Journal of Computational Information Systems 7: 12 (2011) 4212-4224
Available at http://www.Jofcis.com

Survey on Heterogeneous Data for Recognizing Threat

Deris Stiawan, Mohd. Yazid Idris, Abdul Hanan Abdullah[†]

Faculty of Computer Science & Information System, University Teknologi Malaysia (UTM), Johor Bahru 81310, Malaysia

Abstract

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. It gives 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. In this paper, we showed a data mining approach to collecting scattered information in routine update regularly from provider or security community. This paper addresses problems and existing theories in possible future research in this field.

Keywords: Relational Markov Networks; Maximum a Posterior; Optimization; Approximate Probabilistic Inference

1. Introduction

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. As an example in the last era, the concept of Business Intelligence used by the company in its comprehensive study of products and services was from its business rivals or record of stock market analysis of trend information. It would happen before deciding to invest. Unfortunately, the main constraint is how to collect and integrate it, when the data is scattered and enlarged. Moreover, with data are those they are not structured, different field as relational data.

In another scenario, [1] describing benefits of CVE compatibility, integrating vulnerability services and tools to provide more complete security provide and alert advisory services, [2] using blacklisting a user and notifying the user of blacklist status, and [3] collecting URL filtering systems for providing a simple and effective way to protect web security. However, it is possible for proposing collection of scattered information in routine update regularly from provider or security community. This data can be useful information to be associated with others. The dataset includes signature identification, rules, policy, pattern, method attack, URL blacklist, update patch, log system, list of virus variance and regular expression, all these will be collected and labeled to identify attack patterns and can be predicted that it would occur. Furthermore, if the future is similar with the past, it may have an opportunity to make predictions and readiness/ prevention.

[†] Corresponding author.
Email addresses: hanan@utm.my (Abdul Hanan Abdullah).

1553-9105/ Copyright © 2011 Binary Information Press
December, 2011

The main contributions of this paper are the enhancement of a learning phase and be a part of the ongoing research. It aims to increase alarm accuracy 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 our correlating parameters approach. Section 4, discusses learning process. Section 5 summarized our conclusions and present additional issues on which research can be continued.

2. Background & Related Work

According to [4], they explore about Internet exploration has drastically changed people's ways of life, interactions among people at a virtual level in several contexts spanning from the professional life to social relationships. Another side, which is an interesting paradigm of emerging the internet in the future context, is an evolution of the Web 2.0 It is aims at integrating web and sensing technologies.

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. According to [5], describes the current trends in DM. 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. From some previous researches [6], [5], [7], and [8], it can be concluded there are some referred to as DM task; (i) classification: record are grouped into some meaningful subclasses, (ii) sequence detection: observation pattern in the data, sequences are determined, (iii) deviation analysis: anomalous instance and discrepancies, and (iv) dependency analysis: potentially interesting dependencies, relationships, associations between the data items are detected.

In some cases, the data sources have to be integrated into DW and DM to help the users to extract meaningful information from the numerous and heterogeneous data sources. Because 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, [9], 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 [10], 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 answered and examined sets of documents to achieve understanding of their structure and to locate relevant information. This is reinforced by subsequent research by [11], 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 addressed these challenges based on some efforts and 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 [12].

2. There are also some effort and problem from [13] and [14] to introduce the concepts hybrid approach effectively with detecting normal usages and malicious activities using heterogeneous data. What makes this solution different from others?

3. How to collect and integrate the information from different structures, data format, label, meta data and variable of data. These data set bulk in information and growing from community or security services.

4. How can these data be converted and integrated 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 standardization process of data definitions and data structures by using a common conceptual schema across a collection of data sources?

Table 1 Proposal Work

Concepts	Methods	Proposal Work
Text Mining	Indexing, Retrieving, Extraction, Clustering	[15],[9],[16],[17],[18],[19],
Syntactical	Sample pseudonymized	[20], [21], [22], [23]
Middleware	Tier level with tree	[24], [25], [26]
XML Document	Hierarchical structure	[27], [28], [29]
	Set similarity join	[30], [31], [32],
Keyword-based	Searching heterogeneous type	[33]
Elaboration	Heterogeneous data sources	[34]
Web mining	Data management	[5],[35],[36]
Collaborate & Integrating	Agent	[37], [38], [39], [40]

With respect work by [13] presenting 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 cannot be effective with new challenge of intrusion threat. However, with respect we improve and expand this opinion would be improved and expanded to our approach. In this approach we used sixteen event parameters from heterogeneous data input. We present sixteen interrelated of information in database for knowledge process are presented. There are several efforts to resolve this matter as showed in Table 1**.** Accordingly, obtaining general pattern with variation of diversity structure, label, and variable of data towards potentially useful knowledge is another part of this research.

In this study, DM is used to perform data collection using history, patterns, and relationships between 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 objectives (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 of security violation.

