TD(λ) 기법을 사용한 지역적이며 적응적인 QoS 라우팅 기법

정회원 한 정 수*

A Localized Adaptive QoS Routing using TD(λ) method

Jeong-Soo Han* ***Regular Members***

요 약

본 논문에서는 TD(temporal differences) 기법을 사용한 localized QoS 라우팅 기법을 제안하였다. 이 기법은 이웃노드로부터 얻어지는 성공 기댓값을 통해 라우팅 정책을 결정하는 기법이다. 이에 본 논문에서는 라우팅 성공 기댓값을 기반으로 한 다양한 탐색기법으로 경로 선택 시 라우팅 성능을 비교 평가하였으며, 특히 Exploration Bonus를 적용한 탐색 기법이 다른 탐색 기법에 비해 더욱 우수한 성능을 보여주고 있는데, 이는 다른 탐색 기법에 비해 네트워크 상황에 더적응적으로 경로를 선택할 수 있기 때문이다.

Key Words : Localized QoS Routing, Reinforcement Learning, Q-Learning, Temporal Differences, Exploration method

ABSTRACT

In this paper, we propose a localized Adaptive QoS Routing using TD method and evaluate performance of various exploration methods when path is selected. Expecially, through extensive simulation, the proposed routing algorithm and exploration method using Exploration Bonus are shown to be effective in significantly reducing the overall blocking probability, when compared to the other path selection method(exploration method), because the proposed exploration method is more adaptive to network environments than others when path is selected.

Ⅰ. 서 론

Global QoS 라우팅 기법과 비교하여 localized QoS 라우팅 기법이 보다 안정적이고, 간단하며, 네트워크 상황에 보다 적응적이라는 것이 제시되어 왔으며[1], [2][3]에서는 proportional sticky routing (psr)이라고 하는 localized QoS 라우팅 기법이 제안되었다. 이러한 psr 기법은 허용 가능한 최대 데이터 blocking 파라미터를 사용하여 각 사이클 당 경로를 따라 전송하게 될 flow의 양을 제어함과 동시에 각 경로에 할당하게 될 데이터 비율을 재조정함으로써 자체 적응력을 유지하는 기법을 제공하고 있다. 그러나 여기에는 세 가지 문제점을 가지고 있다. 첫째, 실제 네트워크에서 트래픽 패턴을 항상 알 수 있는 것은 아니다. 두 번째로, blocking 확률을 계산하기 위해 경로 상의 정확한 정보를 알아야 한다는 것이다. 마지막으로, 비록 데이터의 패턴과 blocking 확률에 대한 계산이 가능할 지라도 경로 전체에 대한 최적화 문제(global optimization problem)를 해결하기 위해 소요되는 시간은 상당히 크다고 할 수 있다[4].

이러한 문제를 해결하기 위해 [4]에서는 전체 네트워크에 대한 정보나 네트워크의 트래픽 패턴을 알지 못해도 가능한 Q-Learning 기반의 경로 선택

* 신구대학 인터넷정보과(jshan@shingu.ac.kr)
논문번호 : KICS2004-11-261 접수일자 : 2004년 11월 4일

기법을 제안했다.

강화학습(Reinforcement Learning) 기반의 라우팅 기법에서는 경로 선택이 전체 blocking 확률에 영향을 미치는 서로 다른 소스 노드에서 서로 다른 시간에 이루어지기 때문에 temporal credit 할당이 필요하게 된다. 이러한 temporal credit 할당 문제는 일반적으로 TD 기반의 알고리즘을 사용하여 해결되고 있다. 이러한 알고리즘 중에서 하나가 TD(λ) 기법을 사용하는 것이다[5]. 또한 경로 선택 시 사용되는 방법으로 탐색(exploration)과 이용(exploitation) 문제에 대한 다양한 기법들이 경험적으로 제시되고 있다[6][7].

