A Network Traffic Prediction Algorithm Based On Data Mining Technique

D. Prangchumpol

Abstract**—**This paper is a description approach to predict incoming and outgoing data rate in network system by using association rule discover, which is one of the data mining techniques. Information of incoming and outgoing data in each times and network bandwidth are network performance parameters, which needed to solve in the traffic problem. Since congestion and data loss are important network problems. The result of this technique can predicted future network traffic. In addition, this research is useful for network routing selection and network performance improvement.

Keywords**—**Traffic prediction, association rule, data mining.

I. Introduction

IN internet system, routing path is important feature for the performance of internet usage. Moreover, router is equipment that managing and sending data through the path. Most of these equipment use shortest path algorithm for communication. These algorithms calculate the shortest way to serve packet to destination but the method dose not concentrate for traffic of the path. Therefore shortest path might not be the best solution because sometime shortest path which is selected from router has overload traffic. This problem may be down the internet performance in organization such as delay for send and receive data or make a lost data.

The network traffic prediction becomes the key of research areas and an issue in the study of network control. Especially, prediction network traffic during the daytime can help to manage route path to enhance network efficiency and know path that not overload. Therefore, the proposed of this research use association rule to predict network traffic by during the daytime and create model to manage the network path for improve highest efficiency.

The paper is structured as follows: related works are summarized in Section II. The framework for network traffic prediction is presented in Section III. Experiment setting association rule analyses are explained in Section IV. Experimental result is explained in Section V. Finally, Section VI describes conclusions and future work.

II. Related Work

Now briefly review prior research about network traffic prediction. The traffic volume, speed and occupancy data have been regarded as important features in traffic control and information management systems. It is possible to develop models to predict and extrapolate the forthcoming traffic conditions based on these traffic features [1]. In general, the number of samples has great influence on the decision-makings. However, in real world the traffic data is complex the data make classical statistical methods inefficient to provide a relatively good decision for the traffic forecasting and control. To overcome this problem, some new algorithms are imperative to analyze mass data and mine useful information. This procedure is the so called data mining technology. Lots of work has been done in traffic forecasting using data mining technology. Many researcher applied data mining technique to predict network traffic [14]-[17]. Some researchers propose to perform clustering and temporal prediction on network-level traffic states of large-scale traffic networks and use a locality preservation constraints based non-negative matrix factorization (LPNMF) [2]. Moreover, Hauser and Scherer adopted clustering approach to manage urban traffic for the first time. Reasonable management scheme was obtained in their study [3]. After that Park et al. employed Genetic Algorithm (GA) to solve the problem of unclean clusters and enhance the precision of the traffic forecasting [4]. Following, the decision trees [5], Artificial Intelligent (AI) algorithms [6], were applied into the field of traffic forecasting management. Most of researches are limited for the purpose of accidents alarms nevertheless very limited work has been done to connect the traffic features to the traffic conditions. The exploration on correlation of various traffic parameters is necessary for traffic forecasting management. An understanding of potential traffic principals is important for correct traffic management decision-making. Although neural network models were developed for digging the associated rules of the ITS database, the data was labeled in advance and the knowledge learning was under a supervised way [7]. This is not realistic in practice because the classes of the data are difficult to determine before the data mining procedure [8]-[13]. More practical tools of finding the hidden knowledge in mass data stares us in the face.

This paper represents new technique to prediction network traffic by using association rule discovery to calculate the relation of time and traffic Level.

D. Prangchumpol is with the Faculty of science and technology, Suan Sunandha Rajabhat University, Dusit, Bangkok 10300 Thailand (phone: +6602-160-1111; e-mail: dulyawit.pr@ssru.ac.th, dulyawit@gmail.com).

