韓國컴퓨터情報學會 論文誌
第16卷 第2號, 2011. 2.

2011-16-2-3-1

정책 기울기 값 강화학습을 이용한 적응적인 QoS 라우팅 기법 연구

한 정 수*

A Study of Adaptive QoS Routing scheme using Policy-gradient Reinforcement Learning

Jeong-Soo Han*

요 약

본 논문에서는 강화학습(RL : Reinforcement Learning) 환경 하에서 정책 기울기 값 기법을 사용하는 적응적인 QoS 라우팅 기법을 제안하였다. 이 기법은 기존의 강화학습 환경 하에 제공하는 기법에 비해 기대 보상값의 기울기 값을 정책에 반영함으로써 빠른 네트워크 환경을 학습함으로써 보다 우수한 라우팅 성공률을 제공할 수 있는 기법이다. 이를 검증하기 위해 기존의 기법들과 비교 검증함으로써 그 우수성을 확인하였다.

▸ Keyword : 지역적 QoS 라우팅, 강화학습, MDP, POMDP

Abstract

In this paper, we propose a policy-gradient routing scheme under Reinforcement Learning that can be used adaptive QoS routing. A policy-gradient RL routing can provide fast learning of network environments as using optimal policy adapted average estimate rewards gradient values. This technique shows that fast of learning network environments results in high success rate of routing. For prove it, we simulate and compare with three different schemes.

▸ Keyword : Localized QoS Routing, Reinforcement Learning, Markov Decision Process, Partially Observable Markov Decision Processes

• 제1저자 : 한정수
• 투고일 : 2010. 10. 15, 심사일 : 2010. 10. 27, 게재확정일 : 2010. 11. 17.
* 신구대학 컴퓨터멀티미디어과 조교수(Dept. of Computer Multimedia, Shingu University)
• 본 연구는 2009년 교육과학기술부 재정지원 교육역량강화사업의 지원으로 수행 되었음.

Ⅰ. 서 론

현재 많은 연구에서 지역적 QoS 라우팅 기법이 전역 QoS 라우팅 기법과 비교하여 보다 안정적이고, 간단하며, 네트워크 상황에 보다 적응적이라는 것이 제시되고 있으며[1][2], 이에 대한 일환으로 [1]에서는 Proportional Sticky Routing(psr) 기법이 소개되었다. 또한, [3]에서는 강화학습과 같은 지능적인 제어 방식을 이용하여 전체 네트워크에 대한 정보나 네트워크의 트래픽 패턴을 알지 못해도 지역적 라우팅이 가능한 Q-학습(Q-Learning) 기반의 경로 선택 기법을 제안했다. 그러나 이 강화학습 방식에서는 에이전트와 연결된 환경이 계속해서 변화하게 되어 현재 최적의 행동(action)이 미래에 그대로 보장되지는 않는다[4]. 이에 환경에 대한 에이전트의 불확실성(Uncertainty)에 대한 확률적 접근이 필요하게 되는데, 이러한 연구가 POMDP(Partially Observable Markov Decision Processes)이다[5].