본 논문에서, 우리는 TD 알고리즘을 사용한 TD(λ)-라우팅 기법과 경로 선택을 위한 다양한 탐색 기법들 간의 성능을 평가하고자 한다. 특히 탐색 기법 중 *Exploration Bonus*를 적용한 기법을 통해 많이 사용하지 않거나 사용한지 오래된 경로에 더 많은 탐색 기회를 제공함으로써 전체적인 네트워크 성능을 향상시킬 수 있는 방법을 제안하고자 한다.

본 논문의 구성으로는 2장에서 강화학습의 기본적인 지식과 localized QoS 라우팅 적용시 문제점을 제시하고 있다. 이에 대한 해결책으로 3장에서는 본 논문에서 제안하는 TD 라우팅 기법과 라우팅 정보 갱신 규칙 그리고 다양한 경로 선택 방법들(ϵ-greedy-policy, Boltzmann Exploration, Exploration Bonus)을 설명하고 있다. 이들에 대한 성능 평가가 4장에서 제시되고 있으며, 5장에서 결론을 맺는다.

Ⅱ. 배경지식

[그림 1]에서 보는 바와 같이 강화학습에서 에이전트(agent)는 행동(action)을 통해 주위 환경과 연결되어진다. 각 상호작용(interaction)에서, 에이전트는 현재 상태 s에서 입력 i를 받게 되고, 에이전트는 다시 출력으로 행동 a를 선택하게 된다. 이렇게 선택된 행동은 상태를 변경시키며, 이때 발생된 상태 전이(state transition) 값이 강화 값(reinforcement value) r를 통해 에이전트와 통신하게 된다. 일반적으로 강화학습 상에서 에이전트의 목적은 할인된 총 강화 값의 합의 평균(the expected total discounted sum of reinforcement value)($E\left[\sum_{t=0}^{\infty}\gamma^t r_t\right]$)이 최대가 되도록 하는 것이다[8].

여기서 r_t는 t단계에서 받은 강화 값을 나타내며, $0 \le \gamma \ge 1$은 할인율을 나타내는데, 이것은 단기적인 보상과 장기적인 보상사이의 상대적인 차이를 조정하기 위해 사용된다. 이러한 모델을 infinite-horizon discounted model이라 한다[6].

Localized 라우팅 기법에서 네트워크 상의 각 노드는 이웃 노드로부터 얻어진 네트워크 상태 정보와 같은 오직 지역적인 정보만을 기반으로 라우팅 결정을 수행하게 된다. 또한, 라우팅 정책의 목적은 지역 정보만을 사용하여 전체적으로 평균 blocking 확률을 최소화시키는 것이다. 그러나 만약 라우팅 기능을 수행하기 위해 단지 지역적인 정보만을 사용한다면 하나의 노드에서 결정된 라우팅 결정이 네트워크 전체 성능에 어떻게 영향을 끼치게 될지 결정하는 것은 매우 어려운 일이다. 이러한 문제 때문에 강화학습에서의 temporal credit 할당 문제가 발생되게 된다. 이러한 문제는 일반적으로 TD 기법을 사용한 알고리즘을 통해 해결하고 있다.

위에서 설명한 infinite-horizon discounted model 기반의 TD 기법을 사용한 알고리즘에는 TD(λ) 기법과 Q-Learning 기법 등이 있는데, 본 논문에서는 TD(λ) 기법을 사용한 localized QoS 라우팅 기법을 제안하고자 한다.

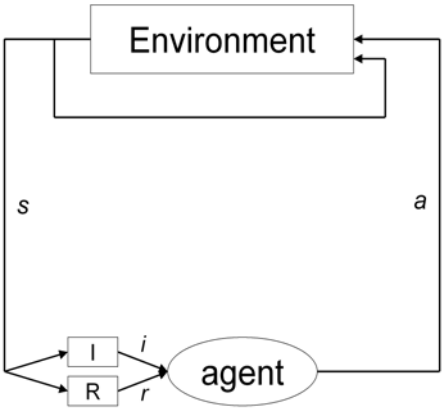

그림 1. 일반적인 강화학습 모델

Ⅲ. 제안 알고리즘

본 논문에서 제안한 TD(λ)-라우팅 기법은 TD(λ) 기법에 기반 한 지역적이고 적응적인 라우팅 알고리즘이다.

