네트워크 트래픽 예측을 위한 시계열 모형의 적합성 검증

정상준*, 김동주**, 권영헌***, 김종근**

A Fitness Verification of Time Series Models for Network Traffic Predictions

Sangjoon Jung*, Dongju Kim**, Younghun Kwon***, Chonggun Kim**

요 약

인터넷의 발달로 네트워크 트래픽은 현저하게 증가되었다. 트래픽의 폭증은 전체 네트워크의 성능에 크게 영향을 미치게 되었으며 트래픽의 관리가 망 관리의 중요한 이슈로 되었다. 본 논문에서는 네트워크 트래픽을 분석하여 효율적인 대응을 수립하기 위해 예측하는 시계열 모형의 적합성을 검증한다. 네트워크 트래픽을 예측하기 위해서는 시간적 흐름에 따라 자료간의 상관 관계를 유추하고, 이 관계를 이용하여 예측을 수행한다. 상관 관계를 유추하는 과정에서 필연적으로 확률적 오류를 포함하게 되는데, 정확한 예측을 위해서는 확률적 오차를 최소화해야 한다. 따라서, 통계학 분야에서 예측 방법으로 널리 쓰이는 시계열 모형인 AR, MA, ARMA, ARIMA 모형을 사용하여 네트워크 트래픽을 예측함과 동시에, 예측하는 과정에서 정확한 예측을 수행할 수 있는지에 대한 적합성을 검증하고자 한다. 적합성 검증은 모형 식별 단계에서 초기 단계인 정상성 가정을 만족하는지의 여부로 판단하며, 정상성 가정은 자기상관함수와 편자기상관함수를 통해 구할 수 있다. 정상성 가정을 만족하지 못하는 모형은 비정상 시계열 자료로 분류되는데 이 경우의 예측은 정확하다고 볼 수 없다. 따라서, 정확한 예측을 수행할 수 있도록 시계열 자료의 정상성 가정을 만족하도록 모형을 분류하는 방안을 제시하고자 한다. 정확한 예측을 수행하면, 네트워크 트래픽을 좀 더 나은 방법으로 관리하며, 예측 결과를 이용하여 동적인 트래픽의 관리가 가능하게 된다.

Key Words : Traffic, Timeseries analysis, Stationary, Network management, Traffic analysis

Abstract

With a rapid growth in the Internet technology, the network traffic is increasing swiftly. As for the increase of traffic, it had a large influence on performance of a total network. Therefore, a traffic management became an important issue of network management. In this paper, we study a forecast plan of network traffic in order to analyze network traffic and to establish efficient correspondence. We use time series forecast models and determine fitness whether the model can forecast network traffic exactly. In order to predict a model, AR, MA, ARMA, and ARIMA must be applied. The suitable model can be found that can express the nature of traffic for the forecast among these models. We determines whether it is satisfied with stationary in the assumption step of the model. The stationary can get the results by using ACF(Auto Correlation Function) and PACF(Partial Auto Correlation Function). If the result of this function cannot satisfy then the forecast model is unsuitable. Therefore, we are going to get the correct model that is to satisfy stationary assumption. So, we proposes a way to classify in order to get time series materials to satisfy stationary. The correct prediction method is managed traffic of a network with a way to be better than now. It is possible to manage traffic dynamically if it can be used.

*경일대학교 교양학부(sjjung@kiu.ac.kr)
**영남대학교 컴퓨터공학과 네트워크연구실(jks4721@yumail.ac.kr, cgkim@yu.ac.kr)
***세경대학 컴퓨터정보통신과 (yhkwon@saekyung-c.ac.kr)
논문번호 : 030179-0428, 접수일자 : 2003년 4월 26일

Ⅰ. 서 론

인터넷을 선두로 하여 전세계적으로 다양한 통신망이 구축, 운용되고 있으며 이들 통신망을 이용하여 다양한 종류의 통신이 서비스되고 있다[1][2]. 이러한 통신 서비스에 대하여 가입자들은 고품질의 서비스를 요구하고 있으며, 이 사항을 충족하기 위해서는 통신망 운용의 관리가 필수적이다[3][4]. 네트워크를 효과적으로 관리하기 위해서는 다양한 방법이 있으나 사용자가 신뢰할 수 있는 성능을 제공하기 위해서는 네트워크 자원들의 운영 형태와 통신 활동의 효율성을 평가하는 방법 외에 네트워크의 트래픽의 추이를 분석하는 방법이 있다[5]. 네트워크의 트래픽을 분석하면 네트워크 침입 등의 징후를 관리자에게 알림으로써 붕괴된 부분을 최대한 빨리 복구하는 체계를 구축하고, 향 후 네트워크의 구성을 확대할 수 있는 근거로 활용할 수 있다[2-5].

오늘날 점점 더 많은 시스템이 네트워크에 직접 연결됨에 따라 보다 많은 트래픽이 발생하게 되었다[1]. 사용자 수의 증가는 네트워크 트래픽을 증가시켰으며 트래픽 증가로 말미암아 네트워크 전체의 성능에 영향을 미치게 되었다. 따라서, 네트워크 성능을 일정한 수준으로 유지하기 위해서는 트래픽의 관리가 필수적이다[1][5]. 트래픽을 효율적으로 관리하기 위해서는 트래픽 분석 작업이 필요하다. 네트워크 트래픽 모니터링은 데이터를 수집하고 분석하는 기능을 제공하고 네트워크 상에 전송되는 패킷을 검사하여 네트워크 관리에 필요한 부가적인 정보를 제공한다[4]. 이 정보는 분석 작업을 통해 좀 더 확장된 정보로 가공되는데 트래픽의 양, 트래픽의 형태, 많은 트래픽을 일으키는 근원지, 병목현상이 발생되는 장소, 최대 트래픽의 양과 최대 트래픽이 지속되는 기간 등으로 가공된다. 트래픽의 분석을 통해 전체 망에서 전송되는 트래픽을 측정하고, 트래픽의 종류를 분석함으로써 중단 없는 서비스를 제공할 수 있는 근거로 사용된다. 트래픽 정보를 관리자에게 제공하여, 관리자가 쉽게 관리하는 방법으로 SNMP를 이용한 네트워크 관리 방법과 서브네트워크에 위치한 모니터링 시스템을 이용하는 방법 등이 있다[2-4].

