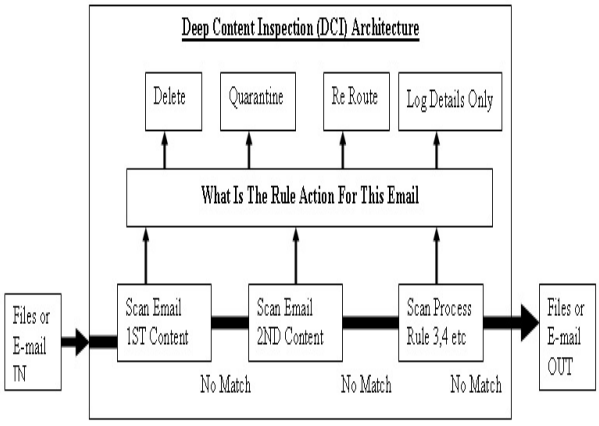

Figure 6. Deep Content Inspection (DCI).

1) Content Matching: works for structured and unstructured data, using keywords, pattern matching, regular expression, file types, file size, file properties, sender, recipient, and network protocol information to detect data loss incident [5]. Content matching, apply action, use Match-Join algorithms is shown Algorithm1. For each two tuples t1 ∋ T1 and t2 ∋ T2 that match, the match-join table J contains there concatenated tuple. Two tuples match if and only if for every common categorical attribute A: t1.A and t2.A are on the same generalization path in the taxonomy tree for A.

2) Learning Method (LM): DLP uses this approach; the basic idea from it is to use machine learning techniques such as Vector space model (SVM) [23], to determine the "confidentiality level" of the scanned email message. Method is the vector space model.

Vectors represent documents, and vector features represent terms and their frequency of appearance. The vectors are used as learning sets to build a probabilistic model, on the basis of which decisions are made whether or not documents are confidential.
This method is efficient for detecting unstructured content in cases where a deterministic technique is difficult to implement and statistical metrics are the best approach available.

Algorithm1 : Match-Join [29].

Input : *T*1 , *T*2
Output : *J* , the match join table of *T*1 and *T*2
01: *J* = Ø
02: **for all** *t*1 ∋ *T*1 ;
03: **for all** *t*2 ∋ *T*2 ;
04: **if** match(*t* 1 , *t* 2) **then**
05: *ApplyAction()* ;
06 t = t1 concatenated t2;
07: *J* = J Join { *t* }
08: **end if** ;
09: **end** for ;
10: **end for** ;
11: **return** *J* ;

F. Comparison between DLP methods

In the first method Content Matching is effective in case identifying all keywords and regular expression, but not effective in case changing document format, otherwise the second method needs to enter huge sample of documents until to increase accuracy issues and reduce the rate of false positive and false negative .The use of any of the previous methods based on its existing policies where these methods do not address the encrypted data and does not address the hidden data within the images, audio and video.
I suggest enhancements on this algorithm to deal with encrypted and hidden data as future work.

V. COMPARISON BETWEEN DLP (DCI) METHOD AND OTHER EXISTING METHODS (DPI).

Through my study of the existing systems and DLP system, each system has advantages and disadvantages, the existing systems are providing protection for networks from the outside and provide periodic reports on the security status of the network and systems and send alarm in case attack occurs, existing systems work with ad hoc approach which don't support centralized approach and without having the ability to content aware so without having the ability to prevent data leakage.

DOI : 01.IJIS.2014.1.2

DLP system works in conjunction with security tools that companies may already have deployed both on endpoint computers (for example, laptops and desktops) and on the network. These may include network and personal firewalls IDS/IPS, antivirus, antispam, encryption and digital rights management tools. The main difference between a DLP system and existing technologies is that DLP systems are content-aware; they are designed to give visibility into where the company's most sensitive data is stored, who has access to it, and where and by whom it is sent outside the company's network. Existing security applications cannot perform this level of monitoring. Additionally, DLP systems must provide comprehensive functionality to prevent this sensitive data from being sent outside the organization through an endpoint computer or through the network.

Disadvantage of the DLP system: DLP system cannot read the encrypted data and the data hidden within images, audio and video, and in implementation stages; because depending on collected data from all business departments may cause weakness in accuracy issues.
Also the Deep Packet Inspection has the ability to inspect the network packet headers. This ability allows firewalls to implement allow/block network access policies since the intent of the packages can be determined where they come from, where they are going, and what ports they are passing through, but inability to comprehend information in that packets. Deep content inspection has ability to understand information in the packets as shown in Figure 7 [36].