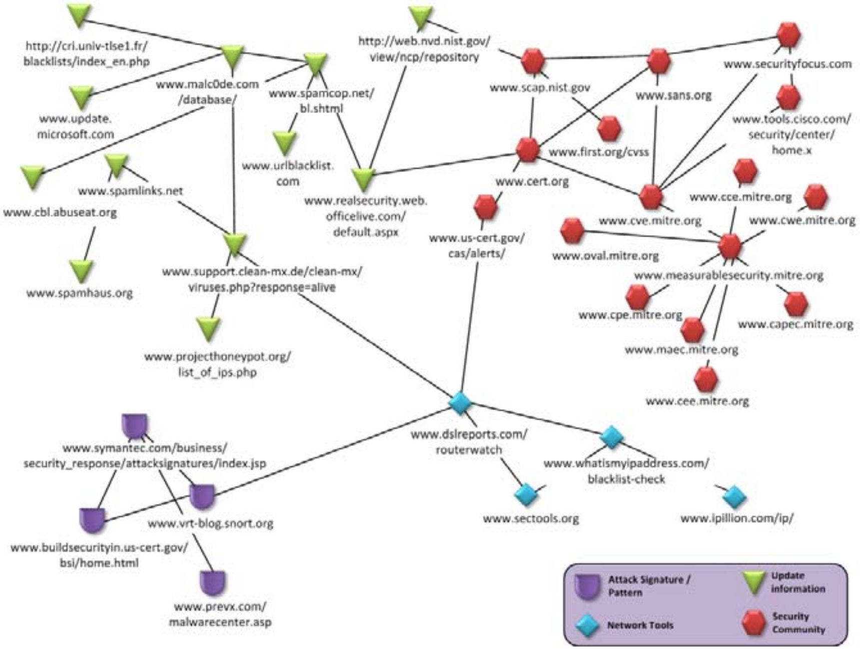


Fig. 1 Interrelated Web from Provider and Security Community.

We use DW to collect scattered information in routine update regularly from provider or security community, and it is illustrated in Figure 1. From our observation, these data can be useful information to be associated with others. The information, increasingly large of volume dataset and multidimensional data has grown rapidly in recent years. The dataset includes signature identification, rules, policy, pattern, method attack, URL blacklist, update patch, log system, list of virus variance and regular expression, all these will be collected and labeled to identify attack patterns and can be predicted that it would occur. These dataset 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 classify it and predict pattern attack future trends. This information is scattered in internet and in the form of text. Unfortunately, the text with complex characteristic has defeated many representation attempts with very rich semantics, However, here is the strength characteristic of the text. From these literates, there are several approaches that might be proposed, as following;

3.1. Text Mining (TM)

Text is everywhere in Internet and has various pieces. It may be relevant and making the connection. For example, there are many update information from security community or vendor network security, it contents vulnerability or patch from new attack (pattern) methods. Many messages are in text form, there are a lot of hidden information that could possibly be extracted from it. In this case, we defined TM as DM on text data. TM is all about previously extracting unknown patterns and associated from large databases.

TM is a multidisciplinary field that includes many tasks such as text analysis, clustering, categorisation, and summarisation. According to some work [17] and [18], 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. However, though datasets can be obtained from it, text is primarily an explicit representation of knowledge in natural language. Correlation between TM and query is for matching the process. In this respect, we should as citied from work [9], they introduced an intrusion detection software component based on text-mining techniques and shown a TM engine called Plato. By using text categorization, it is capable of learning the characteristics of both normal and malicious user behaviour from the log entries generated by the web application server. In other hands, work by [16] used indexing, query and retrieval operation in video and text document to related facets and/or subfacets as well as their conceptual representations, based on theory indexing, we can identification to analyzing content using a patterns.

3.2. Syntactical

Performed work by [21], we cite Syntactical concept, it is one of as an alternative solution that can be used to resolve this matter. This approach using Sample pseudonymized for describing and declaring a log. In this approach, they are using syntactical concepts in a sample syslog audit record, which is pseudonymity-layer audit record contains the following fields (data and the application layer an identifier). The mainly problem in this methods is quota of storing the logging and the syslog application protocol is traditionally implemented over UDP, refers to RFC 3014 and RFC 3195, this method has vulnerability in UDP port 514.

An approach to balancing these interests is replacing identifying features in audit data with pseudonyms. In security area, especially in intrusion detection field using pseudonymized audit data, it is capable of analyzing the pseudonymized audit data. It is identified there are some efforts in this work [20], they experimented to have problem solving in this solution to the pseudonym efficiency problem, which is used to label different groups of users, [22], present how to allow a service provider to issue anonymous without involving third parties in the authorization framework, and [23], they proposed encode additional information into pseudonyms that are used in location tracking systems and stored in data logs, it is also has been confirmed by proposal work [21].

3.3. Middleware

According to [26], they use multiversion concurrency control for concurrent control for concurrent access to index structure. In this study brings database closer to analytical process by allowing read-only transaction to execute without any need for synchronization with reading/writing transactions. Initially we assume this approach can convert and integrating[integrate] this data into information, and subsequently into knowledge by using tree created or middleware strategies. The identification of this approach matches problems in mark (a) as shown] in Figure 2 and this can combine with other approaches. Concept tree and middleware has long had been used for web mining techniques for several times. However, whilst latter solution of the author has successfully solved the problem used this approach, [24] and [25], they provide complementary contributions to related work on Web information sharing and querying, which is focus on providing support for achieving effective and efficient access of data.

