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 discovery, 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 does 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 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.
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:

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.

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 \( T_1, T_2, \ldots, T_{24} \) be time. However, this research restricts the RHS as follows. Let \( L_1, L_2, L_3, L_4, L_5 \) 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 \( \text{support} \) and \( \text{confidence} \) as follows:

\[
\text{conf}(LHS, RHS) = \frac{\text{count}(LHS, RHS)}{\text{count}(LHS)} \quad (1)
\]

\[
\text{sup}(LHS, RHS) = \frac{\text{count}(LHS, RHS)}{\text{count}(All)} \quad (2)
\]

\[
\text{conf}(LHS, RHS) = \frac{\text{count}(LHS, RHS)}{\text{count}(LHS)} \quad (3)
\]

\[
\text{sup}(LHS, RHS) = \frac{\text{count}(LHS, RHS)}{\text{count}(All)} \quad (4)
\]

\[
\text{conf}(LHS, RHS) = \frac{\text{count}(LHS, RHS)}{\text{count}(LHS)} \quad (5)
\]

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 level low.
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 → Level is the probability that network traffic has both time and Level
- Confidence of time → 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.
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 trend of network. In addition the model can be considered to manage the selection of best path to enhance the network performance.

REFERENCES