III. Framework for Network Traffic Prediction

A framework for predict network traffic is follows in Fig. 1. The study consists of 4 main steps as illustrated as follow:

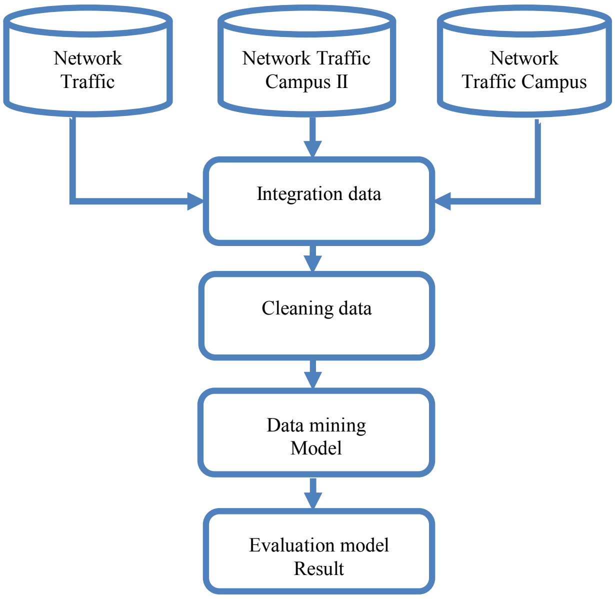
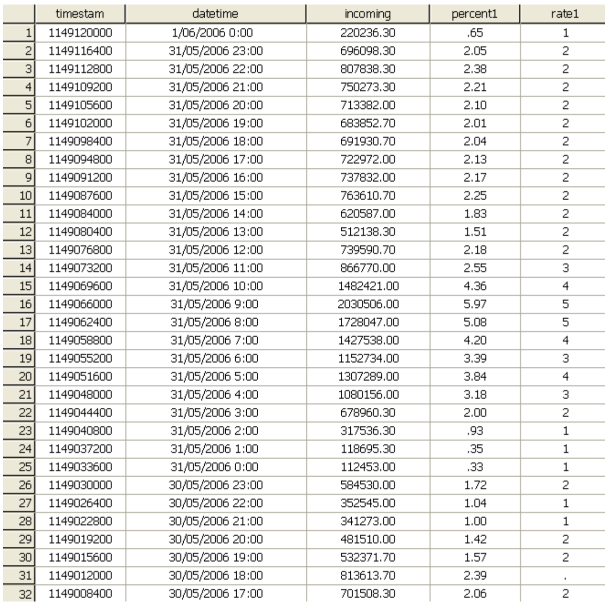


Fig. 1 Framework for network traffic prediction method

1) *Integration Data:* This process collected network traffic data of 3 campuses from Rajabhat University. The data set contain on unixtime, incoming data and percent of usage data. An example was shown Fig. 2.



	timestam	datetime	incoming	percent1	rate1
1	1149120000	1/06/2006 0:00	220236.30	.65	1
2	1149116400	31/05/2006 23:00	696098.30	2.05	2
3	1149112800	31/05/2006 22:00	807838.30	2.38	2
4	1149109200	31/05/2006 21:00	750273.30	2.21	2
5	1149105600	31/05/2006 20:00	713382.00	2.10	2
6	1149102000	31/05/2006 19:00	683852.70	2.01	2
7	1149098400	31/05/2006 18:00	691930.70	2.04	2
8	1149094800	31/05/2006 17:00	722972.00	2.13	2
9	1149091200	31/05/2006 16:00	737832.00	2.17	2
10	1149087600	31/05/2006 15:00	763610.70	2.25	2
11	1149084000	31/05/2006 14:00	620587.00	1.83	2
12	1149080400	31/05/2006 13:00	512138.30	1.51	2
13	1149076800	31/05/2006 12:00	739590.70	2.18	2
14	1149073200	31/05/2006 11:00	866770.00	2.55	3
15	1149069600	31/05/2006 10:00	1482421.00	4.36	4
16	1149066000	31/05/2006 9:00	2030506.00	5.97	5
17	1149062400	31/05/2006 8:00	1728047.00	5.08	5
18	1149058800	31/05/2006 7:00	1427538.00	4.20	4
19	1149055200	31/05/2006 6:00	1152734.00	3.39	3
20	1149051600	31/05/2006 5:00	1307289.00	3.84	4
21	1149048000	31/05/2006 4:00	1080156.00	3.18	3
22	1149044400	31/05/2006 3:00	678960.30	2.00	2
23	1149040800	31/05/2006 2:00	317536.30	.93	1
24	1149037200	31/05/2006 1:00	118695.30	.35	1
25	1149033600	31/05/2006 0:00	112453.00	.33	1
26	1149030000	30/05/2006 23:00	584530.00	1.72	2
27	1149026400	30/05/2006 22:00	352545.00	1.04	1
28	1149022800	30/05/2006 21:00	341273.00	1.00	1
29	1149019200	30/05/2006 20:00	481510.00	1.42	2
30	1149015600	30/05/2006 19:00	532371.70	1.57	2
31	1149012000	30/05/2006 18:00	813613.70	2.39	.
32	1149008400	30/05/2006 17:00	701508.30	2.06	2

