(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2020년01월08일
(11) 등록번호 10-2063643
(24) 등록일자 2020년01월02일

(51) 국제특허분류(Int. Cl.)
H04B 7/155 (2006.01) *G06N 20/00* (2019.01)
H04B 17/336 (2014.01)
(52) CPC특허분류
H04B 7/15592 (2013.01)
G06N 20/00 (2019.01)
(21) 출원번호 10-2019-0107494
(22) 출원일자 2019년08월30일
심사청구일자 2019년08월30일
(56) 선행기술조사문헌
KR1020140103797 A*
KR1020190086133 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
세종대학교 산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
송형규
경기도 성남시 분당구 중앙공원로 17, 한양아파트 320동 303호
백민재
서울특별시 노원구 동일로230가길 15, 102동 1906호 (상계동, 상계우방아파트)
나유진
경기도 용인시 수지구 광교마을로 2, 4306동 1903호
(74) 대리인
특허법인태백

전체 청구항 수 : 총 10 항 심사관 : 신상길

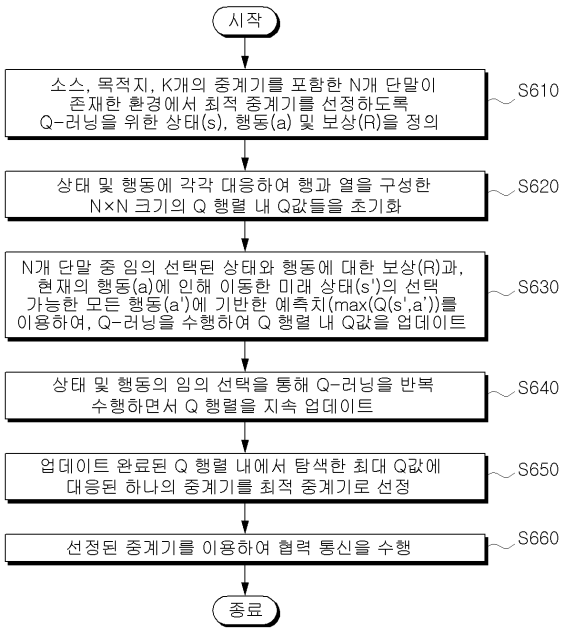
(54) 발명의 명칭 **무선 협력 통신 시스템을 기반으로 하는 Q-러닝 기반의 중계기 선정 장치 및 그 방법**

(57) 요 약

본 발명은 무선 협력 통신 시스템을 기반으로 하는 Q-러닝 기반의 중계기 선정 장치 및 그 방법에 관한 것이다. 본 발명에 따르면, 소스 단말, 목적지 단말, K개의 중계기를 포함한 N개 단말로부터 임의 선택된 n(n=1,…,N)번째 송신측 단말을 상태(s_n)로, 해당 상태(s_n)에서 임의 선택된 m(m=1,…,N)번째 수신측 단말을 행동(a_m)으로, m번

(뒷면에 계속)

대 표 도 - 도6



째 수신측 단말에서의 SNR 지표를 해당 상태(s_n)에서 취한 행동(a_m)에 따른 즉각적인 보상(R; R(s_n,a_m))으로 정의하는 단계, 상태 및 행동에 각각 대응하여 행과 열 성분을 구성한 N×N 크기의 Q 행렬 내 원소들인 Q값들을 초기화하는 단계, N개의 단말 중에서 임의 선택된 상태(s_n)와 행동(a_m)에 대한 즉각적인 보상(R(s_n,a_m))과, 현재 취한 행동(a_m)으로 인해 이동한 미래 상태(s_n')의 선택 가능한 모든 행동(a_m')에 대응하는 Q값들 중 최대치(max(Q(s_n',a_m'))를 이용하여 Q-러닝을 수행하여, Q 행렬 내의 Q(s_n,a_m) 값을 업데이트하는 단계, 상태 및 행동의 임의 선택을 통해 Q-러닝을 반복 수행하면서 Q 행렬을 지속 업데이트하는 단계, 및 업데이트가 완료된 Q 행렬 내에서 탐색한 최대 Q값에 대응된 하나의 중계기를 최적 중계기로 선정하는 단계를 포함하는 중계기 선정 방법을 제공한다.

본 발명에 따르면, 다수의 중계기가 존재하는 환경에서 Q-러닝을 이용하여 최적 중계기를 선정함으로써 시스템의 신뢰성 및 비트 요율 성능을 높일 수 있다.

(52) CPC특허분류
H04B 17/336 (2015.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711075702
부처명 과학기술정보통신부
연구관리전문기관 정보통신기획평가원
연구사업명 대학ICT연구센터지원사업
연구과제명 지능형 비행로봇 융합기술 연구
기 여 율 1/1
주관기관 세종대학교 산학협력단
연구기간 2018.06.01 ~ 2021.12.31

명 세 서

청구범위

청구항 1

무선 협력 통신 시스템을 기반으로 하는 Q-러닝 기반의 중계기 선정 방법에 있어서,

소스 단말, 목적지 단말, K개의 중계기를 포함한 N개 단말(N=K+2)로부터 임의 선택된 n(n=1,…,N)번째 송신측 단말을 상태(s_n)로, 해당 상태(s_n)에서 임의 선택된 m(m=1,…,N)번째 수신측 단말을 행동(a_m)으로, m번째 수신측 단말에서의 SNR 지표를 해당 상태(s_n)에서 취한 행동(a_m)에 따른 즉각적인 보상(R; $R(s_n,a_m)$)으로 정의하는 단계;

상기 상태 및 행동에 각각 대응하여 행과 열 성분을 구성한 N×N 크기의 Q 행렬 내 원소들인 Q값들을 초기화하는 단계;

상기 N개의 단말 중에서 임의 선택된 상태(s_n)와 행동(a_m)에 대한 즉각적인 보상($R(s_n,a_m)$)과, 현재 취한 행동(a_m)으로 인해 이동한 미래 상태(s_n')의 선택 가능한 모든 행동(a_m')에 대응하는 Q값들 중 최대치($\max(Q(s_n',a_m'))$)를 이용하여 Q-러닝을 수행하여, 상기 Q 행렬 내의 $Q(s_n,a_m)$ 값을 업데이트하는 단계;

상기 상태 및 행동의 임의 선택을 통해 Q-러닝을 반복 수행하면서 Q 행렬을 지속 업데이트하는 단계; 및

업데이트가 완료된 Q 행렬 내에서 탐색한 최대 Q값에 대응된 하나의 중계기를 최적 중계기로 선정하는 단계를 포함하며,

상기 N개 단말 중 n,m=1인 단말은 상기 소스 단말, n,m=N인 단말은 상기 목적지 단말, 나머지 N-2개는 상기 K개의 중계기이며,

상기 즉각적인 보상인 R은 상기 선택된 행동 및 상태에 따라 N×N 가지로 존재하고 $R(s_n,a_m)=R(n,m)$로 설정되되, N×N 개의 즉각적인 보상 중에서, 송신측 단말과 수신측 단말이 동일한 경우(n=m=i)의 보상값 R(i,i)과, 상기 소스 단말과 상기 목적지 단말 간 직경로에 해당한 경우의 보상값 R(1,N), 그리고 상기 목적지 단말(n=N)이 송신측인 경우의 보상값 R(N,i)은 아래의 수학식과 같이 모두 '0'의 값으로 설정되는 중계기 선정 방법:

$$R(i,i)=R(1,N)=R(N,i)=0$$

여기서, i={1,…N}이다.

청구항 2

청구항 1에 있어서,

상기 즉각적인 보상 R은 SNR을 기반으로 아래 수학식에 의해 결정되는 중계기 선정 방법:

$$R=\frac{SNR_m}{d\rho}$$

여기서, SNR_m는 m번째 수신측 단말에서의 SNR 값, d는 n번째 송신측 단말과 m번째 수신측 단말 사이의 거리, ρ는 자유 공간 경로 손실을 나타낸다.

청구항 3

삭제

청구항 4

청구항 1에 있어서,

상기 중계기가 송신측(n=2,…,N-1)이고 상기 목적지 단말이 수신측(m=N)인 경우의 보상값 R(i,N)(이때, i≠1, N)은,

보상 R의 수학식에 설정 가중치가 추가로 가산된 값을 사용하는 중계기 선정 방법.

