A Novel P2P traffic Prediction Algorithm Based on Hybrid Model

Han Zhi-jie¹, Wang Ru-chuan², Duan Xiao-yang³

¹College of Computer
University of Science and Technology of China
Hefei, China
²College of Computer
Nanjing University of Posts and Telecommunications
Nanjing, China
³School of Computer and Information Engineering
Henan University
Kaifeng, China

Abstract

The increasing P2P network traffic on the Internet has led to the problem of network congestion. In the consequence of the diversification of the P2P traffic and protocol, research on the management of P2P traffic has had many problems needed to resolve. P2P traffic Prediction is an important problem in the P2P traffic management. Based on the P2P traffic characters, this thesis presents a novel P2P traffic model, giving a traffic prediction algorithm based on wavelet-analysis, and proved the accuracy of the algorithm. Simulation has experiment figures that the algorithm has a high prediction precision and superior real-time performance.

Keywords: P2P; Traffic Prediction; Wavelet; Kalman Filter

1. Introduction

Recent years, because of its huge preponderance, P2P applications attract a lot of users, and with all these P2P applications diffused, the traffic component on the Internet has been much different from the past. Research indicates that P2P traffic makes about 80% of the total Internet traffic. This causes some problems, which mostly behave in these terms: (1) engrossing too much bandwidth, refrains other applications; (2) some P2P software could get through the existed fire wall and security agent, opens a loophole of an enterprise network from inside, and then various virus will get into the enterprise easily,
meantime, the private data will be leaked; (3) the harmful content in the P2P traffic may pollute the network; (4) some resource that has intellectual property is diffused without restraint. All these problems make it much necessary for us to do management to the P2P traffic.

The core issue of P2P traffic management is the traffic prediction, but there are almost no achievements in that area. In essence, P2P traffic is the superposition of linear traffic and nonlinear traffic. For linear prediction, its theoretical basis is that traffic in the network has some linear characteristics, but most net traffic has the nature of self-similarity, multi-construct, outburst continuity \[1\] \[2\] \[3\], which are obvious nonlinear characters \[4\], so linear prediction can not ensure the accuracy. Nonlinear prediction such as wavelet-analysis has become a research focus. But for its high computational complexity, non-recursiveness, it does not support on-line prediction, so it does not have so much practicality \[5\].

Literature \[6\] gives a linear traffic prediction model, which introduces state equation and measurement equation, so it can efficiently treats system noise and measurement noise, and improves the precision of prediction. But for its limitation, the model can not describe the characters such as self-similarity, multi-construct, and multi-scale properties of the net. Literature \[7\] gives a traffic prediction algorithm based on wavelet-analysis, which adapts multi-scale analysis, could efficiently treat the non-linear part of the net traffic, so it improves prediction precision. But because of the limitation of the wavelet itself, the algorithm can only make off-line prediction, lacking real-time function.

Based on the study on the literatures above, and combining with the characteristics of file-sharing P2P traffic, this thesis gives a P2P traffic prediction algorithm that based on wavelet-analysis. It makes different treatments to the linear part and non-linear part of the net traffic \[8\] \[9\], in which the linear part is treated with Kalman Filter, and the non-linear part is treated with wavelet-analysis. Because the wavelet has the properties of random walk, by the method of describing target-state wavelet transform coefficients as the Kalman Filter state variables of its system equation, we can combine the Kalman Filter system equation and multi-scale analysis method, then gives a P2P traffic prediction algorithm that not only real-time and recursive but also has the multi-scale analysis function.

This thesis has the below structure: in the second section, related work will be introduced; in the third section, discussing traffic model in the P2P file-sharing system; the forth section gives a traffic prediction algorithm based on wavelet-Kalman Filter; the fifth section analyzes the correctness and performance of the algorithm; the sixth section gives the simulation experiment; the last section is the summarization.

