EPJ Web of Conferences **68**, 00026 (2014)
DOI: 10.1051/epjconf/20146800026

Frequency Count Attribute Oriented Induction of Corporate Network Data for Mapping Business Activity

Lukas Tanutama[1,a]

[1]Bina Nusantara University, Computer Engineering, Jakarta 11480, Indonesia

Abstract. Companies increasingly rely on Internet for effective and efficient business communication. As Information Technology infrastructure backbone for business activities, corporate network connects the company to Internet and enables its activities globally. It carries data packets generated by the activities of the users performing their business tasks. Traditionally, infrastructure operations mainly maintain data carrying capacity and network devices performance. It would be advantageous if a company knows what activities are running in its network. The research provides a simple method of mapping the business activity reflected by the network data. To map corporate users' activities, a slightly modified Attribute Oriented Induction (AOI) approach to mine the network data was applied. The frequency of each protocol invoked were counted to show what the user intended to do. The collected data was samples taken within a certain sampling period. Samples were taken due to the enormous data packets generated. Protocols of interest are only Internet related while intranet protocols are ignored. It can be concluded that the method could provide the management a general overview of the usage of its infrastructure and lead to efficient, effective and secure ICT infrastructure.

1 Introduction

The nature of contemporary communication activities are global, independent of geographical location and to certain extent should be time independent as well. The public, customers and partners from any location can reach a company provided it is connected to the Internet. To enhance its business performance a company should be able to obtain information from any repository, as well as provide and disseminate its information. Web based technology enables the convenience of performing business transactions by corporate users, customers, and any other interested parties. Business transactions are not limited to only buying and selling but also accessing information, disseminating information, and any other related business activities. Web technology supports sharing of resources among other applications, databases.

The key enabler is Internet connection. Internet connection is the backbone infrastructure of contemporary business entities. Internet provides freedom on how network components and network are configured to suit a company's need. The needs should cover reliability, accessibility, and security of the infrastructure. Reliability is part of internal IT support task. The major task is to ensure that disturbances as alerted by network devices or complaints lodged by users are handled timely. Proper fault management of the infrastructure provides high reliability. Accessibility concerns the speed of accessing external resources. Traditionally infrastructure capacity or bandwidth is considered as the parameter that determines the accessibility quality. Various tools are available for measuring bandwidth usage, most notably MRTG. Accessibility is one determining factor in a company's competitiveness. Business environment demands that reliable Internet connection must be available 7x24 [1]. Included in accessibility is availability of the line to Internet and its response time. Generally IT network operations increases bandwidth if accessibility quality decreased.

To maintain the effectiveness and efficiency of the line to the Internet it is advisable to know the traffic, in this case, the transactions flowing in it. The transactions reflected users' activities. They could initiate transactions or they could be handling transactions coming from external parties. The knowledge on what takes place in the Internet line could lead to better decisions on how to maintain and improve accessibility. The output of the Gateway Router, which is the network device to reach Internet, is the aggregate of corporate traffic to and from Internet. Data packets are generated and received by users via Internet. Network data constitutes of data packets flowing into or out of the network. Network data contains data items that are processed by network devices for their operations. Transactions and information are transported to the recipients. Network devices only processed data items that are relevant for their operations such source IP address, destination IP address, and protocol. Some data packets have payload that is intended for the recipient.

[a] Corresponding author: lukast12@binus.edu

Network devices ignore the payload as it has no role in network operations.

In the network operations, there are numerous exchanges of data packets between a sender and its recipient. The exchange of data packets is following a procedure known as protocol. A party initiates a transaction invokes a protocol. The party can be a user of the network or external ones contacting network users. In this research the data items within a data packet that are of interest is the protocol itself. A protocol is considered reflecting what a user intended to do or responding to a certain requested activity. This paper described the information discovered in applying data mining to the network data. The frequency of a protocol could provide description on which transactions comply with corporate operations and which transactions are beyond the company's business operations. Generally, without data mining certain activity beyond the company's normal operations will not be discovered. The knowledge of the major protocols invoked will enhance the understanding of protocol usage pattern. The company's knowledge on the usage of the major Internet protocols can enable the adaptation of its network infrastructure operations according to the characteristics of protocol usage within the network. The information and knowledge can then support network management for improving its infrastructure and avoid the need to increase capacity as the only solution to keep a network accessible, reliable and secure.

