IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 6, Ver. III (Nov – Dec. 2014), PP 57-59
www.iosrjournals.org

Data Trawling and Security Strategies

Venkata Karthik Gullapalli[1], Aishwarya Asesh[2]
[1](School of Computing Science and Engineering, VIT University, India)
[2](School of Computing Science and Engineering, VIT University, India)

Abstract: *The amount of data in the world seems increasing and computers make it easy to save the data. Companies offer data storage by providing cloud services and the amount of data being stored in these servers is increasing rapidly. In data mining, the data is stored electronically and the search is automated or at least augmented by computer. As the volume of data increases, inexorably, the proportion of it that people understand decreases alarmingly. This paper presents the data leakage problem arises because the services like Facebook and Google store all your data unencrypted on their servers, making it easy for them, or governments and hackers, to monitor the data.*

I. Introduction

Data mining is defined as the process of discovering patterns in data. The process must be automatic or semiautomatic. The patterns discovered must be meaningful in that they lead to some advantage, usually an economic one. The data is invariably present in substantial quantities. Data mining is about solving problems by analyzing data already present in databases. The World Wide Web is becoming an important medium for sharing the information related to wide range of topics. According to most predictions, majority of the human information will be available on the web in five years. The building blocks of data mining is the evolution of a field with the confluences of various disciplines, which includes database management systems(DBMS), Statistics, Artificial Intelligence(AI), and Machine Learning(ML) [3]. Given a truly massive amount of data, the challenge in data mining is to unearth hidden relationships among various attributes of data and between several snapshots of data over a period of time [7]. These hidden patterns have enormous potential in prediction and personalization [7].

II. Data Leakage Problems

Major organizations still leave users passwords vulnerable Password vulnerabilities ought to be a rarity. Well-known and easily-followed techniques exist for generating, using and storing passwords that should keep both individuals and organizations safe. Yet in 2012 we saw one massive password breach after another, at a slew of high profile organizations. Russian cybercriminals posted nearly 6.5 million LinkedIn passwords on the Internet. Teams of hackers rapidly went to work attacking those passwords, and cracked more than 60% within days. That task was made simpler by the fact that LinkedIn hadn't "salted" its password database with random data before encrypting it. Dating website eHarmony quickly reported that some 1.5 million of its own passwords were uploaded to the web following the same attack that hit LinkedIn. Form spring discovered that the passwords of 420,000 of its users had been compromised and posted online, and instructed all 28 million of the site's members to change their passwords as a precaution. Yahoo Voices admitted that nearly 500,000 of its own emails and passwords had been stolen. Multinational technology firm Philips was attacked by the rootbeer gang. The gang walked away with thousands of names, telephone numbers, addresses and unencrypted passwords. IEEE, the world's largest professional association for the advancement of technology, left a log file of nearly 400 million web requests in a world-readable directory. Those requests included the usernames and plain text passwords of nearly 100,000 unique user's.

In an attempt to ascertain Cloud Computing reliability, 11,491 news articles on cloud computing-related outages from 39 news sources between Jan 2008 and Feb 2012 – effectively covering the first five years of cloud computing - were reviewed [1]. During this period, the number of cloud vulnerability incidents rose considerably. For instance, the number of cloud vulnerability incidents more than doubled over a four year period, increasing from 33 in 2009 to 71 in 2011. A total of 172 unique cloud computing outage incidents were uncovered, of which 129 (75%) declared their cause(s) while 43 (25%) did not. As cloud computing matures into mainstream computing, transparency in the disclosure of outages is imperative.

The scope for data leakage is very wide, and not limited to just email and web. We are all too familiar with stories of data loss from laptop theft, hacker break-ins, and backup tapes being lost or stolen, and so on. How can we defend ourselves against the growing threat of data leakage attacks via messaging, social engineering, malicious hackers, and more? Many manufacturers have products to help reduce electronic data leakage, but do not address other vectors. This paper aims to provide a holistic discussion on data leakage and its prevention, and serve as a starting point for businesses in their fight against it.