2. Related work

By now, because of the diversification of P2P protocol, the difficulties to tell P2P traffic form Web traffic and the lack of efficient measure method, there are still not so many achievements in the research area of P2P traffic measurement and P2P traffic characteristic analysis. Foreign countries also have different results on the research of P2P traffic properties. In order to construct an effective P2P traffic model, literature \[1\]-\[4\] work on the net traffic measurement. The result of the measurement indicates that P2P traffic have the properties of regionality, periodicity and self-simulation, meantime, the P2P traffic influences the statistical properties of total traffic in the high-speed IP backbone networks. According to the analysis of P2P traffic in our country ’s backbone net, literature [] points out that as a kind of application, in the term of total traffic, P2P still has property of long dependence, but for specific application, the long dependence is different, however, the long dependence of aggregate traffic is determined by the magistral traffic. While based on the P2P measurement in the environment of France telecom networks, literature \[2\] makes an indication that in certain conditions, P2P whole traffic will lose its property of long dependence. The literature then divides P2P traffic into signaling traffic and actual traffic, between which, the signaling traffic mainly includes such as TCP connection, querying request, querying answer; while actual traffic performs different characteristics according to different P2P
application systems. Literature [1] for the first time abstracts P2P total traffic as a computing process of the M/G/∞ Queueing System., which divides net traffic into three parts: short-time-traffic, ACK long-time-traffic, data micro-long-traffic. Rate process of convergence traffic, \( \{ X_n \} \), is represented as:

\[
X_n = X_n^e + X_n^m + e_n
\]  

(1)

\( \{ X_n^e \} \) represents the data micro-long-traffic, that is the key factor of the total traffic, and is non-steady variable; \( \{ X_n^m \} \) is the rate process of short-time-traffic, which influences low frequency region, has the characteristic of periodicity, and could be ignored. \( \{ e_n \} \) is the rate process of ACK long-time-traffic, and it can be described as a white noise that has a constant power spectral density.

Assuming \( P(n) \) is the activity probability of \( n \) source peers at time \( t \), and \( x(t) \) is a birth-death process that has \( N+1 \) states, then the state transition from \( t_n \) to \( t_{n+1} \) could be described by the difference equation:

\[
\frac{dP_n(t)}{dt} = (N+1-n)\lambda P_{n-1}(t) + (n+1)\mu P_{n+1}(t) - [(N-n)\lambda + n\mu]P_n(t)
\]  

(2)

By means of generation function, we can give a traffic control equation based on Kalman Filter:

\[
y(k+1) = Cy(k) + DU
\]  

(3)

\[
y(k+1) = E[y(t_n, t_{n+1} \mid x(t))], C = \frac{\gamma}{\lambda + \mu}[1 - e^{-(\lambda + \mu)t}], \quad D = \frac{\gamma}{\lambda + \mu}[(\lambda + u)T - (1 - e^{-(\gamma + u)t})], \quad U = \frac{N\lambda}{\lambda + \mu}
\]

The equation above is tenable for the mean value of traffic, but for the random variable \( y(k+1) \), observation noise \( v(k+1) \) is needed to consider. So, we can give the linear traffic prediction model:

\[
x(k+1) = x(k) + Bu(k)
\]

\[
y(k+1) = Cx(k+1) + v(k+1), \quad (B = D/C).
\]

\[
\tilde{x}(k) = [1 - K(k)C]\tilde{x}(k-1) + K(k)y(k) + [1 - K(k)C]Bu(k-1)
\]  

(4)

In which,

\[
\begin{align*}
\hat{x}(k) &= \tilde{x}(k) + K(k)[y(k) - C(\tilde{x}(k))] \\
K(k) &= \tilde{P}(k)C/[\tilde{P}(k)C^2 + R(k)] \\
\tilde{P}(k) &= [1 - K(k)C]\tilde{P}(k)
\end{align*}
\]