네트워크의 효율적인 관리를 위해 트래픽을 측정하고 분석하는 방법이 중요하지만, 이보다 진일보한 방법으로 특정 시스템에서의 트래픽을 예측하는 것이 있다. 트래픽의 예측을 수행하게 되면, 특정 시스템에서의 통신 부하를 예측하게 됨으로써, 서버 또는 링크를 보호할 수 있게 되고, 사이버 시위 등과 같은 불규칙한 트래픽 폭주를 대비할 수 있어 능동적인 성능 관리를 수행할 수 있다[16-18]. 유사 연구로는 누적된 HTTP 패킷을 기반으로 Web 네트워크 트래픽의 양을 측정하는 방법과 실제 전화 교환기 상의 트래픽 이용률 예측 알고리즘이 제시되기도 하였다[2][5]. 이 논문들은 시계열 모형을 이용하여 예측 값을 구하고 예측값과 실제값 간의 오차가 얼마인지에 대해 살펴본 논문이다. 이것은 예측 알고리즘을 적용하여 과거에 측정된 값들 간의 관계를 함수로 구하여 미래에 발생할 값을 구한 것으로 볼 수 있다[12-15]. 이것은 자기 유사성(Self-Similarity)과 같은 반복적인 규칙 또는 과거 값들간의 관계를 구해 현 시점에서 미래에 발생할 값을 구한 것으로, 미래의 값을 예측하기 위해 적용하는 함수에 확률적인 오차를 배제함으로써 예측된 결과의 신뢰성에 의문점을 내재하고 있다.

따라서, 본 논문에서는 트래픽 예측의 한 방법으로 시계열 모형을 이용하여 트래픽을 예측하는 것이 가능한지에 대해 살펴본다. 그리고, 모형을 적용하기 위해 포함되는 확률적인 오차를 최소화하여 정확한 결과를 얻을 수 있는 형태로 가공하는 방법을 제안한다. 이 예측 모형은 트래픽이 발생하는 시점에서 곧바로 예측값을 얻는 실시간 형태로 적용한다.

Ⅱ. 관련 연구

2.1 시계열 예측 모형

시계열 자료는 시간의 흐름에 따라 관측된 일련의 결과이다[6]. 시계열분석은 과거 시계열 자료의 패턴이 미래에도 지속적으로 유지된다는 가정 하에서, 현재까지 수집된 자료들을 분석하여 미래를 예측하는 것으로 시계열자료가 생성된 시스템 또는 확률과정을 모형화하여 시스템 또는 확률 과정을 이해하고 제어할 수 있도록 하는 것이다. 즉, 관측된 과거의 시계열로부터 규칙적인 패턴을 발견하고 미래에도 그러한 패턴이 계속 될 것이라는 기대 하에 그 패턴을 모형화하여 미래 예측하는 것이 시계열 모형에 의한 예측이다[6-8]. 이 때 과거의 값들이 미래에도 같은 확률 과정으로 유지된다는 가정을 만족하느냐가 매우 중요한 데, 모형을 수립했을 때 이 가정을 만족해야만 정확한 예측 모형으로 인정될 수 있다[6]. 또한, 가정을 만족하더라도 모형의 가정에는 필연적으

로 오차항을 포함하는데, 이 오차를 얼마나 최소화할 수 있는가가 중요한 이슈이다. 먼저 시계열 모형의 종류를 살펴보고, 시계열 모형을 적용하기 위해 필요한 가정을 알아보고, 이 가정을 만족하기 위해 필요한 것을 살펴본다.

(1) 자기회귀모형(AR, AutoRegressive Model)

현시점 t에서의 시계열 Z'_t는 p개의 과거값들의 가중합과 이들로 설명되지 않는 부분인 오차항 a_t의 선형결합으로 표현된다. 자기회귀모형은 시계열 자체에 대한 회귀 형태를 취하는 모형으로 일반 p차 AR과정을 따른 $\{Z'_t\}$는 다음과 같이 나타낸다.

$$Z'_t = \phi_1 Z'_{t-1} + \phi_2 Z'_{t-2} + \cdots + \phi_p Z'_{t-p} + a_t$$

(2) 이동평균모형(MA, MovingAverage Model)

이동평균 모형은 시계열 $\{Z'_t\}$가 시계열 자체에 대한 회귀형태를 띠고 있는 자기회귀 과정과는 달리 현재와 과거의 백색잡음들의 가중선형결합으로 표현되는 모형으로, 일반 q차 MA과정은 다음과 같이 나타낸다.

$$Z'_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$$

(3) 자기회귀이동평균 모형(ARMA)

어떤 시계열 데이터의 현재값 Z'_t가 자신의 과거값들 Z_{t-1}, Z_{t-2}, Z_t와, 오차항 a_t, 그리고 과거의 오차항들 a_{t-1}, a_{t-2}, $\cdots$, a_{t-q}에 의해 나타낼 수 있을 때, 이 시계열 데이터를 자기회귀 이동평균 모형이라 한다. 이때 가장 긴 AR의 차수 p와 MA의 차수 q를 ARMA모형의 차수라 하고 ARMA(p, q)로 나타낸다.

$$Z'_t = \phi_1 Z'_{t-1} + \phi_2 Z'_{t-2} + \cdots + \phi_p Z'_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$$

ARMA 모형은 p차의 AR(p) 모형과 q차의 MA(q) 모형의 합으로 나타나있기 때문에 두 모형의 성질을 공유한다. ARMA모형의 안정성은 AR부분에서 결정되고 가역성은 MA부분에 의하여 결정된다. 즉, 안정성의 조건을 만족하는 ARMA 모형은 MA(∞) 모형으로 변환될 수 있고, 가역성의 조건을 만족하는 ARMA 모형은 AR(∞)모형으로 표현될 수 있다.

(4) 자기회귀누적이동평균모형(ARIMA, AutoRegressive Integrated Moving Average Model)

시계열 $\{Z'_t\}$의 d차 차분한 시계열 $\{W_t = (1-B)^d Z_t\}$이 AR차수가 p, MA 차수가 q인 ARMA(p, q) 모형을 갖는다면 시계열 $\{Z'_t\}$는 차수가 (p, d, q)인 자기회귀누적이동평균(ARIMA) 모형을 갖는다고 한다.

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + u_t - \theta_1 u_{t-1} - \theta_2 u_{t-2} - \cdots - \theta_q u_{t-q}$$

$$\phi(L) Z_t = (L) u_t \text{ 에서}$$

단, $\phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$

$\theta(L) = 1 - \theta_1 L - \cdots - \theta_q L^q$

즉, $\phi(L) \nabla^d Z_t = (L) u_t$

2.2 자기상관함수(ACF, AutoCorrelation Function)와 편자기상관함수(PACF, Partial AutoCorrelation Function)

시계열 자료를 기반으로 예측하기 위해서는 추론 과정이 필요하다. 이 추론 과정을 단순화시키기 위해 약간의 합리적인 가정이 필요한데 이것이 정상성의 가정이다. 정상성(Stationary) 가정이란 시계열의 확률적 성질이 시간에 따라 변하지 않는다는 것이다[6-8]. 즉, 시계열의 평균과, 분산이 시점 t에 관계없이 일정하고, 공분산은 두 시점의 시차 k에만 영향을 받는다. 그림 1은 정상성 가정을 만족하는 경우와 그렇지 않은 경우를 보여주고 있다.