3.4. XML Documents

XML has emerged as the leading language for representing and exchanging data not only on the Web, but also in general enterprise. Furthermore, the popularity of XML as a data exchange format become the right choice for the global scheme in data integration application. Thus, tools are required to mediate between XML queries and heterogeneous data sources to integrate data in XML. XML was conceived for tagging text data. Essentially, it provides support for exchanging text documents on the web. There are some effort to integrated DM, with Query solution by [32], idea is got to use XML for this problem, it is helpful for providing data services which accomplish data integration tasks across heterogeneous data sources. According to some reported works by [27], [28] and [31], from these approaches, it is assume that XML query can be considered importantly in order to discover and proceed knowledge in heterogeneous data.

Several queries in mapping algorithm have been proposed to translate XML queries into SQL. Although there are several queries in mapping algorithm is clearly described by [30], they described id-based query algorithm XML to SQL. [29], and also proposed to grouping heterogeneous data based on ontology approaches. They cope with employing onthologies that are controlled vocabularies for specific problem domains.

3.5. Keyword Based

According to [33] , they describes solution based on Keyword-based, a large number of heterogeneous types of data from diverse data sources, but having no means of managing and searching them in a convenient and unified fashion. Unfortunately, this effort is just for searching mechanism in desktop with query keyword scattered across multiple data unit for retrieval section in file contents. In order to achieve unified representation of several heterogeneous data units, an appropriate data structure is essential. There are some operations, such as: (i) selection, (ii) fragment join, (iii) pairwase fragment join, and (iv) powerset fragment join. Performed work by [31] also suggest the same approach, but with some XML and XHTML improvement.

3.6. Elaboration

Some efforts use an informational context to combine concept of data source, therefore giving rising to many possible combinations, work by [34], they present the new system to avoid searching such large space of possible concept combinations. Complex concept mining in this approach represented data type and combination of concepts to which it is associated. It is assumed that this approach is an innovative solution for the discovery of complex matches between database sources and the mechanism can be used to improvement content analyzer our approach.

3.7. Web Mining

As the mentioned above, DM is the process of posing queries and extracting information previously unknown from large quantities of data. In this case, heterogeneous data must be managed effectively. Several technology tools have to work together to effectively mine data on the Web, mining tools to predict trends and activities on the Web. Referring to [35], they describe the first major application area is content mining. The hyperlinks and anchors in a page are part of that page's text, and in a semantically marked-up page they are page elements in the same way that text is, and work by [36], they classify Web Mining into: (i) web content mining, (ii) web structure mining, and (iii) web usage mining. Finally, resume work by [5], an area that is becoming increasingly important is the semantic web, Web mining can also be used to improve web enthologies. Therefore, Web Mining are closely with e-commerce, business intelligence and intrusion detection for counter terrorism.

3.8. Collaborate & Integration

In the concept of collaborating & integrating with agent, these methods essentially process that function on behalf of other processes and other. We analysing, the agent method development of the Web in the early 1990s and has grown rapidly since the crawler system used in search engines [37] and [38], Agents may be self-describing; they may be decentralized, distributed, autonomous, and heterogeneous. Various agent architectures have been proposed [39], [40]. In this case, we can use agent methods to retrieving data, sorting data and filtering data for the future work and challenges problem for the future research.

4. Exploratory Our Approach

Complexity is one of the important aspects of automatic inference in knowledge-based system, any information and knowledge representation paradigm. 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 [9] performed work in 2007, describes intrusion detection software based on text-mining techniques with using text categorisation, it is capable of learning the characteristics of both normal and malicious user behaviour form the generated log entries, and [19], 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 [18] and [41], 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. Text mining can work with unstructured datasets such as full-text documents, HTML files, emails, etc. It is a multidisciplinary field that includes many tasks such as text analysis, clustering, categorization, and summarization. Additionally [17], describe text mining task include text categorization, text clustering, concept/entity extraction, document summarization, and entity relation modeling.

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.

In this initial research, increasingly true alarm rate and classification packet is mainly focuses using knowledge learning method. Figure 2, show our approach, (a) potential useful information and collection of text, (b) database collecting with text mining, (c) knowledge based on correlation data and learning behaviour-based, and (d) classification (normal / suspicious) to increase accuracy/precision: (i) predict pattern attack future trends, (ii) prevent before attack comes to network and (iii) detection security threat.