(5)

and \( B = D/C \)

\[
\begin{align*}
C &= \frac{\gamma}{\lambda + \mu}[1 - e^{-(\lambda + \mu)t}] \\
D &= \frac{\gamma}{\lambda + \mu}[(\lambda + \mu)T - (1 - e^{-(\lambda + \mu)t})] \\
\mu &= \frac{N\lambda}{\lambda + \mu}
\end{align*}
\]  

(6)
\[ Q(k), R(k) \] are the system noise and measurement noise, \( \tilde{P}(k), \hat{P}(k-1) \) are the prediction error and estimation error, \( \tilde{u}(k) \) is mean square deviation of system noise, \( K(k) \) is filter gain, \( \lambda \) is mean square exponent of ON time, \( \mu \) is mean square exponent of OFF time, \( \gamma \) is the packet generation rate, which is a constant here, \( T \) is measured time interval.

The prediction accuracy of the model is determined by its state equation, however, the state equation can not treat the self-similarity and the multi-scale property of real net traffic. What is more important, the state equation is linear, and can not describe the true non-linear property of the net traffic. Beside this, for the model just linearly treats the net traffic, it increases the computational complexity, and decreases the prediction accuracy. While literature [7] gives a traffic prediction model based on wavelet-analysis:

\[
X(t) = \text{Approx}_j(t) + \sum_{j=1}^{J} \text{det} \, a_{j,l}(t) = \sum_{k} a^{j}_k(k) \phi_{j,k}(t) + \sum_{j=1}^{j} \sum_{k} d^{j}_k(k) \varphi_{j,k}(t) \tag{7}
\]

\( \phi_{j,k}(t) \) is scaling function, \( \varphi_{j,k}(t) \) is wavelet function, \( J \) is finest scale and we have equations:

\[
\begin{align*}
     a^{j}_k(k) &= \langle X, \phi_{j,k} \rangle \\
     d^{j}_k(k) &= \langle X, \varphi_{j,k} \rangle 
\end{align*}
\tag{8}
\]

\[
\begin{align*}
     \phi_{j,k}(t) &= 2^{-j/2} \phi_{0}(2^{-j}t - k), \, k \in Z \\
     \varphi_{j,k}(t) &= 2^{-j/2} \varphi_{0}(2^{-j}t - k), \, k \in Z 
\end{align*}
\tag{9}
\]

Wavelet-analysis is multi-scaled, so it could efficiently analyzes the net traffic that has the properties of multi-scale and self-simulation, and at the same time, the traffic prediction method based on wavelet-analysis could effectively improve prediction accuracy. But, for the wavelet-analysis itself is not recursive [5], and has a high computational complexity, it does not support on-line prediction. Besides, the wavelet has a boundary effect, which increases complexity of the algorithm.

3. File-sharing P2P traffic prediction model

Generally, the resources shared in the File-sharing P2P network are video files and software, and the total traffic is made of the traffic generated in the stage of searching resource (signaling traffic) and the traffic generated in the stage of downloading resource (data traffic). According to the existed research on P2P traffic prediction [1-6], in the backbone net, P2P traffic has some constant ratio relation with the total traffic in the internet, and the packet size of different P2P protocols is relatively fixed. Based on the formula (1) in literature [1], we construct the P2P traffic model:

\[
X(t) = A(t) + B(t) + V(t) \tag{10}
\]
In which, \( X(t) \) is total traffic of the network, \( A(t) \) is the stationary part of P2P net, \( B(t) \) is periodic variation part, \( V(t) \) is random variation part with statistical characteristic. \( A(t) \) and \( B(t) \) could be considered as stationary part wholly, and be merged as \( S(t) \), so:

\[
X(t) = S(t) + V(t)
\]  

Doing discrete wavelet transformation to \( S(t) \), we can get:

\[
S(t) = \sum_{m=1}^{2^L} b_{c,s,m}^L \phi_{L,m}(t) + \sum_{j=L}^{M-1} \sum_{m=1}^{2^j} b_{d,s,m}^j \psi_{j,m}(t)
\]  

(12)