3.1 TD(λ) 라우팅

본 논문에서 제안한 라우팅 모델을 사용하기 위해 [표 1]에서와 같이 네트워크 라우팅 항목과 TD(λ)

기법 상의 항목에 대한 상호 연결이 필요하다.

요청을 받은 현재 상태 s인 노드는 $V_d^{t,n}(s,a)$ 값을 가지고 있다. 해당 노드로 도착한 요청은 미리 정해진 경로를 통해 전송하게 되고, 모든 요청에 대한 이러한 값들은 일반적인 값 함수를 갱신하기 위해 누적이 되고 이용될 것이다. 각 갱신 단계에 대해, 계속된 상태들 사이에서 (1)과 같은 TD 기법에 따라 차이(difference)인 $\delta_{s,a}^{t,n}$ 값이 얻어진다.

$$\delta_{s,a}^{t,n} = r_s(a) + \gamma V_d^{t,n}(s',a') - V_d^{t,n}(s,a) \qquad (1)$$

여기서, γ은 감소율로 알려진 학습 상수이다. $V_d^{t,n}(s',a')$이 다음 상태 s'에서 발생된 실질적인 측정값이기 때문에, 상태 s가 방문될 때마다 그 측정값은 $r_s(a)+\gamma V_d^{t,n}(s',a')$에 가깝게 갱신된다.

표 1. 기네트워크 라우팅 항목과 TD(λ) 기법과의 비교

네트워크 라우팅 항목	TD(λ) 기법 항목
각 네트워크 노드	에이전트
목적지로의 경로들 집합	행동들($a \in A$)
요청을 수신한 노드	상태(s)
선택된 경로 상에 요청에 전송되었을 경우 얻게 되는 값	강화값($r_s(a)$, 1:수락, 0: 거절)
소스 노드에서 목적지 d로 전송하기 위해 해당 경로로 전송할 때 n번 시도에 따른 t회 성공 시 기대되는 값	값 함수($V_d^{t,n}(s,a)$)
지역 라우팅 정책	최적화 정책($V^*(s)$)

3.2 라우팅 정보 갱신 규칙

요청에 대한 연결 결과에 따라, 노드 s상에서 s'로의 라우팅 정보를 갱신하는 것이 필요하다. 값 함수(value function)를 갱신하기 위해서는 (1)의 $\delta_{s,a}^{t,n}$과 함께 다음과 같은 수식이 적용된다.

$$V_d^{t,n}(s,a) = V_d^{t,n}(s,a) + \alpha\delta_{s,d}^{t,n} e_d(s,a) \qquad (2)$$

여기서, α는 학습율이며, $e_d(s,d)$는 에이전트가 탐색 과정에서 선택한 (상태-행동)쌍이 얼마나 좋은가에 대한 평가를 나타내는 적합도(eligibility trace)이다. 최적화 정책($V^*(s)$)와 적합도($e_d(s,d)$)는 다음과 같이 수행된다.

$$V^*(s) = argmax_{a \in A}\left\{V_d^{t,n}(s,a)\right\} \qquad (3)$$

$$e_d(s,a) = \begin{cases} \gamma\lambda e_d(s,a)+1 & \text{if } V_d^{t,n}(s,a) = V^*(s) \quad or \quad elected \\ \gamma\lambda e_d(s,a) & otherwise \end{cases} \qquad (4)$$

여기서, $\gamma\lambda(0<\gamma>1, 0<\lambda>1)$는 감소율(decay factor)을 의미하며, 만일, 현재 상태에서 선택 가능한 (상태-행동)쌍들 중에서 가장 큰 $V_d^{t,n}(s,a)$ 값을 갖는 (상태-행동) 쌍을 선택한 경우 적합도를 이전 상태에서 선택한(상태-행동) 쌍에 대한 적합도보다 1만큼 증가시키고, 그렇지 않은 경우 적합도를 $\gamma\lambda$씩 감소시키는 역할을 한다.