As can be seen from Figure 2, in marking (a) is potential useful information and collection of text, from preliminary observed there are many inter-related information about security violation [1], [42], [43], [44], [45], [46], [47] and [48]. However, as we discussed in above, collecting data from provider, community or security services, which is text categorization task involves several sequential steps such as pre-processing of the documents, feature selection, dimensional reduction, document indexing, and inductive classifier learning. We decide to composite data with distinctive type of algorithm: (i) probabilistic, (ii) profile-based, (iii) pattern-based. Furthermore, in mark (b), phase for finding an approach in knowledge-based system, any information and knowledge representation paradigm. Text mining and DM are inherently hard problem in term of computational complexity. This method depends on the input information that has been collected in a database. The information in the database comes from a variety of information collected and stored from time to time. In some cases, the new types of attacks based on previous patterns, especially the attacks from malicious threat, on knowledge process, performed composite and combining the data residing on the database to be sorted, queries and reused as input. The learning process occurs to combine and choose quickly by comparing the fit of the data in the database.

We identified the problem in collecting information from different structure, label, and variable of data, shown in Figure 3. 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. For most databases with a large number of fields, the features-selection task is possibly the most critical in the data preprocessing stage. We shown in Table 2 are descriptions attributes from sixteen parameters.

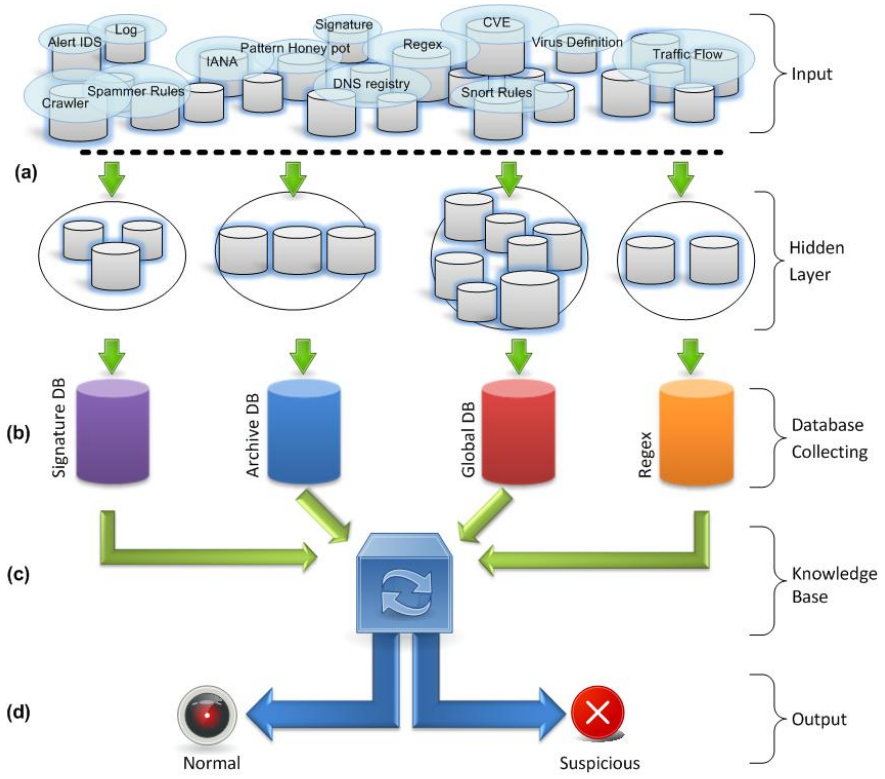


Fig. 2 A Knowledge Approach

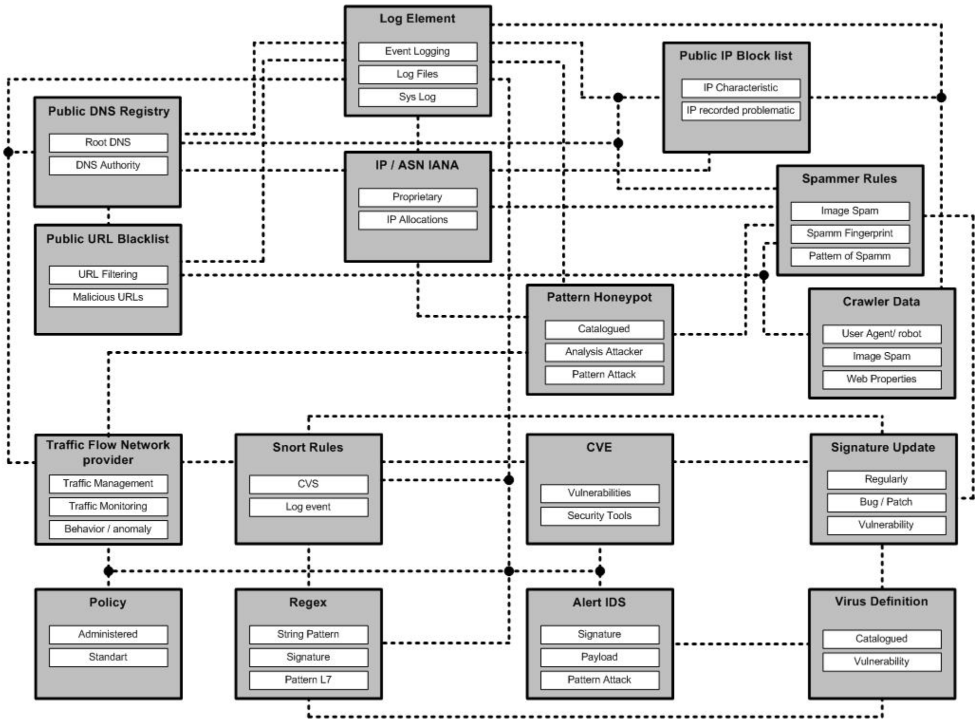
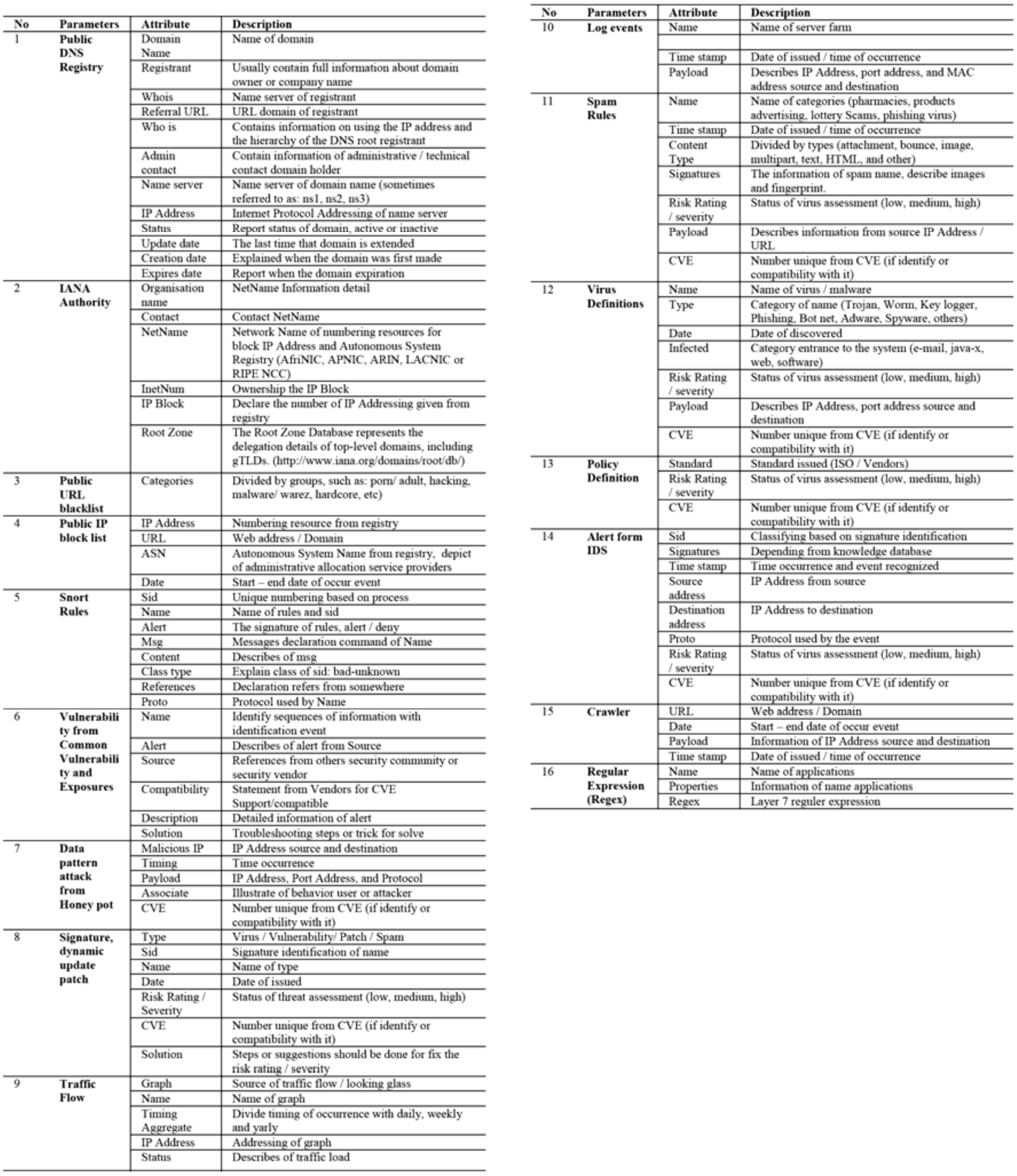


