ISSN 1392 – 1215 ELECTRONICS AND ELECTRICAL ENGINEERING 2009. No. 5(93)

ELEKTRONIKA IR ELEKTROTECHNIKA

T180 TELECOMMUNICATIONS ENGINEERING

TELEKOMUNIKACIJŲ INŽINERIJA

Some Aspects of Traffic Analysis used for Internet Traffic Prediction

G. Rutka

Faculty of Electronics and Telecommunication, Riga Technical University
Riga, Azenes str. 12-317, LV-1048, phone:+371 29627600, e-mail: gundega.rutka@rtu.lv

Introduction

It is very important for the resource distribution and control to predict the network traffic's future varying trend. Via the effectively observed value of some time series in time t, we can predict its future value in time t+τ , and thus build the optimized foundation for the resource distribution and control.

Traditionally, autoregressive moving average (ARMA) model has been used for network traffic prediction, which has difficulties in setting up its parameter values and dealing with non-linear time series. Another complexity is related with self-similar nature of network traffic makes high accurate prediction difficult. The parameter fitting procedure is time consuming. The goal is to forecast future traffic variations as precisely as possible, based on the measured traffic history.

Internet Traffic Simulation via Self- Similarity

The process is self-similar if its statistical behavior is independent of the time-scale. This means that averaging over equal periods of time the statistical characteristics of the process does not change. This is a mathematical concept of the self-similarity.

One of the first attempts to describe long range dependent (LRD) traffic exploited Fractional Brownian Motion (FBM) models, whose Gaussian nature helps in the study of the queuing behavior. However, FBM models present a restrictive correlation structure that fails to capture the short-term correlation of real traffic and its rich scaling behavior.

An M/G/∞ queue with service time with infinite variance is used in [1], [2] to model video sources. All these traffic models deviate considerably from classical Markovian models, which, however, continue to be widely used for performance evaluation purposes [3], [4], [5], [6]. In these works, the Markov Modulated Poisson Process (MMPP) is considered as the best Markov process to emulate LRD [4] and scale invariance [3] (multifractality in particular), though in [5], [6] it is correctly pointed out that any MMPP cannot exhibit LRD in a mathematically proper way, i.e., it is always possible to find a time lag above which an MMPP correlation decays exponentially. On the contrary, we need a model which is equally good for capturing short range dependence (SRD) and LRD processes. We will explore fractional ARIMA (FARIMA) time-series model for capturing LRD as well as SRD [7].

ARMA, ARIMA and seasonal ARMA

Box and Jenkins developed the ARMA model which is the combination of an autoregressive (AR) model and the moving average (MA) model. The ARIMA (p,d,q) model combines the Auto Regressive (AR) and Moving Average (MA) models developed earlier with differencing factor that removes in trend in the data or time series. This time series model is a kind of statistical model broadly used in network traffic analysis [8], [9], [10].

The order of the ARMA model in discrete time t is described by two integers (p,q), that are the orders of the AR and MA parts, respectively. The general expression for an ARMA-process y(t) is the following:

$$y(t)=\sum_{i=1}^{p}a(i)\cdot X(t-i)+\sum_{i=1}^{q}b(i)\cdot\varepsilon(t-i)+c+\varepsilon(t)\,, \qquad (1)$$

where p – the order of the AR part of the ARMA model; $a_1,a_2,\ldots a_p$ – the coefficients of the AR part of the model (of the recursive linear filter); q – the order of the MA part of the ARMA model; $b_1,\ldots b_q$ – the coefficients of the MA part of the model (of the non-recursive linear filter); X(t) – elements of the (input) white noise; ε(t) – output uncorrelated errors.