이제까지 강화학습과 POMDP 연구 모두가 행동 선택에 있어서 측정된 값을 기반으로 한 'greedy' 방식의 값 함수(value function) 접근방법을 사용했다. 이 방식은 함수 근사값(function approximation)을 사용하는 방식으로 다음과 같은 문제점들을 가지고 있다. 첫째, 이 방식은 주로 결정적인 정책(deterministic policies)을 찾는데 사용된다. 하지만, 최적화된 정책은 주로 특정 확률을 갖는 행동들을 찾는 경우가 많기 때문에 결정적인 정책을 찾는 방법은 최적의 솔루션이 될 수 없다[6]. 두 번째, 행동에 대한 측정된 값이 작으면 선택되어질 수도 아닐 수도 있다는 것이다. 이것은 값 함수 접근방법을 사용하는 알고리즘의 정확성에 심각한 문제를 발생시키게 된다. 이에 본 논문에서는 정책(행동 선택)을 위해 값 함수를 근사화하는 대신 기대 보상값의 기울기 값을 정책에 적용함으로써 근사화하는 방법인 정책 기울기 값 강화학습(policy-gradient RL) 기법을 사용하고자 한다. 이 기법에서는 정책 파라미터(θ)가 정책에 대한 보상값(ρ), 즉 기대 보상값에 대한 기울기 값에 비례하여 갱신될 수 있으며 $\Delta\theta \approx \alpha\frac{\partial\rho}{\partial\theta}$로 표현할 수 있다는 것이다. 이는 값 함수을 사용하는 기법과 달리 θ값의 작은 변화가 정책 자체에 그다지 큰 영향을 미치지 않는다는 점이 장점이다. 이 점은 기대 보상값에 대한 기울기 값의 근사값들에 대한 편차를 줄임으로써 빠른 학습을 제공할 수 있다는 것이다[6][7]. 이러한 상황은 네트워크 상황에 연결하면 네트워크 트래픽 상황에 대한 빠른 학습으로 연결될 수 있으며, 이를 통해 각 노드에서 빠른 QoS 라우팅이 가능해 질 수 있다. 이렇게 강화학습 환경과 네트워크 환경을 연결하여 성능을 분석한 노력들은 [3][5]들에서 잘 볼 수 있다. 마지막으로 논문에서 제시한 기법을 시뮬레이션을 통해 그 우수성을 검증하고자 한다.

Ⅱ. 정책 기울기 값 강화학습(policy-gradient RL)

MDP(Markov Decision Process)는 복잡한 POMDP 문제를 해결하기 위한 기초를 제공한다. MDP는 에이전트(agent)와 환경(environment)과의 상호관계와 이에 따른 강화 값(reinforcement value)을 통하여 에이전트의 행동을 개선해 나가는 방법으로서 환경에 대한 정확한 사전 지식 없이 학습 및 적응성을 보장할 수 있는 방법이다. 일반적으로 MDP는 각 시간 $t\in\{0,1,2,\ldots\}$에서 상태(state)들은 $s_t\in S$, 행동(action)들은 $a_t\in A$, 보상(reward)들은 $r_t\in R$로 표현한다. 환경에 대한 동적인 모습은 상태전이 확률 $P^a_{ss'}=P_r\{s_{t+1}=s'|s_t=s,a_t=a\}$로 표현이 가능하며, 기대 보상값은 $R^a_s=E\{r_{t+1}|s_t=s,a_t=a\},\forall s,s'\in S,a\in A$과 같이 표현할 수 있다. 또한, 각 시간에 에이전트의 정책 결정 과정은 $\pi(s,a,\theta)=P_r\{a_t=a|s_t=s,\theta\},\forall s\in S,a\in A, where\ \theta\in R^l$로 표현한다. 여기서 π는 θ에 대해서 미분가능하다고 가정함으로써 $\frac{\partial\pi(s,a)}{\partial\theta}$가 성립된다.

에이전트 목표, 즉 정책은 평균 보상값을 수식화함으로써 표현할 수 있는데, 평균 보상값은 상당 기간의 평균 기대 보상값으로써 (1)과 같이 $\rho(\pi)$로 표현할 수 있다.

$$\rho(\pi)=\lim_{n\to\infty}\frac{1}{n}E\{r_1+r_2+\ldots.+r_n|\pi\}=\sum_s d^\pi(s)\sum_a\pi(s,a)R^a_s, \quad \ldots\ldots (1)$$

여기서 $d^\pi(s)=\lim_{t\to\infty}P_r\{s_t=s|s_0,\pi\}$는 π와 s_0상에서, 즉 최적화 정책이나 초기 상태값과는 독립적인 값을 갖는다. 이를 토대로 각 상태와 행동 값을 통해 에이전트 정책을 (2)와 같이 표현할 수 있다.

$$Q^\pi(s,a)=\sum_{t=1}^{\infty}E\{r_t-\rho(\pi)|s_0=s,a_0=a,\pi\},\ \forall s\in S,a\in A \quad \ldots\ldots (2)$$

따라서, 정책 기울기 값 강화학습 기법에서 다음과 같은 결과를 얻을 수 있으며, 이를 최적화 정책에 적용할 수 있다.

$$\frac{\partial\rho}{\partial\theta}(s,a)=\sum_s d^\pi(s)\sum_a\frac{\partial\pi(s,a)}{\partial\theta}Q^\pi(s,a) \quad \ldots\ldots (3)$$

이에 대한 검증은 [6]에 잘 나와 있다.

