account of being route of dense data, resulting in efficiency of usage; such as delay in data receiving / sending and loss of partial data.

According to the aforementioned problems, if we can forecast the natures of internet usage quantity in the semester opening and closing period which each period of each day has data quantity on the routes, we can use them as a model as applied to be used in considering and managing the efficient use of the routes and assisting in reducing the route quantity in sending the data in the network systems further. The previous researches have applied the rules of analysis of the data relation in forecasting the data traffics; as a result, this Research aims to study and compare the efficiency based on the data mining techniques in order to forecast the traffics on the routes used in receiving / sending the data in the network systems in order to be used as guidelines for consideration for option to use the routes to reach the maximum efficiency further.

The paper is structured as follows: related works are summarized in Section 2. The framework for network traffic prediction is presented in Section 2. Experiment setting association rule analyses are explained in Section 4. Experimental result is explained in Section 5. Finally, Section 6 describes conclusions and future work.

2. Related Work

Literature reviewed 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 (Wen and Lee, 2005). In general, the number of samples has great influence on the decision-makings. However, in real world the traffic data is complex, only use statistical methods inefficient to provide a relatively good decision for the traffic forecasting and control. To overcome this problem, 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 (Berry and Linnoff, 2004) , (Ng'ambi, 2002) , (Feamste and Rexford 2004) , (Han and Kamber, 2001). Artificial Intelligent (AI) algorithms (Jia, Yang, Kong, and Lin, 2006) were applied into the field of traffic forecasting management. Moreover, Hauser and Scherer adopted clustering approach to manage urban traffic for the first time. Reasonable management scheme was obtained in their study (Hauser and Scherer, 2001). After that Park *et al.*(2003) employed Genetic Algorithm (GA) to solve the problem of unclean clusters and enhance the precision of the traffic forecasting. 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 (Raahemi et al., 2008). This is not realistic in practice because the classes of the data are difficult to determine before the data mining procedure (Li et al., 2010, 2011, 2012) More practical tools of finding the hidden knowledge in mass data stares us in the face. In

addition decision tree is based on the sample's variable values. There are different methods available for prediction that classification with decision trees (Xu and Lin, 2009) , (Howe ,and et al., 2005) , (Knab et al., 2006) , (Ganti et al., 1999), Gehrke, et al., 1998).

Previous work, researcher use only time to predict traffic level (Prangchumpol ,2013). and use association rule for prediction network traffic in education institutes (Prangchumpol ,2013). Therefore this research represents comparison techniques to prediction network traffic by using the relation of semester, time and traffic level.

3. Framework for Network Traffic Prediction

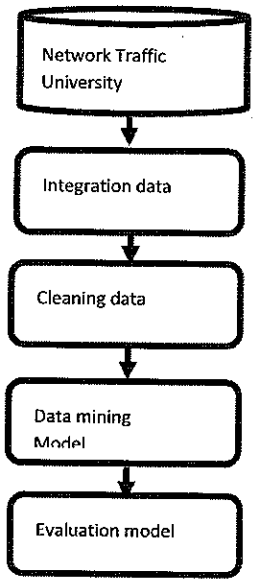
Framework for predict network traffic is follows in Fig. 1. The study consists of 4 main steps as illustrated as follow:

Integration Data: This process collected network traffic data from Rajabhat University. The data set contain on unixtime, incoming data and percent of usage data.

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

3) Data Mining Model: In this step association rule and decision tree model are built. The relations during semester time and traffic level are analyzed for two data mining technique. These techniques were described in section 4.

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



Framework for network traffic prediction method

4. Experimental Setting

Two data mining techniques which are association rule and decision tree are compared to find best technique for predict network traffic. The dataset is collected from Rajabhat University

include in network traffic, time and academic semester. Three data were integrating and cleaning. Moreover 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, and Level5 is high level network traffic.

