Research on Heteregeneous Data for Recognizing Threat

Deris Stiawan, Abdul Hanan Abdullah,
Faculty of Computer Science & Information System, Universiti Teknologi Malaysia
deris@unsri.ac.id

Mohd. Yazid Idris
Faculty of Computer Science & Information System, Universiti Teknologi Malaysia

Keywords: The information increasingly large of volume dataset and multidimensional data has grown rapidly in recent years. Inter-related and update information from security communities or vendor network security has present of content vulnerability and patching bug from new attack (pattern) methods. It given a collection of datasets, we were asked to examine a sample of such data and look for pattern which may exist between certain pattern methods over time. There are several challenges, including handling dynamic data, sparse data, incomplete data, uncertain data, and semistructured / unstructured data. In this paper, we are addressing these challenges and using data mining approach to collecting scattered information in routine update regularly from provider or security community.

Abstract: Heterogeneous data, Intrusion prevention and prediction, Data Mining

1. INTRODUCTION

Currently, much of the information is now in textual form, this information can be correlate and appropriate for solving problem on a particular problem. This could be data from the web, library data, logging, and past information that are stored as archives, these data can form a pattern of specific information.

In this case, the information increasingly large of volume dataset and multidimensional data has grown rapidly in recent years. In another scenario, (Martin 2001) describes benefits of CVE compatibility, integrating vulnerability services and tools to provide more complete security provide and alert advisory services, (Tsang et al. 2009) using blacklisting a user and notifying the user of blacklist status, and (Zhou et al. 2010) collecting URL filtering systems for provide a simple and effective way to protect web security.

However, it is possible for propose collecting scattered information in routine update regularly from provider or security community. This data can be useful information to be associated with other. The data set includes signature identification, rules, policy, pattern, method attack, URL blacklist, update patch, log system, list variant of virus and regular expression, all this will be collected and labeled to identify attack patterns and can predict it that would occur. Furthermore, if the future is similar to the past, we may have an opportunity to make predictions and readiness/ prevention.

The main contributions this paper is the enhancement of a learning phase and is part of the research have being done, which aim to increasingly accuracy alarm in detection and prevention system. The remaining of the paper is structured as follows: In Section 2 we present and briefly discuss background and related work. Section 3 proposes analysis problem. Section 4, discusses our approach. Section 5 summarized our conclusions and present additional issues on which research can be continued.

2. BACKGROUD & RELATED WORK

Data Mining (DM) is an integration of multiple technologies, these include database management, data warehouse (DW), statistics, Machine Learning (ML), decision support, visualization, and parallel computing. In this approach for finding decision function, classification function and regression function, it is adequate to use DM approach with supervised learning. DM is the process of posing queries and extracting information previously unknown from large quantities of data.

In some case, the data sources have to be integrated into DW, DM helps the users to extract meaningful information from the numerous and heterogeneous data sources. The data libraries could have different semantics and syntax, it will be difficult to extract useful information. Sophisticated DM tools are needed for this purpose.

In other hands, (Adeva et al. 2007), they introduces an intrusion detection software component based on text mining techniques, using text categorization. This approach is capable to learning the characteristic of both normal and malicious user behavior from the log entries generated by the web application server Text mining refers to the discovery of non-trivial, previously unknown and potentially useful knowledge from a collection of text. Currently, Text Mining (TM) has become an inevitable part in information retrieval, around 80% of the information stored in computer consist of text and digital files. According to work by (Lopes et al. 2007), they framework for visual text mining to support exploration of both general structure and relevant topics within a textual document collection, in this effort, they have answer and examine sets of documents to achieve understanding of their structure and to locate relevant information. This is reinforced by subsequent research by (Zhang et al. 2008), they argued text classification, namely text categorization, is defined as assigning predefined categories to text documents, where documents can be news stories, technical reports, web pages, and categories are most often subjects or topics, but may also be based on style (genres), pertinence, etc.