3.3 경로 선택 방법

강화학습에서는 상태에서 많은 행동 중에 최적의 보상을 얻을 수 있는 행동을 선택하는 방법을 TD(λ)-라우팅 상에서 전송하게 될 요청에 대한 경로 선택에 적용하기로 한다. 강화학습에서는 ϵ-greedy-policy과 Boltzmann Exploration 방식을 주로 사용한다[6]. 첫 번째 방식인 ϵ-greedy-policy는 항상 가장 높은 기대 보상값을 갖는 행동을 선택하는 것으로 (3)과 (4)를 사용할 수 있다. 또 다른 방식인 Boltzmann Exploration은 확률 p를 가지고 가장 높은 기대 보상값을 갖는 행동을 랜덤하게 선택하는 방식이다. 이러한 경우에 (4)와 함께 값 함수($V_d^{t,n}(s,a)$)가 확률 분포에 따라 확률적으로 행동을 선택하기 위해 사용된다.

$$P(a|s) = \frac{e^{\frac{V_d^{t,n}(s,a)}{T}}}{\sum_{b \in A} e^{\frac{V_d^{t,n}(s,b)}{T}}} \qquad (5)$$

(5)에서 T는 행동 선택의 임의성(randomness) 정도를 제어하는 온도 변수(temperature variable)로서 T값에 따라 그 값이 작으면 이용을, 값이 크면 탐색을 수행하게 된다[6].

이러한 기법들의 문제는 초기의 계속된 경로 선택에 의해 누적된 $V_d^{t,n}(s,a)$ 값과 다른 경로의 값과의 차이를 보상하기 위한 시간이 소요된다는 점이다.

마지막으로 Sutton[8]이 제안한 Exploration Bonus 기법을 적용하는 것이다. 이 방식은 선택된 횟수가 적거나 오래 전에 선택된 행동에 보너스를 적용하는 것이다. 본 논문에서는 예측 오차가 큰 상태에 높은 우선순위를 제공하는 방식으로 다음과 같은

방식으로 Exploration Bonus인 $B(s)$를 계산하는 방식을 제안한다.

① 먼저 현재 상태를 저장한다 : $V_{old} = V(s,a)$.
② (2)를 사용하여 상태 값 $V(s,a)$를 갱신한다.
③ 갱신 전 상태와의 오차를 계산한다:
$\Delta_a^s = \left| V_{old} - V(s,a) \right|$
④ 오차가 최대인 행동 b를 선택한다 :
$B(s) = max_{b \in A}\{\Delta_b^s\}$

이 기법에서 만약 잘못된 선택일 경우 다른 행동들에게 계속해서 적용하게 된다. 이 방식은 잘못된 경로 선택에 대한 보상 시간 없이 즉시 다른 경로를 선택할 수 있다는 장점을 가지고 있다.

[표 2]는 본 논문에서 제안한 TD(λ)-라우팅 알고리즘과 세 가지 경로 선택 방식을 설명하고 있다.

Ⅳ. 성능 평가

[그림 2]는 본 논문에서 제안한 TD(λ)-라우팅 기법의 성능 평가를 위해 사용된 네트워크 토폴로지들을 보여주고 있다. 여기서 사용하는 성능 평가 환경은 [1][2][3]에서 사용된 시뮬레이션 환경을 그대로 사용하고 있다. 따라서 다음과 같은 가정을 사용한다. 먼저 모든 회선은 무방향성이고, 각 방향으로 똑같은 C unit의 대역폭을 갖는다. 네트워크에 도착한 연결 요청은 1 unit 대역폭을 요구한다고 가정하자. 연결 요청은 소스 노드에 λ를 갖는 포아송 프로세스를 따르며, 목적지는 소스 노드를 제외한 모든 노드로부터 랜덤하게 선택된다. 연결 요청에 대한 지속시간은 $1/\mu$를 갖는 지수 분포를 따른다. 네트워크 부하는 $\rho = \lambda N h / \mu L C$를 정의한다. 여기서 N은 소스 노드의 총수이며, L은 회선의 총수, h는 평균적으로 모드 소스-목적지 쌍에서 연결 요청 당 평균 홉 수를 나타낸다. 또한 시뮬레이션에서 사용된 파라미터들을 C=20, μ=60 sec으로 정의한다. 소스 노드 상의 평균 도착율 λ은 네트워크 부하에 의존한다.