On August 31, 2014, a collection of almost 200 private pictures of various celebrities and these images were believed to have been obtained via a breach of Apple's cloud services suite iCloud. It has been reported that a collection of 5 million Gmail addresses and passwords have been leaked. Recently, usernames and passwords of Dropbox users have leaked online. The usernames and passwords were unfortunately stolen from other services and used in attempts to log in to Dropbox accounts. New and sophisticated techniques that have been developed in the area of data mining (also known as knowledge discovery), can aid in the extraction of useful information from the web [8]. The correct solution would be to encourage the creation and use of de-centralized and end-to-end encrypted services that do not store all your data in one place. For the sake of national security and to protect the privacy of its citizens, India should develop its own social media platforms.

III. Data Leakage Prevention (DLP)

The use of data—particularly data about people—for data mining has serious ethical implications, and practitioners of data mining techniques must act responsibly by making themselves aware of the ethical issues that surround their particular application [6]. The explosion of online social networking (OSN) in recent years has caused damages to organizations due to leakage of information. Peoples' social networking behavior, whether accidental or intentional, provides an opportunity for advanced persistent threats (APT) attackers to realize their social engineering techniques and undetectable zero-day exploits. APT attackers use a spear-phishing method that targeted on victim organizations through social media in order to conduct reconnaissance and theft of confidential proprietary information. OSN has the most challenging channel of information leakage and provides an explanation about the underlying factors of employees leaking information via this channel through a theoretical lens from information systems. OSN becomes an attack vector of APT owing to Peoples' social networking behavior, and finally, recommends security education, training and awareness (SETA) for organizations to combat these threats. Various data mining techniques (Induction, Compression and Approximation) and algorithms developed to mine the large volumes of heterogeneous data stored in the data warehouses [3].

Data leak prevention (DLP) is a suite of technologies aimed at stemming the loss of sensitive information that occurs in enterprises across the globe. By focusing on the location, classification and monitoring of information at rest, in use and in motion, this solution can go far in helping an enterprise get a handle on what information it has, and in stopping the numerous leaks of information that occur each day. DLP is not a plug-and-play solution. The successful implementation of this technology requires significant preparation and diligent ongoing maintenance. Enterprises seeking to integrate and implement DLP should be prepared for a significant effort that, if done correctly, can greatly reduce risk to the organization. Those implementing the solution must take a strategic approach that addresses risks, impacts and mitigation steps, along with appropriate governance and assurance measures.

3.1 Data Anonymization – Removing Personally Identifiable Information From Data Sets

A K-anonymized dataset has the property that each record is in distinguishable from at least others. Even simple restrictions of optimized anonymity are NP-hard, leading to significant computational challenges. New approach methods can be evolved exploring the space of possible anonymization that tames the combinatory of the problem, and develop data-man- agreement strategies to reduce reliance on expensive operations such as sorting. A desirable feature of protecting privacy through - anonymity is its preservation of data integrity. Despite its intuitive appeal, it is possible that non-integrity preserving approaches to privacy (such as random perturbation) may produce a more informative result in many circumstances. Indeed, it may be interesting to consider combined approaches, such as -anonymizing over only a subset of potentially identifying columns and randomly perturbing the others. A better understanding of when and how to apply various privacy-preserving methods deserves further study. Optimal algorithms will be useful in this regard since they eliminate the possibility that a poor outcome is the result of a highly sub-optimal solution rather than an inherent limitation of the specific technique.

3.2 De-Identification And Linking Data Records

Common strategies for de-identifying datasets are deleting or masking personal identifiers, such as name and social security number, and suppressing or generalizing quasi-identifiers, such as date of birth and zip code.

3.3 Dlp Implementation Challenges

User resistance for change is the most difficult obstacle which has to be handled with greatest care. Training workshops and seminars must be held on regular basis to infuse confidence in them for adopting DLP procedures. The effectiveness of DLP solution must be closely monitored to iron out any issues if they arise during implementation. Likewise, over-optimism also needs to be checked upon as people tend to get carried away and get over dependent on the DLP technology. Policy and procedure framework must be properly documented and accordingly implemented.