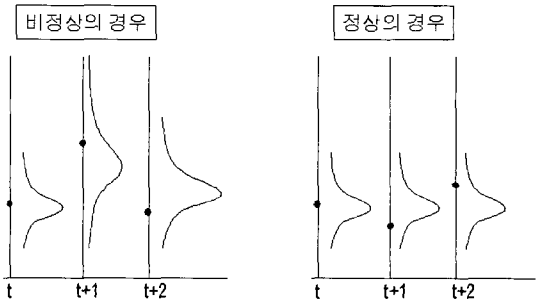


그림 1. 정상성

그림에서 보면 일반적인 확률과정과 정상 확률과정의 차이를 알 수 있는데 정상 확률 과정의 경우 각 시점의 분포가 평균과 분산이 동일하다. 이와 같이

과거 시간을 통하여 변하지 않고 그대로 남아있는 성질을 이용하는데, 그 이유는 미래를 예측하고자 하는 사람들에게는 상당히 안정되게 남아 있는 과거 자료에 대한 일정한 추세나 패턴을 이용하려고 하기 때문이다. 이것은 시간에 따라 변하지 않는 시계열의 어떤 함수가 필요하다는 것이다. 이러한 안정성의 가정을 정상조건이라 한다[6-8].

위에서 언급한 정상성 가정을 만족하는지의 여부를 판단하기 위해서는 자기상관함수와 편자기상관함수를 구해 그 결과값을 보고 판단한다. 먼저, 자기상관함수란 상관 계수와 비슷한 개념으로 상관 계수는 두 개의 서로 다른 특성을 나타내는 변수들 사이에 존재하는 상호작용 관계를 나타내며, 자기상관함수는 동일한 변수를 시점을 달리하여 관찰했을 때, 시점에 따라 다른 관찰 값들 사이에 존재하는 상호작용 관계를 나타내는 것이다. 즉, 동일한 변수에서 얻어진 연속적인 관측 값들 사이의 상호 연관 관계를 나타내는 척도이다. 자기상관함수는 다음과 같이 구할 수 있다. 정상시계열 $\{Zt\}$가 주어졌을 때 평균 μ와 자기공분산함수 γ_k를 다음과 같이 구한다.

$\hat{\mu} = \sum_{t=1}^{n} \frac{Z_t}{n} = \overline{Z}$ 일 때,

$\hat{\gamma_k} = \frac{1}{n} \sum_{t=1}^{n-k} (Z_t - \overline{Z})(Z_{t+k} - \overline{Z})$를

구할 수 있으며, 따라서 표본자기상관함수(Sample AutoCorrelation Function : SACF)는

$$\hat{\rho_k} = \frac{\hat{\gamma_k}}{\hat{\gamma_0}} = \frac{\sum_{t=1}^{n-k} (Z_t - \overline{Z})(Z_{t+k} - \overline{Z})}{\sum_{t=1}^{n} (Z_t - \overline{Z})^2}$$ 으로

추정된다.

편자기상관함수는 구하고자 하는 연속적인 2개의 시계열 자료에서 상관계수를 구하는데 있어서, 두 변수를 제외한 모든 변수의 영향을 제거하고 두 변수 사이의 순수한 상관 계수를 의미한다. 정상시계열 $\{Zt\}$에서 시차가 k인 편자기상관함수는 회귀식

$$Z_t = \phi_{k1} Z_{t-1} + \phi_{k2} Z_{t-2} + \cdots + \phi_{kk} Z_{t-k} + a_t$$

에서 k번째 회귀계수 ϕ_{kk}를 의미하는데, 이는 다음과 같이 표현할 수 있다.

$$\phi_{kk} = Corr[Z_t, Z_{t-k} | Z_{t-1}, Z_{t-2}, \cdots, Z_{t-k+1}]$$

즉, $Z_{t-1}, Z_{t-2}, \cdots, Z_{t-k+1}$영향을 제거시킨 후 Z_t와 Z_{t-k}의 순수한 상관계수를 의미한다.

시계열 자료에서 정상성을 만족하기 위해서는 위의 두 가지 상관함수를 구하여야 알 수 있다. 구해진 자기상관함수와 편자기상관함수 모두 각 시차에서 신뢰한계(표준편차의 2배)내에 존재하고 있을 경우 자기상관함수와 편자기상관함수의 성질을 만족한다고 볼 수 있다[6].

AR 모형을 식별하기 위해 자기상관함수와 편자기상관함수의 특성을 정리하면 다음 표와 같다. 표 1에서 보는 바와 같이 AR 모형을 식별하기 위해서는 자기상관함수는 지수적으로 감소하고, 편자기상관함수에서 유의한 값을 갖는 시차를 결정하면 자기회귀모형을 식별할 수 있다.

표 1. 시차에 따른 AR모형의 일반적 특징

시차	ACF (자기상관함수)	PACF (편자기상관함수)
AR(1)	시차가 증가함에 따라 지수적으로 감소하는 모습을 보임	시차 1에서만 유의한 값을 가지며 나머지 시차에서는 신뢰한계내에 들어와 시차 1 이후에 절단되는 모습을 보임
AR(2)	모든 시차에서 0이 아닌 값을 가지며, 지수적으로 감소하는 형태를 보임	시차 1과 2에서만 값을 가지고 있으며, 시차가 3 이상의 경우에는 0의 값을 갖는 절단된 형태
AR(p)	지수곡선 또는 sine 곡선의 형태를 그리며 시차 k가 커짐에 따라 급격히 감소	시차 p까지는 유의한 값을 보이며, p+1 이후 시차에서는 0의 값을 가짐

MA 모형을 식별하기 위해 자기상관함수와 편자기상관함수의 특성을 정리하면 표 2와 같다. 표 2에서 보는 바와 같이 MA 모형을 식별하기 위해서는 편자기상관함수는 지수적으로 감소하고, 자기상관함수에서 유의한 값을 갖는 시차를 결정하면 이동평균모형을 식별할 수 있다.