청구항 5

청구항 1에 있어서,

상기 Q 행렬 내의 $Q(s_n,a_m)$ 값은 아래 수학식에 의해 업데이트되는 중계기 선정 방법:

$$New\ Q(s_n,a_m) = Q(s_n,a_m) + \alpha[R(s_n,a_m) + \gamma \cdot \max Q(s_n',a_m') - Q(s_n,a_m)]$$

여기서, New $Q(s_n,a_m)$는 업데이트된 $Q(s_n,a_m)$ 값, α(0<α<1)는 학습률, γ(0<γ<1)는 할인 계수(discount factor)를 나타낸다.

청구항 6

청구항 1에 있어서,

상기 상태와 행동의 임의 선택 시에 Decaying ε-greedy 알고리즘을 적용하여 시간이 경과할수록 무작위 선택 행동의 확률을 감소시키는 중계기 선정 방법.

청구항 7

협력 통신 시스템을 위한 Q-러닝 기반의 중계기 선정 장치에 있어서,

소스 단말, 목적지 단말, K개의 중계기를 포함한 N개 단말(N=K+2)로부터 임의 선택된 n(n=1,…,N)번째 송신측 단말을 상태(s_n)로, 해당 상태(s_n)에서 임의 선택된 m(m=1,…,N)번째 수신측 단말을 행동(a_m)으로, m번째 수신측 단말에서의 SNR 지표를 해당 상태(s_n)에서 취한 행동(a_m)에 따른 즉각적인 보상(R; $R(s_n,a_m)$)으로 정의하는 설정부;

상기 상태 및 행동에 각각 대응하여 행과 열 성분을 구성한 N×N 크기의 Q 행렬 내 원소들인 Q값들을 초기화하는 초기화부;

상기 N개의 단말 중에서 임의 선택된 상태(s_n)와 행동(a_m)에 대한 즉각적인 보상($R(s_n,a_m)$)과, 현재 취한 행동(a_m)으로 인해 이동한 미래 상태(s_n')의 선택 가능한 모든 행동(a_m')에 대응하는 Q값들 중 최대치($\max(Q(s_n',a_m'))$)를 이용하여 Q-러닝을 수행하여, 상기 Q 행렬 내의 $Q(s_n,a_m)$ 값을 업데이트하되, 상기 상태 및 행동의 임의 선택을 통해 Q-러닝을 반복 수행하면서 Q 행렬을 지속 업데이트하는 학습부; 및

상기 업데이트가 완료된 Q 행렬 내에서 탐색한 최대 Q값에 대응된 하나의 중계기를 최적 중계기로 선정하는 결정부를 포함하며,

상기 N개 단말 중 n,m=1인 단말은 상기 소스 단말, n,m=N인 단말은 상기 목적지 단말, 나머지 N-2개는 상기 K개의 중계기이며,

상기 즉각적인 보상인 R은 상기 선택된 행동 및 상태에 따라 N×N 가지로 존재하고 $R(s_n,a_m)=R(n,m)$로 설정되되, N×N 개의 즉각적인 보상 중에서, 송신측 단말과 수신측 단말이 동일한 경우(n=m=i)의 보상값 R(i,i)과, 상기 소스 단말과 상기 목적지 단말 간 직경로에 해당한 경우의 보상값 R(1,N), 그리고 상기 목적지 단말(n=N)이 송신측인 경우의 보상값 R(N,i)은 아래의 수학식과 같이 모두 '0'의 값으로 설정되는 중계기 선정 장치:

$$R(i,i) = R(1,N) = R(N,i) = 0$$

여기서, i={1,…N}이다.

청구항 8

청구항 7에 있어서,

상기 즉각적인 보상 R은 SNR을 기반으로 아래 수학식에 의해 결정되는 중계기 선정 장치:

$$R = \frac{SNR_m}{d\rho}$$

여기서, SNR_m는 m번째 수신측 단말에서의 SNR 값, d는 n번째 송신측 단말과 m번째 수신측 단말 사이의 거리, ρ는 자유 공간 경로 손실을 나타낸다.

청구항 9

삭제

청구항 10

청구항 7에 있어서,

상기 중계기가 송신측(n=2,…,N-1)이고 상기 목적지 단말이 수신측(m=N)인 경우의 보상값 R(i,N)(이때, i≠1, N)은,

보상 R의 수학식에 설정 가중치가 추가로 가산된 값을 사용하는 중계기 선정 장치.

청구항 11

청구항 7에 있어서,

상기 Q 행렬 내의 $Q(s_n,a_m)$ 값은 아래 수학식에 의해 업데이트되는 중계기 선정 장치:

$$New\ Q(s_n,a_m) = Q(s_n,a_m) + \alpha[R(s_n,a_m) + \gamma \cdot \max Q(s_n',a_m') - Q(s_n,a_m)]$$

여기서, New $Q(s_n,a_m)$는 업데이트된 $Q(s_n,a_m)$ 값, α (0<α<1)는 학습률, γ (0<γ<1)는 할인 계수(discount factor)를 나타낸다.

청구항 12

청구항 7에 있어서,

상기 학습부는,

상기 상태와 행동의 임의 선택 시에 Decaying ε-greedy 알고리즘을 적용하여 시간이 경과할수록 무작위 선택 행동의 확률을 감소시키는 중계기 선정 장치.

발명의 설명

기 술 분 야

[0001] 본 발명은 무선 협력 통신 시스템을 기반으로 하는 Q-러닝 기반의 중계기 선정 장치 및 그 방법에 관한 것으로서, 보다 상세하게는 복수의 중계기가 존재하는 환경에서 Q-러닝 알고리즘을 기반으로 최적의 중계기를 선정하여 시스템의 성능을 높일 수 있는 Q-러닝 기반의 중계기 선정 장치 및 그 방법에 관한 것이다.

배 경 기 술

[0002] 송신기 및 수신기에서 다중 안테나를 사용하여 통신하는 MIMO 무선 통신 시스템은 높은 대역폭 이득과 높은 데이터 전송률을 얻는다. 하지만, 이러한 MIMO 시스템에서 단말(UE; User Equipment)의 비용과 크기의 제약이 존재한다면 단말 안에 다중의 안테나를 설치하는 것은 불가능하다.

[0003] 이러한 문제를 해결하기 위한 협력 통신 기법은 송신기와 수신기 사이에서 중계기를 이용하여 가상의 MIMO 환경을 구현할 수 있다. 협력통신 기법은 크게 AF(Amplify and Forward) 기법 및 DF(Decoded and Forward) 기법으로 나뉜다.

[0004] AF 기법은 중계기에서 수신되는 신호의 전력을 증폭시켜 재전송하는 기법으로, 구현은 간단하지만 수신단에서

신호의 전력을 정규화하고 증폭시키는 과정에서 잡음의 증폭으로 인하여 성능의 열화를 초래한다. DF 기법은 중계기에서 원 신호를 복조 후 변조를 한 뒤 전송하는 기법으로, AF 기법에 비해 연산량 측면에서 복잡한 기법이지만 대다수의 통신 단말에 변복조기와 부호복호화기가 탑재되어 있어 현실적으로 구현 가능한 기법이라 할 수 있다.

[0005] 하지만, 이와 같은 협력 통신 기법에서 다수의 중계기가 존재하는 환경에서 모든 중계기를 이용하여 신호를 전송하는 것은 불필요한 자원의 낭비를 초래한다. 따라서 필요한 중계기를 적절히 선정해야 하며, 적절한 중계기를 선택하지 못한다면 전체 시스템의 성능을 저하시킬 수 있다.

[0006] 따라서 다수의 중계기가 존재하는 환경에서 송신기와 수신기 사이에서 최적의 중계기를 선정하기 위한 기법이 필요하다. 기존의 중계기 선정 기법으로는 무작위 중계기 선정 기법(Random Relay Selection), 문턱값 기반 중계기 선정 기법(Threshold-based Relay Selection), 그리고 조화평균 중계기 선정 기법(Best Harmonic Mean(BHM) Relay Selection)이 존재한다.