단, $Q^{\pi}(s,a)$는 단순히 알지 못하고, 측정해야 하는데, 이를 위해서는 실제적인 보상값을 토대로 계산이 가능하다. 즉, $R_t = \sum_{k=1}^{\infty} r_{t+k} - \rho(\pi)$으로 계산이 가능하다.

III. 정책 기울기 값 강화학습 기반의 QoS 라우팅 기법

지역적 라우팅 기법은 네트워크 상태를 정확히 알지 못하는 상태에서 단지 송신지에서 유지하는 정보만을 의존하여 라우팅하게 되며, 이러한 상황은 결국 강화학습 하에서 에이전트가 환경에 대한 관찰을 통해 결정하게 되는 방식과 연결될 수 있다. 즉, 에이전트는 첫째, 매 시도(t)시에 자신의 상태($s^t(n) \in S$)에 대한 정보와 두 번째, 각 행동을 취할 시 비용에 대한 정보(R_a^s), 세 번째 최종 목적지 상태에 대한 정보, 마지막으로 전이 확률에 대한 확률분포에 대한 정보들은 알 수 있지만 전체 네트워크 상태 정보를 알 수 없기 때문에 특정 행동에 대한 성공 확률은 알 수 없다. 이러한 문제로 인해 지역적 라우팅 문제를 강화학습으로 적용할 수 있다. 또한 네트워크 상황에 적응적 라우팅 기법은 환경에 대한 관찰과 이에 대한 갱신을 통한 에이전트 대응방식과 연결될 수 있다.

3.1 강화학습 기반의 라우팅 모델

본 논문에서 제안한 라우팅 모델을 사용하기 위해 [표 1]에서와 같이 네트워크 라우팅 항목과 강화학습상의 항목에 대한 상호 연결이 필요하다.

이 기법에서의 정책은 네트워크에서 라우팅하기 위한 정책으로 표현되며, 정책의 입력으로는 노드를, 출력으로는 해당 경로로의 선택, 그리고 가중치로는 정책 파라미터로 표현할 수 있다.

요청을 받은 현재 상태 s인 노드는 $\rho(\pi)$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. 따라서, 각 노드에서는 (2)에서 제시한 $Q^{\pi}(s,a)$를 따라 라우팅하게 되며, 이에 대한 결과로 R_a^s를 얻게 될 것이다. 이에 각 노드는 자신의 $\rho(\pi)$값을 갱신한 결과인 $\delta_{s,a}$를 얻을 수 있으며, (4)와 같이 표현할 수 있다.

$$\delta_{s,a} = R_a^s + \gamma \frac{\partial \rho}{\partial \theta}(s',a') - \frac{\partial \rho}{\partial \theta}(s,a) \quad \cdots\cdots (4)$$

여기서, Ÿ은 감소율로 알려진 학습 상수로써, 이는 강화학습에서 사용한다. $\frac{\partial \rho}{\partial \theta}(s',a')$는 다음 상태 s'에서 발생된 실질적인 측정값이기 때문에, 상태 s가 방문될 때마다 그 측정값은 $R_a^s = \gamma \frac{\partial \rho}{\partial \theta}(s',a')$에 가깝게 갱신된다.