Additional, in most of Thailand, semester is split into three terms: Semester1 is started in early June and ends in late September or early April, Semester2 is started in November and ends in early March, and Semester3 is started in mid-March and ends mid-to-late May.

Let *S1, S2, S3, S4* be semester where *S4* is not semester and *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. The example of training data set is showed as follows:

EXAMPLE OF TRAINING DATA

Semester	Time	Traffic Level
Semester1	24:00 AM	1
Semester1	01:00 AM	1
.....	.....	.....
Semester1	09:00 AM	3
Semester1	10:00 AM	4
Semester1	11:00 AM	3
Semester1	12:00 PM	5
Semester1	13:00 PM	4
Semester1	14:00 PM	3
Semester1	15:00 PM	3

A. Association rule

Association rule model use the relationship in the form of $LHS \rightarrow RHS$ is applied for extracting rules. The extracted rules for LHS are based on duration of semester and 1-hour periods of time. Therefore, a rule $S_i, T_j \rightarrow L_k$ 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:

such as *conf* $Semester, Time \rightarrow Level$

$$= \frac{count(Semester, Time\ and\ Level)}{count(Semester, Time)} \quad (1)$$

such as *sup* $Semester, Time \rightarrow Level$

$$= \frac{count(Semester, Time\ and\ Level)}{count(All)} \quad (2)$$

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.

Decision tree

A decision tree has a flowchart like structure. In this structure each internal node does a test on a certain property and each branch of this node show a result of the test; however a leaf node indicates a class. The top node is called the root of tree. This classification can be prediction of new situations and also decision making about them. Generally in a decision tree the root is on the top and leaf nodes are situated in the lowest level.

In our application, we select Information Gain, which will be explained in following subsection, in the role of selection criterion for the learner. Pruning during induction is based on the minimal number of two instances in leaves. This measure is based on information theory, which indicates required information to classify a given record of data. The expected information to encode possible class label of an arbitrary record of a training set in bits is given by Y of an arbitrary record of a training set in bits is given in Equation (3).

$$H(Y) = -\sum_{i=1}^{k} P(Y = C_i) \log(P(Y = C_i)) \quad (3)$$

Where $P(Y=C_i)$ is the nonzero probability that the record belongs to a class C_i. A log function to the base two is used, because the information is encoded in bits. $H(Y)$ is also known as the entropy of the data. This parameter gets a high value if class label Y has uniform distribution in training set and low value if its distribution varies. The conditional entropy $H(Y/x_a)$ is the expected information required to classify a record based on some known attribute x_a :

$$H(Y / x_a) = -\sum_{u \in val(x_a)} P(x_a = u) H(Y / x_a = u) \quad (4)$$

Information gain is defined as the difference between the original information requirement (3) and the new requirement after obtaining the value of x_a (4). That is

$$I(x_a) = H(Y) - H(Y / x_a) \quad (5)$$

In other words, $I(x)$ reveal how much information would be gained by splitting on . We like to do splitting on the attribute that would produce partitions that are more pure and the amount of information still required to finish classifying their records is minimal. Therefore, it is sufficient to

choose the attribute with the highest information gain and using it as a splitting attribute on the current node in Decision Tree.

Data Set

The data set was collected from Rajabhat University during one year for every day. The data set transaction about 87,600 records.

Model Evaluation

The performances of two models were tested. In general, the data is divided into a training data set and a test data set.

Data obtained in all semesters are used to train the model while data acquired for 30 days in June are used to test the performance of the model. Note that the ratio of the training set and testing set is 60:40.

5. Experimental Result

The result of association rule has overall accuracy equal 98.74 % [2] and the output of decision tree is represented by ID3 in Table II.