3. ANALYSIS PROBLEM

While several work have been proposed, there are several challenges for solving these problems, including handling dynamic data, sparse data, incomplete data, uncertain data, and semistructured / unstructured data. We have addressing these challenges based on some effort problems from previously work;

1) The problems are not fully defined in advance. Grammars will have to be modified to take account of new data. This is not easy: the addition of just one new example can completely alter a grammar and render worthless all the work that has been expended in building it, declared by (Witten et al. 1999).

2) There also some effort and problem from (Singhal 2007) and (Junqi & Zhengbing 2008) to introduce the concepts hybrid approach effectively with detecting normal usages and malicious activities using heterogeneous data. Furthermore, what makes this solution different from others?

3) How to collecting and integrating information from different structure, data format, label, meta data and variable of data. These data set bulk in information and growing from community or security services?

4) How we can convert and integrating this data into information, and subsequently into knowledge.

5) How to extract the relationships, and then correlate data source to run on the new environment if the data sources could be based on complex structure and many relationships?

6) Is it true to integrate data for the process of the standardization data definitions and data structures by using a common conceptual schema across a collection of data sources?

With respect work by (Singhal 2007) present four data source with multiple audit streams from diverse cyber sensor: (i) raw network traffic, (ii) netflow data, (iii) system call, and (iv) output alert from IDS. Unfortunately, we assume this method can not effective with new challenge of intrusion threat. However, with respect we improve and expand this opinion to our approach, in this approach we use sixteen event parameters from heterogeneous data input. We present sixteen interrelated of information in database for knowledge process. Accordingly, obtaining general pattern with variation diversity structure, label, and variable of data to potentially useful knowledge is another part of this research.

In this study, DM is used to perform data collection using history, patterns, and relationships to classification and estimation of attack in stream network. This is due to hybrid system receive data from many different sources and it is expected that a hybrid system has the potential to detect sophisticated attacks that involve multiple networks with the information from multiple sources. As a mentioned above, Learning technique from DM can be solution for research objective (i) prediction of attack pattern, (ii) identification from anomaly habitual activity, (iii) estimation normal activity based on habitual activity, (iv) classification attack / suspicious packet, (v) mapping habitual-activity, and (vi) early prevention security violation.

We use DW to collecting scattered information in routine update regularly from provider or security community, we illustration in **Figure 1**. From our observation, these data can be useful information to be associated with other. The information, increasingly large of volume dataset and multidimensional data has grown rapidly in recent years. The data set includes signature identification, rules, policy, pattern, method attack, URL blacklist, update patch, log system, list variant of virus and regular expression, all this will be collected and labeled to identify attack patterns and can predict it that would occur. These data set bulk in information and growing from community or security services. Therefore, there is a critical need of data analysis system that can automatically analyze the data to classification it and predict pattern attack future trends. This information is scattered in internet and in the form of text. Unfortunately, text is complex characteristic that has defeated many representation attempts with very rich semantics, However, here is the strength characteristic of the text.