Fig. 3 A Correlate and Integrating

Table 2 Event Parameters



No	Parameters	Attribute	Description
1	**Public DNS Registry**	Domain Name	Name of domain
		Registrant	Usually contain full information about domain owner or company name
		Whois	Name server of registrant
		Referral URL	URL domain of registrant
		Who is	Contains information on using the IP address and the hierarchy of the DNS root registrant
		Admin contact	Contain information of administrative / technical contact domain holder
		Name server	Name server of domain name (sometimes referred to as: ns1, ns2, ns3)
		IP Address	Internet Protocol Addressing of name server
		Status	Report status of domain, active or inactive
		Update date	The last time that domain is extended
		Creation date	Explained when the domain was first made
		Expires date	Report when the domain expiration
2	**IANA Authority**	Organisation name	NetName Information detail
		Contact	Contact NetName
		NetName	Network Name of numbering resources for block IP Address and Autonomous System Registry (AfriNIC, APNIC, ARIN, LACNIC or RIPE NCC)
		InetNum	Ownership the IP Block
		IP Block	Declare the number of IP Addressing given from registry
		Root Zone	The Root Zone Database represents the delegation details of top-level domains, including gTLDs. (http://www.iana.org/domains/root/db/)
3	**Public URL blacklist**	Categories	Divided by groups, such as: porn/ adult, hacking, malware/ warez, hardcore, etc)
4	**Public IP block list**	IP Address	Numbering resource from registry
		URL	Web address / Domain
		ASN	Autonomous System Name from registry, depict of administrative allocation service providers
		Date	Start – end date of occur event
5	**Snort Rules**	Sid	Unique numbering based on process
		Name	Name of rules and sid
		Alert	The signature of rules, alert / deny
		Msg	Messages declaration command of Name
		Content	Describes of msg
		Class type	Explain class of sid: bad-unknown
		References	Declaration refers from somewhere
		Proto	Protocol used by Name
6	**Vulnerability from Common Vulnerability and Exposures**	Name	Identify sequences of information with identification event
		Alert	Describes of alert from Source
		Source	References from others security community or security vendor
		Compatibility	Statement from Vendors for CVE Support/compatible
		Description	Detailed information of alert
		Solution	Troubleshooting steps or trick for solve
7	**Data pattern attack from Honey pot**	Malicious IP	IP Address source and destination
		Timing	Time occurrence
		Payload	IP Address, Port Address, and Protocol
		Associate	Illustrate of behavior user or attacker
		CVE	Number unique from CVE (if identify or compatibility with it)
8	**Signature, dynamic update patch**	Type	Virus / Vulnerability/ Patch / Spam
		Sid	Signature identification of name
		Name	Name of type
		Date	Date of issued
		Risk Rating / Severity	Status of threat assessment (low, medium, high)
		CVE	Number unique from CVE (if identify or compatibility with it)
		Solution	Steps or suggestions should be done for fix the risk rating / severity
9	**Traffic Flow**	Graph	Source of traffic flow / looking glass
		Name	Name of graph
		Timing Aggregate	Divide timing of occurrence with daily, weekly and yarly
		IP Address	Addressing of graph
		Status	Describes of traffic load
10	**Log events**	Name	Name of server farm
		Time stamp	Date of issued / time of occurrence
		Payload	Describes IP Address, port address, and MAC address source and destination
11	**Spam Rules**	Name	Name of categories (pharmacies, products advertising, lottery Scams, phishing virus)
		Time stamp	Date of issued / time of occurrence
		Content Type	Divided by types (attachment, bounce, image, multipart, text, HTML, and other)
		Signatures	The information of spam name, describe images and fingerprint.
		Risk Rating / severity	Status of virus assessment (low, medium, high)
		Payload	Describes information from source IP Address / URL
		CVE	Number unique from CVE (if identify or compatibility with it)
12	**Virus Definitions**	Name	Name of virus / malware
		Type	Category of name (Trojan, Worm, Key logger, Phishing, Bot net, Adware, Spyware, others)
		Date	Date of discovered
		Infected	Category entrance to the system (e-mail, java-x, web, software)
		Risk Rating / severity	Status of virus assessment (low, medium, high)
		Payload	Describes IP Address, port address source and destination
		CVE	Number unique from CVE (if identify or compatibility with it)
13	**Policy Definition**	Standard	Standard issued (ISO / Vendors)
		Risk Rating / severity	Status of virus assessment (low, medium, high)
		CVE	Number unique from CVE (if identify or compatibility with it)
14	**Alert form IDS**	Sid	Classifying based on signature identification
		Signatures	Depending from knowledge database
		Time stamp	Time occurrence and event recognized
		Source address	IP Address from source
		Destination address	IP Address to destination
		Proto	Protocol used by the event
		Risk Rating / severity	Status of virus assessment (low, medium, high)
		CVE	Number unique from CVE (if identify or compatibility with it)
15	**Crawler**	URL	Web address / Domain
		Date	Start – end date of occur event
		Payload	Information of IP Address source and destination
		Time stamp	Date of issued / time of occurrence
16	**Regular Expression (Regex)**	Name	Name of applications
		Properties	Information of name applications
		Regex	Layer 7 reguler expression

5. Conclusion & Future Works

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.

Acknowledgement

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

References

[1] R.A. Martin, “Managing Vulnerabilities in Networked Systems,” *Computer*, vol. 34, 2001, pp. 32-38.

[2] P.P. Tsang, A. Kapadia, C. Cornelius, and S.W. Smith, “Nymble : Blocking Misbehaving Users in Anonymizing Networks,” *IEEE Transaction Dependable and secure computing*, 2009, pp. 1-15.