\( \phi_{L,m}(t) \) is scale function, \( \psi_{j,m}(t) \) is wavelet function, \( 2^L \) is finest scale, and

\[
b_{c,s,m}^L = <S(t), \phi_{L,m}(t)>, \quad b_{d,s,m}^j = <S(t), \psi_{j,m}(t)>
\]  

(13)

According to the self-simulation property of file-sharing P2P network, we treat the wavelet-analysis coefficient, call the treatment process BWT, and define the scale factor \( u(k) = \frac{b_{c,s,m}^L}{b_{d,s,m}^j} \), ( \( u(k) \geq 1 \)), \( u(k) \) is normal distribution with mean square of 0, whose probability distribution function is defined as \( N(j) \), then the process BWT is:

In the finest scale \( j \), random variation \( u(j) \), \( b c_j \) are generated according to \( N(j) \), we can get:

\[
bd_j = \frac{b d_j}{u(j)};
\]

Repeating the above process until \( j = 0 \), finally, we can get:

\[
WB = \begin{bmatrix} b c_0^0 & b d_0^0 & b d_0^1 & b d_0^1 \\
bd_0^2 & b d_0^2 & b c_2^2 & \cdots & b c_{2^{M-1}}^M \end{bmatrix}
\]

From the model above,

\[
Z(k) = W^T \bullet WB + V(k)
\]  

(14)

\( W_k^T \) is inverse coefficient matrix of scale function \( \phi_{M,j}(t) \), based on this we can give KF function:

\[
\begin{bmatrix}
X(k) = \Phi(k,k-1)X(k-1) + J(k) \\
Z(k) = C(k)X(k) + V(k)
\end{bmatrix}
\]  

(15)

In which, transfer matrix \( \Phi(k,k-1) \) is \( I_{N \times N} \), \( I \) represents for the unit matrix, \( J(k), V(k) \) are white noise model, \( X(k) = WB \).
4. P2P traffic prediction algorithm

4.1 Algorithm

Based on the constructed P2P traffic model, we give the algorithm for P2P traffic. In order to describe conveniently, we define:

\[ Z(k) = [z(k_1), z(k_2), \ldots, z(k_N)]^T \]  (8)

\[ Z_1^k = [Z^T(1), Z^T(2), \ldots, Z^T(k)] \]  (9)

\( Z(k) \) is the sets of measured series \( z(k_1), z(k_2), \ldots, z(k_N) \) in \( K \) periodic, \( Z_1^k \) is the measured value series set at all time points in periodic 1, 2, \ldots, \( k \). Main processes of the algorithm are as the below:

1. Establishing measurement equations as (7);
2. Getting initial value = \( \hat{X}, P_0, K_0 \);
   
   In the \( (k-1) \) periodic, we get \( \hat{X}(k-1 \mid k-1), P(k-1 \mid k-1) \); and calculate \( \hat{X}(k \mid k-1), P(k-1 \mid k-1) \), in which, \( \hat{X}(k \mid k-1) = E\{X(k) \mid Z_1^{k-1}\} \),

\[ P(k \mid k-1) = E\{[X(k) - \hat{X}(k \mid k-1)][X(k) - \hat{X}(k \mid k-1)]^T\} \]  (11)

Then, doing wavelet reconstruction to \( \hat{X}(k \mid k-1) \), getting prediction value of the \( K \) periodic \( S(k) \),

\[ \hat{S}(k \mid k-1) = E\{S(k) \mid Z_1^{k-1}\} \]  (12)