Fig. 2 Example of network traffic data

2) *Cleaning Data:* In this process is convert data type to suitable format for data mining model.

3) *Data Mining Model:* An association rule model is built in this step. The relation during time and traffic level is analyzed. This technique is recommending based on similarity and were describe in Section IV.

4) *Evaluation Model and Result:* This step concludes and analysis rules.

IV. Experimental Setting and Association Rule Analysis

An association rule which is a data mining technique is selected to predict network traffic in university. Network traffic is categorized into 5 levels:

Level1 is low level network traffic
Level2 is medium low level network traffic
Level3 is medium level network traffic
Level4 is medium high level network traffic
Level5 is high level network traffic

Here, the levels of access are assumed to be uniformly distributed. The relationship in the form of $LHS \rightarrow RHS$ is applied for extracting rules. The extracted rules for LHS are based on 1-hour periods of time.

Let T1, T2,…, T24 be time. However, this research restricts the RHS as follows. Let L1, L2, L3, L4, L5 be the levels of user access for the RHS that can be predicted based on the term on the LHS. Therefore, a rule $T_j \rightarrow L_k$ is created. where L_k occurs most frequently in the rows.

For each rule of the form $LHS \rightarrow RHS$, define the *supp* and *conf* as the *support* and *confidence* as follows:

$$conf\,(LHS, RHS) = \frac{count(LHS, RHS)}{count(LHS)} \tag{1}$$

such as $conf\ time \rightarrow Level$

$$= \frac{count\big(time\ \ and\ \ Level\big)}{count\big(time\big)} \tag{2}$$

$$\sup(LHS, RHS) = \frac{count(LHS, RHS)}{count(All)} \tag{3}$$

such as $\sup time \rightarrow Level$

$$= \frac{count(time\ \ and\ \ Level)}{count(All)} \tag{4}$$

$$= 9AM \rightarrow Level2 \tag{5}$$

An example, the relation of network traffic was showed in (5). The rule can explain that at 9 AM university has network traffic in Level 2 or medium low level.

TABLE I
PREDICTION MODEL OF NETWORK TRAFFIC WITH CONFIDENCE AND SUPPORT

No.	Rule			Conf (%)	Sup (%)
1	12:00 AM	⇒	Level1	63.64	1.55
2	01:00 AM	⇒	Level1	59.09	1.23
…..	……………..			……	……
10	09:00 AM	⇒	Level1	54.55	2.15
11	10:00 AM	⇒	Level2	77.27	1.34
12	11:00 AM	⇒	Level2	50	0.28
13	12:00 PM	⇒	Level1	100	0.56
14	13:00 PM	⇒	Level1	100	0.56
15	14:00 PM	⇒	Level1	72.73	2.76
16	15:00 PM	⇒	Level3	54.55	3.2
…..	……………..			…..	…..

The equation shows the rules for predicting network traffic at 9 AM. Confidence and support value are used for rule selections. Because plenty of rules are generated, some simple concerns in rule selections include:

1) Select the rule with maximum confidence.
2) Select the rule with maximum support if confidence value is equal.
3) Select the rule that happens first when confidence and support values are equal.

From Table II, the rule explains:

- Support of $time \rightarrow Level$ is the probability that network traffic has both *time* and *Level*
- Confidence of $time \rightarrow Level$ is probability that *Level* given that the content appear in *time*.