표 2. 시차에 따른 MA모형의 일반적 특징

시차	ACF(자기상관함수)	PACF (편자기상관함수)
MA(1)	시차가 1인 경우에만 일정한 값을 가지고 시차 2 이후에는 0의 값을 갖는 절단 형태를 띠고 있음	1의 부호에 따라 지수적으로 감소하는 형태를 띰
MA(2)	시차 2까지 유의한 값을 보이며, 시차 3이후에는 0의 값을 갖는 절단 형태를 띠고 있음	지수적으로 감소하거나 점차적으로 진폭이 줄어드는 sine 곡선의 형태를 나타냄
MA(q)	시차 q까지 유의한 값을 가지고, 시차 q+1 이후에는 0으로 절단되는 형태	지수곡선 또는 sine 곡선의 형태를 그리며 시차 k가 커짐에 따라 급격히 감소하는 형태

ARMA 모형의 식별하기 위해 자기상관함수와 편자기상관함수의 특성은 표 3과 같이 요약할 수 있다.

표 3. 시차에 따른 ARMA모형의 일반적 특징

시차	ACF(자기상관함수)	PACF (편자기상관함수)
ARMA(1, 1)	AR(1)모형과 MA(1)모형의 자기상관함수를 결합한 형태로, 시차 1의 자기상관함수에는 MA(1)모형의 모수가 포함되어 있지만 시차 2부터는 AR(1) 모형과 같이 지수적으로 감소하는 형태를 보임	AR(1)모형과 MA(1)모형의 편자기상관함수를 결합한 형태로, 시차 2부터 MA(1) 모형의 편자기상관함수의 형태와 같이 지수적으로 감소하거나 부호가 바뀌면서 감소하는 모습을 보임

ARIMA 모형의 식별을 식별하기 위해서는 앞에서 언급한 AR, MA, ARMA 모형을 제외한 모형으로 분류할 수 있다. AR, MA, ARMA 모형은 정상성 가정을 만족하는 경우에 사용되는 시계열 모형, 즉 정상 시계열 모형이다. 자기상관함수와 편자기상관함수를 이용하여 정상성 가정을 만족하지 못하는 경우, 차분이나 로그 변환을 통해 시계열 자료를 변환하는데, 변환되는 시계열 모형을 비정상 시계열 모형이라고 하고, 이 때 적용되는 모형이 ARIMA 모형이다. 따라서, ARIMA 모형은 추세를 가지거나 분산이 일정하지 않는 시계열 자료가 대부분이다.

정리하면 자기상관함수는 관측된 시계열이 시간에 따라 독립인지 아닌지 독립이 아니라면 시간에 어떻게 종속되어 있는지를 파악하는데 이용하며, 또한 모형진단에서 잔차들에 대한 자기상관함수를 통해서 백색잡음의 성질(평균은 0, 분산은 일정하며, 랜덤하게 분포)을 만족하는지를 판단하는데 이용된다. 편자기상관함수는 구하고자 하는 연속적인 2개의 시계열 자료에서 상관계수를 구하는데 있어서, 두 변수를 제외한 모든 변수의 영향을 제거하고 두 변수 사이의 순수한 상관 계수를 의미한다. 자기상관함수와 편자기상관함수를 구하는 이유는 정확한 예측을 위해서이다. 즉, 예측을 하기 위해 사용되는 과거값들의 상관관계가 미래를 예측하는 예측하는데 정확한 상관관계로 나타나는가에 대한 문제이다. 이것은 자기 유사성(Self-similarity)에서 가지는 한계점을 의미한다. 자기 유사성은 일정하게 반복되는 패턴을 서로 다른 레벨의 집합에서도 반복된다는 점을 이용하는 예측이다[12-15]. 동일하게 반복될 경우에는 문제가 발생하지 않지만, 비슷하거나 확률적 오류를 포함할 경우의 예측은 오차를 포함하게 된다. 실제 자기 유사성을 추출하기 위해 구하는 Hurst Parameter는 1이 아닌 경우가 대부분이다[12-15]. 여기에서 1은 정확하게 반복된다는 것을 의미한다. 자기상관함수와 편자기상관함수는 이런 확률적 오류를 최소화하여 보다 정확한 예측을 하기 위해 필요한 함수이다.

Ⅲ. 네트워크 트래픽의 실시간 예측

3.1 실험 네트워크 환경 및 시계열 데이터의 수집

네트워크 트래픽 예측을 위해 시간별 네트워크 트래픽 데이터를 수집한다. 시계열 자료를 얻기 위해서는 시간대별 트래픽 양을 얻게 되며, 이 데이터를 수집하기 위한 네트워크 환경은 다음 그림과 같다.

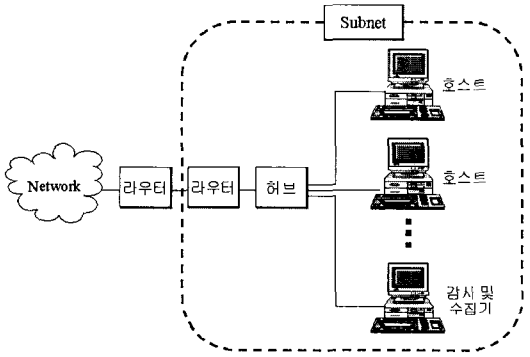


그림 2. 시계열 데이터 수집을 위한 네트워크 환경

그림 2에서 보이는 바와 같이 데이터를 수집하기 위해 서브네트워크 내에 트래픽을 자동으로 수집하기 위한 트래픽 수집기를 설치한 컴퓨터를 허브에 연결한다[12]. 서브네트워크는 영남대학교 컴퓨터공학과 네트워크 연구실을 이용하는데 이 서브넷은 외부 네트워크를 이용하는 호스트의 수가 20대 이상으로 구성되어 있다. 데이터 수집을 실시간으로 한 이유는 누적된 데이터의 예측은 누적된 시간 만큼에 대한 미래, 즉 예측된 값이더라도 과거 시간에 종속되는 값이다. 따라서, 현재의 시간에 필요한 예측값을 구하기 위해서는 시간에 의존적인 시계열 데이터를 구할 필요가 있으며, 이를 실시간 데이터로 분류한다.

시계열 데이터는 2001년 7월부터 2002년 8월까지 수집한 패킷의 총량인데 시간 단위로 수집하였다. 장기간에 걸쳐 패킷을 수집하는 동안 정전이나 네트워크 장애, 바이러스가 발생한 경우에는 수집할 수 없었는데, 정확한 예측 모형을 추출하기 위해서 한시간이라도 누락된 데이터일 경우 하루, 일주일, 한 달에 해당하는 전체 데이터를 모형 수립 단계에서 제외하였다.