표 1. 라우팅 항목과 강화학습 항목 연결
Table 1. The mapping of routing element and RL

라우팅 항목	강화학습 기법 항목
각 네트워크 노드	에이전트
목적지로의 경로들 집합	행동들($a \in A$)
요청을 수신한 노드	상태($s \in S$)
지역 라우팅 정책	최적화 정책($Q^{\pi}(s,a)$)
선택된 경로 상에 요청에 전송되었을 경우 얻게 되는 값	보상값(R_a^s)
요청을 받은 각 노드가 가지고 있는 평균 기대 보상값	평균 기대 보상값($\rho(\pi)$)

3.2. 라우팅 정보 갱신

요청에 대한 연결 결과에 따라, 노드 s상에서 s'로의 라우팅 정보를 갱신하는 것이 필요하다. 이는 앞서 언급한 바와 같이 기대 보상값에 대한 근사값들의 편차를 줄임으로써 빠른 학습 즉, 빠른 라우팅 정보갱신을 이루고자 한다. 이는 결과적으로 최적화 정책 즉, $Q^{\pi}(s,a)$에 적용되어 라우팅 경로 선택 시 결정적 역할을 하게 된다.

값 함수를 갱신하기 위해서는 (4)의 $\delta_{s,a}$과 함께 (5)와 같은 수식이 적용된다.

$$Q^{\pi}(s,a) = Q^{\pi}(s,a) + \alpha \delta_{s,a} \quad \cdots\cdots (5)$$

여기서, α는 학습율이며, $e_d(s,a)$는 에이전트가 탐색 과정에서 선택한 (상태-행동)쌍이 얼마나 좋은가에 대한 평가를 나타내는 적합도(eligibility trace)이다. 최적화 정책($Q^{\pi}(s,a)$)와 적합도($e_d(s,a)$)는 (6)과 같이 수행된다.

$$e_d(s,a) = \begin{cases} \gamma\lambda e_d(s,a) + 1 & \text{if } elected \\ \gamma\lambda e_d(s,a) & otherwise \end{cases} \quad \cdots\cdots (6)$$

여기서, $\gamma\lambda (o < \gamma < 1, 0 < \gamma)$는 감소율(decay factor)을 나타내는 학습상수이며, 만일, 현재 상태에서 선택 가능한 (상태-행동)쌍들 중에서 가장 큰 $\rho(s,a)$ 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. 다중 경로 탐색 기법

정책 기울기 값 강화학습 라우팅 모델에서 사용할 다중 경로에 대해 본 논문에서는 S.Banerjee가 제안한 SSP(Single-Sink Program) 알고리즘[8]을 변경한 SEMA 알고리즘을 제안한다. SEMA 알고리즘은 Dijkstra의 최단 거리 알고리즘을 반복적으로 사용하여 송신지에서 목적지까지의 최소 가중치를 갖는 여러 개의 edge-disjoint 경로들을 찾는 것이다. 즉, 기존에 방문한 노드와 회선은 배제하는 방법을 반복적으로 사용함으로써 disjoint한 경로를 찾을 수 있으며, 이는 여러 개의 disjoint경로를 계속해서 찾을 수 있다는 장점을 가지고 있다.

그래프 $G=(V,E)$는 각 회선 $(u,v)\in E$에 비용 $c(u,v)$를 갖는 방향성 그래프로 하고 $|v|=n, |E|=m$, 로 정의할 때 SEMA 알고리즘은 [표 2]와 같다.

SEMA 알고리즘은 edge-disjoint한 다중 경로를 찾기 위해 $\theta(n,m)$의 Dijkstra 알고리즘을 여러 번 실행하게 되므로 순차적인 $\theta(n,m)$ 시간으로 해결할 수 있다.

[표 3]는 본 논문에서 제안한 정책 기울기 값 강화학습 라우팅 알고리즘 방식을 설명하고 있다.