For p,q,P,Q>0 and s>0 we say that a time series $\{X_t\}$ is a multiplicative seasonal ARMA model (SARMA (p,q) × $(P,Q)_s$) if (see (2)):

$$\Phi(B^s)\varphi(B)X_t=\Theta(B^s)\,\theta(B)W_t\,, \qquad (2)$$

where

$$\Phi(B^s)=1-\sum_{j=1}^{P}\Phi_j B^{js} \qquad (3)$$

and

$$\Theta(B^s)=1+\sum_{j=1}^{Q}\Theta_j B^{js}\,. \qquad (4)$$

For several time series containing different length number and statistics we build the ARMA, ARIMA and SARMA network traffic models.

Model Validation

If the model is validated on the same data set from which it was estimated, the fit always improves as the flexibility of the model structure increases. We need to compensate for this automatic decrease of the loss functions. There are several approaches for this. Probably the best known technique is Akaike's Final Prediction Error (FPE) criterion and his closely related Information Theoretic Criterion (AIC). Both simulate the cross validation situation, where the model is tested on another data set.
The AIC is formed as

$$\mathrm{AIC} = \log(V(1+2\frac{d}{N})), \tag{5}$$

where d – the total number of estimated parameters, N – the length of the data record, V – the loss function (quadratic fit) for the structure in question.

$$\mathrm{FPE} = \frac{1+\frac{d}{N}}{1-\frac{d}{N}} V. \tag{6}$$

Research

Our research is emphasized on analysis of ARMA, ARIMA and SARMA use for prediction of different internet traffic. Internet traffic is a wide concept. Internet traffic can contain different type of data gathered by different features, for example, flow metrics (duration, packet-count, total bytes), packet inter- arrival time (mean, variance), website access statistics ect. In our experiments we use website access statistics (application level). Traffic data is taken from website http://freestats.com/ - website access statistics for different time periods, collected for 7 days (*7d*) and 28 days (*28d*). Another data trace is simulated using FARIMA model using the setting parameters for the Hurst parameter H and the number of observations N: H=0.362 and N=168 (*Farima7d)*, H=0.50 and N=672 *(Farima28d)*. In our experiments we analyze traffic sources as given in Table 1.

Table 1. Summary of the traffic data used in the study

No	Trace name	Observations	Step
1.	7d	168	1h
2.	28d	672	1h
3.	Farima7d (simulated)	168	1h
4.	Farima28d (simulated)	672	1h

For statistical analyses we use program package "MATLAB p6.5", MATLAB scripts for ARMA, ARIMA and SARMA analysis [11].

The steps in our analysis can be summarized as follows:
1. Testing for stationary of the time series.
2. Seasonality detection.
3. Identification of the order of the AR component and the MA component from the autocorrelation plots of the stationary series.
4. The model parameter estimation and validation.
5. Estimation of the prediction accuracy.

Review of Studied Cases

Firstly, we analyse the autocorrelation function (ACF) of our experimental data (see Fig.1).

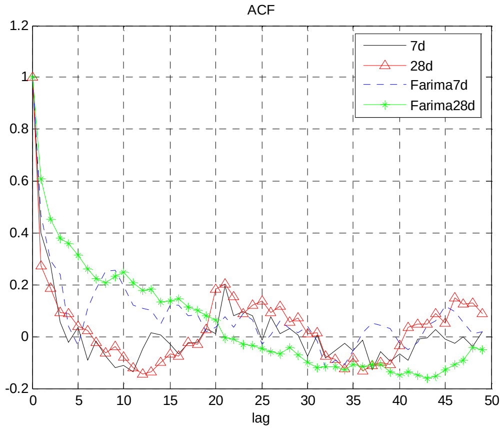

Fig. 1. ACF for traffic traces

The ACF doesn't give a clear concept of the data periodicity. For this reason we analyze periodogram (see Fig.2) of our experimental data to identify the period (seasonality) if it exists.