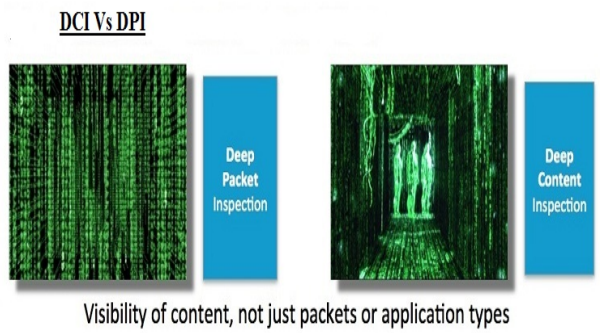

Figure 7. DCI Vs DPI.

VI. Related Work

In recent years, the challenge of dealing with the malicious insider has been acknowledged, and several methods have been proposed for solving this problem. Data leakage is one of the main goals of a malicious insider, and therefore most of the methods proposed for insider threat detection are also applicable for detecting and preventing data leakage.
Initially, Maybury et al [25] presented the results of a collaborative study involving a characterization and analysis of the methods used to counter malicious insiders in the U.S. intelligence community. The study proposes a generic model of malicious insider behavior, distinguishing motives, actions, and associated observables. Several prototype techniques were developed for providing early warning of malicious insider activity, including the use of honey tokens, network traffic profiling, and knowledge-based algorithms for structured analysis and data fusion. Hong et al [24] surveyed proposed methods for detecting insider attack in the research literature, including host-based user profiling based on features such as database and file system accesses, system calls, and OS commands; network based detection; and use of honey pots. Franqueira et al [26] distinguished between internal insiders and external insiders.
Mun et al [28] proposed the use of an intrusion detection system for detecting insider attackers. The proposed system is based on assigning grades and privilege levels to users and security levels to documents and monitoring user access to documents.

VII. Future Work

I suggest working on solving the problems of the existing system, such as work to develop classifiers algorithms or integrate current products that are able to read the encrypted data and the data hidden within images, audio and video, and dealing with content that is not a plain text, such as binary files.
Algorithm suggested provides file-unzipping capabilities to interpret a file when the content is obscured several levels down; for example, when an Excel spreadsheet is embedded in a zipped Word file. And policies enhanced in collecting data from business department. using DLP to protect data in cloud computing (Private and Public), this needs more enhancements in the design, policies and procedures to control all the data and application as required.
Mobile devices and particularly smart phones are expected to become the main computerized devices that members of the organization use and will be used in the future. Because smart phones are used to access the organization's confidential data such as emails and documents, it is expected that they will be used to leak information accidentally and intentionally. There have been several attempts to extend the organization's security perimeter into smart phones. There is a need for future research to find new approaches to give members of the organizations the access to confidential information through their smart phones, and on the other hand to prevent this information from leaking through the smart phone intentionally or accidentally.

VIII. Conclusions

This paper describes the importance of the information regards to companies and the seriousness of corporate data leakage. This paper studied the current systems used to protect data and the DLP system in terms of their components and methods used within them, and the differences between them. Showing the difference between existing systems and DLP system, as well as the importance of covering the shortage existed in current DLP systems such as of developing policies, Integration with other systems such as encryption, audio, video and images, so it is recommended to provide a scientific solutions to the problem of data leakage and mitigation.

References

[1] Simon Liu, Rick Kuhn , " Data Loss Prevention " , IEEE,USA ,2010.
[2] Preeti Raman, Hilmi Güneş Kayacık, and Anil Somayaji ,"Understanding Data Leak Prevention",ANNUAL SYMPOSIUM ON INFORMATION ASSURANCE (ASIA) Journal , JUNE 7-8,2011.
[3] ISACA , "Data Leak Prevention" ,vailable, http ://www.isaca.org/Knowledge- Center/Research/ResearchDeliverables/Pages/Data-Leak-Prevention.aspx.
[4] Tomoyoshi Takebayshi ,Hiroshi Tsuda , takayuki Hasebe ," Data Loss Prevention Technologies " , FUJITSU SCI,2010.