[0007] 무작위 중계기 선정 기법은 다수의 중계기 중에서 하나의 중계기를 무작위 선정하는 방식으로, 세 가지 중 가장 간단한 기법에 해당한다. 하지만, 해당 기법은 중계기와 목적지 단말에서의 수신 신호의 SNR이나 채널을 전혀 고려하지 않기에, 채널 상황이 좋지 않은 중계기가 선정된 경우 협력 통신의 성능이 열화된다.

[0008] 문턱값 기반 중계기 선정 기법은 준 최적의 중계기 선정 기법으로, 문턱값을 어떻게 정의하느냐에 따라 중계기 선정이 달라지고 성능이 상이해진다. 일반적으로는 소스 단말과 각 중계기 사이의 채널 계수를 평균한 값을 문턱값으로 결정한 후, 채널 계수가 문턱값보다 큰 것으로 판단된 중계기 중에서 하나를 선정하여 협력 통신을 수행한다. 하지만, 이러한 방식은 모든 채널 상태 정보를 고려하는 것이 아닌, 소스 단말과 중계기 간의 채널만을 고려하므로 성능이 열화된다.

[0009] 조화평균 중계기 선정 기법은 최적의 중계기 선정 기법으로, 모든 중계기의 조화 평균을 구하기 위해 채널 상태 정보를 필요로 한다. 다시 말해, 채널 상태 정보를 이용하여 소스 단말-중계기, 중계기-목적지 단말 간 조화 평균을 구하고 그 정보를 소스 단말로 피드백하여 중계기를 선택한다. 이때, 소스 단말에서는 모든 중계기에서 구한 조화평균의 최댓값을 이용하여 중계기를 선정한다.

[0010] 이와 같은 조화평균 중계기 선정 기법은 최적의 중계기를 선정하므로 통신 성능이 가장 우수하지만, 소스 단말이 모든 채널에 대한 정보를 알아야 하므로, 구현의 복잡도 및 연산량이 높고 실제 환경에서는 적용되기 어렵다.

[0011] 따라서 상술한 기존 기법의 단점들을 극복하면서 기존의 기법과 유사한 성능을 도출할 수 있는 새로운 중계기 선정 기법이 요구된다.

[0012] 본 발명의 배경이 되는 기술은 한국공개특허 제2017-0103285호(2017.09.13 공개)에 개시되어 있다.

발명의 내용

해결하려는 과제

[0013] 본 발명은 복수의 중계기가 존재하는 환경에서 Q-러닝 알고리즘을 기반으로 최적의 중계기를 선정하여 수신 신호의 SNR 성능과 시스템의 신뢰성을 향상시킬 수 있는 무선 협력 통신 시스템을 기반으로 하는 Q-러닝 기반의 중계기 선정 장치 및 그 방법을 제공하는데 목적이 있다.

과제의 해결 수단

[0014] 본 발명은, 무선 협력 통신 시스템을 기반으로 하는 Q-러닝 기반의 중계기 선정 방법에 있어서, 소스 단말, 목적지 단말, K개의 중계기를 포함한 N개 단말(N=K+2)로부터 임의 선택된 n(n=1,…,N)번째 송신측 단말을 상태(s_n)로, 해당 상태(s_n)에서 임의 선택된 m(m=1,…,N)번째 수신측 단말을 행동(a_m)으로, m번째 수신측 단말에서의 SNR 지표를 해당 상태(s_n)에서 취한 행동(a_m)에 따른 즉각적인 보상(R; $R(s_n,a_m)$)으로 정의하는 단계와, 상기 상태 및 행동에 각각 대응하여 행과 열 성분을 구성한 N×N 크기의 Q 행렬 내 원소들인 Q값들을 초기화하는 단계와, 상기 N개의 단말 중에서 임의 선택된 상태(s_n)와 행동(a_m)에 대한 즉각적인 보상($R(s_n,a_m)$)과, 현재 취한 행동(a_m)으로 인해 이동한 미래 상태(s_n')의 선택 가능한 모든 행동(a_m')에 대응하는 Q값들 중 최대치($\max(Q(s_n',a_m'))$)를 이용하여 Q-러닝을 수행하여, 상기 Q 행렬 내의 $Q(s_n,a_m)$ 값을 업데이트하는 단계와, 상기

상태 및 행동의 임의 선택을 통해 Q-러닝을 반복 수행하면서 Q 행렬을 지속 업데이트하는 단계, 및 업데이트가 완료된 Q 행렬 내에서 탐색한 최대 Q값에 대응된 하나의 중계기를 최적 중계기로 선정하는 단계를 포함하는 중계기 선정 방법을 제공한다.

[0015] 또한, 상기 즉각적인 보상 R은 상기 선택된 행동 및 상태에 따라 N×N 가지로 존재하며, SNR을 기반으로 아래 수학식에 의해 결정될 수 있다.

[0016] $$R = \frac{SNR_m}{d\rho}$$

[0017] 여기서, SNR_m는 m번째 수신측 단말에서의 SNR 값, d는 n번째 송신측 단말과 m번째 수신측 단말 사이의 거리, ρ는 자유 공간 경로 손실을 나타낸다.

[0018] 또한, 상기 N개 단말 중 n,m=1인 단말은 상기 소스 단말, n,m=N인 단말은 상기 목적지 단말, 나머지 N-2개는 상기 K개의 중계기이며, 상기 즉각적인 보상인 $R(s_n,a_m)=R(n,m)$로 설정되되, N×N 개의 즉각적인 보상 중에서, 송신측 단말과 수신측 단말이 동일한 경우(n=m=i)의 보상값 R(i,i)과, 상기 소스 단말과 상기 목적지 단말 간 직경로에 해당한 경우의 보상값 R(1,N), 그리고 상기 목적지 단말(n=N)이 송신측인 경우의 보상값 R(N,i)은 아래의 수학식과 같이 모두 '0'의 값으로 설정될 수 있다.

[0019] $$R(i,i) = R(1,N) = R(N,i) = 0$$

[0020] 여기서, i={1,…N}이다.

[0021] 또한, 상기 중계기가 송신측(n=2,…,N-1)이고 상기 목적지 단말이 수신측(m=N)인 경우의 보상값 R(i,N)(이때, i≠1,N)은, 보상 R의 수학식에 설정 가중치가 추가로 가산된 값을 사용할 수 있다.

[0022] 또한, 상기 Q 행렬 내의 $Q(s_n,a_m)$ 값은 아래 수학식에 의해 업데이트될 수 있다.

[0023] $$New\ Q(s_n,a_m) = Q(s_n,a_m) + \alpha[R(s_n,a_m) + \gamma \cdot \max Q(s_n',a_m') - Q(s_n,a_m)]$$

[0024] 여기서, New $Q(s_n,a_m)$는 업데이트된 $Q(s_n,a_m)$ 값, α (0<α<1)는 학습률, γ (0<γ<1)는 할인 계수(discount factor)를 나타낸다.

[0025] 또한, 상기 상태와 행동의 임의 선택 시에 Decaying ε-greedy 알고리즘을 적용하여 시간이 경과할수록 무작위 선택 행동의 확률을 감소시킬 수 있다.

[0026] 그리고, 본 발명은, 협력 통신 시스템을 위한 Q-러닝 기반의 중계기 선정 장치에 있어서, 소스 단말, 목적지 단말, K개의 중계기를 포함한 N개 단말(N=K+2)로부터 임의 선택된 n(n=1,…,N)번째 송신측 단말을 상태(s_n)로, 해당 상태(s_n)에서 임의 선택된 m(m=1,…,N)번째 수신측 단말을 행동(a_m)으로, m번째 수신측 단말에서의 SNR 지표를 해당 상태(s_n)에서 취한 행동(a_m)에 따른 즉각적인 보상(R; $R(s_n,a_m)$)으로 정의하는 설정부와, 상기 상태 및 행동에 각각 대응하여 행과 열 성분을 구성한 N×N 크기의 Q 행렬 내 원소들인 Q값들을 초기화하는 초기화부와, 상기 N개의 단말 중에서 임의 선택된 상태(s_n)와 행동(a_m)에 대한 즉각적인 보상($R(s_n,a_m)$)과, 현재 취한 행동(a_m)으로 인해 이동한 미래 상태(s_n')의 선택 가능한 모든 행동(a_m')에 대응하는 Q값들 중 최대치($\max(Q(s_n',a_m'))$)를 이용하여 Q-러닝을 수행하여, 상기 Q 행렬 내의 $Q(s_n,a_m)$ 값을 업데이트하되, 상기 상태 및 행동의 임의 선택을 통해 Q-러닝을 반복 수행하면서 Q 행렬을 지속 업데이트하는 학습부, 및 상기 업데이트가 완료된 Q 행렬 내에서 탐색한 최대 Q값에 대응된 하나의 중계기를 최적 중계기로 선정하는 결정부를 포함하는 중계기 선정 장치를 제공한다.