[3] Z. Zhou, T. Song, and Y. Jia, “A High-Performance URL Lookup Engine for URL Filtering Systems,” *IEEE ICC 2010*, 2010, pp. 1-5.

[4] L. Atzori, A. Iera, and G. Morabito, “The Internet of Things: A survey,” *Computer Networks*, vol. 54, Jun. 2010, pp. 2787-2805.

[5] B. Thuraisingham, *web data mining and applications in business intelligence and counter terrorism*, CRC Press, USA, 2003.

[6] R. Lehn, “Data warehousing tool ’ s architecture : From multidimensional analysis to data mining . Multidimensional analysis : an example,” 1997, pp. 636-643.

[7] S.-yun Wu and E. Yen, “Expert Systems with Applications Data mining-based intrusion detectors,” *Expert Systems With Applications*, vol. 36, 2009, pp. 5605-5612.

[8] K.C. Nalavade and B.B. Meshram, “Intrusion Prevention Systems : Data Mining Approach,” *ACM Proceedings of the International Conference and Workshop on Emerging Trends in Technology*, 2010, pp. 211-214.

[9] J.G. Adeva, J. Manuel, and P. Atxa, “Intrusion detection in web applications using text mining,” *Engineering Applications of Artificial Intelligence*, vol. 20, 2007, pp. 555-566.

[10] A.A. Lopes, R. Pinho, F.V. Paulovich, and R. Minghim, “Visual text mining using association rules,” *Computers & Graphics*, vol. 31, 2007, pp. 316-326.

[11] W. Zhang, T. Yoshida, and X. Tang, “Knowledge-Based Systems Text classification based on multi-word with support vector machine,” *Knowledge-Based Systems*, vol. 21, 2008, pp. 879-886.

[12] I.H. Witten, Z. Bray, M. Mahoui, and B. Teahan, “Text mining: a new frontier for lossless compression,” *Proceedings DCC 1999 Data Compression Conference*, 1999, pp. 198-207.

[13] A. Singhal, *Data Warehousing and Data Mining Techiques for Cyber Security*, Advance in Information Security Springer, 2007.

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

[15] T.Z. and F.J.D. Sholom M. Weiss, Nitin Indurkhya, *Text Mining Predictive Methods for Analyzing Unstructured Information*, Springer, 2005.

[16] M. Belkhatir and M. Charhad, “A Conceptual Framework for Automatic Text-Based Indexing and Retrieval in Digital Video Collections,” *Database and Expert Systems Applications 18th International Conference, DEXA 2007, Regensburg, Germany, September 3-7, 2007. Proceedings*, N.R.A.G.P. Roland Wagner, ed., Springer US, 2007, pp. 392-403.

[17] H.M. Al Fawareh, S. Usoh, W. Rozaini, and S. Osman, “Ambiguity in Text Mining,” *Proceedings of the International Conference on Computer and Communication Engineering 2008*, 2008, pp. 1172-1176.

[18] D. Sanchez, M.J. Mart, I. Blanco, and C.J.D. Torre, “Text Knowledge Mining : An Alternative to Text Data Mining,” *Knowledge Creation Diffusion Utilization*, 2008.

[19] H. Abe and S. Tsumoto, “Detection of Trends of Technical Phrases in Text Mining,” *IEEE International Conference on Granular Computing*, 2009, pp. 7-12.

[20] E. Lundin and E. Jonsson, “Anomaly-based intrusion detection: privacy concerns and other problems,” *Computer Networks*, vol. 34, Oct. 2000, pp. 623-640.

[21] U. Flegel, *Privacy-Respecting Intrusion Detection*, Springer, 2007.

[22] G. Bianchi, M. Bonola, V. Falletta, F. Proto, and S. Teofili, “The SPARTA pseudonym and authorization system,” *Science of Computer Programming*, vol. 74, Sep. 2008, pp. 23-33.

[23] S.G. Weber, “Harnessing Pseudonyms with Implicit Attributes for Privacy-Respecting Mission Log Analysis,” *2009 International Conference on Intelligent Networking and Collaborative Systems*, IEEE, 2009, pp. 119-126.

[24] C.-min Wang, H.-min Chen, G.-cher Lee, S.-te Wang, and S.-fong Hong, “A Tree-Structured Persistence Server for Data Management of Collaborative Applications,” *19th International Conference on Advanced Information Networking and Applications (AINA 05) Volume 1 (AINA papers)*, Ieee, 2005, pp. 503-506.

[25] B. Benatallah, M. Hacid, H. Paik, C. Rey, and F. Toumani, “Towards semantic-driven, flexible and scalable framework for peering and querying e-catalog communities☆,” *Information Systems*, vol. 31, Jun. 2006, pp. 266-294.

[26] W. Binder, A. Mosincat, S. Spycher, I. Constantinescu, and B. Faltings, “Multiversion concurrency control for the generalized search tree,” *Database and Expert Systems Applications 18th International Conference, DEXA 2007, Regensburg, Germany, September 3-7, 2007. Proceedings*, 2008, pp. 172-181.