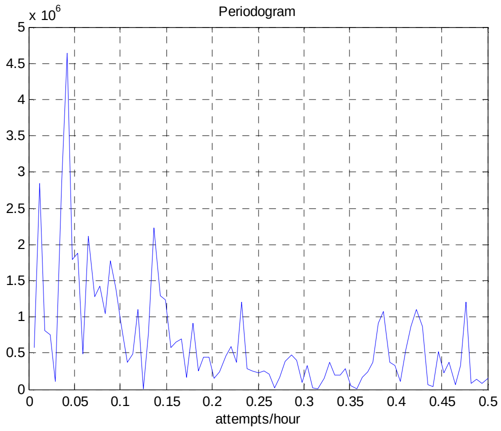

Fig. 2. Periodogram for trace "7d"

Fig. 2 shows the periodogram of trace "*7d*". We can see that:

– There is a peak when the frequency f is about 0.0119 which is called the main frequency. From this we can we can get the periodicity T=1/f=1/0.0119=84 and infer that the period of this network traffic is 84.

–There is a second peak when the frequency f is about 0.042 which is called second harmonic. This second harmonic period of this network traffic is 24 (24 hours or 1 day).

– There is a third peak when the frequency f is about 0.14 which is called third harmonic. The period of this peak is 7.

The period estimated from the periodogramms of different data traces are as follows:

- for trace “*7d*” – T=7, T=24, T=84;
- for trace “*28d*” – T=24;
- for trace “*Farima7d*” – T=24, T=28;
- for trace “*Farima28d*” – T=112.

Accordingly to the estimates period of seasonality we can build SARMA model for each data trace.

In our experiments with SARMA models we seta as input values: N – number of time series (observations); T – period of seasonal part; T_{pred} – prediction period (default T_{pred}=T); S_{slide} – number of seasons to slide between estimation (default S_{slides}=2); T_{est} – validation period, estimated:

$$T_{est} = \frac{N}{T} - 1 . \qquad (7)$$

Taking into account estimated seasonal part of the traffic traces, we calculate the validation period T_{est} . For real trace “*7d*” we select period of T=7, 12, 24. For real trace “*28d*” we select period T=24. For example, for the traffic trace “*7d*” of seasonal part T=7, the validation period in this case is T_{est}=168/7-1=23.

The calculated validation periods of different data traces are as follows:

- for trace “*7d*” – T_{est} =6, T_{est} =13, T_{est} =23;
- for trace “*28d*” – T_{est} =27;
- for trace “*Farima7d*” – T_{est} =5, T_{est} =6;
- for trace “*Farima28d*” – T_{est} =5.

The best results for real and simulated traces using ARMA, ARIMA and SARMA models are summarized in the Table 2 and the Table 3. The best result in the Table 2 and the Table 3 is marked in bold.

Table 2. Best AIC and FPE for real traces

Trace Model	***7d***	***28d***
ARMA	Model: (p,q)=(2,1) AIC: 8.25 FPE: 3840	Model: (2,2) AIC: 8.45 FPE: 4697
ARIMA	Model: (p,d,q)=(2,1,1) AIC: 8.19 FPE: 3618	Model: (p,d,q)=(1,1,2) AIC: 8.43 FPE: 4587
SARMA	Model: (2,0) ×$(1,1)_7$ AIC: 8.05 FPE: 3167 Model: (1,0) ×$(1,1)_{12}$ AIC: 8.05 FPE: 3179 **Model: (2,2) ×$(1,1)_{24}$** **AIC: 7.10** **FPE: 1433**	**Model: (1,0) ×$(0,0)_{24}$** **AIC: 8.31** **FPE: 4055**

Table 3. Best AIC and FPE for simulated traces

Trace Model	***Farima7d***	***Farima28d***
ARMA	Model (p,q)=(2,2) AIC: -2.33 FPE: -7.91	Model (p,q)=(2,1) AIC: -9.09 FPE: -0.0913
ARIMA	Model: (p,d,q)=(1,1,2) AIC: -2.11 FPE: -8.09	Model: (p,d,q)=(1,1,2) AIC: -6.77 FPE: -0.0934
SARMA	**Model: (0,1) ×$(1,1)_{24}$** **AIC: -1.31** **FPE: 0.316** Model: (0,1) ×$(1,1)_{28}$ AIC: -3.74 FPE: -0.026	**Model: (0,1) ×$(1,1)_{112}$** **AIC: -4.96** **FPE: -0.0071**

As we see in the Table 2 and the Table 3, the best result for simulated traces is achieved with SARMA model.