표 2. 다중 경로 탐색 기법(SEMA) 알고리즘
Table 2. Shortest Edge-disjoint Multi-path searching Algorithm

SEMA(Shortest Edge-disjoint Multi-path searching Algorithm)
초기화] 찾고자 하는 다중 경로 집합 $\delta=\{\varnothing\}$ 단계 1] 최단거리 계산(I) 송신지 s에서 Dijkstra 알고리즘을 통해 최단 거리 트리 T를 생성하고, T상의 s에서 목적지 v로의 최단경로를 P_1으로 정한다. 또한, s에서 각 노드 x의 비용을 $C(s,x)$로 정의한다. 단계 2] 비용 재계산 및 그래프 수정 G상의 모든 회선(a,b)의 비용은 $C'(a,b)=C(a,b)+C(s,a)-C(s,b)$로 재계산된다. 이때, T에 속하는 모든 회선은 0으로 계산된다. 또한, P_1에 속하는 G상의 회선들의 방향을 반대로 구성하여 새로운 그래프 G_v를 생성한다. 단계 3] 최단거리 계산(II) G_v 상의 s에서 v로의 최단거리를 계산하고 이를 P_2로 정한다. 단계 4] 최단경로들 생성 $\{P_1\cup P_2\}-\{P_1\cap P_2\}$ 결과인 P_1'과 P_2'를 생성하고(이들이 disjoint shortest path이다), 다중 경로 집합 $\delta=\{P_1', P_2'\}+\delta$를 계산한다. 단계 5] 그래프 축소 G상에서 δ에 포함하는 회선들을 제외한 새로운 그래프 G'를 생성하고 이를 $G=G'$으로 재설정한다. 단계 6] 반복 송신지 s에 연결된 회선이 없을 때까지 단계 1]을 반복 수행한다.

IV. 시뮬레이션 및 결과분석

본 절에서는 논문에서 제안하는 정책 기울기 값 강화학습 기법을 사용한 QoS 라우팅 기법에 대한 성능을 알아보고자 한다. 이 기법은 네트워크 환경에 대한 정보를 빠르게 획득하고 이를 라우팅 정책에 적용함으로써 정확한 라우팅을 제공할 수 있다. 이를 검증하기 위해 기존의 다양한 기법들과 네트워크 부하에 따른 라우팅 성공률 즉, 서비스 성공률을 네트워크 부하와 시뮬레이션 시간으로 비교하여 살펴보기로 한다. 특히, 여기서 비교하는 기법들은 본 논문에서 제안하는 policy-gradient RL routing과 [1]와 [5]에서 제시한 기법들인 RL 기법을 이용한 TD Routing(Temporal Difference routing) 기법과, 앞서 서론에서 언급한 psr 기반의 Localized QoS Routing 기법이다.

표 3. 정책 기울기 값 강화학습 라우팅 알고리즘
Table 3. policy-gradient RL routing Algorithm

policy-gradient RL routing Algorithm
step 1. 네트워크 상의 모든 상태와 행동들에 대해 $Q^{\pi}(s,a)=0, e(s,a)=0$로 초기화 step 2. 초기 상태(s) 선택 Repeat { step 3. SEMA에 따라 행동(a)와 다음 상태(s')를 선택 $greedy\ policy:\ Q^{\pi}(s,a)$ 값이 최대인 행동 선택 $Q^{\pi}(s)=\arg max_{a\in A}\{Q^{\pi}(s,a)\}$ step 4. 연결 요청에 대한 step 3 선택에 대한 강화 값(R_a^s) 획득 $R_a^s=\begin{cases}1 & \text{if } request\ is\ accepted\\ 0 & \text{if } request\ is\ rejected\end{cases}$ step 5. 다음 상태(s')와의 차이값 $\delta_{s,a}$과 적합도 $e_d(s,a)$계산 $\delta_{s,a}=R_a^s+\gamma\frac{\partial\rho}{\partial\theta}(s',a')-\frac{\partial\rho}{\partial\theta}(s,a)$ $e_d(s,a)=\gamma\lambda e_d(s,a)+1$ step 6. 현재 상태(s) 라우팅 정보 갱신 $Q^{\pi}(s,a)=Q^{\pi}(s,a)+\alpha\delta_{s,a}$ step 7. 선택되지 못한 다른 행동에 대한 적합도 적용 $e_d(s,a)=\gamma\lambda e_d(s,a)$ } Until 최종 상태(d) 도착

4.1 시뮬레이션 환경

[그림 2]는 본 논문에서 제안한 알고리즘들의 성능 평가를 위해 사용된 네트워크 토폴로지들을 보여주고 있다. 여기서 사용하는 성능 평가 환경은 [1][5]에서 사용된 시뮬레이션

