RESEARCH ARTICLE OPEN ACCESS

Implementation of Naive Bayes as a Quality of Service Determination on Traffic Network Communication Protocol

Sudarma, M.*, Pramana, D.H **

* Computer System and Informatics, Department of Electrical Engineering Faculty of Engineering, Udayana University, Bali - Indonesia

** Department of Information System, School of Information Technology and Computer (STMIK STIKOM BALI),

ABSTRACT

The utilization of communication model based on computer network technology is a technology that has been widely used. Communication model on computer network, using communication protocol which fits in with standard communication type that widely known as network protocol. The number of ports which identify network protocol according to IANA is numbering 65,536 ports. The utilization of network protocol in communication of computer network, is sometimes demanding communication priority which often known as QoS (Quality of Service). Referred communication priority is based on the number of communication model utilization which using network protocol as in HTTP protocol.

The analysis of computer network traffic is one of the methods to find out the utilization of computer network communication protocol, so that can be a basis of priority setting (QoS). The utilization of Naïve Bayes is used to classify communication protocol in computer network. The utilization of wireshark application is used as network capture tool and the result achieved is classifying network protocol used as a setting reference of QoS.

***Keywords* -** Network protocol, QoS, Network capture

I. INTRODUCTION

The right classification for internet traffic is very important to be done especially in the case of designing network architecture design, network management and network security[1]. The classification being conducted is based on the number of communication activity types. Communication activity in computer network is regulated in communication process using network protocol. The development of network protocol which identified into port number is developed based on the utilization of still developing communication model and standardized internationally. The amount of port numbers set in communication process is 65,536[2].

The number of network protocol utilization in a communication sometimes demanding the use of communication priority such as throughput quality, delay time, reliability and communication security[3]. The utilization of service priority is often called in the term of QoS. The basis of providing QoS priority is by analyzing of a Network Traffic. Network Traffic or Internet Traffic is data communication traffic in a network marked with one set of statistical flow by the implementation of a structured pattern[4]. The referred structure pattern is information from the header of communication information data. Wireshark application is a reliable application in case of network traffic capture[5]. The outcome of network traffic capture consists of network traffic record from communication transaction running in computer network.

The utilization of Naïve Bayes method in the research conducted is used as classification method for network traffic. The outcome of classification process will be used as a reference in determining to provide QoS for Network protocol which often used in network communication..

II. THEORETICAL BASE

2.1. Naïve Bayes Classification

Bayes method is a statistical approach to perform induction inference in issue of classification. This method is using conditional probability as its basis. Naïve Bayes Classification Method is based on Bayes theorem, with an assumption that the effect of attribute value in certain class is not dependent on the value of other attributes. This assumption is often called as “independent feature model”. The probability for conditional model classifier is as follows:

$$P(C|F_1,\dots F_n) \quad \dots\dots(1)$$

The conditional above is a dependent class variable C with a small amount of outcome or class, depends on several variables of F1 until Fn features. So that the writing of Bayes theory is:

$$P(C|F_1,\dots F_n) = \frac{p(C)p(F_1,\dots F_n|C)}{p(F_1,\dots F_n)} \quad \dots\dots(2)$$

or

$$posterior = \frac{prior\ x\ likelihood}{evidence} \quad \dots\dots(3)$$

Naïve conditional independent assumption has played a role. Considering that each Fi feature is conditionally independent to the other feature Fj for j ≠ i. It means that:

$$p(Fi|C,Fj) = p(Fi|C) \qquad \dots\dots(4)$$

for i ≠ j, so that joint model can be stated as:

$$p(C|F1,\dots,Fn) = p(C)\,p(F1|C)\,p(F2\,|C)\,p(F3\,|C)\dots$$

$$= p(C)\prod_{i=1}^{n} p(Fi|C)$$

It means that under independent assumption above, conditional distribution of variable class C can be stated like this:

$$P(C|F_1,\dots F_n) = \frac{1}{z}\,p(C)\prod_{i=1}^{n} p(Fi|C) \qquad \dots\dots(5)$$

where Z (evidence) is a scale factor depended only on F1,…Fn, namely a constant if the value of feature variable is known.