2 Previous Works

The research of Li, Shue, and Lee [7], [8] were based on network data, whereas most researchers paid more attention to the required bandwidth to access the Internet. Their research objective was customer service management [7] and support of the development of service management strategy [8]. Li, Shue, and Lee [8] applied Attribute-Oriented Induction (AOI) data mining approach for data generalization. The traffic pattern obtained can be used by the Internet Service Provider (ISP) to provide the requested service. The traffic pattern can also be used to develop network management strategy such as classification of service according to the traffic pattern.

Due to the big number of data exchanged in the network and its tremendous growth, conventional data analysis is time consuming for discovering hidden information and knowledge. Growth of data demands proper methods to find the hidden information and knowledge. The method that is suitable to discover information from big data is data mining [2], [3], [4], [5], and [9]. Data mining consists of a number of process and software that can support the process. Han and Kamber [6] provided theoretical foundation of data mining while Witten [11] describes the theory supported with a research software tool Waikato Environment Knowledge Analyzer (WEKA). Data mining is a step to obtain information and knowledge from big data [6]. The basis of data mining algorithm and techniques are coming from [6], [10]. Bose and Chen [12] used data mining of call records for their study of classifying groups of communication service usage such as international call, roaming in, and some other transaction activities of a customer. They concluded that data mining enabled them to examine the customer traffic pattern according to their usage, revenue generated, roaming out, user behaviour pattern and service subscribed. The published work of the researchers showed that mining communication technical data could lead to the understanding of a customer and the possibility of custom designed value added services that would benefit the parties involved in communication notably the originator, the recipient and the service provider.

Many studies showing how data mining techniques can successfully extract information to provide technical solutions have been published. Alampalayam and Kumar [13] used data mining to make predictive security model. Baldi et al. [14] applied data mining techniques for effective and scalable traffic analysis. Data mining is one of the favourite techniques for security related applications. Papers of Clifton and Gengo [15] proposed the development of custom intrusion detection filters using data mining, Golnabi et al. [16] showed the analysis of firewall policy rules using data mining techniques.

Han and Kamber [6] theoretically defines that interesting pattern comply with a number of criteria such as human understanding, validity and benefits. A pattern has some parameters for evaluating its value. Apart from objective valuation, there could be subjective valuations that are dependent on human perception. An interesting pattern for somebody does not necessarily attract somebody else. Data mining needs supporting tools to provide patterns according to predetermined criteria or agreed parameters [2], [4], and [9].

3 Methodology

The research design is generally based on CRISP-DM or Cross Industry Standard Process for Data Mining. Data mining process for network data is dynamic. Data coming from the network devices is basically standardized, but there are parts that depend on other factors such as protocols invoked. Data understanding starts with data thorough study of the collected data. In this research, the collected data are the interpretation of Wireshark as the selected protocol analyzer. The collected data of interest are contained in data packets that is flowing to and from the Internet. The data packets reflect the customer's activities. Each packet data consists of two parts. The first part contains data concerning the network itself such as source address, destination address, application requested, application response and other protocol related data. This network part contains technical information necessary for proper operations of network equipment. The second part contains data belonging to the corporate. This data is normally confidential and sometimes being rendered unintelligible due to encryption.

The next step is data preparation. It includes all activities needed to construct the data that will be fed into

the mining process. Data preparation means the selection, integration, and redundancy cleaning of collected data. The object of this research is Internet network traffic data that flows from corporate network to its Internet access Service Provider. Network data collected is from the network device using a commonly available data-sniffing tool called Wireshark. Data preparation enables the process of running of the chosen data mining tool. The decision of which data would be used for analysis is based on several criteria, such as relevance to the objective, quality and technical constraints [17]. Data preparation is time consuming and error prone. There should be proper reasons of selecting data or attributes. In this research, the attribute of interest is protocol. A protocols shows what task a user intended to perform.