RESULT OF DECISION TREE

Time	Semester	Traffic Level
24.00 AM	All	1
01.00 AM	All	1
02.00 AM	All	1
03.00 AM	All	1
04.00 AM	All	1
05.00 AM	All	1
06.00 AM	All	1
07.00 AM	All	2
08.00 AM	Semester1	3
08.00 AM	Semester2	3
08.00 AM	Semester4	2
09.00 AM	Semester1	3
09.00 AM	Semester2	5
09.00 AM	Semester4	2
10.00 AM	Semester1	4
10.00 AM	Semester2	4
10.00 AM	Semester4	3
11.00 AM	Semester1	3
11.00 AM	Semester2	3
11.00 AM	Semester4	3
12.00 AM	Semester1	5
12.00 AM	Semester2	5
12.00 AM	Semester4	3
13.00 AM	Semester1	4
13.00 AM	Semester2	5
13.00 AM	Semester4	3

Time	Semester	Traffic Level
14.00 AM	Semester1	3
14.00 AM	Semester2	5
14.00 AM	Semester4	4
15.00 AM	Semester1	3
15.00 AM	Semester2	5
15.00 AM	Semester4	2
16.00 AM	Semester1	4
16.00 AM	Semester2	5
16.00 AM	Semester4	2
17.00 AM	Semester1	3
17.00 AM	Semester2	5
17.00 AM	Semester4	2
18.00 AM	All	3
19.00 AM	Semester1	2
19.00 AM	Semester2	2
19.00 AM	Semester4	1
20.00 AM	Semester1	2
20.00 AM	Semester2	2
20.00 AM	Semester4	1
21.00 AM	All	1
22.00 AM	All	1
23.00 AM	All	1

This summary is presented in Fig. 4, along with some important statistical parameters. One of it is *Kappa statistic*, a measure of agreement between two individuals, with a 0.9885 value; other parameters are *mean absolute error*- a quantity used to measure how close forecasts or predictions are to the eventual outcomes, *root mean squared error* - a good measure of the model's accuracy, *root relative squared error* -the average of the actual values, *relative absolute error* - similar to the relative squared error.

```
=== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances         119               99.1667 %
Incorrectly Classified Instances         1                0.8333 %
```

Evaluation on training set.