3.2 시계열 분석을 위한 예측 방법

네트워크 트래픽을 예측하기 위해 앞에서 언급한 바와 같이 시계열 예측 모형을 적용한다. 트래픽을 예측하기 위해서는 정상성 가정이 필수적이다. 정상성 가정을 만족하면 합리적인 모형이 수립되고, 이 모형을 근거로 예측된 자료가 정확한 예측 결과라고 볼 수 있다. 따라서, 네트워크 트래픽의 정확한 양을 예측하기 위해서는 정상성 가정을 만족하는지의 여부를 조사한다. 정상성 가정을 만족하는 경우 정상 시계열로 분류하고, 정상성 가정을 만족하지 못하는 경우 비정상 시계열로 분류한다. 이 가정을 만족하는지의 여부는 자기상관함수와 편자기상관함수를 통해 구할 수 있다. 만약 정상성 가정을 만족하지 못한다면, 차분이나 로그변환을 수행한다.

3.3 시계열 모형의 구축

트래픽을 분석하기 위해 가장 먼저 모형 식별 통계량을 이용하여 잠정적인 모형을 선택한다. 그리고, 선택된 모형을 식별하기 위해 정상성 가정을 만족하는지 여부를 조사한다. 만약, 정상성 가정을 만족하지 못하는 경우에는 예측 모형으로 사용하기 부적합하다. 선택된 모형의 적합성을 진단하여 모형이 부적합한 경우에는 모형을 수정하여 만족스러운 모형이 선택될 때까지 반복한다. 이 모형 구축 절차를 그림 3에서 보인다.

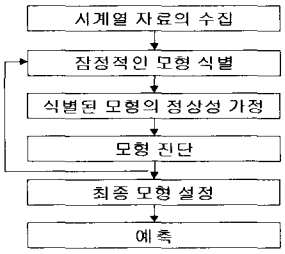


그림 3. 시계열 모형의 구축

정상성 가정을 수행하기 위해서는 자기상관함수와 편자기상관함수를 구하는데 정확한 값을 구하기 위해 통계학 분야에서 시계열 분석을 위해 많이 사용하는 SPSS 10.0.7을 사용한다. 이는 데이터로부터의 객관적인 결과를 얻기 위해서이다.

3.4 시계열 모형의 가정

예측 모형을 식별하기 위해 정상성 가정을 수행한다. 자기상관함수와 편자기상관함수의 결과가 가정을 만족하는 신뢰구간에서 유의한 값을 가지는지 여부를 조사한다. 정상성 가정을 만족하는 예측 모형이 발견된다면, 트래픽의 정확한 예측을 수행할 수 있다. 그림 4는 1년간 수집된 트래픽의 자기상관함수와 편자기상관함수의 그래프를 보이고 있다.

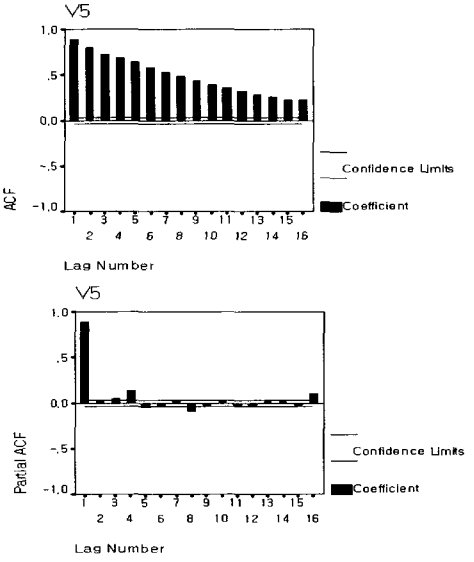


그림 4. 년간 트래픽의 자기상관함수와 편자기상관함수

위의 그림이 나타나는 바와 같이 자기상관함수와 편자기상관함수를 통해 시계열 모형을 식별할 수 없다는 것을 알 수 있다. 특정 시차까지 유의한 값을 나타내어야 모형을 식별할 수 있지만, 위 그래프는 특정 시차가 아니라 임의적인 시차에서 유의한 값이 나타나므로 정상성 가정을 만족하는 시계열 모형이 존재하지 않는다고 할 수 있다. 이는 과거에 측정된 값들간의 상관 관계를 이용하여 미래의 값을 예측함이 있어서, 적용 가능한 상관관계라고 볼 수 없다는 것을 의미한다. 즉, 과거의 상관관계는 현시점에서 미래에 발생하는 값과의 상관관계에 영향을 미치지 않는다는 것을 의미한다. 따라서, 연간 총 누적 트래픽을 근거로 트래픽을 예측할 경우에는 정상 시계열 자료라고 볼 수 없다. 자기 유사성에 근거한 예측은 정상 시계열 일 경우에만 가능하다. 과거 값들 간의 관계가 정확하게 반복되는 경우에 적용이 가능하다. 또한, 상관 관계를 정확히 반복적으로 적용하더라도 동일한 패턴의 반복을 적용함에 따라 필연적으로 발생하는 확률적 오차를 고려하지 않고는 정확한 예측 값을 구할 수 없다.

정상 시계열 모형을 구할 수 없다는 것은 정상성 가정, 즉 시간이 지남에 따라 평균과 분산이라는 확률적인 성질이 일정하지 않다는 것을 의미한다. 정상 시계열이 아닌 경우 시계열 자료를 차분 변환과 로그변환을 수행하게 된다. 차분 변환을 하는 이유는 평균을 일정하게 하기 위해서이고, 로그 변환을 수행하는 이유는 분산을 일정하게 하기 위해서이다. 그림 5는 차분 변환과 로그 변환을 수행한 결과 자기상관함수와 편자기상관함수를 구한 결과이다.

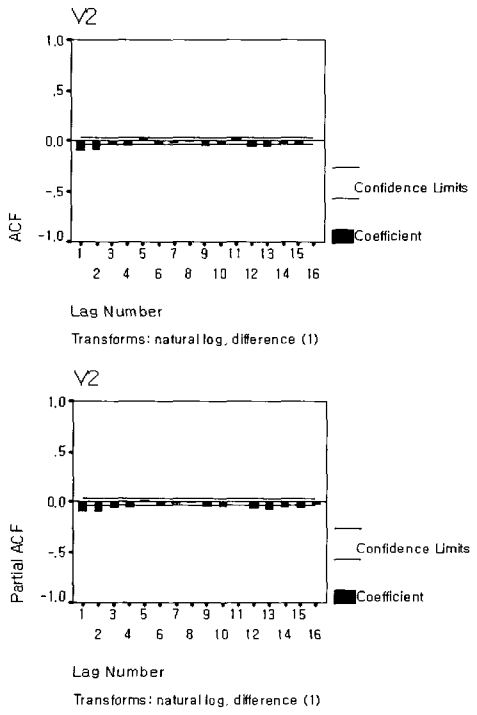


그림 5. 차분, 로그변환을 수행한 후의 자기상관함수와 편자기상관함수

위의 그림에서 보는 바와 같이 차분 변환과 로그 변환을 해 보았지만 유의한 해석을 얻을 수 없다. 따라서, 1년간 수집된 자료에서는 비정상 시계열 모형인 ARIMA 모형을 구할 수 없다. 따라서, 장기간에 수집된 데이터를 이용하여 트래픽을 예측하는 것은 합리적이지 않다는 것을 알 수 있다.