Data cleaning ensure the data quality. Data cleaning should be carefully performed in order not to distort the data itself. Distorted data could render data mining result irrelevant [17]. As network data was collected directly from a network, in this case the Gateway Router, the collected data was practically clean. It can then be transformed to a format that is acceptable to the data mining tool. Within the transformation step, the relevant attributes can be selected. In this research, the only attribute of interest is protocols invoked. Prior to data mining process redundant data were eliminated. Redundant data are network data with the same source address, destination address and protocol. Data reduction was to obtain a reduced representation of the data set**.** Samples of network data were taken due to the amount of data that are flowing through the line. The size of a 5-minute data sample from a fully loaded 2 Mbps line could reach 600 Mb. The collected data for research were snapshots or data samples from its daily Internet activities. To obtain a valid profile, it is necessary to have snapshots based on day of the week and time of day and within a snapshot window. The snapshot window can be defined according to length of time allocated for taking samples or maximum file size to store the samples. The selected snapshot window depends on the network device feature. Data set collected that is based on day of the week may generate the profile of that day. Data collections were scheduled according to time slots plan on weekdays. Samples were taken following the planned time slot and sampling window of 5 minutes. It is also beneficial to find the effect of time of the day on the daylong activities. In this research, the corporate activity profile would be uncovered based on its frequently invoked protocols. .

The data mining method used is based on Attribute Oriented Induction (AOI) approach. AOI approach is one of the techniques in descriptive data mining. According to Han and Kamber (2001), descriptive data mining presents information that consists of important properties of the collected data. The major parameter for understanding communication traffic call attempts; hence, the AOI approach is also modified slightly. The modified approach is named as Frequency Count AOI. The descriptive information obtained could then be analyzed from different angles and perspectives. In descriptive data mining, summarized data in concise, descriptive terms may provide an overall picture of a subject of interest. Han and Kamber (2001) called descriptive data mining as concept description that generates descriptions for characterization and comparison. Concept description is closely tied to generalization. Network data contains information at primitive concept level such as protocol. Summarization from a large data will lead to data generalization. In AOI, the generalization is performed based on the number of distinct values of each attribute in the relevant data of interest.

4 Results and Discussions

The network data set selected was collected for three working days and five sampling window time slot. Each sampling window was 5 minutes. A sampling window was the length of time that network data was collected. Even using a short window time, the number of data is more than 25,000 records or samples within five minutes. Data mining process utilizes data modified AOI that is named Frequency Count Attribute Oriented Induction (FC-AOI). The attribute used is protocol. Table 1 shows the average number of records, average distinct records, and average percentage of distinct records from total generated records. The frequency or counts are counted within the sampling window time of the sampling days.

Table 1 Average Transaction and Distinct Transactions

Item	Average Total Recordss	Average Distinct Records	Percentage
Day-1	139,513	2021	1.68%
Day-2	377,849	4834	1.29%
Day-3	452,400	9035	2.63%

Total records are the total number of data packets the network users generate within the sampling window. Distinct transactions count is the result of removing redundant records due to protocol procedure and protocol frame size. Protocol procedure produced several records that have the same source address, destination address and protocol attributes. Large payload results in large number of similar records. The raw data have embedded information that could differentiate the records. An initial record could generate large number of subsequent records due to the protocol procedure and number of payload bytes involved Table 1 showed that only a small percentage of the records in this corporate line are distinct records. The small number of distinct records could mean that the payload is big. Technically, few records generated substantial number of records.

The next investigation concerns the activities that generated the data packets. The attribute that determined users' activities is the protocol. The data mining process finds the major protocols were invoked. Table 2 shows five major protocols and its percentage from total records. The five major protocols contributed 90% of the sampled records average. TCP, HTTP, and UDP were the dominant protocols.