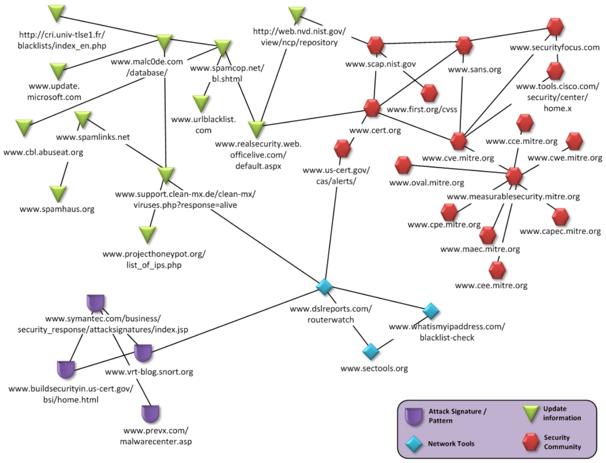
http://cri.univ-tlse1.fr/blacklists/index_en.php
www.malc0de.com/database/
www.spamcop.net/bl.shtml
http://web.nvd.nist.gov/view/ncp/repository
www.update.microsoft.com
www.urlblacklist.com
www.spamlinks.net
www.cbl.abuseat.org
www.realsecurity.web.officelive.com/default.aspx
www.spamhaus.org
www.support.clean-mx.de/clean-mx/viruses.php?response=alive
www.projecthoneypot.org/list_of_ips.php
www.scap.nist.gov
www.sans.org
www.securityfocus.com
www.tools.cisco.com/security/center/home.x
www.first.org/cvss
www.cert.org
www.us-cert.gov/cas/alerts/
www.cve.mitre.org
www.cce.mitre.org
www.cwe.mitre.org
www.oval.mitre.org
www.measurablesecurity.mitre.org
www.cpe.mitre.org
www.capec.mitre.org
www.maec.mitre.org
www.cee.mitre.org
www.symantec.com/business/security_response/attacksignatures/index.jsp
www.vrt-blog.snort.org
www.buildsecurityin.us-cert.gov/bsi/home.html
www.prevx.com/malwarecenter.asp
www.dslreports.com/routerwatch
www.whatismyipaddress.com/blacklist-check
www.sectools.org
Attack Signature / Pattern
Update information
Network Tools
Security Community

Figure 1: Interrelated web from provider and security community.

TM is a multidisciplinary field that includes many tasks such as text analysis, clustering, categorisation, and summarisation. According to some work (Al Fawareh et al. 2008) and (Sanchez et al. 2008), they have clearly described for addressing issues of ambiguity in natural language texts, and have presented a technique for resolving ambiguity problem in extracting an entity from texts. This text data mining approach has proved to be very useful in many applications.

4. OUR APPROACH

Text mining and DM are inherently hard problem in term of computational complexity. An interesting and summary some previously work using text mining help solve problem in security attack. From (Abe & Tsumoto 2009), uses method of detecting trends of technical term on importance indices using three sub processes: (i) technical term extraction in a corpus, (ii) importance indices calculation, (iii) trend detection.

We may have an opportunity to make prediction future threat from past experiences, these scenario called text categorization, making a prediction requires more that a lookup of past experience. Furthermore, for prediction, a pattern must be found in past experience that will hold in the future, leading to accurate result on new, unseen examples. As the basis of this approach are (Sanchez et al. 2008) and (Romero & Ventura 2007), text mining is concerned with obtaining new, non-trivial, and potentially useful knowledge for text repositories stored in computers, and almost all text mining approaches existing in the literature, that have been shown to be very useful in practice, are based on induction.

In the context of TM, information retrieval is one the main problems; the more general approach, a complete document will have many words and it is unlikely that it will completely match a stored document. Instead of an exact match, we try to find the closets matches to the stored documents. The proposed system has the following handling steps;

1) Take an unstructured document and automatically fill in the value of a spreadsheet. For example: information attack pattern from CVE in XML format data. Meanwhile, from security community (http://www.us-cert.gov/cas/techalerts/) have information infiltrating a botnet via Internet Relay Chat (IRC). Wherefore, when the information is unstructured, such as that found in a collection of documents, then a separate process is needed to extract data from an unstructured format.

2) Create pseudonymized for describe and declare a log of event parameters

3) The partitioning document is divided by time, not randomly. We assume this mechanism can closely simulate the prediction of future events before inside to system.

4) Document Standardization, once the documents are collected, There are several variations with different formats available, depending on when the document was generated, some of them using the ASCII format, CSV or format as images.

We identified the problem in collecting information from different structure, label, and variable of data, shown in **Figure 2**. Here data can refer to heterogeneous data, is a set bulk in information and growing from provider, community or security services. Therefore, there is a critical need of data analysis system that can automatically analyze the data to classification it and predict pattern attack future trends.