3.5 분류에 근거한 예측 모형의 수립

일년간 수집된 데이터를 이용하여 시계열 예측 모형을 구축하는 것은 사실상 어렵다는 것을 알 수 있다. 또한 6개월을 단위로 해서 위와 같이 자기상관함수와 편자기상관함수를 구해 보았으나, 마찬가지로 정상성 가정을 만족하는 시계열 모형을 식별할 수 없다. 따라서, 정상성 가정을 만족하는 모형을 구하려면 장기간이 아닌 단기간 자료를 활용해야 한다는 것을 알 수 있다. 그래서 월별, 주간별, 일별로 분류하여 분류된 데이터를 기반으로 예측 모형을 식별한다.

Ⅳ. 네트워크 트래픽의 예측 방법에 따른 검증

정확한 예측 모형을 수립하기 위해 트래픽의 손실된 부분을 제외하고 월별, 주별, 일별로 분류하여 이를 토대로 자기상관함수와 편자기상관함수를 구한다. 이 두 함수의 결과값을 보고 시계열 모형을 식별한다.

4.1 월별 분류에 의한 분석

(1) 분석기간

분석기간은 트래픽의 손실이 조금이라도 발생한 달은 제외하였다. 정전이나, 네트워크의 장애가 하루라도 발생한 달은 제외하여 2001년 11월, 2001년 12월, 2002년 3월, 2002년 7월에 해당하는 4 개의 데이터를 이용하여 자기상관함수와 편자기상관함수를 구한다.

(2) 분석 결과

모형의 식별을 위해 자기상관함수와 편자기상관함수를 이용한다. 자기상관함수와 편자기 상관함수에서 특정 시차까지 유의한 값을 가지는 모형을 찾아 모형으로 판단한다. 표 4는 월별 분류에 의한 분석 결과를 보이고 있다.

표 4 월별 분류에 의한 분석 결과

모형	AR	MA	ARMA	ARIMA	모형 식별 불가능
분석 결과	3	0	0	0	1

4번의 분석 항목 중에서 3번은 모형식별이 가능하고, 그 모형은 AR 모형을 따른다는 것을 확인할 수 있다. 그러나 2001년 12월 자료는 모형 식별이 불가능하다. 모형 식별이 불가능하다는 것은 정상성 가정을 만족하지 못하고, 정상성 가정을 만족하도록 자료 자체의 변환을 시도하였으나 변환된 자료로도 어떠한 모형 판단의 근거를 찾기가 어렵다고 볼 수 있다.

그림 6은 AR(1) 모형을 따르는 자기상관함수와 편자기상관함수의 예를 보이고 있다.

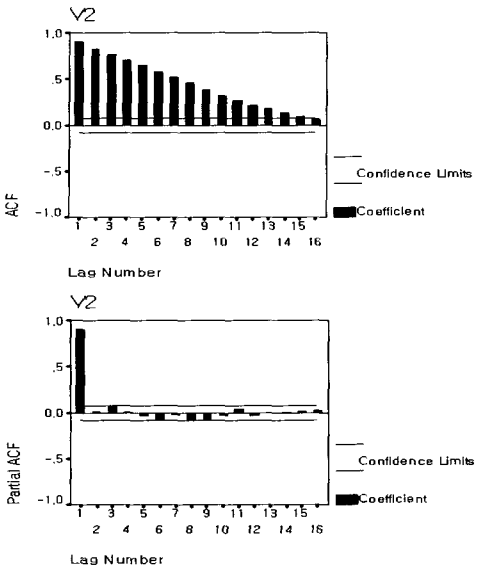


그림 6. 월별 분류 자료가 정상성 가정을 만족하는 경우

4.2 주간별 분류에 의한 분석

(1) 분석기간

주간별 분류에 의한 분석기간은 2001년 10월 28일부터 2002년 8월 3일까지이다. 데이터의 정확한 분석을 위해, 데이터 수집에서 한시간이라도 빠진 데이터가 있을 경우에는 하루 전체를 삭제하였고, 하루가 빠졌을 경우에는 일주일 전체를 삭제하였다. 따라서, 분석기간 중에, 1주일 데이터가 고스란히 남아 있는 경우에만 트래픽 분석을 수행한다. 22주의 데이터를 근거로 주별 분류에 의한 분석을 수행한다.

(2) 분석 결과

모형의 식별을 위해 자기상관함수와 편자기상관함수를 이용한다. 표 6은 분석 결과를 보이고 있다.

표 5. 주간별 분류에 의한 분석 결과

모형	AR	MA	ARMA	ARIMA	모형 식별 불가능
분석 결과	6	1	1	0	14

이 표에서 볼 수 있듯이 22번의 항목 중에서 14번은 모형 식별이 불가능하다는 것을 볼 수 있다. 그리고, AR모형은 6회, MA와 ARMA모형은 각 1회가 나타났다.

그림 7은 AR(1)을 따르는 모형의 자기상관함수와 편자기상관함수의 예를 보이고 있다. 자기상관함수에서는 지수적으로 감소하고, 편자기상관함수에서는 시차가 1인 경우에는 유의한 값을 가지고, 2이후에는 유의한 값을 가지지 않는다는 것을 볼 수 있다. 이 그림을 통해 시계열 자료가 AR(1) 모형을 따른다는 것을 확인할 수 있다.

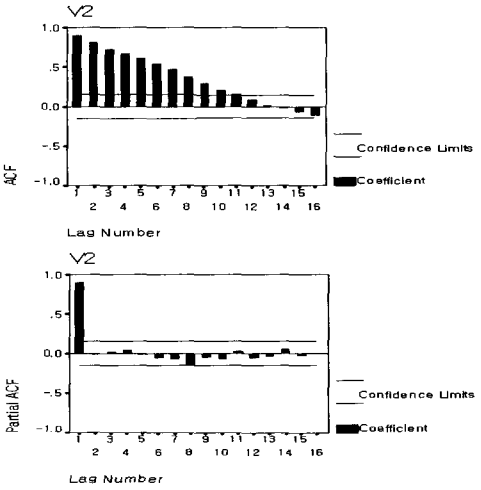


그림 7. 주간별 분류 자료가 정상성 가정을 만족하는 경우

4.3 일별 분류에 의한 분석

(1) 분석 기간

일별 분류에 의한 모형을 식별하기 위해 2001년 10월 28일부터 2001년 8월 3일까지의 데이터를 기준으로 분석을 수행한다. 일별 분류에 의해 분석하기 위해 수집된 데이터가 매우 방대하여, 360회 이상의 데이터를 분석하기에는 시간적인 어려움이 따라 150회로 제한한다. 주별 분석에 사용되었던 날짜를 근거로 주별 분석과 일별 분석의 상관관계를 확인하고자 주별 분석 일자와 동일한 날짜로 제한한다.