발명의 효과

[0027] 본 발명에 따르면, 복수의 중계기가 존재하는 환경에서 Q-러닝 알고리즘을 기반으로 최적의 중계기를 선정함으로써 높은 신뢰성을 얻을 수 있으며 비트 오율 성능을 향상시키고 복잡도를 줄일 수 있다. 또한, 본 발명에서 제안한 기법은 기존의 기법과 동일한 성능을 얻으면서 구현 복잡도 및 연산량과 오버헤드를 줄일 수 있다.

도면의 간단한 설명

[0028] 도 1은 본 발명의 실시예를 위한 협력 통신 시스템의 구성을 나타낸 도면이다.

도 2는 본 발명의 실시예에 적용되는 Q-러닝 모델을 설명한 도면이다.

도 3은 도 2의 Q-러닝 모델을 통해 시간에 따라 업데이트되는 Q-테이블을 설명하는 도면이다.

도 4는 본 발명의 실시예에 따른 최적 중계기 선정 기법의 개념을 설명하는 도면이다.

도 5는 본 발명의 실시예에 따른 무선 협력 통신 시스템을 기반으로 하는 Q-러닝 기반의 중계기 선정 장치의 구성을 나타낸 도면이다.

도 6은 도 5를 이용한 최적 중계기 선정 방법을 설명하는 도면이다.

도 7은 K=4, N=6인 협력 통신 시스템에서 최적 중계기를 선정하는 예시를 설명하기 위한 도면이다.

도 8은 도 7에 대해 획득한 보상 행렬을 예시한 도면이다.

도 9는 본 발명의 실시예에서 Q-러닝 전의 초기화된 Q 행렬을 예시한 도면이다.

도 10은 본 발명의 실시예에 따라 Q 행렬을 업데이트하여 최적 중계기를 선정하는 과정을 예시한 도면이다.

도 11은 본 발명의 기법과 기존의 기법의 SNR 대비 BER 성능 그래프를 나타낸 도면이다.

발명을 실시하기 위한 구체적인 내용

[0029] 그러면 첨부한 도면을 참고로 하여 본 발명의 실시 예에 대하여 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자가 용이하게 실시할 수 있도록 상세히 설명한다.

[0030] 도 1은 본 발명의 실시예를 위한 협력 통신 시스템의 구성을 나타낸 도면이다.

[0031] 도 1에 나타낸 바와 같이, 본 발명의 실시예는 소스 단말(S;Source)과 목적지 단말(D;Destination) 사이에 다수의 중계기(R;Relay)가 존재하는 시스템 환경을 고려한다. 시스템 내 노드의 수는 소스 단말(S), K개 중계기(R_1~R_K), 그리고 목적지 단말(D)을 포함하여 총 N개(N=K+2)를 가진다. 도 1에서 R_m은 K개 중계기(R_1~R_K) 중 선정된 m번째 중계기로 가정한다.

[0032] 본 발명의 실시예는 최적의 중계기를 선정하기 위하여 Q-러닝 알고리즘을 이용한다. 이하에서는 본 발명의 상세한 설명에 앞서 이러한 Q-러닝 알고리즘을 간단히 설명한다.

[0033] 도 2는 본 발명의 실시예에 적용되는 Q-러닝 모델을 설명한 도면이다. Q-learning 기법은 일반적으로 에이전트(Agent), 환경(Environment), 상태(State), 행동(Action), 그리고 보상(Reward)으로 이루어진 강화학습(Reinforcement Learning) 알고리즘이다. 학습이 완료되면 최적의 정책을 얻을 수 있는데, 여기서 정책이란 최적의 중계기를 찾는 규칙이다. 이러한 정책은 행동에 대한 보상을 최대화하는 규칙이다. 다시 말해, Q-러닝의 목적은 최적의 정책을 찾는 것이라고 할 수 있다.

[0034] 이때, 상태를 $S=\{s_1,\cdots,s_n\}$, 그리고 행동을 $A=\{a_1,\cdots,a_m\}$이라고 한다면, 에이전트는 주어진 환경에서 행동을 취하고 환경으로부터 피드백을 받아 최적의 정책을 찾는다. 여기서 n,m은 각각 주어진 상태와 행동의 인덱스를 나타낸다. 이때, Q-러닝의 Q-함수의 업데이트 과정은 아래와 같이 이루어지게 된다.

수학식 1

[0035] $$Q(s,a)\leftarrow Q(s,a)+\alpha[R(s,a)+\gamma \cdot \max Q(s',a')-Q(s,a)]$$

[0036] 여기서, R(s,a)는 현재 상태 s에서 행동 a를 취하여 얻는 즉각적인 보상을 나타내며, 현재 행동을 취함으로써 발생하는 미래의 보상은 Q값(Q-values) 즉 Q(s',a')을 통해 얻을 수 있다. 여기서 s'는 현재 상태 s에서 행동 a를 취했을 때 이동한 다음(미래) 상태를 나타낸다. a'는 다음 상태에서 취할 수 있는 행동을 의미한다.

[0037] γ은 할인 계수(discount factor)로써 현재와 미래 보상에 대한 가중치를 나타낸다. 이러한 γ는 0과 1 사이

에서 결정되며, 0에 가까울수록 현재 보상에 대해서 만족하고 1에 가까울수록 미래에 받을 보상에 대한 기댓값이 높아진다.

[0038] α는 학습률(learning rate)로, 현재와 미래의 정보 중 어떤 정보에 더 의존하여 학습하는지를 나타낸다. 마찬가지로 α는 0과 1 사이에서 결정되며 0에 가까울수록 현재 정보에 의존하여 학습하고 1에 가까울수록 미래에 얻는 정보에 따라 학습이 진행된다.

[0039] 학습이 진행될수록 Q값이 업데이트되고 결과적으로 주어진 환경에서 취할 수 있는 최적의 행동을 찾을 수 있게 된다. 요약하면, 이러한 학습 과정의 목적은 Q값으로 이루어진 Q-테이블(Q-table)을 업데이트하고, 최종 업데이트 결과를 이용하여 환경으로부터 보상의 최댓값을 얻을 수 있는 행동을 취하는 것이다. 이때의 행동은 다음과 같이 표현할 수 있다.

수학식 2

[0040] $$a = \arg\max(Q(s,a))$$

[0041] 이러한 수학식 2는 Q를 최대화하는 행동을 나타낸다.

[0042] 도 3은 도 2의 Q-러닝 모델을 통해 시간에 따라 업데이트되는 Q-테이블을 설명하는 도면이다. 도 3에 도시된 것과 같이 Q-테이블은 Q값을 저장하는 역할을 하며, 각 시간 t에 따른 Q값의 업데이트(변화) 추이를 알 수 있다.

[0043] 학습 초기인 t=0에서는 일반적으로 모든 Q값들이 0으로 초기화된다. 즉, Q-테이블은 처음(t=0)에 모두 0으로 초기화되어 있고 학습이 진행됨에 따라 업데이트된 Q값들이 저장된다. $Q(a,s_N)$은 상태 a에서 행동 a_N을 취했을 때의 Q값을 나타낸다.

[0044] Q값은 수학식 1의 Q-function에 따라 각각의 시간에 해당하는 상태에서 취할 수 있는 행동마다 다르게 업데이트된다. 또한 Q값은 보상(R)을 어떻게 정의하는지에 따라 달리 얻어진다. 즉, Q-러닝에서 Q값은 보상값 R에 의존하며 각 상태와 행동에 따른 보상 R을 정의한 R-matrix(R행렬)를 이용하여 업데이트 될 수 있다.

[0045] 이와 같이, Q-테이블 내에서 시간 t에서의 Q값들은 해당 시간에 임의 선택된 상태 s와 행동 a으로 인하여 얻어지는 보상 R을 통하여 얻어진다. 또한, 각 상태와 각 행동을 행과 열로 정의하게 되면 해당 Q값들은 하나의 Q-matrix(Q 행렬)로 표현 가능하다.