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        1
                 1         0          1         1         1          1        2
                 1         0.01       0.95      1         0.974      1        3
                 1         0          1         1         1          1        4
                 0.941     0          1         0.941     0.97       1        5
Weighted Avg.    0.992     0.002      0.992     0.992     0.992      1
```

Detailed accuracy by class.

In Fig. 3, contains information regarding the detailed accuracy by class. Here are detailed information concerning the next statistical parameters:

- *TP Rate (True positive rate):* the report of the positive instances classified as positive. There have been classified as positive 94.1% of the positive instances from class 5, and 100% from the first to forth class. The best percentage is for the last class which means that all the positive instances were classified as positive.
- *Precision:* the number of correctly classified instances divided by the whole classified instances number. For example, the precision value is 0.95 for class 3, is 1 for class 1, class 3, class 4, and class 5.
- *Recall*: the same with TP Rate
- *FP Rate (False Positive Rate)*: the report of the negative classified instances as positive. In our example, for class 3 this report value is 0.01, meaning only 1 % from the negative instances have been classified as positive.
- *F-Measure* is a measure of a test's accuracy

This information can be used in prediction work. Table III demonstrates the percentage accuracy of Association Rule and Decision Tree. The accuracy of Association Rule has 98.74% and Decision Tree has accuracy 99.16%. Therefore the model from Decision Tree by using time, Semester and network traffic level can use for prediction quality of network traffic and this model outperform use only time factor.

COMPARAION OF ACCURACY PERCENTAGE FOR DATA MINING MODEL

Data mining Model	***Accuracy Percentage (%)***
Association Rule	98.74%
Decision Tree	99.16%

6. Conclusion and Future Works

This Research is to compare the efficiency of 2 models used in forecasting the traffics in the network system for educational institutes or universities in order to study the model as appropriate for managing the network systems for the educational organizations in each semester. It is found that using the Decision Tree Technique can forecast the trend of reducing the traffics in the manner which is more efficient than the Association Rule. Use of the Association Rule Discovery Technique for forecasting these traffics can be used as guidelines for considering, managing, and opting to use the routes to meet the maximum efficiency during semester opening and closing period and can be used as a factor to develop the data receiving/sending efficiency more than ever.

ACKNOWLEDGMENT

The authors would like to thank Suan Sunandha Rajabhat University for scholarship support.

REFERENCES

A Berry, M.J.,and Linnoff, G. S. (2004) .Data Mining Techniques for Marketing, Sale and Customer Relationship Management. *New York: Wiley Publishing*.

Feamste, N. ,and Rexford, J. (2004). 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.

Ganti, Gehrke, V., Ramakrishnan, R. ,and Loh. , W.-Y. (1999). BOAT - Optimistic Decision Tree Construction. *Proc. ACM SIGMOD'99*,Philadephia, USA.169-I 80.

Gehrke, J. , Ramakrishnan, R. ,and Ganti, V.(1998). RainForest – A Framework for Fast Decision Tree Construction of Large Datasets. *Proc. VLDB '98,* New York, USA. 416- 427.

Han, J. ,and Kamber, M. (2001). Data Mining Concepts and Techniques. USA: Morgan Kaufman.

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

Howe, Nicholas R., Rath, Toni M. ,and Manmatha, R. (2005). Boosted Secision Trees for Word Recognition in Handwritten Document Retrieval. *SIGIR*, 377- 383.

Knab, P. , Pinzger, M. ,and Bernstein, A. (2006) .Predicting Defect Densities in Source Code Files with Decision Tree Learners. *MSR*, -125.

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

Li, Z. , Yan, X. , Yuan, C., Zhao, J. ,and Peng, Z. (2011). 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, 17-24.

Li ,Z., Yan, X. , Yuan, C. , Peng, Z. ,and L. Li. (2011). 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, 2589-2607.

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

Li, Z., Yan. X., Guo, Z., Liu,P., Yuan, C. ,and Peng, Z. (2012).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, 1-15.

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

Ng'ambi, D. (2002). Pre_empting User Questions through Anticipation-Data Mining FAQ Lists," in Proc. of SAICSIT, 101-109.

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

Prangchumpol, D. (2013). A Network Traffic Prediction Algorithm Based on Data Mining Technique. *International Conference on Computer Communications and Networks Security (ICCCNS 2013),* Oslo, Norway, July 22-23.