TD(λ)-라우팅 기법에서 사용된 파라미터의 가정으로는 먼저 감소율 γ=0.9이며, 학습율 α=0.8 그리고 온도변수 T=0.1로 정한다. 또한 연결 요청을 전송할 사용 가능한 경로(feasible path)는 [1]에서와 같이 주 경로(primary path)와 보조 경로(alternative

표 2. 제안한 TD(λ) 라우팅 알고리즘

step 1. 네트워크 상의 모든 상태와 행동들에 대해 $V(s,a) = 0,\ e(s,a) = 0$로 초기화
step 2. 초기 상태(s) 선택
Repeat {
  step 3. 경로 선택 방식에 따라 행동(a)와 다음 상태(s')를 선택
    3.1 ϵ-greedy-policy : $V_d^{t,n}(s,a)$ 값이 최대인 행동 선택
    $V^*(s) = argmax_{a \in A}\{V_d^{t,n}(s,a)\}$
    3.2 Boltzmann Exploration : 확률 p를 갖는 행동을 랜덤하게 선택
    $P(a|s) = \frac{e^{\frac{V_d^{t,n}(s,a)}{T}}}{\sum_{b \in A} e^{\frac{V_d^{t,n}(s,b)}{T}}}$
    3.3 Exploration Bonus : $B(s)$인 행동을 선택
    - $V_{old} = V(s,a)$
    - $V(s,a)$ 갱신
    - $\Delta_a^s = \left| V_{old} - V(s,a) \right|$
    - $B(s) = max_{b \in A}\{\Delta_b^s\}$
  step 4. 연결 요청에 대한 step 3 선택에 대한 강화값($r_s(a)$) 획득
    $r_t(a) = \begin{cases} 1 & \text{if } request\ is\ accepted \\ 0 & \text{if } request\ is\ rejected \end{cases}$
  step 5. 선택된 다음 상태(s')와의 차이값 $\delta_{s,a}^{t,n}$과 적합도 $e_d(s,a)$계산
    $\delta_{s,a}^{t,n} = r_s(a) + \gamma V_d^{t,n}(s',a') - V_d^{t,n}(s,a)$
    $e_d(s,a) = \gamma\lambda e_d(s,a) + 1$
  step 6. 현재 상태(s) 라우팅 정보 갱신
    $V_d^{t,n}(s,a) = V_d^{t,n}(s,a) + \alpha\delta_{s,d}^{t,n} e_d(s,a)$
  step 7. 선택되지 못한 다른 행동에 대한 적합도 적용
    $e_d(s,a) = \gamma\lambda e_d(s,a)$
} Until 최종 상태(d) 도착

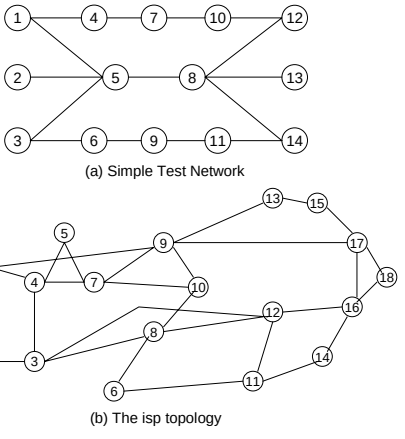

그림 2. 시뮬레이션 네트워크

path)를 사용하였다. 마지막으로 Exploration Bonus 기법은 초기에 ϵ-greedy-policy 기법을 사용하다가 연결 요청에 대한 blocking이 발생했을 때 적용하게 된다.