[0046] 그리고 수학식 1과 같은 Q-러닝 알고리즘을 시간에 따라 반복하게 되면 도 3과 같이 매 시간마다 Q값이 계속하여 업데이트된다. Q값은 현재 상태에서의 보상 값보다 다음 상태의 보상 값이 더 크면 업데이트된다. 이렇게 학습이 완료되면 최대 Q값을 얻을 수 있으며, 이를 이용하여 보상을 최대화하는 정책을 얻을 수 있다.

[0047] 여기서, 이러한 Q-러닝 방식은 학습을 진행하면서 최적의 Q값을 얻을 수 있지만 학습 과정에서 문제점이 존재한다면 준 최적의 값을 얻을 수 있다. 대표적인 문제점으로 탐사와 탐험 문제(Exploration and Exploitation)가 있다. 학습 과정에서 탐사만 일방적으로 진행할 경우 최적보다는 준 최적의 값을 가져올 확률이 높으므로, 이를 해결하기 위해 탐험을 진행해서 가끔 다른 행동을 취하게 함으로써 최적의 값을 찾을 수 있게 해주는 탐험 알고리즘을 사용한다.

[0048] 이러한 탐험 알고리즘은 대표적으로 ε-greedy 알고리즘이 있다. 이는 ε의 확률만큼 무작위 행동을 수행한다. 하지만 어느 정도 학습이 진행되었을 때도 ε의 확률만큼 무작위 행동을 한다면 이것이 오히려 문제가 될 수 있다. 따라서 본 발명의 실시예는 이를 해결하고자 학습이 진행될수록 ε의 값이 작아지는 Decaying ε-greedy 방식을 사용하여 학습을 진행한다.

수학식 3

[0049] $$a_t = \begin{cases} p, & \epsilon\text{의 확률일 경우} \\ \pi(s_t), & 1-\epsilon\text{의 확률일 경우} \end{cases}$$

[0050] 수학식 3은 ε의 확률만큼 무작위 행동을 하였을 때 얻을 수 있는 두 가지 행동에 대한 결과를 나타난다. p는

시간 t에서 임의로 행동한 결과를 나타내는 랜덤 변수이고, 1-ε의 확률로 행동하게 된다면 행동은 곧 곧 정책(π)이 된다. Q-러닝의 목적이 최적의 정책을 찾는 것이므로 위와 같은 행동이 지속된다면 최적의 정책을 구할 수 있다.

[0051] 도 3과 같이, 본 발명의 실시예는 에이전트가 Q-function에 따라서 학습하고 Decaying ε-greedy 알고리즘에 따라 행동을 하게 되면 환경은 그에 대한 결과로 다음 상태에 대한 정보와 보상 값을 준다. 이렇게 한번 행동을 취하면 Q-테이블에 Q-매트릭스 즉, 본 발명에서 제안한 보상 값에 의존한 Q값이 업데이트되어 저장된다. 이러한 과정이 반복되어 학습이 진행되고 학습이 완료되면 Q-값이 가장 큰 중계기가 최적의 중계기로 선정된다.

[0052] 도 4는 본 발명의 실시예에 따른 최적 중계기 선정 기법의 개념을 설명하는 도면이다. 도 4와 같이, 최적의 중계기 선정 기법은 Q-러닝을 위한 상태, 행동 그리고 보상(Reward)에 대한 파라미터를 정의하여 학습을 진행한다.

[0053] 이를 위해, 먼저 환경 설정에서 상태, 행동 그리고 보상 값을 어떻게 정의할지 결정한다(S410). 그리고, 현재 위치한 상태에서 학습을 시작하고 상태 및 행동을 업데이트하고(S420), 환경에 의해 업데이트된 상태와 보상 값을 관찰하고 저장한다(S430). 이러한 과정은 학습 종료 여부를 판단하여(S440), 학습이 종료되기 전까지 반복된다. 그리고 학습이 종료되면 업데이트된 Q-값을 이용하여 최적 중계기를 선정하여 협력 통신을 수행한다(S450).

[0054] 이러한 Q-러닝 기반의 중계기 선정 방식은 자가 학습을 통해 중계기를 선정하므로 기존의 중계기 선정 방식의 복잡도 및 연산량 그리고 오버헤드를 줄일 수 있다.

[0055] 이하에서는 본 발명의 실시예에서 제안하는 Q-러닝을 이용한 최적 중계기 선정 기법을 상세히 설명한다.

[0056] 도 5는 본 발명의 실시예에 따른 무선 협력 통신 시스템을 기반으로 하는 Q-러닝 기반의 중계기 선정 장치의 구성을 나타낸 도면이고, 도 6은 도 5를 이용한 최적 중계기 선정 방법을 설명하는 도면이다.

[0057] 도 5 및 도 6을 참조하면, 본 발명의 실시예에 따른 Q-러닝 기반의 중계기 선정 장치(100)는 설정부(110), 초기화부(120), 학습부(130) 및 결정부(140)를 포함한다.

[0058] 본 발명의 실시예에 따른 중계기 선정 장치(100)는 소스 단말에 포함되어 동작할 수도 있고 협력 통신 네트워크 내 존재하는 각 단말에 포함되어 중계기의 선택이 필요한 상황에서 동작하도록 구현될 수도 있다. 또한, 중계기 선정 장치(100)는 각 단말과 통신하는 기지국이나 액세스 포인트에 포함되어 구동할 수도 있다.

[0059] 먼저, 설정부(110)는 도 1과 같이 소스 단말, 목적지 단말, K개의 중계기를 포함한 N개 단말(N=K+2)이 존재한 환경에서 최적 중계기를 선정하도록 Q-러닝을 위한 상태(s), 행동(a) 및 보상(R)을 정의한다(S610).

[0060] 이를 위해, 설정부(110)는 N개 단말로부터 임의 선택된 n(n=1,…,N)번째 송신측 단말을 '상태'(s_n)로, 해당 상태(s_n)에서 임의 선택된 m(m=1,…,N)번째 수신측 단말을 '행동'(a_m)으로, m번째 수신측 단말에서의 SNR 지표를 해당 상태(s_n)에서 취한 행동(a_m)에 따른 '즉각적인 보상'(R; $R(s_n,a_m)$)으로 정의한다.

[0061] 이때, Q-러닝을 위한 '상태' 및 '행동'의 정의에서, 송신측 단말의 인덱스 n이 1에서 N까지 가능하다는 것은 N개 단말 모두가 Q 러닝 시에는 송신측 단말로 사용 가능하다는 것이고, 수신측 단말의 인덱스 m이 1에서 M까지 가능하다는 것 또한 N개 단말 모두가 Q 러닝 시에는 수신측 단말로 사용 가능하다는 것이다.

[0062] 이로부터 알 수 있는 것은, 실제 협력 통신 시에는 소스 단말(S)은 송신기 역할, 목적지 단말(D)은 수신기 역할, 그리고 중계기(R)는 소스 단말(S)로부터 받은 신호를 목적지 단말(D)로 전달하는 중계기 역할을 수행하지만, Q-러닝 시에는 이들 역할이 한정되지 않는다는 것이다.

[0063] 물론, 이러한 경우 실질적으로 협력 통신에서 불필요한 경로에 대한 상황도 Q-러닝 과정에 포함되지만, 본 발명의 실시예는 해당 상황에서의 즉각적인 보상(R)을 0으로 미리 세팅해 둠으로써, 추후에 해당 상황을 선택할 가능성을 상쇄시킨다. 이는 추후 도 8과 같은 N×N 크기의 R 행렬을 통하여 더욱 상세히 설명할 것이다.

[0064] 이하의 본 발명의 실시예에서, 총 N개 단말 중 n='N'인 단말과 m='1'인 단말은 소스 단말(S), n='N'인 단말과 m='N'인 단말은 목적지 단말(D), 나머지 n,m={2,…,N-1}인 단말은 K개의 중계기(R)를 의미하는 것을 가정한다.

[0065] 그리고, 본 발명의 실시예에서 정의된 Q-러닝의 상태와 행동을 기반으로, 추후 Q 러닝 과정에서 선택된 임의 상태 및 임의 행동은 다음의 수학식 4와 같이 간략히 나타내어 진다.