Table 2 Major Transaction Protocols

Protocol	Percentage
TCP	53.3%
HTTP	15.5%
UDP	15.45
POP	3.6%
SMB	2.3%

As the network data samples are coming from a corporate network, the users' invoked or handled protocols could be considered as business related. No abnormal activity was discovered.

Table 1 shows that the number of distinct records is only a small fraction of the total records. As a protocol can generate more than one transaction record, it is not sufficient to find corporate network the activities from total records. It is interesting to investigate the users' initiated protocols that generated the records.. Mining the distinct protocol network data resulted in Table 3.

Table 3 Major Distinct Protocols

Protocol	Percentage
TCP	29.4%
BT-uTP	16.0%
DNS	13.6%
UDP	13.3%
HTTP	10.3%

Table 3 showed different set of protocols that dominates the network data records. Table 3 shows the major protocols that were responsible for generating total records. Table 3 uncovered a protocol that showed the existence of a protocol that is considered not related to normal business task. The protocol is BT-uTP. This protocol is a peer-to-peer protocol to download files that could compromise corporate network security. It could also expose the company vulnerable to legal litigation. In addition bandwidth capacity could become insufficient as peer to peer downloads are bandwidth consuming. Most likely majority of transactions is due to heavy payload of peer to peer connections. Further investigation could be the future research.

The mining results showed that data mining could uncover hidden incidents. The indications could also identify the user or users generating or accepting non-company related transactions. The management could take precautions or decide the next course of action. Regular mining of network data should be performed due to the dynamic change brought about by the implementations of actions resulted from previous discovery. This research concerned only the mapping of corporate user activity as reflected through the protocol attribute only. More corporate network activity could be mapped using the same set of data but different attributes.

References

[1] L.M.Applegate, R.D. Austin,. D.L. Soule.). *Corporate Information Strategy and Management*, (Mc.Graw-Hill. 2009)

[2] Berson, Smith, & Thearling. *Building DM Applications for CRM.* (McGraw-Hill, 2000. Retrieved from http://www.ii.edu.mk/predmeti/Inteligentni%20sistemi/Predavanja/dm%20techniques.pdf)

[3] U.M.Fayyad, G..Piatetsky-Shapiro, G., P. Smyth, AI Magazine,.**17**, 3 (1996).

[4] B. Gangopadhyay, A. Arsenio, C. Antunes, *Data Engineering and Management*. (Lecture Notes in Computer Science. pp. 101 - 108, Springer Berlin / Heidelberg,2012)

[5] G. Goth, Comm.of The ACM, **53**, 11 (2010)

[6] J.Han, M. Kamber, *Data mining: concepts and techniques* (Morgan Kaufman, 2001)

[7] S.T.Li, L.Y. Shue, S.F.Lee, Expert Systems with Applications, **30**, 621-632 (2006).

[8] S.T.Li, L.Y. Shue, S.F.Lee Expert Systems with Applications, **35,** 739-754 (2008)

[9] M.J. Martín-bautista , M. Vila, V.H. Escobar-jeria, Proceedings Of The IADIS, pp 73-76 (2008)

[10] G.M. Weiss, B.D. Davison, B.D. (2010) *Handbook of Technology Management* (John Wiley and Sons, 2010).

[11] I.H.Witten, F. Eibe, *Data Mining: Practical Machine Learning Tools and Techniques*. (Morgan Kaufmann, 2005).

[12] I.Bose, X. Chen, Elect.Commerce Research and Applications, **9** 3 197-208 (2010).

[13] S.P. Alampalayam, A. Kumar, Proc. IEEE GLOBECOM, 5062, 2004.

[14] M. Baldi, E. Baralis, F, Risso, Proc. IEEE IM, 105 - 118, (2005).

[15] C. Clifton, G. Gengo, IEEE MILCOM Military Communications Conference Proceedings **1**, 440-443, (2000).

[16] K. Golnabi, R.K. Min, L. Khan, E. Al-Shaer, Proc. IEEE NOMS, 305 - 315, (2006).

[17] C.Shearer, Journal Of Data Warehousing, **5**, 4, 13-22, (2000).