환경을 그대로 사용하고 있다. 따라서 다음과 같은 가정을 사용한다. 먼저 모든 회선은 무방향성이고, 각 방향으로 똑같은 $C\,unit$의 대역폭을 갖는다. 네트워크에 도착한 연결 요청은 $1\,unit$ 대역폭을 요구한다고 가정하자. 연결 요청(평균 도착율)은 소스 노드에 k를 갖는 포아송 프로세스를 따르며, 목적지는 소스 노드를 제외한 모든 노드로부터 랜덤하게 선택된다. 연결 요청에 대한 지속시간은 $\frac{1}{\mu}$를 갖는 지수 분포를 따른다. 네트워크 부하는 $\rho = \frac{\lambda N h}{\mu L C}$로 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 $C=20$, $\mu=60\,\mathrm{sec}$로 정의한다. 소스 노드 상의 평균 도착율 k는 성능 파라미터로 사용한다. 또한, 감소율 $\gamma\lambda$=0.9를 사용한다.

또한 연결 요청을 전송할 사용 가능한 경로는 위에서 살펴본 다중 경로 찾기 알고리즘(SEMA)을 통해 주 경로(primary path)와 보조 경로(alternative path)들로 분류하여 사용한다. 마지막으로 라우팅 경로 선택은 greedy-기법을 적용하는 것을 원칙으로 한다.

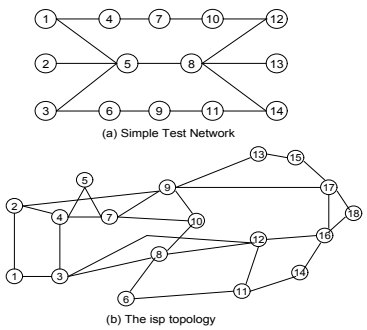

그림 1. 시뮬레이션 네트워크
Fig 1. Simulation Network

4.2 결과분석

사용자 서비스가 요청되면 에이전트 상에서 적절한 행동(경로)이 목적지에 도달 때까지 선택된다. 목적지에 도달하기 위한 행동들을 선택하기 위해서는 수식 (5)가 수행되어야 하며, 이를 통해 테스트 네트워크 환경 상에서 적절한 행동이 선택된다. 이러한 모델은 에이전트가 각 시도에 대한 선택된 행동이 유효한지 그렇지 않은지를 판단함으로써 갱신된다. 마지막으로 목적지에 도달하게 되면 기대값 $\delta_{s,a}$ 값이 갱신되며, 결과적으로 $\rho(\pi)$ 값이 갱신되게 된다.

[그림 2]에서는 [그림 1-a] Simple Test Network상에서 네트워크 부하에 따른 서비스 블럭킹 확률을 보여주고 있다. 네트워크 부하가 0.32까지는 세 가지 기법이 모두 비슷한 성능을 보이고 있지만 그 이 후에는 성능의 차이를 볼 수 있다. 먼저 psr 기법은 데이터의 패턴과 블럭킹 확률에 대한 계산이 가능할 지라도 경로 전체에 대한 최적화 문제(global optimization problem)를 해결하기 위해 소요되는 시간은 상당히 크기 때문에 다른 기법들에 비해 좋은 성능을 보이고 있지 못하다. 하지만, TD routing기법과 policy-gradient RL routing 기법은 미리 정해진 경로를 통해 자신만의 노드에서 결정되어진 값을 토대로 라우팅 되기 때문에 비교적 좋은 성능을 보이고 있다. 더욱이, 본 논문에서 제안하는 policy-gradient RL routing 기법은 네트워크 트래픽 상황에 대한 빠른 판단과 이를 라우팅 정책에 적용하기 때문에 제일 좋은 성능을 보이고 있다. 특히, 부하값이 0.36 ~ 0.56 사이에 가장 좋은 성능을 보여주고 있는데, 이는 다른 기법에 비해 policy-gradient RL routing 기법이 갖는 장점(네트워크 상황에 대한 빠른 수렴)으로 인해 더 좋은 성능을 보이고 있는 것으로 파악된다. 세 가지 기법 모두 부하가 0.7에 가까워질 수록 블럭킹 확률이 80% 이상 높아지는 것을 볼 수 있는데, 이는 네트워크 대역폭의 70%에 달하는 트래픽이 발생 할 경우 대역폭의 확장과 같은 물리적인 증설을 요구하기 때문으로 파악된다.