Table I shows the total association prediction model for network traffic with confidence and support values.

The performance of this model is tested. In general, the data are divided into a training data set and a test data set.

Data obtained in November for 30 days are used to train the model while data acquired for 15 days in December are used to test the performance of the model. Note that the ratio of the training set and testing set is 60:40. Moreover Table II shows an example result of association rule discovery at 23.00 PM for all days on November.

V. EXPERIMENTAL RESULT

The rules which have confidence value were appeared in Table III. In addition, the accuracy of the relation was computed by divide the data set in training set and testing set. The results of the test model were shown in the T and F (where T is the test model accuracy, F is the model with the error). This table shows the result of the accuracy.

TABLE II
CONFIDENCE AND SUPPORT VALUE OF ASSOCIATION RULE AT 23.00 PM

Rule			*Conf (%)*	*Sup (%)*
23.00 PM	⇒	Level2	32.26	1.34
23.00 PM	⇒	Level2	32.26	1.34
23.00 PM	⇒	Level2	32.26	1.34
23.00 PM	⇒	Level2	32.26	1.34
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level2	32.26	1.34
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level2	32.26	1.34
23.00 PM	⇒	Level2	32.26	1.34
23.00 PM	⇒	Level2	32.26	1.34
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level2	32.26	1.34
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level2	32.26	1.34
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level1	67.74	2.82
23.00 PM	⇒	Level1	67.74	2.82

TABLE III
THE ACCURACY OF CONFIDENCE TRAINING VALUE AND CONFIDENCE TESTING VALUE

Rule No.	*Rule*			*Conf (%) Train*	*Conf (%) Test*	*Result*
1	0:00 AM	⇒	Level1	77.27	88.89	T
2	23:00 PM	⇒	Level1	59.09	88.89	T
3	22:00 PM	⇒	Level1	50.00	77.78	T
4	21:00 PM	⇒	Level1	54.55	55.56	T
5	20:00 PM	⇒	Level1	40.90	66.67	T
6	19:00 PM	⇒	Level2	45.45	66.67	T
7	18:00 PM	⇒	Level2	63.64	77.78	T
8	17:00 PM	⇒	Level2	59.09	66.67	T
9	16:00 PM	⇒	Level2	86.36	88.89	T
10	15:00 PM	⇒	Level2	72.73	77.78	T
11	14:00 PM	⇒	Level2	54.55	55.56	T
12	13:00 PM	⇒	Level2	59.09	66.67	T
13	12:00 PM	⇒	Level2	59.09	77.78	T
14	11:00 AM	⇒	Level2	54.55	55.56	T
15	10:00 AM	⇒	Level2	31.82	55.56	T
16	9:00 AM	⇒	Level2	22.72	55.56	T
17	8:00 AM	⇒	Level2	27.27	44.44	T
18	7:00 AM	⇒	Level3	18.18	33.33	T
19	6:00 AM	⇒	Level2	36.36	33.33	F
20	5:00 AM	⇒	Level3	36.36	55.56	T
21	4:00 AM	⇒	Level3	45.45	33.33	T
22	3:00 AM	⇒	Level2	40.91	44.44	T
23	2:00 AM	⇒	Level1	54.55	66.67	T
24	1.00 AM	⇒	Level1	77.27	100	T

Rule No. 1 can describe that at 0:00 AM has traffic in Level1. The confidence value of training set is 77.27% and confidence value of test set is 88.89%. Therefore the first order of accuracy equal to T.

Rule No. 19 can describe that at 6:00 AM has traffic in Level2. The confidence value of training set is 36.36% and confidence value of test set is 33.33%.Therefore the testing model for this rule has accuracy equal to F.

The result of this experiment has overall accuracy equal 95.83% therefore the model from association rule by using time and network traffic level can use for prediction quality of network traffic.

VI. Conclusion and Future Works

Efficiency and speed of the network traffic depends on many different factors. A traffic volume in each network path is one factor that will affect for network usage. This research was investigated the relationship of the volume of traffic on the network and found that associations rule techniques can apply to predict tend of network. In addition the model can be considered to manage the selection of best path to enhance the network performance.