DOI : 01.IJIS.2014.1.2

[5] Michael Hart, Pratyusa Manadhata2,and Rob Johnson ," Text Classification for Data Loss Prevention", Stony Brook University, August 6, 2011.
[6] Hilmi A. Lahoud, PhD (ABD), Xin Tang, PhD,"Information Security Labs in IDS/IPS for Distance Education", SIGITE'06, October 19–21, 2006, Minneapolis, Minnesota, USA.
[7] Dhiren & Maulik, Hardik Joshi, Bhadresh K,Patel, "Towards Application Classification with vulnerability signature for IDS/IPS",SecurIT'12,Augest,17-19,2012,Kollam,Kerala,India.
[8] Tejinder Aulakh, "INTRUSION DETECTION AND PREVENTION SYSTEM: CGI ATTACKS", December 2009, San José State University.
[9] Yanfang Ye, Tao Li, Shenghuo Zhu, Weiwei Zhuang, Egemen Tas, Umesh Gupta, Melih Abdulhayoglu "Combining File Content and File Relations for Cloud Based Malware Detection", KDD'11, August 21–24, 2011, San Diego, California, USA.
[10] Simon A. Williamson, Pradeep Varakantham,, Debin Gao, Ong Chen Hui,"Active Malware Analysis using Stochastic Games", 11th International Conference on Autonomous Agents and Multiagent Systems – Innovative Applications Track (AAMAS 2012), Conitzer, Winikoff.
[11] Joanna Rutkowska,"Introducing Stealth Malware Taxonomy", November 2006, COSEINC.
[12] Robert Drum 2006, IDS AND IPS PLACEMENT FOR NETWORK PROTECTION, http://www.infosecwriters.com/text_resources/pdf/IDS_Placement_RDrum.pdf
[13] Nawal A. Elfeshawy • Osama S. Faragallah,"Divided two-part adaptive intrusion detection system", Published online: 13 June 2012 , Springer Science+Business Media, LLC 2012.
[14] Prathaben Kanagasingham,"Data Loss Prevention", 2008,SANSInstitute InfoSec .
[15] Gene Allen, Founder; ByStorm Software ,"Data Loss Prevention vs. Security Solutions and Your Data", ByStorm Software .
[16] Junchen Jiang, Yi Tang, Bin Liu, Yang Xu , Xiaofei Wang, "Skip Finite Automaton: A Content Scanning Engine toSecure Enterprise Networks" Global Telecommunications Conference , 2010 IEEE .
[17] Qihua Wang, Hongxia Jin,"Data Leakage Mitigation for Discretionary Access Control in Collaboration Clouds", SACMAT'11, June 15–17, 2011, Innsbruck, Austria.
[18] Powell Hamilton ,"Data Loss Prevention Program -WP ",Foundstone
[19] Wikipedia http://en.wikipedia.org/wiki/Data_loss_prevention_software
[20] ADP , http://www.adp.com/about-us.aspx
[21] ADP , http://www.adp.com/~/media/Corporate%20Overview/ADP-Corporate-Overview.ashx
[22] Ted Holland, "Understanding IPS and IDS:Using IPS and IDS together for Defense in Depth", 2004, SANSInstitute InfoSec.
[23] Cohen, W.W., and Singer, Y. 1999. Context-sensitive learning methods for text categorization.ACM Transactions on Information Systems (TOIS), 17(2), 11–17.
[24] Hong, J., Kim, J., and Cho, J. 2010. The trend of the security research for the insider cyber threat. International Journal of Future Generation Communication and Networking.
[25] Maybury, M,Analysis and detection of malicious insiders. Proceedings, 2005 International Conference on Intelligence Analysis, 2005.
[26] Franqueira,V., Cleeff, A., Eck, P., and Wieringa, R. 2010. External insider threat: a real security challenge in enterprise value webs. Proceedings, 5th International Conference on Availability, Reliability and Security.
[27] Salem, B.M., Heshkop, S., and Stolfo, S.J. 2008. A survey of insider attack detection eesearch. Insider Attack and Cyber Security- Beyond the Hacker, Springer.
[28] Mun, H., Han, K., Yeun, C.Y., and Kim, K. 2008. Yet another intrusion detection system against Insider Attacks. Proceedings, Symposium on Cryptography and Information Security .
[29] Erez Shmuelia,c, Tamir Tassab,c, Raz Wassersteina, Bracha Shapiraa, Lior Rokach, Limiting Disclosure of Sensitive Data in Sequential Releases of Databases,2012
[30] Mohammad A. Faysel , and Syed S. Haque Towards Cyber Defense: Research in Intrusion Detection and Intrusion Prevention Systems , July 2010
[31] Wikipedia,http://en.wikipedia.org/wiki/WikiLeaks
[32] Frost & Sullivan.World Data Leakage Prevention Market . Technical Report ND34D-74,Frost & Sullivan, United States. 2008.
[33] SearchSecurity , http://searchsecurity.techtarget.com/definition/security-information-and-event-management-SIEM.
[34] Sarang Dharmapurikar Praveen Krishnamurthy Todd Sproull John Lockwood, Deep Packet Inspection using Parallel Bloom Filters.
[35] wikipedia,http://en.wikipedia.org/wiki/Deep_content_inspection.
[36] wedgenetworks , http://www.wedgenetworks.com/resources/technology/deep-content-inspection-with-wedgeos.html.

DOI : 01.IJIS.2014.1.2