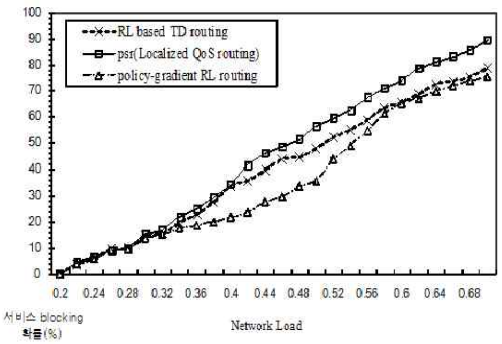

그림 2. Simple Test Network상의 네트워크 부하에 따른 서비스 블럭킹 확률
Fig 2. Service blocking Probability on Simple Test Network according to network load

[그림 3]은 [그림 1-b] ISP network상에서 네트워크 부하에 따른 서비스 블럭킹 확률을 보여주고 있다. 이 결과는 [그림 2]와 비슷한 결과를 보여주고 있다. 하지만 [그림 1-b] ISP network가 [그림 1-a] Simple Test Network 보다 detour와 같은 다른 경로들을 많이 갖고 있기 때문에 [그림

2] 결과보다는 블럭킹 확률이 더 낮은 것을 볼 수 있었다. 특히, 부하값이 0.28이상에서 [그림 2]보다 우수한 성능을 보여주고 있으며, 0.32부터 다른 기법에 비해 우수한 성능을 보여주고 있다.

[그림 4]는 [그림 1-b] ISP network상에서 시뮬레이션 시간에 따른 서비스 블럭킹 확률을 보여주고 있다. 이 결과는 세 가지 기법이 얼마나 빠르게 네트워크 상황을 판단하여 라우팅하는지에 대한 성능을 보여주고 있다. 즉, 시뮬레이션 시간이 길어질 수록 psr기법보다는 RL 기법을 사용하고 있는 TD routing 기법과 policy-gradient RL routing 기법이 더 나은 성능을 보여주고 있다. 특히, 평균 기대 보상값을 통한 정책 값을 결정하는 policy-gradient RL routing 기법이 단순히 보상값을 토대로 정책을 결정하는 TD routing 기법에 비해 보다 우수한 성능을 보이고 있음을 알 수 있었다. 특히 25분이 경과되는 시점에서는 다른 기법들에 비해 우수한 성능을 보이고 있다.

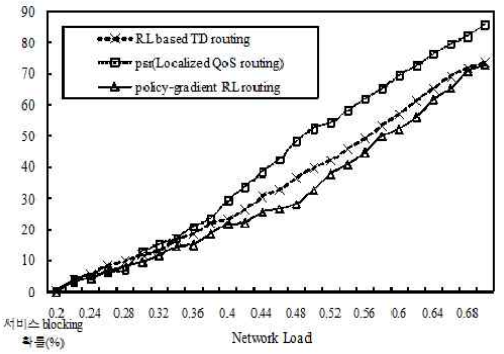

그림 3. ISP network상의 네트워크 부하에 따른 서비스 블럭킹 확률
Fig 3. Service blocking Probability on ISP network according to network load

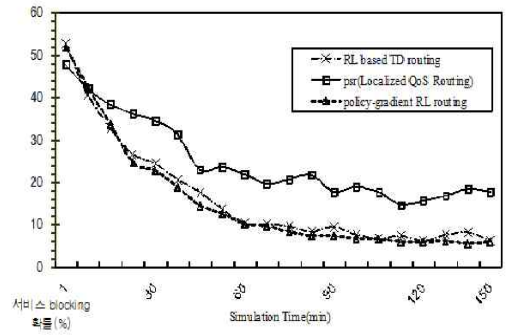

그림 4. ISP network상의 시뮬레이션 시간에 따른 서비스 블럭킹 확률
Fig 4. Service blocking Probability on ISP network according to simulation time