The model of this form far more easily to be managed, since they are divided into class prior p(C) and independent probability distribution p(Fi□C). If there is k classes and if model for each p(Fi□C = c) can be stated in parameter form, then appropriate Bayes naïve model has the parameter of (k – 1) + n r k. In practice, often k = 2 (binary classification) and r = 1 (Bernoulli variable as a feature) in general, so that the number of Naïve Bayes parameter model is 2n + 1, where n is the number of binary features used for classification and prediction.

2.2. Quality of Service (QoS)

QoS in Computer Network is referred to providing quality priority for communication process occurred in computer network. The providing of quality is based on priority setting. QoS utilization is often implemented on communication protocol in computer network which often called as Network protocol. In general the final purpose of QoS is to provide a better and planned network service with dedicated bandwidth jitter and latency which under control and increasing characteristic loss[6].

2.3. Network Traffic

Measurement and analysis of network traffic is important to do in order to get the knowledge regarding network traffic characteristic[7]. Normally a network traffic data has information such as:

a. IP Address: IP Address is often known as computer address. This computer address serves as a computer identity in a network communication. This address is divided into two parts which is as source and destination identity. IP Source Address is the source address which is identical with the sender when data communication process occurred. Whereas IP Destination Address is the destination address that identical with the data receiver in data communication process.
b. Protocol: Protocol is the rule applied in data communication process which the processing is identified based on the type of service. Each protocol running will be named according to the process performed in communication process of computer network. Examples of communication protocols including tcp, udp, http, ftp, icmp protocol, etc.
c. Length: Length is the size of data which running in computer network. The common size used in network traffic is in the form of byte.

Network traffic itself can be displayed in the form of raw data (data in the form of traffic record as in a result) or in the finished form (as in the form of graph).

2.4. Wireshark

Wireshark is one of many Network Analyzer tools that widely used by Network Administrators to analyze their network performances. Wireshark is widely preferable due to its interface which using Graphical User Interface (GUI) or graphic display. Wireshark is used for network troubleshooting, software analysis and communication protocol development, and in education. Wireshark is widely used by network administrators to analyze their network performance. Wireshark is able to capture data/information which passing over a network that being observed in the form of network traffic. The benefits of using Wireshark application are as follows:

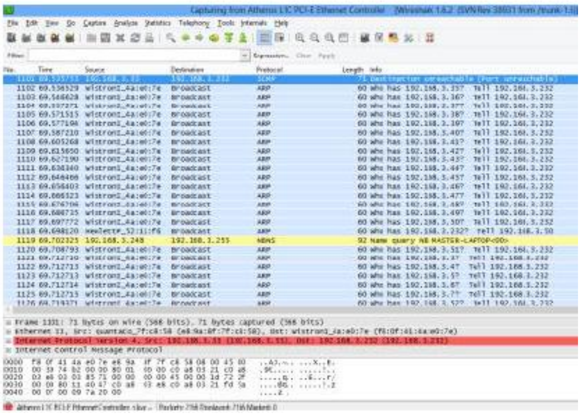
a. Capturing packet of information or data sent and received in computer network.
b. Knowing the activities that occurring in computer network.
c. Knowing and analyzing our computer network performance such as access speed/data sharing and network connection to internet.
d. Observing the security of our computer network.

Some information that can be captured by wireshark tool as network traffic information among others are time elapse (the time recorded in certain period), source address (of data sender, constitutes of IP address or mac address), destination address (of data sender, constitutes of IP address or mac address), protocol (service running in computer network), length (data size that being sent), and info (additional information of each service running in computer network).

III. METHODOLOGY

3.1. Network Capturing

The capturing of Network Traffic is using wireshark application. Network Traffic Capture is performed by capturing the traffic. The capturing of network traffic which being performed results in approximately up to tens of millions of traffic records. But the number of records yielded each day is not equal. The inequality of the number of traffic records is due to the inequality of communication models in computer network performed by the user. The model of network traffic capture from wireshark is like the following figure.