수학식 4

[0066] $$s_n = n \ \ (n = 1, \cdots, N)$$
$$a_m = m \ \ (m = 1, \cdots, N)$$

[0067] 수학식 4에 의하면, 소정 상태(s_n)에서 취한 소정 행동(a_m)에 따른 즉각적인 보상 R값인 $R(s_n,a_m)=R(n,m)$으로 표현 가능하다. 물론, 해당 상태와 행동에 대응한 Q 값 역시 $Q(s_n,a_m)=Q(n,m)$로 표현될 수 있다.

[0068] 본 발명의 실시예는 최적 중계기를 선정하기 위한 즉각적인 보상 R을 구체적으로 제안한다. 즉각적인 보상 R을 어떻게 정의하는지에 따라 Q값이 바뀌게 되고 정책이 바뀌게 되어 선정되는 중계기가 달라지기 때문이다.

[0069] 이러한 '즉각적인 보상' R은 아래 수학식 5와 같이 m번째 수신측 단말에서의 SNR을 기반으로 결정될 수 있다.

수학식 5

[0070] $$R = \frac{SNR_m}{d\rho}$$

[0071] 여기서, SNR_m는 m번째 수신측 단말에서의 SNR 값이고, d는 n번째 송신측 단말과 m번째 수신측 단말 사이의 거리이고, ρ는 자유 공간 경로 손실을 나타낸다.

[0072] 이러한 즉각적인 보상 R은 어떤 상태 s_n에서 어떠한 행동 a_m을 취한 경우에 대한 보상이므로, 임의 선택된 상태와 임의 선택된 행동의 모든 조합을 고려하면 총 N×N 가지 존재한다.

[0073] 따라서 N×N 가지의 즉각적인 보상은 수학식 6과 같이 하나의 R 행렬(R_M)로 표현될 수 있다.

수학식 6

[0074] $$R_M = \begin{bmatrix} R(1,1) & R(1,2) & \cdots & R(1,N) \\ R(2,1) & R(2,2) & \cdots & R(2,N) \\ \vdots & \vdots & \ddots & \vdots \\ R(N,1) & R(N,2) & \cdots & R(N,N) \end{bmatrix}$$

[0075] 이와 같이, 수학식 6은 선택된 상태와 행동의 조합에 따라 수학식 5를 통해 연산한 N×N 개의 R값 원소들로 구성된다. 간단히, $R_M=\{R(1,1),\cdots,R(n,m),\cdots,R(N,N)\}$로 표현될 수 있다.

[0076] 앞서 설명한 바와 같이, $R(s_n,a_m)=R(n,m)$로 정의되므로, 예를 들어 R(1,2)는 n=1, m=2로서 상태 s_1에서 행동 a_2을 취하여 상태 s_2로 이동한 경우에 대한 즉각적인 보상 $R(s_2,a_2)$을 나타낸다.

[0077] 여기서, n='1'인 단말은 소스 단말(S)이고, m='2'인 단말은 K개의 중계기 중에서 1번 중계기(R_1)에 해당한다. 따라서, R(1,2)이란, 소스 단말(S)에서 1번 중계기(R_1)로 신호를 송신할 경우에, 1번 중계기(R_1)의 수신 신호로부터 얻은 수학식 5에 따른 보상 값을 나타낸다. 물론, R(1,2) 연산 시에는, 수학식 5에 소스 단말(S)과 제1 중계기(R1) 간의 거리, 자유공간 경로 손실, 그리고 수신 신호의 SNR 값이 대입된다.

[0078] 다음은 도 7과 같이 소스 단말, 목적지 단말, 그리고 4개의 중계기를 포함하는 K=4, N=6 인 협력 통신 시스템에서의 R 행렬을 구성하는 방법을 설명한다.

[0079] 도 7은 K=4, N=6인 협력 통신 시스템에서 최적 중계기를 선정하는 예시를 설명하기 위한 도면이다.

[0080] 도 7에서 소스 단말-목적지 단말 간 거리 및 서로 다른 중계기 간 수직 거리는 모두 1로 정규화되어 있다고 가

정한다. 이때 상태 s_1 부터 s_N 까지 각 상태에서 취할 수 있는 행동은 원 상태(원래 상태)로 돌아오는 행동을 포함하여 a_1 부터 a_N 까지 존재한다.

[0081] 이하에서는 도 7의 시스템에서 1번 중계기(R1), 즉 상태 s_2에 해당하는 중계기가 최적 중계기로 선택 되어지는 과정에 대한 실시예를 들어본다.

[0082] 도 8은 도 7의 시스템에서 본 발명의 실시예의 기법으로 획득한 보상 행렬을 예시한 도면이다.

[0083] 여기서, 소스 단말-중계기-목적지 단말 간 수신 SNR은 0부터 5까지 임의로 분포시키고, 소스 단말-중계기, 중계기-목적지 단말 간 거리는 0부터 1까지 임의로 분포시키며, 편의상 중계기-중계기 간 거리는 0.2로 가정하며, 자유공간 경로손실 $\rho=2$로 설정하였다.

[0084] 도 8과 같이 행렬 R은 R(1,1) 부터 R(N,N) 까지의 N×N 개의 보상 값들을 원소로 함을 알 수 있다. 이때, 송신측 단말과 수신측 단말이 동일한 경우(n=m=i)의 보상값 R(i,i)과, 소스 단말과 목적지 단말 간 직경로에 해당한 경우의 보상값 R(1,N), 그리고 목적지 단말(n=N)이 송신측인 경우의 보상값 R(N,i)은 아래의 수학식 7과 같이 모두 '0'의 값으로 설정된 것을 알 수 있다. 이때, i={1,…N}이다.

수학식 7

[0085] $$R(i,i)=R(1,N)=R(N,i)=0$$

[0086] 이와 같이, 소스 단말-목적지 단말 간 직경로, 원 상태로 돌아오는 행동, 목적지 단말에서 다시 다른 중계기 및 소스 단말로 돌아오는 행동에 대한 보상값으로 0을 부여한다.

[0087] 그리고, 중계기(n=2,…,N-1)가 송신측이고 목적지 단말(n=N)이 수신측인 경우의 보상값 R(i,N)(이때, i≠1,N)은, 수학식 5의 R 값에 설정 가중치(예를 들어, 100)가 추가로 가산된 것을 알 수 있다. 즉, 목적지 단말에 도착하는 중계기에 대해 가중치 100을 추가로 부가한다.

[0088] 상술한 바와 같이 보상 값을 정의하는 이유는 다음과 같다. 본 실시예에서 정의한 Q-러닝의 '상태' 및 '행동'은 시스템 내의 모든 N개 단말을 대상으로 하며, 이를 통해 학습을 수행할 경우, 실제 협력 통신에서는 전혀 불필요한 경로 즉, 동일 단말에서 동일 단말로 신호가 전송되는 상황(소스 단말→소스 단말, 중계기→중계기, 목적지 단말→목적지 단말)과, 중계기 경유 없이 소스 단말에서 목적지 단말로 직접 신호가 전송되는 상황(소스 단말→목적지 단말), 그리고 목적지 단말에서 소스 단말이나 중계기로 신호가 전송되는 상황(목적지 단말→소스 단말, 목적지 단말→중계기)도 Q-러닝 과정에 포함되게 된다. 여기서, 본 실시예의 경우 상술한 불필요한 상황들에서의 즉각적인 보상(R)을 0으로 미리 세팅해 둠으로써 추후 해당 상황을 선택할 가능성을 상쇄시킬 수 있다.

[0089] 본 발명의 실시예는 상술한 바와 같은 방법으로 Q-러닝을 위한 상태(s), 행동(a) 및 즉각적인 보상(R)을 각각 정의해 두고, 상술한 보상 값을 전제하여 학습을 진행시킨다.

[0090] 이후, 초기화부(120)는 Q-러닝에 앞서, 상태 및 행동에 각각 대응하여 행과 열 성분을 구성한 N×N 크기의 Q 행렬 내 원소들인 Q값들을 초기화한다(S620).

[0091] 도 9는 본 발명의 실시예에서 Q-러닝 전의 초기화된 Q 행렬을 예시한 도면이다.