IV. Ideology And Reasons

This paper motivates the future work in this area through a review of the field and related research questions. Specifically, this paper defines the data leak prevention problem, describe current approaches, and outline potential research directions in the field. As a part of this discussion, this paper explores the idea that while intrusion detection techniques may be applicable to many aspects of the data leak prevention problem, the problem is distinct enough that it requires its own solutions. A social network is a social structure made up of individuals or organizations called nodes, which are connected by one or more specific types of interdependency, such as friendship, common interest, and exchange of finance, relationships of beliefs, knowledge or prestige. A cyber threat can be both unintentional and intentional, targeted or non-targeted, and it can come from a variety of sources, including foreign nations engaged in espionage and information warfare, criminals, hackers, virus writers, disgruntled employees and contractors working within an organization. Social networking sites are not only to communicate or interact with other people globally, but also one effective way for business promotion. Investigation and study can be made on the cyber threats in social networking websites. After going through the paper which states an amassing history of online social websites one can classify their types and also discuss the cyber threats, suggest the anti-threats strategies and visualize the future trends of such hoppy popular websites. Social networking sites spread information faster than any other media. Over 50% of people learn about breaking news on social media. 65% of traditional media reporters and editors use sites like Facebook and LinkedIn for story research, and 52% use Twitter. Social networking sites are the top news source for 27.8% of Americans, ranking close to newspapers (28.8%) and above radio (18.8%) and other print publications (6%). Twitter and YouTube users reported the July 20, 2012 Aurora, CO theater shooting before news crews could arrive on the scene, and the Red Cross urged witnesses to tell family members they were safe via social media outlets. In the same breath one could argue that social media enables the spread of unreliable and false information. 49.1% of people have heard false news via social media. On Sep. 5, 2012 false rumors of fires, shootouts, and caravans of gunmen in a Mexico City suburb spread via Twitter and Facebook caused panic, flooded the local police department with over 3,000 phone calls, and temporarily closed schools.

V. Conclusion

It is a big task to mine the data with more accuracy and processing time. The research we develop using rule induction along with Association rule mining algorithm in data mining is beneficial in terms of accuracy and processing time [5]. Further research aspects can be done on these aspects of data leakage in India and how these issues can be solved by applying the correct prevention technique or correct implementation of data handling methods. A Person's Private information can be among its most valuable assets. DLP solutions can offer a multifaceted capability to significantly increase a person's ability to manage risks to their key information assets. However, these solutions can be complex and prone to disrupt other processes and organizational culture if improperly or hurriedly implemented. Careful planning and preparation, communication and awareness training are paramount in deploying a successful DLP program amongst the common population of India or the Entrepreneur groups.

References

Journal Papers:

[1] Brijesh Kumar Baradwaj, and Saurabh Pal, Mining Educational Data to Analyze Students Performance, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 2, No. 6, 2011.

[2] V. Sangamithra, T. Kalaikumaran and S. Karthik, Data mining techniques for detecting the crime hotspot by using GIS, ISSN 2278 - 1323, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 1, Issue 10, December 2012.

[3] Amit Kapoor, Data Mining: Past, Present and Future Scenario, ISSN 2278-6856, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 1, January – February 2014.

[4] Naveeta Mehata, and Shilpa Dang, Data Mining Techniques for Identifying the Customer Behavior of Investment in Stock Market in India, ISSN 2277 3622, International Journal of Marketing, Financial Services & Management Research, Vol.1 Issue 11, November 2012.

[5] Kapil Sharma, Sheveta Vashisht, Heena Sharma, Richa Dhiman, and Jasreena Kaur Bains, A Hybrid Approach Based On Association Rule Mining and Rule Induction in Data Mining, ISSN: 2231-2307 International Journal of Soft Computing and Engineering (IJSCE), Volume-3, Issue-1, March 2013.

Books:

[6] Ian H. Witten, Eibe Frank, and Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kaufmann Publishers, Burlington, MA 01803, USA).

Chapters in Books:

[7] N R Srinivasa Raghavan, Data mining in e-commerce: A survey, (Sadhana Vol. 30, Parts 2 & 3, April/June 2005) 275–289.

Theses:

[8] Minos N. Garofalakis, Rajeev Rastogi, S. Seshadri, and Kyuseok Shim, Data Mining and the Web: Past, Present and Future, Bell Laboratories.