Fig. 3 shows the prediction values against real values for the best results for traffic trace “7d” using SARMA (2,2) ×$(1,1)_{24}$. Fig. 4. represents ACF and PACF of residuals for traffic trace “7d” using SARMA (2,2) ×$(1,1)_{24}$.

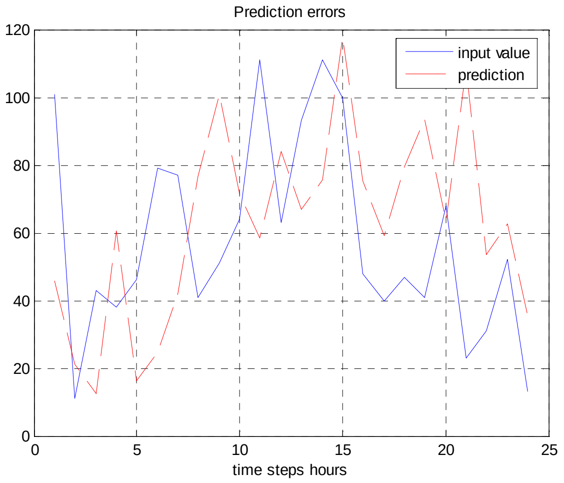

Fig. 3. Prediction curve for trace “*7d*” with SARMA (2,2) ×$(1,1)_{24}$

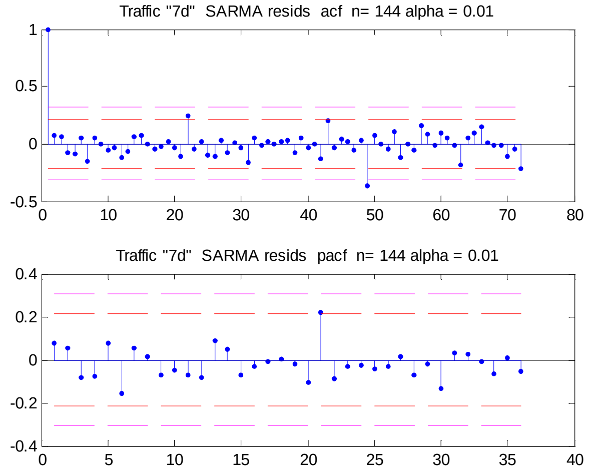

Fig. 4. ACF and PACF of residuals for trace “*7d*” with SARMA (2,2) ×$(1,1)_{24}$

Conclusions

We performed experiments with real traffic and simulated traffic to study the feasibility of the proposed steps on modeling and prediction. An important aspect is the seasonality of the traffic traces. In cases when we detect seasonality, the prediction model with the less AIC is SARMA. We found that the relative error between prediction values and actual values are less than 0.3.

Research must be continued for deeper studies of internet traffic analysis. Our research has started with traffic analysis in the application level containing less than 700 observations. Interesting statistics could be achieved in other network layers according to the OSI. These circumstances may be significant for time series prediction purpose planning server capacity for longer time period.