[0092] 이러한 도 9는 도 7의 K=4, N=6인 시스템에 대한 초기화된 Q 행렬을 나타낸다. 상태 및 행동은 N개 존재하며 학습이 시작되기 전에 Q행렬은 0으로 초기화 되어 있다.

[0093] 도 9의 Q 행렬에서 N개의 상태(s_1~s_6)는 행 성분에 대응하고, N개의 행동(a_1~a_6)은 열 성분에 대응한다. 따라서 Q 행렬 내 원소들은 N개의 상태(s_n)와 행동(a_m) 각각의 조합에 대응하여 Q(1,1) 부터 Q(N,N) 까지의 N×N 개의 Q값들을 원소로 한다. 수학식 4의 원리에 따르면, Q 행렬은 간단히 Q={Q(1,1),…,Q(n,m),…,R(N,N)}로 표현된다.

[0094] 다음, 학습부(130)는 N개의 단말 중에서 임의 선택된 상태(s_n)와 행동(a_m)에 대한 즉각적인 보상($R(s_n,a_m)$)과, 현재 취한 행동(a_m)으로 인해 이동한 미래 상태(s_n')의 선택 가능한 모든 행동(a_m')에 대응하는 Q값들 중 최대치

($\max(Q(s_n',a_m'))$)를 이용하여 Q-러닝을 수행하여, Q 행렬 내의 $Q(s_n,a_m)$ 값을 업데이트한다(S630).

[0095] 여기서, Q 행렬 내의 $Q(s_n,a_m)$ 값은 아래 수학식 8에 의해 업데이트된다. 이는 수학식 1에 기반한 것이다.

수학식 8

[0096] $$New\ Q(s_n,a_m) = Q(s_n,a_m) + \alpha[R(s_n,a_m) + \gamma \cdot \max Q(s_n',a_m') - Q(s_n,a_m)]$$

[0097] 여기서, New $Q(s_n,a_m)$는 업데이트된 $Q(s_n,a_m)$ 값, α ($0<\alpha<1$)는 학습률, γ ($0<\gamma<1$)는 할인 계수(discount factor)를 나타낸다.

[0098] 그리고 학습부(130)는 상태 및 행동의 임의 선택을 통해 Q-러닝을 반복 수행하면서 Q 행렬을 지속 업데이트한다(640). 여기서 물론, 학습부(130)는 학습 과정에서 상태와 행동의 임의 선택 시에 Decaying ε-greedy 알고리즘을 적용하여 시간이 경과할수록 무작위 선택 행동의 확률을 감소시키도록 한다.

[0099] 이처럼, 학습부(130)는 매 시간마다, 상태와 행동의 임의 선택에 따른 즉각적인 보상 $R(s_n,a_m)$과, 현재 취한 행동으로 인해 발생 가능의 선택 가능한 모든 행동으로부터 도출한 예측값 $\max(Q(s_n',a_m'))$을 이용하여 Q값을 업데이트한다. 또한 매 시간별로 Q-러닝을 통해 업데이트된 Q 행렬은 Q-Table에 각 시간에 따라 저장된다.

[0100] 이후, 결정부(140)는 업데이트가 완료된 Q 행렬 내에서 최대 Q값을 탐색하고, 탐색한 최대 Q값에 대응된 하나의 중계기를 최적 중계기로 선정한다(S650). 이에 따라, 소스 단말(S)과 목적지 단말(D)은 선정된 선택된 최적 중계기를 이용하여 협력 통신을 수행한다(S660).

[0101] 도 10은 본 발명의 실시예에 따라 Q 행렬을 업데이트하여 최적 중계기를 선정하는 과정을 예시한 도면이다.

[0102] 초기에 Q 행렬은 0으로 초기화되어 있다. 학습을 시작하면 무작위 행동을 진행하게 된다. 예를 들어, 도 10의 왼쪽 아래 그림처럼 상태 s_1에서 시작하여 상태 s_5로 이동하는 행동 a_5에 대한 Q-값을 업데이트 하게 된다.

[0103] 수학식 8과 도 10을 이용하면, 상태 s_1와 행동 a_5에 대응한 Q값 즉, $Q(s_1,a_5)$는 아래 수학식 9과 같이 계산될 수 있다. 이때, α=0.8, γ=0.9로 가정하였다.

수학식 9

[0104] $$\begin{aligned} New\ Q(s_1,a_5) &= Q(s_1,a_5) + 0.8\times[R(s_1,a_5) + 0.9\times\max(Q(s_5,a_1),\cdots,Q(s_5,a_6)) - Q(s_1,a_5)] \\ &= 0 + 0.8\times((1.666 + 0.9\times 0) - 0) \\ &= 1.3333 \end{aligned}$$

[0105] 상태 s_1에서 시작하여 상태 s_5로의 Q값($Q(s_1,a_5)$)은 현재의 보상 값 $R(s_1,a_5)$, 그리고 다음 상태 s_5에서 가질 수 있는 모든 행동($a_1,a_2,\cdots,a_6$)에 대한 Q값들 ($Q(s_5,a_1),Q(s_5,a_2),\cdots,Q(s_5,a_6)$) 중에서 최대 Q값을 통하여 업데이트 된다.

[0106] 초기 Q값 들은 모두 0이다. 즉, 도 9에서 $Q(s_1,a_5)=Q(1,5)$이므로 $Q(s_1,a_5)$은 0이고, 마찬가지로 $Q(s_5,a_1)$ 내지 $Q(s_5,a_6)$ 값도 모두 0이므로 $\max(Q(s_5,a_1),\cdots,Q(s_5,a_6))=0$이 된다. 그리고, $R(s_1,a_5)$ 값은 도 8에서 $R(1,5)=1.666$임을 알 수 있다.

[0107] 따라서, 수학식 9에 해당 값들을 대입하면 $Q(s_1,a_5)$ 값은 1.3333으로 업데이트된다. 즉, New $Q(s_1,a_5)=1.3333$이며 업데이트된 Q 행렬은 도 9의 왼쪽 아래 그림과 같다.

[0108] 이와 같이, 초기 Q값은 모두 0이므로 첫 번째 행동에서의 현재 Q-값과 다음 상태에서 가능한 모든 행동에 대한 Q값은 0이 된다. 따라서 이때의 Q값은 α와 보상 값 R에 대해서만 의존하여 업데이트 된다.

[0109] 이러한 Q-러닝 과정을 반복하여 Q-값이 업데이트 되면 오른쪽 아래 그림처럼 발전되며, 최종적으로는 오른쪽 위

그림처럼 수렴하게 된다. 오른쪽 위 그림은 학습이 완료되었을 때의 Q 행렬을 나타낸다.

[0110] 정확화게는 오른쪽 위 그림은 최대 Q값에 대해 100이 나오도록, 최대 Q값을 이용하여 Q값들을 정규화한 결과에 해당한다. 그 결과, 상태 s_2에서 s_6 으로 가는 행동에 대한 Q값인 Q(2,6)이 최대화 된 것을 확인할 수 있다.

[0111] 여기서, Q(2,6)의 경우, n=2, m=6으로서, 인덱스 n=2인 1번 중계기(R_1)에서 인덱스 m=6인 목적지 단말(D)로 신호를 전송하는 경우에 대응한다. 이를 통해 목적지 단말(D)에 도달하기까지 거치는 중계기 중 1번 중계기(R_1)를 거치는 것이 최적이라는 것을 확인할 수 있다. 이에 따라, 소스 단말과 목적지 단말은 1번 중계기를 이용하여 협력 통신을 수행하면 된다.

[0112] 본 발명에서 제안한 기법과 기존의 기법의 BER(Bit Error Rate) 성능 비교를 위한 모의 실험을 실시하였다.

[0113] 모의실험은 OFDM 시스템 기반의 DF 기법의 협력통신을 수행하였고 128개의 부반송파를 사용하였다. 보호구간(CP; Cyclic Prefix)의 길이는 32이고 변조 방식은 QPSK(Quadrature Phase Shift Keying) 방식을 사용하였다. 채널은 7개의 다중 경로를 가지는 Rayleigh 페이딩 채널로 실시하였으며 소스-데스티네이션 간 거리는 1로 정규화 하였고 각 중계기는 랜덤하게 분포시켰다. 전송 전력은 모두 1로 정규화 했으며 할인 계수(discount factor)는 0.8, 학습률(learning rate)은 0.8, epsilon ε은 0.9로 정의하였다.