V. 결론

본 논문은 네트워크 상의 전체 상태정보에 대한 사전 지식 없이 지역적으로 라우팅할 수 있는 새로운 라우팅 기법을 제안하였다. 제안된 라우팅 기법은 강화학습 환경 하에서 기대 보상값에 대한 기울기 값을 정책에 반영하는 정책 기울기 값 강화학습 라우팅 기법을 통해 네트워크 상황을 보다 빠르고 정확하게 그 때 그 때 적응할 수 있는 적응적 라우팅 기법을 함께 제안하였다. 또한, 다중경로 탐색을 위해 SEMA 알고리즘을 제안했다.

강화학습 기법을 사용하는 라우팅 기법은 각 노드에서 네트워크에 대한 전체 상태정보를 알 수 없다는 지역적 라우팅 환경 하에서 적용하기 위한 기법으로 사용되고 있으며, 강화학습 환경하에서 정책 기울기 값 강화학습 라우팅 기법은 경로 선택으로 주어지는 보상값을 토대로 기대 보상값을 계산하고 이를 정책에 반영함으로써 보다 빠른 판단을 할 수 있게 도와준다. 더욱이 기대 보상값의 기울기 값을 반영함으로써 보다 최적화된 라우팅 정책 값을 도출할 수 있으며, 이를 통해 우수한 성능을 검증할 수 있었다.

또한, 본 논문에서 제안한 기법을 검증하기 위해 세 가지 기법들을 비교 검증하였는데, 각 기법들의 특성에 맞는 결과를 도출할 수 있었다. 즉, psr routing 기법은 최적화 정책을 도출하기 위한 시간적인 문제점을, TD routing은 단순한 보상값을 토대로 라우팅 결정을 하여 그 성능 면에서 제안하는 기법보다 낮은 성능을 보여주고 있다. 특히, 시뮬레이션 시간이 길어짐에 따라 네트워크 환경 변화를 빠르게 판단하는 policy-gradient RL routing기법이 다른 기법보다 우수함을 볼 수 있었다.

참고문헌

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

[2] 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, pp.

24-28, June 2003.
[3] Yvn Tpac Valdivia, Marley M, Vellasco, Marco A. Pacheco "An Adaptive Network Routing Strategy with Temporal Differences," Inteligencia Artificial, Revista Lberoamericana de Inteligencia Aritificial, No. 12, pp. 85-91, 2001.
[4] Jeongsoo Han, "Network-Adaptive QoS Routing Using Local Information," APNOMS 2006, LNCS 4238, pp. 190-199, 2006.
[5] Leslie Pack Kaelbling, Michael L. Littman, Andrew W.Moore, "Reinforcement Learning:A Survey," Journal of Artificial Intelligence Research 4, pp. 237-285, 1996
[6] Richard S. Sutton etc, "*Policy Gradient Methods for Reinforcement Learning with Function Approximation*," Advances in Neural Information Processing System, pp. 1057-1063, MIT Press 2000.
[7] Gregory Z. Grudic, Vijay Kumar, "Using Policy Gradient Reinforcement Learning on Automous Robot Controllers," IROS03, Las Vagas, US, October, 2003.
[8] S.Banerjee, R.K. Ghosh and A.P.K Reddy, "Parallel algorithm for shortest pairs of edge-disjoint paths," Journal Parallel Distrib. Comput. pp. 165-171, 1996.

저 자 소 개

한 정 수

1997 : 성균관대학교 공학석사.
2003 : 성균관대학교 공학박사.
2003 - 현재 : 신구대학 컴퓨터멀티미디어과 조교수
관심분야 : QoS라우팅, 네트워크 트래픽 관리 등
E-mail : jshan@shingu.ac.kr