Prangchumpol, D. (2014). Improving the Performance of Network Traffic Prediction for Academic Organization by Using Association Rule Mining. *International Conference on the Innovative Computing Technology*, Luton, UK.

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

Wen, Y. and Lee, T. (2005). Fuzzy data mining and grey recurrent neural network forecasting for traffic information systems. *Proc. IEEE International Conference on Information Reuse and Integration*. 356-361.

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

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

Comparison of Data Mining Techniques for Network Traffics Prediction in Educational Institutes

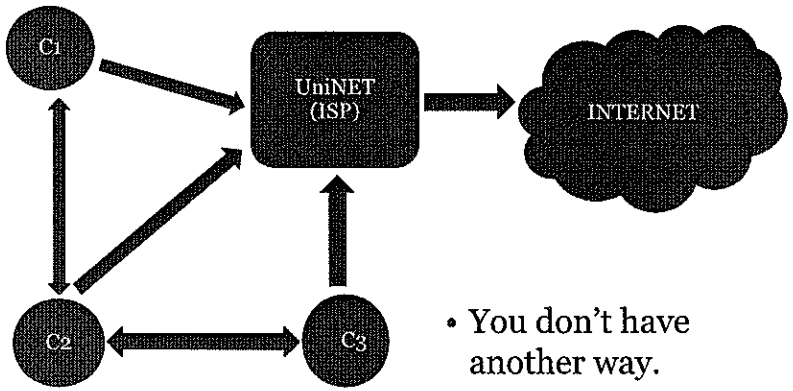
Dulyawit Prangchumpol

Faculty of science and technology,
Suan Sunandha Rajabhat University, Thailand

Introduction

Nowadays, internet system plays significant role in every organization. To be more specific, there is massive increase of internet usage in academic organizations because internet users have more access to communication devices.

Introduction



- You don't have another way.

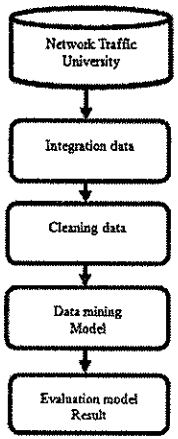
or

- You have choice to go.

Objectives

This research aimed to study and compare the efficiency based on the data mining techniques in order to forecast the traffics on the routes used in receiving / sending the data in the network systems in order to be used as guidelines for consideration for option to use the routes to reach the maximum efficiency further.

Framework for network traffic prediction



Experimental Setting

Two data mining techniques which are
- Association rule
- Decision tree

are compared to find best technique for predict network traffic.

Association Rule

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

Association Rule

Additional, in most of Thailand, semester is split into three terms:
- Semester1 is started in early June and ends in late September or early April
- Semester2 is started in November and ends in early March.

Semester3 is started in mid-March and ends mid-to-late May.

Confidence and Support

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

such as $conf\ Semester, Time \rightarrow Level$

$$= \frac{count(Semester, Time\ and\ Level)}{count(Semester, Time)} \quad (2)$$

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

such as $sup\ Semester, Time \rightarrow Level$

$$= \frac{count(Semester, Time\ and\ Level)}{count(All)} \quad (4)$$

Association Rule with conf and supp

No	Rule	Conf (%)	Sup (%)
1	Semester1,24:00 AM ⇒ Level1	65.64	1.55
2	Semester1,01:00 AM ⇒ Level1	59.09	1.23
.....	.....	.....	.....
10	Semester1,09:00 AM ⇒ Level3	82.55	2.15
11	Semester1,10:00 AM ⇒ Level4	49.32	1.34
12	Semester1,11:00 AM ⇒ Level3	60	0.28
13	Semester1,12:00 PM ⇒ Level5	89.34	1.52
14	Semester1,13:00 PM ⇒ Level4	87.44	2.33
15	Semester1,14:00 PM ⇒ Level3	100	2.76
16	Semester1,15:00 PM ⇒ Level3	100	3.2
.....	.....	.....	.....

Association Rule with conf and supp

- 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.

Dicission tree

- In our application, we select Information Gain, which will be explained in following subsection, in the role of selection criterion for the learner. Pruning during induction is based on the minimal number of two instances in leaves.

Decision tree

- The expected information to encode possible class label of an arbitrary record of a training set in bits is given by Y of an arbitrary record of a training set in bits is given in Equation :

$$H(Y) = -\sum_{i=1}^{k} P(Y = C_i)\log(P(Y = C_i))$$

$$H(Y/x_a) = -\sum_{u \in val(x_a)} P(x_a = u)H(Y/x_a = u)$$

$$I(x_a) = H(Y) - H(Y/x_a)$$

Data Set

The data set was collected from Rajabhat University during one year for every day. The data set transaction about 87,600 records.

Model Evaluation

- The performances of two models were tested. In general, the data is divided into a training data set and a test data set.
- Data obtained in all semesters are used to train the model while data acquired for 30 days in June are used to test the performance of the model. Note that the ratio of the training set and testing set is 60:40.

Experimental Result

Data mining Model	Accuracy Percentage (%)
Association Rule	98.74%
Decision Tree	99.16%

Conclusion

This Research is to compare the efficiency of 2 models used in forecasting the traffics in the network system for educational institutes or universities in order to study the model as appropriate for managing the network systems for the educational organizations in each semester.

It is found that using the Decision Tree Technique can forecast the trend of reducing the traffics in the manner which is more efficient than the Association Rule.

Thank you for your attention