[27] I. Manolescu, D. Florescu, and D. Kossmann, “Answering XML queries on heterogeneous data sources,” *VLDB Conference*, 2001.

[28] G. Gardarin, A. Mensch, T.T. Dang-ngoc, and L. Smit, “Integrating Heterogeneous Data Sources with XML and XQuery,” *Database and Expert Systems Applications, 2002*, 2002, pp. Database and Expert Systems Applications, 2002. Pr.

[29] T. Amagasa, L. Wen, and H. Kitagawa, “Proximity Search of XML Data Using Ontology and XPath Edit Similarity,” *Database and Expert Systems Applications 18th International Conference, DEXA 2007*, Springer, 2007, pp. 298-307.

[30] M. Atay, A. Chebotko, S. Lu, and F. Fotouhi, “XML-to-SQL Query Mapping in the Presence of Multi-valued Schema Mappings and Recursive,” *Database and Expert Systems Applications*, Springer, 2007, pp. 603-616.

[31] C. Kit, T. Amagasa, and H. Kitagawa, “OLAP Query Processing for XML Data in RDBMS,” *Databases for Next Generation Researchers, 2007. SWOD 2007*, Ieee, 2007, pp. 7-12.

[32] M.V. Cappellen, W. Cordewiner, and C. Innocenti, “Data Aggregation , Heterogeneous Data Sources and Streaming Processing : How Can XQuery Help ?,” *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 2008, pp. 1-8.

[33] S. Pradhan, “LNCS 4653 - Towards a Novel Desktop Search Technique,” *Database and Expert Systems Applications 18th International Conference, DEXA 2007, Regensburg, Germany, September 3-7, 2007. Proceedings*, W. Roland, R. Norman, and G. Pernul, eds., Springer-Verlag, 2007, pp. 192-201.

[34] Y.B. Idrissi and J. Vachon, “A Context-Based Approach for the Discovery of Complex Matches Between Database Sources,” *Database and Expert Systems Applications 18th International Conference, DEXA 2007, Regensburg, Germany, September 3-7, 2007. Proceedings*, W. Roland, N. Revell, and G. Pernul, eds., Springer-Verlag, 2007, pp. 864-873.

[35] G. Stumme, a Hotho, and B. Berendt, “Semantic Web MiningState of the art and future directions,” *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 4, Jun. 2006, pp. 124-143.

[36] W. Yong-gui, “Research on Semantic Web Mining,” *Computer Design and Applications (ICCDA), 2010 International*, 2010, pp. 67-70.

[37] D. Eichmann, “Ethical Web agents,” *Computer Networks and ISDN Systems*, vol. 28, Dec. 1995, pp. 127-136.

[38] Y.-shan Chang, H.-chun Hsieh, S.-ming Yuan, and W. Lo, “An Agent-based Search Engine based on the Internet Search Service on the CORBA,” *Distributed Objects and Applications*, 1999, pp. 26-33.

[39] N. Gibbins, “Agent-based Semantic Web Services,” *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 1, Feb. 2004, pp. 141-154.

[40] J. Chen and W. Liu, “A Framework for Intelligent Meta-search Engine Based on Agent,” *Third International Conference on Information Technology and Applications (ICITA 05)*, IEEE, 2005, pp. 276-279.

[41] C. Romero and S. Ventura, “Educational data mining : A survey from 1995 to 2005,” *Expert Systems with Applications*, vol. 33, 2007, pp. 135-146.

[42] N. Carey, G. Mohay, and A. Clark, “Attack Signature Matching and Discovery in Systems Employing Heterogeneous IDS,” *Computer*, 2003.

[43] C. Câmpeanu and S. Yu, “Pattern expressions and pattern automata,” *Information Processing Letters*, vol. 92, 2004, pp. 267-274.

[44] M. Shimamura and K. Kono, “Using Attack Information to Reduce False Positives in Network IDS,” *Proceedings of the 11th IEEE Symposium on Computers and Communications (ISCC 06)*, vol. 06, 2006.

[45] T. Dutkevych, A. Piskozub, and N. Tymoshyk, “Real-Time Intrusion Prevention and Anomaly Analyze System for Corporate Networks,” *IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems Technology and Application*, 2007, pp. 599-602.

[46] S. Barnum, “Attack Patterns: Knowing Your Enemy in Order to Defeat Them,” *Conference BlackHat DC 2007*, 2007.

[47] L.-chyau Wuu, C.-hsiang Hung, and S.-fong Chen, “Building intrusion pattern miner for Snort network intrusion detection system,” *Journal of Systems and Software*, vol. 80, 2007, pp. 1699-1715.

[48] C.-tien D. Lo, Y.-gang Tai, K. Psarris, and S. Antonio, “Hardware Implementation for Network Intrusion Detection Rules with Regular Expression Support,” *Design*, 2008, pp. 1535-1539.