[그림 2]의 (a)와 같은 작은 네트워크에서는 세 가지 경로 선택방식이 거의 비슷한 성능을 보이고 있다. [그림 3]에서 보는 바와 같이 초기 50% 이상의 blocking 확률은 모든 소스 노드에서 주 경로를 사용하여 발생하게 되는 현상이며, 차츰 시간이 지남에 따라 Exploration Bonus기법이 다른 기법에 비해 3%의 성능향상(잘못된 경로에 대한 $V_d^{t,n}(s,a)$ 보상의 불필요)과 함께 모두 그 수치가 10%대로 현저하게 감소되는 것(다른 주 경로 또는 보조 경로 선택을 통한 감소)을 볼 수 있다.

[그림 2]의 (b)와 같은 ISP 네트워크의 성능은 [그림 4]와 같이 그 차이가 현저하다. [그림 2]의 (a)와 비교하여 더 많은 사용가능한 경로들을 가지고 있으나, 서로 다른 소스 노드들에서 연결 요청들이 발생되어 초기 blocking 확률은 60% 이상이 되고 있다. 그러나 세 가지 방식이 각각 적용되면 새로운 경로를 찾게 되어 20%대로 떨어지게 된다. 하지만 Exploration Bonus 방식은 그 값이 일정하게 유지되는데 반해, 다른 두 가지 방식은 파형형태의 모양을 갖는다. 파형의 위 부분은 현재의 잘못된 경로에 대한 $V_d^{t,n}(s,a)$ 값과 찾고자 하는 새로운 경로에 대한 $V_d^{t,n}(s,a')$ 값의 차이에 의해 발생되는 현상이며, 이는 확률적으로 경로를 선택하는 Boltzmann Exploration 방식보다 ϵ-greedy-policy 방식이 더 오래 지속되는 현상을 볼 수 있다.

[그림 3]과 [그림 4]에서 보는 바와 같이 네트워크 크기와 소스 노드의 수에 상관없이 Exploration Bonus 기법이 다른 경로의 빠른 선택을 통해 전체적인 성능이 우수함으로 알 수 있다.

[그림 5]는 [그림 2]의 (b) 네트워크 상에서 네트워크 부하에 따른 성능을 보여주고 있다. 앞서 살펴본 바와 같이 ϵ-greedy-policy 방식과 Boltzmann Exploration 방식은 비슷한 성능을 보여주고 있다. 이는 경로 선택 시 가장 높은 $V_d^{t,n}(s,a)$ 값에 많이 의존하기 때문이다. 이에 비해 Exploration Bonus 방식은 네트워크 부하가 늘어날수록 훨씬 좋은 성능을 보여주고 있다. 이는 경로 선택 방식이 네트워크 상태에 훨씬 더 적응적으로 반응할 수 있기 때문이다.

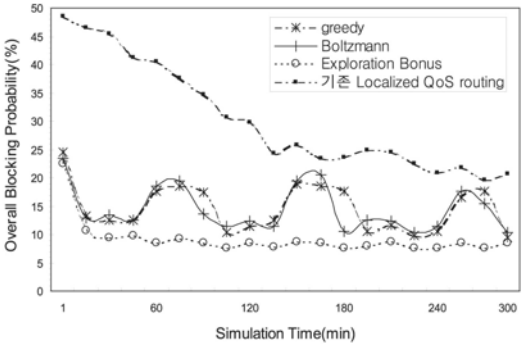

그림 3. 그림 2의 (a)네트워크에 대한 세 가지 경로선택 방식과 기존의 Localized QoS routing 기법에 대한 전체 blocking 확률

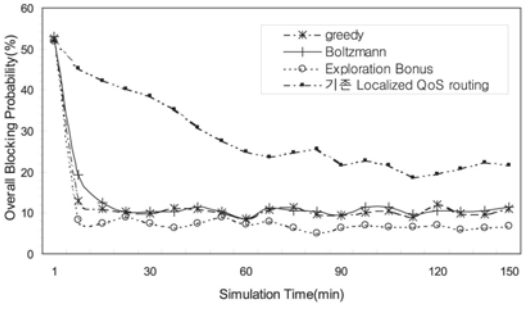

그림 4. 그림 2의 (b)네트워크에 대한 세 가지 경로선택 방식과 기존의 Localized QoS routing 기법에 대한 전체 blocking 확률