References

1. **Krunz M., Makowski A.** A Source Model for VBR Video Traffic Based on M/G/infinity Input Processes // Proc. of IEEE Infocom'98. – 1998. – P. 1441–1448.
2. **Zukerman M., Neame T., Addie R. G.** Internet Traffic Modeling and Future technology Implications // Proc. of IEEE Infocom'03. – 2003. – P.1–4.
3. **Horvath A., Telek M.** A Markovian Point Process Exhibiting Multifractal Behavior and its Application to Traffic Modeling // Proc. of 4-th Int. Conf. on Matrix-Analytic Methods in Stochastic Models MAM4. – 2002. – P.14–18.
4. **Andersen A. T., Nielsen B. F.** A Markovian Approach for Modeling Packet Traffic with Long-Range Dependence // IEEE JSAC. – 1998. – Vol.16, No. 5. – P. 719–732.
5. **Robert S. and Le Boudec J. Y.** On a Markov Modulated Chain Exhibiting Self-Similarities Over Finite Timescale // Performance Evaluation. – Elsevier. – 1996. – Vol. 27–28. – P. 159–173.
6. **Robert S., Le Boudec J. Y.** New Models for Pseudo Self-Similar Traffic // Performance Evaluation. – 1997. – Vol. 30, No. 1–2. – P. 57–68.
7. **Ghaderi M.** On the relevance of self-similarity in network traffic prediction // Tech. Rep., CS-2003-28, School of Computer Science. – University of Waterloo, Waterloo. – 2003.
8. **Zhigang J., Qiyan C., Huajie G.** Dial-up User Models and Traffic Prediction // IEEE Proceedings "TENCON 2004". – November, 2004. – P. 636–639.
9. **El Hag H. M. A., Sharif S. M.** An adjusted ARIMA model for internet traffic // IEEE Proceedings "AFRICON" 2007. – Sept., 2007. – P. 1–6.
10. **Yantai S., Minfang Y., Jiakun L., Yang. W.** Wireless traffic modeling and prediction using seasonal ARIMA models // IEEE Conference on Communications 2003. – May, 2003. – P. 1675–1679.
11. Internet source: www.stat.unc.edu/faculty/hurd/stat185Data/progdoc.html

Received 2009 02 11

G. Rutka. Some Aspects of Traffic Analysis used for Internet Traffic Prediction // Electronics and Electrical Engineering. – Kaunas: Technologija, 2009. – No. 5(93). – P. 7–10.

An overview of ARMA models used for traffic prediction is presented. The prediction of time series plays a very important role for network planning. For this reason we build the network prediction models using ARMA, ARIMA and seasonal ARMA (SARMA) model. Analyzing these models we can underline the significance of seasonal component detection. SARMA is the best prediction model for time series containing seasonal component. Ill. 4, bibl. 11 (in English; summaries in English, Russian and Lithuanian).

Г. Рутка. Некоторые аспекты анализа прогнозирования потоков передачи данных в интернете // Электроника и электротехника. – Каунас: Технология, 2009. – № 5(93). – С. 7–10.

Дано представление об использовании ARMA моделей (метода) при прогнозировании нагрузки сети (трафика). Прогнозирование временных последовательностей играет очень важную роль при планировании сетей. В связи с этим мы разработали модели (методы) прогнозирования сетей, используя ARMA, ARIMA и периодическую модель ARMA. Проведя анализ вышеупомянутых моделей (методов), мы можем подчеркнуть важность нахождения промежутков времени. SARMA является наилучшим методом для прогнозирования промежутков времени, содержащих периодическую составляющую. Ил. 4, библ. 11 (на английском языке; рефераты на английском, русском и литовском яз.).

G. Rutka. Kai kurie interneto duomenų srautų prognozavimo analizės aspektai // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2009. – Nr. 5(93). – P. 7–10.

Pateikta duomenų srautams prognozuoti naudojamų ARMA modelių analizė. Prognozuoti laikines sekas labai svarbu planuojant tinklus. Dėl to tinklo duomenų srautų prognozavimo modeliai buvo sudaromi naudojant ARMA, ARIMA bei sezoninį ARMA (SARMA) modelį. Atlikus šių modelių analizę galima teigti, kad sezoninis komponentų detektavimas turi didelę reikšmę, o prognostinis modelis SARMA geriausiai tinka sezoninį komponentą turinčioms sekoms prognozuoti. Il. 4, bibl. 11 (anglų kalba; santraukos anglų, rusų ir lietuvių k.).