Capturing from Atheros L1C PCI-E Ethernet Controller [Wireshark 1.8.2 (SVN Rev 38931 from /trunk-1.8)]
File Edit View Go Capture Analyze Statistics Telephony Tools Internals Help
Filter:
Expression... Clear Apply

```
No.  Time      Source             Destination    Protocol Length Info
1101 69.535733 192.168.3.33       192.168.3.232  ICMP     71 Destination unreachable (Port unreachable)
1102 69.536529 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.35?  Tell 192.168.3.232
1103 69.546628 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.36?  Tell 192.168.3.232
1104 69.557271 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.37?  Tell 192.168.3.232
1105 69.571515 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.38?  Tell 192.168.3.232
1106 69.577194 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.39?  Tell 192.168.3.232
1107 69.587210 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.40?  Tell 192.168.3.232
1108 69.605268 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.41?  Tell 192.168.3.232
1109 69.615650 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.42?  Tell 192.168.3.232
1110 69.627190 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.43?  Tell 192.168.3.232
1111 69.636340 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.44?  Tell 192.168.3.232
1112 69.646466 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.45?  Tell 192.168.3.232
1113 69.656403 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.46?  Tell 192.168.3.232
1114 69.666323 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.47?  Tell 192.168.3.232
1115 69.676706 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.48?  Tell 192.168.3.232
1116 69.686715 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.49?  Tell 192.168.3.232
1117 69.697772 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.50?  Tell 192.168.3.232
1118 69.698120 HewlettP_52:11:f6  Broadcast      ARP      60 Who has 192.168.3.232?  Tell 192.168.3.50
1119 69.702325 192.168.3.248      192.168.3.255  NBNS     92 Name query NB MASTER-LAPTOP<00>
1120 69.708793 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.51?  Tell 192.168.3.232
1121 69.712710 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.3?  Tell 192.168.3.232
1122 69.712713 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.4?  Tell 192.168.3.232
1123 69.712713 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.5?  Tell 192.168.3.232
1124 69.712714 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.6?  Tell 192.168.3.232
1125 69.712715 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.7?  Tell 192.168.3.232
1126 69.719371 Wistroni_4a:e0:7e  Broadcast      ARP      60 Who has 192.168.3.52?  Tell 192.168.3.232

⊞ Frame 1101: 71 bytes on wire (568 bits), 71 bytes captured (568 bits)
⊞ Ethernet II, Src: Quantaco_7f:c8:58 (e8:9a:8f:7f:c8:58), Dst: Wistroni_4a:e0:7e (f8:0f:41:4a:e0:7e)
⊞ Internet Protocol Version 4, Src: 192.168.3.33 (192.168.3.33), Dst: 192.168.3.232 (192.168.3.232)
⊞ Internet Control Message Protocol

0000  f8 0f 41 4a e0 7e e8 9a  8f 7f c8 58 08 00 45 00   ..AJ.~.. ...X..E.
0010  00 39 74 b2 00 00 80 01  00 00 c0 a8 03 21 c0 a8   .9t..... .....!..
0020  03 e8 03 03 85 71 00 00  00 00 45 00 00 1d 72 2f   .....q.. ..E...r/
0030  00 00 80 11 40 47 c0 a8  03 e8 c0 a8 03 21 fd 5a   ....@G.. .....!.Z
0040  00 0f 00 09 7a 20 00                               ....z .
```

Figure 1. The capturing of network traffic.

3.2. Data Filtering

Data Filtering is performed by selecting the data which will be used as a classification counting process. Record format from the result of network capture constitutes raw data which is processed and filtered using Microsoft SQL Server 2008. Counting is conducted to calculate the same traffic records. In data filter, field being counted later using Naïve Bayes method only employs Protocol, Length and total counting fields. From the information of the three fields above, it can be used as a reference in QoS setting by analyzing the number of counting.

3.3. Naive Bayes Classification

Network traffic data which is being mapped in Naïve Bayes method as a classification class is as follows:

a. Protocol types in class label including the types of ARP, DHCP, DHCPv6, DNS, HTTP, ICMP, ICMPv6, IGMP, MDNS, MNDP, NBNS, NTP, SSDP, SSHv2, TCP.
b. Length Ranges are 0 – 32, 33 – 64, 65 – 128, 129 – 256, 257 – 512, 513 – 1024, 1025 – 2048.
c. Counting Ranges are 0 – 500, 501 – 1000, 1001 – 1500, 1501 – 2000, 2001 – 2500, 2501 – 3000, x > 3000.
d. QoS priorities as the setting of network traffic result including low, intermediate, and high priority.