References

[1] Y. Wen, and T. Lee, "Fuzzy data mining and grey recurrent neural network forecasting for traffic information systems," in *Proc. IEEE International Conference on Information Reuse and Integration*, pp. 356-361.

[2] Y. Hand, and F. Moutarde, "Analysis of Network-level Traffic States using Locality Preservative Non-negative Matrix Factorization," in *Proc. 14th IEEE Intelligent Transport Systems Conference (ITSC'2011),* Washington : United States,2011.

[3] T. Hauser, and W. Scherer, "Data mining tools for real time traffic signal decision support and maintenance," in *Proc. IEEE International Conference on Systems*, 2001, 3: 1471-1477.

[4] B. Park, D. Lee, and H. Yun, "Enhancement of time of day based traffic signal control," in *Proc. IEEE International Conference on Systems*, 2003,4: 3619-3624.

[5] Xu, P., and S. Lin, "Internet traffic classification using C4.5 decision tree,". *J. Softw.*, vol.20(10), pp. 2692-2704, 2009.

[6] L. Jia, L. Yang, Q. Kong, and S. Lin, " Study of artificial immune clustering algorithm and its applications to urban traffic control," *Int. J. Inform. Technol.*, 2006, vol.12, pp.1-9.

[7] B. Raahemi, A. Kouznetsov, A. Hayajneh, and P. Rabinovitch, "Classification of peer-to-peer traffic using incremental neural networks (fuzzy ARTMAP)," in *Proc. IEEE Canadian Conference on Electrical and Computer Engineering*, 2008, pp. 719-724.

[8] Z. Li, X. Yan, C. Yuan, J. Zhao ,and Z. Peng, "The fault diagnosis approach for gears using multidimensional features and intelligent classifier," *Imeche. Sem. Worldwide*, vol.41, pp. 76-86, 2010.

[9] Z. Li, X. Yan, C. Yuan, J. Zhao, and Z. Peng, "Fault detection and diagnosis of the gearbox in marine propulsion system based on bispectrum analysis and artificial neural networks," *J. Mar. Sci. Appl.*, ,vol.10, pp. 17-24, 2011.

[10] Z. Li, X. Yan, C. Yuan, Z. Peng, and L. Li, "Virtual prototype and experimental research on gear multi-fault diagnosis using wavelet-autoregressive model and principal component analysis method," *Mech. Syst. Signal Pr .*, vol. 25, pp.2589-2607, 2011.

[11] Z. Li, X. Yan, Y. Jiang, L. Qin ,and J. Wu, "A new data mining approach for gear crack level identification based on manifold learning," *Mechanika*,vol 18, pp.29-34, 2012.

[12] Li, Z., X. Yan, Z. Guo, P. Liu, C. Yuan ,and Z. Peng, "A new intelligent fusion method of multi-dimensional sensors and its application to tribo-system fault diagnosis of marine diesel engines," *Tribol. Lett.*, vol.47, pp. 1-15,2012.

[13] Li, Z., X. Yan, C. Yuan, and Z. Peng, "Intelligent fault diagnosis method for marine diesel engines using instantaneous angular speed," *J. Mech. Sci. Technol.*, vol. 26(8), pp. 2413-2423, 2012.

[14] M.J.A Berry, and G. S. Linnoff, "Data Mining Techniques for Marketing, Sale and Customer Relationship Management," New York: Wiley Publishing, 2004.

[15] D. Ng'ambi "Pre_empting User Questions through Anticipation-Data Mining FAQ Lists," in *Proc. of SAICSIT*,2002,pp.101-109.

[16] N. Feamste, and J. Rexford, "Network-Wide BGP Route Prediction for Traffic Engineering". a Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA Internet and Networking Systems, AT&T Labs. Research, Florham Park, NJ, USA, 2004.

[17] J. Han, and M. Kamber, "Data Mining Concepts and Techniques," USA : Morgan Kaufman,2001.