(2) 분석 결과

표 8은 일별 분류에 의한 자기상관함수와 편자기상관함수를 통해 모형 식별한 결과를 보이고 있다.

표 6. 일별 분류에 의한 분석 결과

모형	AR	MA	ARMA	ARIMA	모형 식별 불가능
분석 결과	105	0	17	16	12

이 표에서 알 수 있듯이 150회 분석 중에서 105회에 해당하는 시계열 자료가 AR 모형을 따른다는 것을 확인할 수 있다. ARMA 모형은 17회, ARIMA모형은 16회, 그리고 모형 식별이 불가능한 경우에는 12회로 나타난다. 이 표를 통해, 일별 분류에 의한 분석 결과는 70%가 AR 모형을 따른다고 볼 수 있다.

그림 8은 AR(1)모형을 따르는 시계열의 자기상관함수와 편자기상관함수의 예를 보이고 있다. 자기상관함수는 Sine 함수 형태로 감소하고 있고, 편자기상관함수는 시차가 1인 경우에는 유의한 값을 가지고, 2이후에는 유의하지 않는 값을 가지므로 AR(1) 모형이라는 것을 알 수 있다.

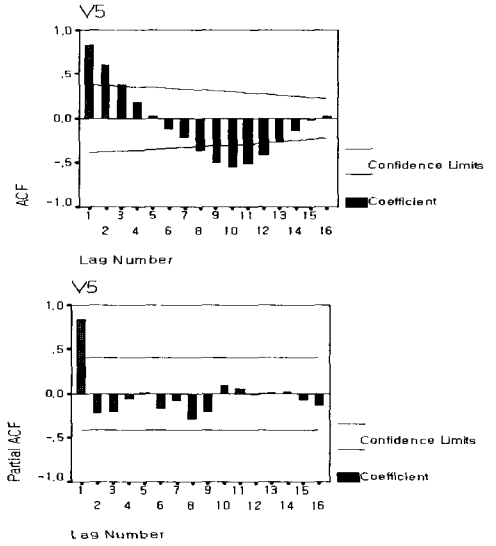


그림 8. 일별 분류 자료가 정상성 가정을 만족하는 경우

4.4 트래픽 분석 결과

네트워크 트래픽을 예측하기 위해 과거에 측정된 자료들간의 상관관계를 이용하여 예측을 수행하였다. 미래의 값을 예측하는 과정에서 과거 값들의 상관관계를 적용하면 확률적 오차를 포함하게 되는데, 이 확률적 오차를 많이 포함하면 예측값에 대한 신뢰도가 낮아진다. 본 논문에서는 통계학 분야에서 예측 방법으로 널리 사용되는 시계열 모형인 AR, MA, ARMA, ARIMA 모형을 사용하여 네트워크 트래픽을 예측하였다. 각 모형의 예측값이 정확한 예측값으로 신뢰할 수 있는지에 대한 적합성을 검증함으로써, 정확한 예측값을 구할 수 있다는 것을 알 수 있다. 적합성 검증은 모형 식별 단계에서 초기 단계인 정상성 가정을 만족하는지의 여부로 판단하였으며, 정상성 가정은 자기상관함수와 편자기상관함수를 통해 구할 수 있었다. 정상성 가정을 만족하지 못하는 모형인 경우를 비정상 시계열 자료로 분류하였고, 이 경우에 예측은 정확하다고 볼 수 없다.

전체 트래픽을 분석해 본 결과 비정상 시계열로 분류되었으며, 이 경우 예측값의 신빙성은 보장할 수 없었다. 따라서, 기간을 조정함으로써 정상시계열을 얻을 수 있었으며, 월별, 주별, 일별로 분류해 보았다. 기간별 분류에 의해 예측 모형을 식별해 본 결과, 4개의 월별 항목 중에서 3번은 AR모형으로 나타났고, 1회는 모형을 식별할 수 없었다. 선정된 항목이 적어 월별로 분류하면 AR 모형으로 나타난다고 확언할 수는 없지만, 일반적으로 월 단위 회귀형태를 가진다고 볼 수 있다. 주별 분류는 22개의 항목 중에서 14개의 항목이 모형을 식별할 수 없었으며 63%이상이 예측 모형을 찾을 수 없다는 결론을 얻었다. 이는 주간별로 수집된 데이터에서는 시계열의 상관성을 찾을 수 없다고 볼 수 있다. 일별 분류는 150개의 항목 중에서 105개의 모형이 AR 모형을 따른다. 일별 분류 항목 중에서 70% 정도가 AR 모형을 따르므로, 일별 분류에 의한 예측 결과는 시계열 자료의 회귀성을 따른다고 볼 수 있다.

이 자료를 근거로 예측 모형을 식별하기 위해서는 일별 분류가 가장 적합하며, 예측 모형은 AR 모형이 많이 나타난다고 볼 수 있다. 실제 트래픽의 예측을 위해서는 시간별로 24개의 항목을 갖도록 시계열 자료를 구성하여 네트워크 트래픽을 예측하는 것이 효과적이라고 볼 수 있다. 따라서, 정확한 예측을 수행할 수 있도록 시계열 자료를 분류하여 정상성 가정을 만족하는 모형으로 일 단위로 분류하는 방법이 가장 효과적이라는 것을 살펴 볼 수 있었다.

V. 결 론

본 연구에서는 네트워크 트래픽을 측정하여 트래픽의 예측값을 구할 수 있는 알고리즘의 적합성 검증에 대해 살펴보았다. 자기상관함수와 편자기상관함수를 이용하여 시계열 모형을 검증함으로써 예측 모형을 이용한 네트워크 트래픽의 예측이 가능하다고

볼 수 있다. 수집된 시계열 자료가 모든 시계열 예측 모형에 적합하다고 볼 수 없으며, 누적된 데이터가 많아질수록 예측 모형을 찾아내기가 어렵다는 것을 알 수 있다. 또한, 날짜별 분류에 의한 분석 방법을 이용할 경우 비교적 효과적인 예측 방법이 된다는 것을 확인하였다. AR 모형의 특징인 회귀적인 형태를 가질 경우 용이하다는 것에서 알 수 있듯이, 네트워크 트래픽 자료는 24시간을 기준으로 회귀 형태를 갖는다는 것을 알 수 있다.