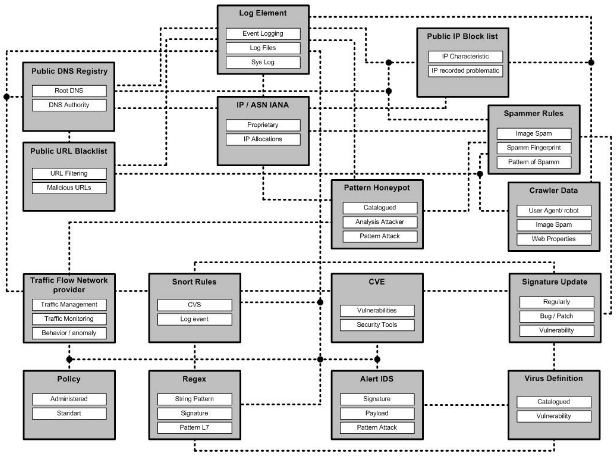

Figure 2. Correspond with parameters

5. CONCLUSION

Data Warehouse (DW) is an important supporting technology for Data Mining (DM), DW is an essentially an integration of various data source for decision support and analysis. There are several advantages to make the TM as a solution emphasis; (1) The engine of TM currently includes functionalities for text categorisation, language identification, text/ document summarisation, text clustering, and similarity analysis, (2) TM can do detecting trend, important indices calculation with information extraction methods, (3) TM allows identify request functionality from different structure text in one paragraphs, it is depending form of text, (4) TM can find the best decision rules.

This approach still needs further exploration in future research mainly query correlation each parameters, using data mining approach is one primary our focus. In the future research can also include more factors to implement our approach in real environment and benchmarking with other IPS software solution to tested effectiveness on accuracy, attack containing, measurement vulnerabilities, and risk/nearness True Positive and False Positive value.

ACKNOWLEDGEMENTS

This research is supported by The Ministry of Higher Education Malaysia and collaboration with Research Management Center (RMC) Universiti Teknologi Malaysia.

6. REFERENCES

Abe, H. & Tsumoto, S., 2009. Detection of Trends of Technical Phrases in Text Mining. *IEEE International Conference on Granular Computing*, pp.7-12.

Adeva, J.G., Manuel, J. & Atxa, P., 2007. Intrusion detection in web applications using text mining. *Engineering Applications of Artificial Intelligence*, 20, pp.555-566.

Al Fawareh, H.M. et al., 2008. Ambiguity in Text Mining. *Proceedings of the International Conference on Computer and Communication Engineering 2008*, pp.1172-1176.

Junqi, W. & Zhengbing, H., 2008. Study of Intrusion Detection Systems (IDSs) in Network Security. *IEEE. Wireless Communications, Networking and Mobile Computing. WICOM 08*, pp.1-4.

Lopes, A.A. et al., 2007. Visual text mining using association rules. *Computers & Graphics*, 31, pp.316-326.

Martin, R.A., 2001. Managing Vulnerabilities in Networked Systems. *Computer*, 34(11), pp.32-38.

Romero, C. & Ventura, S., 2007. Educational data mining : A survey from 1995 to 2005. *Expert Systems with Applications*, 33, pp.135-146.

Sanchez, D. et al., 2008. Text Knowledge Mining : An Alternative to Text Data Mining. *Knowledge Creation Diffusion Utilization*.

Singhal, A., 2007. *Data Warehousing and Data Mining Techiques for Cyber Security* 31st ed., Advance in Information Security Springer.

Tsang, P.P. et al., 2009. Nymble : Blocking Misbehaving Users in Anonymizing Networks. *IEEE Transaction Dependable and secure computing*, pp.1-15.

Witten, I.H. et al., 1999. Text mining: a new frontier for lossless compression. *Proceedings DCC 1999 Data Compression Conference*, pp.198-207.

Zhang, W., Yoshida, T. & Tang, X., 2008. Knowledge-Based Systems Text classification based on multi-word with support vector machine. *Knowledge-Based Systems*, 21(8), pp.879-886.

Zhou, Z., Song, T. & Jia, Y., 2010. A High-Performance URL Lookup Engine for URL Filtering Systems. *IEEE ICC 2010*, pp.1-5.