3. In the \( k \) periodic, we get the value of \( Z(k) \) in turn, For \( k = 1 \) to \( N \)
   
   Begin
   
   1. Updating \( \hat{X}_c(k \mid k_1, k_2, \ldots, k_{i-1}) \) with \( z(k_i) \), then based on \( Z_1^{k-1} \) and observation information, we can get the estimated value of state \( X(k) \), and the relative covariance matrix of estimation error:

\[ \hat{X}_c(k \mid k_1, k_2, \ldots, k_i) = E\{X(k) \mid Z_1^{k-1}, z(k_1), z(k_2), \ldots, z(k_i)\} \]  (13)

\[ P(k \mid k_1, k_2, \ldots, k_i) = E\{\hat{X}(k \mid k_1, k_2, \ldots, k_i) - X(k) \mid k_1, k_2, \ldots, k_i\} \]  (14)

2. Doing wavelet reconstruction to \( \hat{X}_c(k \mid k_i) \), we can get estimated value of \( S(k) \), which is based on \( Z_1^{k-1} \) and the observation information \( z(k_1), z(k_2), \ldots, z(k_i) \).

\[ \hat{S}(k \mid k_1, k_2, \ldots, k_i) = E\{S(k) \mid Z_1^{k-1}, z(k_1), z(k_2), \ldots, z(k_i)\} \]  (15)

End
So, based on \( Z_{i}^{k-1} \) and the observation information \( z(k_{1}), z(k_{2}), \ldots, z(k_{l}) \), we can make prediction of \( s(k_{i}) \) \((l = i + 1, \cdots, N)\), the calculation equation are shown as (33), (34):

\[
\hat{s}(k_{i} | k_{1}, k_{2}, \cdots, k_{l}) = E\{s(k_{i}) | Z_{i}^{k-1}, z(k_{1}), z(k_{2}), \cdots, z(k_{l})\}, \quad l = i + 1, \cdots, N
\]

\[
\tilde{X}(k | k_{1}, k_{2}, \cdots, k_{l}) = X(k) - \hat{X}_{e}(k | k_{1}, k_{2}, \cdots, k_{l}) \quad (1 \leq i \leq N)
\]

According to steps(3), (4), finally, we can get the estimated value of state \( X(k) \) based on \( Z_{1}^{k} \), and the relative covariance matrix of estimation error, which is:

\[
\hat{X}(k | k) = \hat{X}_{e}(k | k_{N}) = E\{X(k) | Z_{1}^{k}\}
\]

\[
P(k | k) = P_{e}(k | k_{N}) \quad \text{and the estimated value of } S(k) \text{ based on } Z_{1}^{k}:
\]

\[
\hat{S}(k | k) = \hat{S}(k | k_{N})
\]

Repeating (3), (4), (5), until all the periodic are treated.

5. Simulation experiment

Using NS2, the experiment firstly generates the topological graph of P2P overlay network, then choosing a peer arbitrarily as traffic prediction node; programming the algorithm above in Matlab6.5; and at last showing the result by graph. For the diversity of P2P business traffic, here, we make simulation towards multi-business traffic. Besides, because of the traffic generated by transferring video files is the domain traffic in P2P net, we assume the P2P traffic has the properties of long range dependence and self-similarity.

In graph 1, abscissa is the statistical time quantum, ordinate is net traffic. In the experiment, we measure the prediction value of net traffic based on our algorithm, and comparing it with the real net traffic. As is shown in graph1, the result indicates our prediction algorithm gets a higher coincidence degree with the real net traffic, which is because we treat towards the non-linear multi-scale properties of the traffic.

Fig.1 P2P network prediction traffic
6. Summary

In the consequence of the diversification of the P2P business and protocol, research on the management of P2P traffic has had many problems needed to resolve. Prediction of the P2P traffic is kernel problem in the P2P traffic management. Based on the existed P2P traffic characters, this thesis structures a P2P traffic model, gives a traffic prediction algorithm bases on wavelet-analysis, and proves the accuracy of the algorithm. Simulation experiment figures that the algorithm a high prediction precision and superior real-time performance.

References


[2] Nadia Ben Azzouna and Fabrice Guillemin France Telecom, Division R&D Impact of peer-to-peer applications on wide area network traffic: an experimental approach 2, Avenue Pierre Marzin, 22300 Lannion, France IEEE Communications Society