The number of classifications which representing classification based on priority are at 105 classes. The following is the examples of label classes based on priority.

Table 1. The sample of label classes.

Protocol	Lenth Range (byte)	Counting	Priority
ARP	0 - 32	0 - 500	Medium
ARP	33 - 64	501 - 1000	Medium
DHCP	257 - 512	2001 - 2500	Medium
DHCPv6	513 - 1024	2501 - 3000	Medium
DHCPv6	1025 - 2048	x > 3000	Medium
DNS	0 - 32	0 - 500	Medium
HTTP	257 - 512	2001 - 2500	High
HTTP	513 - 1024	2501 - 3000	High
HTTP	1025 - 2048	x > 3000	High
ICMP	0 - 32	0 - 500	Medium
ICMPv6	513 - 1024	2501 - 3000	Low
ICMPv6	1025 - 2048	x > 3000	Low
IGMP	0 - 32	0 - 500	Medium
IGMP	33 - 64	501 - 1000	Medium
MDNS	1025 - 2048	x > 3000	Medium
MNDP	0 - 32	0 - 500	Low
NBNS	1025 - 2048	x > 3000	Medium
NTP	0 - 32	0 - 500	Medium
SSHv2	513 - 1024	2501 - 3000	Medium
TCP	513 - 1024	2501 - 3000	High

From the classification classes above we perform Naïve Bayes counting on sample data of the Network Traffic records.

Table 2. Sample data of network traffic.

Protocol	Length	Counting	Priority
HTTP	1430	4788	?

Sample data calculation to P is as follows:

- P (Low) : 33/105 = 0.3142857.
- P (Intermediate) : 60/105 = 0.57142857.
- P (High) : 12/105 = 0.11428571.

P calculation to Protocol Names:

- P value (HTTP ! Low) : 2/33 = 0.06060606.
- P value (HTTP ! Intermediate) : 2/60 = 0.03333333.
- P value (HTTP ! High) : 3/12 = 0.25.

P calculation to the Length:

- P value (1025 – 2048 ! Low) : 4/33 = 0.12121212.
- P value (1025 – 2048 ! Intermediate) : 6/60 = 0.1.
- P value (1025 – 2048 ! High) : 5/12 = 0.41666667.

P calculation to the Counting:

- P value (x > 3000 ! Low) : 4/33 = 0.12121212.
- P value (x > 3000 ! Intermediate) : 6/60 = 0.12121212.
- P value (x > 3000 ! High) : 5/12 = 0.41666667.

The calculation of Low Posterior: 0.3142857 x 0.06060606 x 0.12121212 x 0.12121212 = 0.000279854804193053

The calculation of Intermediate Posterior: 0.57142857 x 0.03333333 x 0.1 x 0.12121212 = 0.000230880204906205

The calculation of High Posterior: 0.11428571 x 0.25 x 0.41666667 x 0.41666667 = 0.00496031735367063

Based on the calculation of Posterior value of each priority class can be seen that High Posterior has the highest value, so the calculation outcome with Naïve Bayes method results in High priority classification for Network Traffic Data in table 2.

IV. RESULTS AND DISCUSION

Calculation trial in the research conducted is to perform calculation for 91 network traffic records which already having filtration in the stage of Data filtering where in raw data the number of traffic records which have not been filtered (to eliminate record duplication) reaching millions of records. In the table above, is mapped into priority graphics as follows:

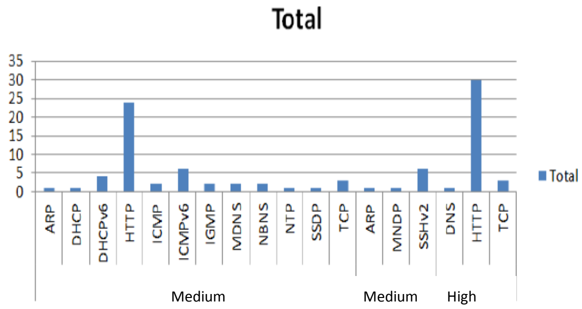


Figure 2. The number of protocols based on priority.