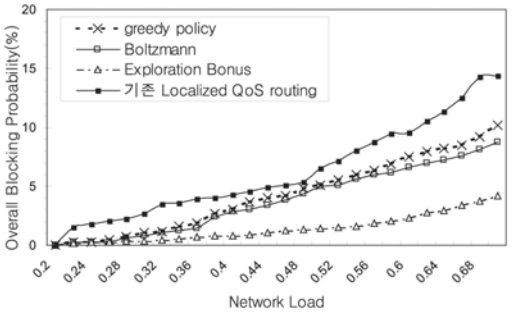

그림 5. 그림 2의 (b)네트워크에 대한 세 가지 경로선택 방식과 기존의 Localized QoS routing 기법에 대한 네트워크 부하에 따른 전체 blocking 확률

Ⅵ. 결 론

본 논문에서는 TD-기법을 사용한 Localized Adaptive QoS 라우팅 방식과 경로 선택시 Exploration Bonus 기법을 적용한 방식을 제안하였다. 이는 소스 노드상의 지역 정보만을 이용하여 QoS 라우팅 방식을 제공하며, 네트워크 부하에 따른 더 적응적인 경로 선택 기법을 제공하고 있다. 특히, 시뮬

레이션 결과를 통해 ϵ-greedy-policy 방식과 Boltzmann Exploration 기법은 $V_d^{t,n}(s,a)$ 값에 강하게 의존하기 때문에 경로 선택 시 비효율적이지만 Exploration Bonus 기법은 그 값 간의 오차에 따른 선택이므로 훨씬 더 네트워크 상황에 적응성이 좋다는 것을 알 수 있었다

참 고 문 헌

[1] X.Yuan and A.Saifee, “Path Selection Methods for Localized Quality of Service Routing”, Technical Report, TR-010801, Dept of Computer Science, Florida State University, July, 2001

[2] Srihari Nelakuditi, Zhi-Li Zhang and Rose P.Tsang, “Adaptive Proportional Routing: A Localized QoS Routing Approach”, In IEEE Infocom, April 2000.

[3] Srihari Nelakuditi, Zhi-Li Zhang, “A Localized Adaptive Proportioning Approach to QoS Routing”, IEEE Communications Magazine, June 2002

[4] Y.Liu, C.K. Tham and TCK. Hui, “MAPS: A Localized and Distributed Adaptive Path Selection in MPLS Networks” in Proceedings of 2003 IEEE Workshop on High Performance Switching and Routing, Torino, Italy, June 2003, pp.24-28

[5] Yvn Tpac Valdivia, Marley M, Vellasco, Marco A. Pacheco “An Adaptive Network Routing Strategy with Temporal Differences”, Inteligencia Artificial, Revista Lbero-americana de Inteligencia Aritificial, No 12(2001), pp. 85-91

[6] Leslie Pack Kaelbling, Michael L. Littman, Andrew W.Moore, “Reinforcement Learning: A Survey”, Journal of Artificial Intelligence Research 4, 1996, pp 237-285

[7] P.Marbach, O.Mihatsch, and J.N.Tsitsiklis, “Call Admission Control and Routing in Integrated Service Networks Using Neuro-Dynamic Programming”, IEEE Journal on Selected Areas in Communications, Vol. 18, No.2, Feb 2000, pp.197-208

[8] Sutton, R.S. “Learning to predict by the method of temporal differences” Machine Learning 3. 1988, pp.9-44

한 정 수 (Jeong-Soo Han) 정회원

1997년 2월 성균관대학교 정보공학과 졸업

1999년 2월 성균관대학교 전기전자및컴퓨터공학부 석사

2003년 2월 성균관대학교 전기전자및컴퓨터공학부 박사

현재 신구대학 인터넷정보과 전임강사

<관심분야> 네트워크 관리, QoS라우팅, 서비스복구 라우팅