현재 네트워크 관리는 트래픽을 감시하여 관리자에게 트래픽의 양을 공지하는데 그치고 있다. 트래픽의 양을 보고 라우팅 경로를 적절히 분산하기 위해서는 관리자가 강제로 라우팅 경로를 변경해야 한다. 관리자가 라우팅 경로를 분산하기 위해서는 라우팅 테이블을 수정함으로써 가능하다. 네트워크 관리 시스템과 라우터간에 SNMP 통신을 수행하여 라우팅 테이블을 갱신할 수 있다. 따라서, 본 논문에서 제안하는 트래픽 수집 및 예측 시스템이 각 라우터와 SNMP 통신을 수행하여 트래픽의 양을 예측하고, 라우팅 테이블의 MIB 정보를 갱신함으로써, 라우팅 경로를 분산할 수 있다. 라우팅 경로의 부하분산을 통해 네트워크 전체의 효율성을 크게 향상시킬 수 있다.

예측 모형인 AR, MA, ARMA모형을 사용하기 위해서는 모수를 추정해야 하는데, 모수 추정은 조건부 최소제곱으로 구할 수 있다. 현재 예측 모형을 구현하여 트래픽을 예측하고 있으며, 향 후 연구로는 예측 결과를 이용하여 라우팅 테이블을 갱신하도록 라우터와 예측 시스템간의 SNMP 통신 모듈을 구현하고자 한다.

참 고 문 헌

[1] Yantai Shu, Zhigang Jin, Lianfang Zhang, Lei Wang, "Traffic Prediction Using FARIMA Models", IEEE International Conference on Communications, pp.891-895. 1999,6.

[2] 홍원택, 안성진, 정진욱, "시계열 분석을 이용한 SNMP MIB-II 기반의 회선 이용률 예측 기법", 한국정보처리학회 논문지 제6권 제9호,

[3] 안성진, 정진욱, "SNMP MIB-II를 이용한 인터네트 분석 파라미터계산 알고리즘에 관한 연구", 한국정보처리학회 논문지 제5권 제8호, pp.2102-2116, 1998,8.

[4] 김동수, 정태명, "실시간 네트워크 관리를 위한 SNMP 확장에 관한 연구", 한국정보처리학회 논문지 제6권 제2호, pp.449-458, 1999,2.

[5] 이강원, 김태윤, "효율적인 통신망 설계를 위한 예측 시스템 설계", 한국정보과학회 논문지, 제25권 제1호, pp.76-82, 1998,1.

[6] 이덕기, "예측방법의 이해", SPSS 아카데미, 1999.

[7] 정동빈, 원태연, "SPSS를 활용한 시계열 자료와 단순화 분석", SPSS 아카데미, 2001.

[8] 최기헌, 이종협, "SAS/ETS를 이용한 시계열 분석과 그 응용", 자유아카데미, 1994.

[9] William Stallings, "SNMP, SNMPv2, SNMPv3, and RMON 1 and 2", Addison Wesley, 1999.

[10] Mark A. Miller, "Managing Internetworks SNM

[11] Kyung Hyu Lee, "An Agent-Manager Scheme P", M&T books, 1998. for the Integrated Transfort Network Management", IEEE International Conference on Communications, pp.1017-1021, 1999,6.

[12] 김동일, 김창호, "Ethernet 트래픽의 장기간 의존성 및 Self-Similiar 트래픽 소스 모델링에 관한 연구", Telecommunication Review, 제11권 6호, 2001, 11~12월.

[13] Wilinger,W., Wilson,D., Taqqu, M. "Self-similar Traffic Modeling for Highspeed Networks", ConneXions, Nov. 1994.

[14] 김창호, 김동일외, "실시간 운영중인 네트웍의 트래픽 특성과 성능 분석에 관한 연구", 추계한국해양통신학회 Proc., pp358-213, 1998.

[15] 김창호, 김동일외, "데이터 트래픽에서의 Self-similar 특성", 춘계해양정보통신학회 Proc., p146-151, 1999.

[16] 정상준, 권영헌, 최혁수, 이정협, 김종근, "실시간 망 관리를 위한 패킷 분석 시스템의 설계 및 구현", 한국멀티미디어학회 춘계학술발표대회 논문집. 2001,5.

[17] 정상준, 권영헌, 최혁수, 김종근, "시계열 분석을 이용한 실시간 네트워크 트래픽 예측 시스템의 설계", 한국정보처리학회 추계학술발표대회 논문집. 2001,10.

[18] 정상준, 최혁수, 권영헌, 임인택, 권은영, 김종근, "실시간 네트워크 트래픽의 예측을 이용한 성능관리 시스템 연구", 한국정보처리학회 춘계학술발표대회 논문집. 2002,4.

정 상 준(Sangjoon Jung) 정회원

1999년 2월 : 영남대학교 통계학과 이학사
2001년 2월 : 영남대학교 컴퓨터공학과 석사
2003년 2월 : 영남대학교 컴퓨터공학과 박사수료
2003년 3월 ~ 현재 : 경일대학교 교양학부 강의전담강사

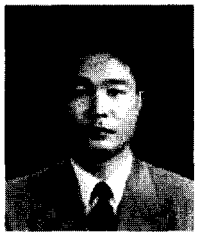
<관심분야> 네트워크 관리, 통신망 성능 분석, 원격교육

김 동 주(Dongju Kim) 정회원

2002년 2월 : 가야대학교 컴퓨터공학과 학사
2002년 ~ 현재 : 영남대학교 대학원 컴퓨터공학과(석사과정)

<관심분야> 네트워크 관리 및 네트워크 보안

권 영 헌(Younghun Kwon) 정회원



1988년 2월 : 부산외국어대학
교 전자계산학과 학사
1990년 2월 : 동아대학교 전자공학과 석사
1998년 2월 : 영남대학교 컴퓨터공학과 박사수료
1998년 3월 ~ 현재 : 세경대학
컴퓨터정보통신과 조교수

<관심분야> 초고속통신망, 통신망 성능분석, 네트워크/트래픽 제어

김 종 근(Chonggun Kim) 정회원

1981년 2월 : 영남대학교 전자공학과 학사
1987년 2월 : 영남대학교 전자공학과 석사
1991년 2월 : (일본)東京電氣通信大學 박사
1996년6월 ~ 1997년7월 : (미국) Virginaia Tech. 방문교수
현재 : 영남대학교 컴퓨터공학전공 교수

<관심분야> 컴퓨터 네트워크, 분산처리, 인터넷응용, 멀티미디어기반 가상강의 시스템