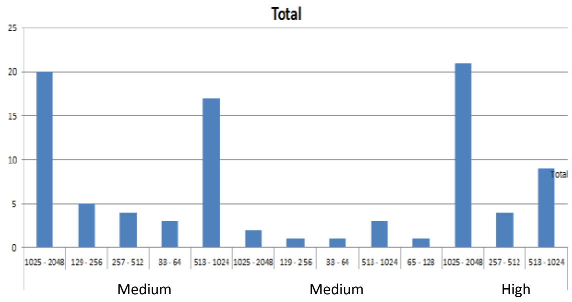


Figure 3. The number of length ranges based on priority.

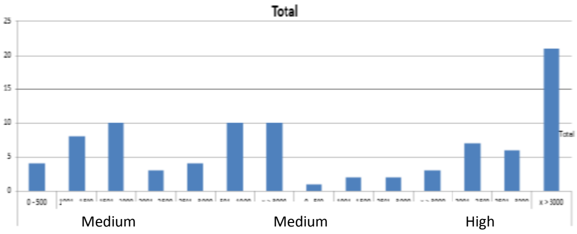


Figure 4. The number of counting range based on priority.

In Figure 2, it can be seen that the result of network traffic classification from protocol side toward the priority explaining that HTTP protocol has an intermediate up to high priority in QoS setting. Figure 3 shows that Length which can be given intermediate priority is for communication which having data size from 1025 – 2048 bytes and 513 – 1024. Meanwhile for high QoS priority is for data transmitting communication in the size of 1025 – 2048 bytes. Figure 3 shows that QoS priority can be given for the same communication process which having many appearances of more than 3000 records. So that in network management implementation, QoS policy can be done based on priority toward protocol, length size and counting (the number of equal activity record data).

V. CONCLUSION

A conclusion Based on the test result so it can be concluded that the utilization of Naïve Bayes method is able to show the classification result based on the using of Protocol, the size of data communication which being transmitted in the form of byte length, and also based on counting (records which often appeared as activities). The classification yielded is a reference for the policy of QoS utilization by a network administrator.

The expansion of the research conducted in this study, can be maximized by the use of Scheduling Process. So that the result of the classification obtained is based on the requirement of QoS priority based on the time pattern of communication activity.

REFERENCES

[1] Yuhai Liu, Zhiqiang Li, Shanqing Guo, Taiming Feng: Efficient, Accurate Internet Traffic Classification using Discretization in Naive Bayes. ICNSC 2008: 1589-1592

[2] IANA Port Numbers, http://www.iana.org/assignments/port/numbers. Diakses tanggal 24 September 2013

[3] Stephen S. Yau, Yin Yin: QoS-Based Service Ranking and Selection for Service-Based Systems. IEEE SCC 2011: 56-63

[4] Jun Zhang, Chao Chen, Yang Xiang, Wanlei Zhou, Yong Xiang: Internet Traffic Classification by Aggregating Correlated Naive Bayes Predictions. 5-15

[5] SecTools.Org: Top 125 Network Security Tools, http://sectools.org. Diakses tanggal 24 September 2013

[6] Fatoni. (2012). Analisis Kualitas Layanan Jaringan Intranet (Studi Kasus Universitas Bina Darma). Diperoleh (tanggal akses 25-9-2013) darihttp://blog.binadarma.ac.id/fatoni/wp-content/uploads /2011/04/Jurnal-QoS.pdf

[7] Jiqing Liu, Jinhua Huang.2010.Broadband Network.

BIBLIOGRAPHY OF AUTHORS

Dr. Ir. Made Sudarma, M.A.Sc.
Computer System and Informatics,
Department of Electrical Engineering,
Udayana University

Dandy Pramana Hostiadi, S.T.
Department of Information System,
School of Information Technology
and Computer (STMIK STIKOM
BALI),