[0114] 도 11은 본 발명의 기법과 기존의 기법의 SNR 대비 BER 성능 그래프를 나타낸 도면이다.

[0115] 도 11을 통해, 기존 기법 중에서 무작위 중계기 선정 기법(Random Relay Selection)은 랜덤하게 중계기를 선정하기 때문에 성능이 낮은 것을 확인 할 수 있다. 또한, 문턱값 기반 중계기 선정 기법(Threshold-based Relay Selection)은 문턱값보다 높은 중계기를 선정하여 협력통신을 수행하기 때문에 무작위 중계기 선정 기법에 비해서는 성능이 좋다. 조화평균 중계기 선정 기법(Best Harmonic Mean(BHM) Relay Selection)은 모든 채널 상태 정보를 고려하여 최적의 중계기를 선정하므로 기존 기법 중에서 성능이 가장 좋다.

[0116] 본 발명의 실시예에 따른 기법의 성능은 기존의 조화평균 중계기 선정 기법과 성능이 동일한 것을 알 수 있다. 그런데 조화평균 중계기 선정 기법은 채널 상태 정보를 송신기로 피드백 해야하는 오버헤드의 부담이 있고, 조화 평균을 모든 중계기에서 구하는 과정에서 연산량의 복잡도가 높아지므로 현실적으로 불가능한 단점이 있다.

[0117] 반면에 본 발명의 실시예에 따른 Q-러닝 기반의 중계기 선정 기법은 Q-러닝 기반의 자가 학습을 통해 중계기를 선정하므로 기존의 중계기 선정 방식의 복잡도 및 연산량 그리고 오버헤드 문제를 줄일 수 있다.

[0118] 이상과 같은 본 발명에 따르면, 복수의 중계기가 존재하는 환경에서 Q-러닝 알고리즘을 기반으로 최적의 중계기를 선정함으로써 높은 신뢰성을 얻을 수 있으며 비트 오율 성능을 향상시키고 복잡도를 줄일 수 있다. 또한, 본 발명에서 제안한 기법은 기존의 기법과 동일한 성능을 얻으면서 구현 복잡도 및 연산량과 오버헤드를 줄일 수 있다.

[0119] 본 발명은 도면에 도시된 실시 예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 다른 실시 예가 가능하다는 점을 이해할 것이다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 특허청구범위의 기술적 사상에 의하여 정해져야 할 것이다.

부호의 설명

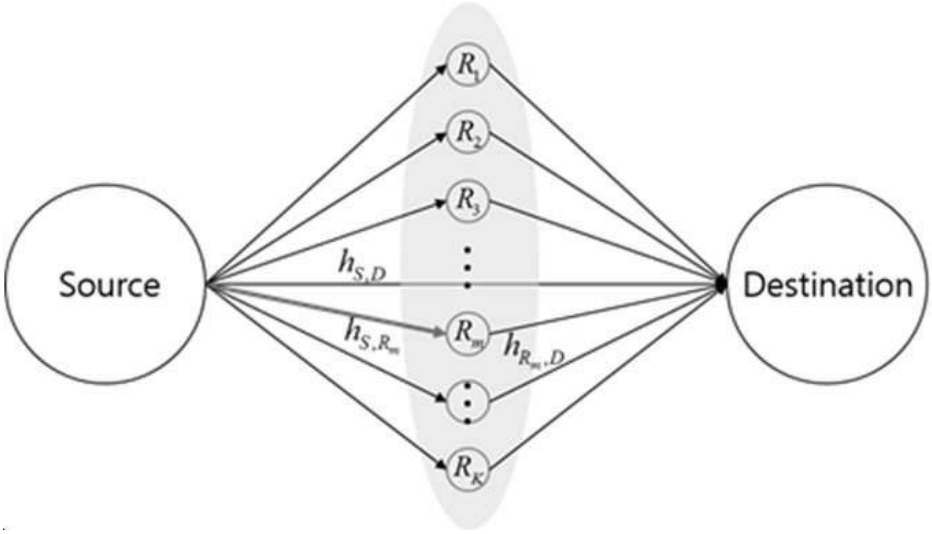
[0120] 100: 중계기 선정 장치 110: 설정부

120: 초기화부 130: 학습부

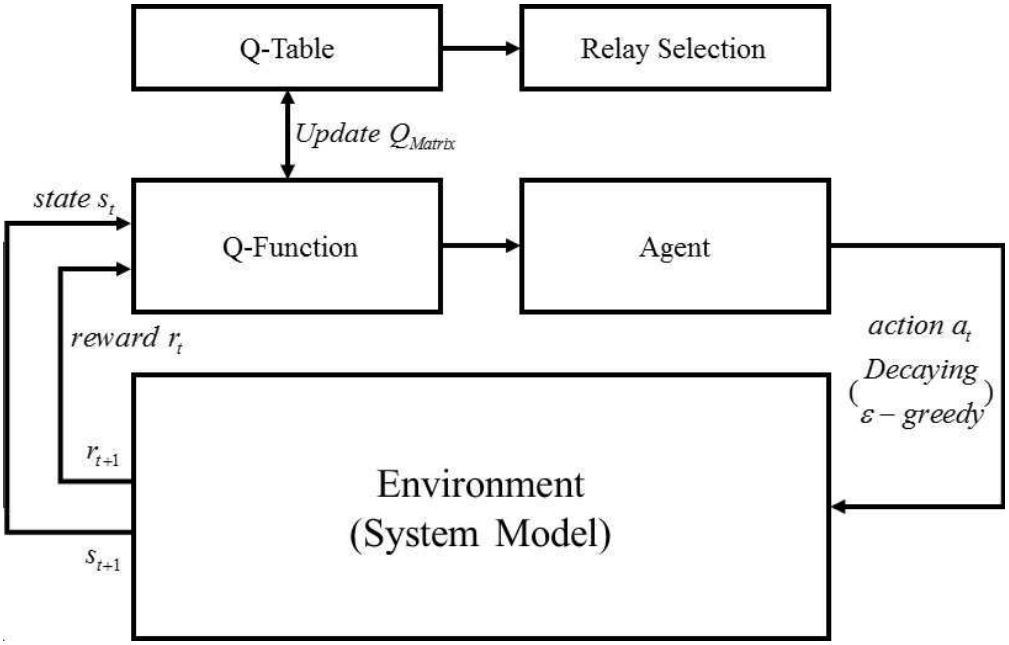
140: 결정부

도면

도면1



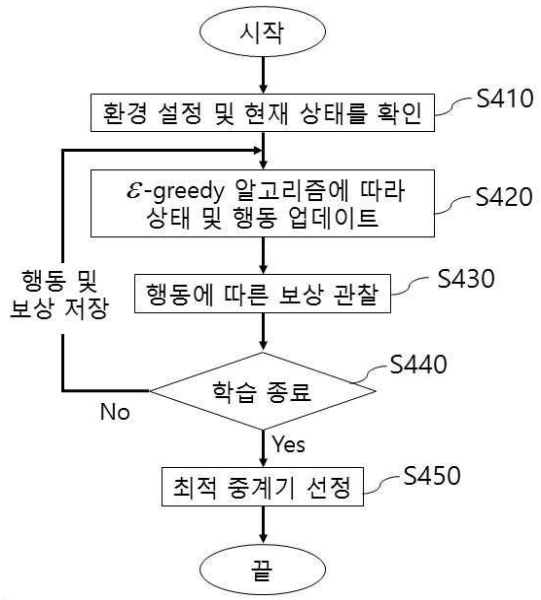
도면2



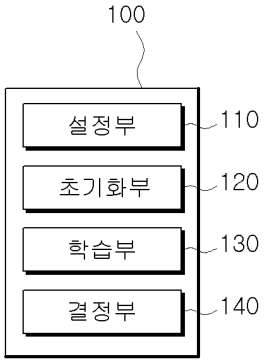
도면3

Time	$Q(s,a_1)$	⋯	$Q(s,a_m)$	⋯	$Q(s,a_N)$
0	0	⋯	0	⋯	0
.	.	⋯	.	⋯	.
.	.	⋯	.	⋯	.
.	.	⋯	.	⋯	.
t	$Q_t(s,a_1)$	⋯	$Q_t(s,a_m)$	⋯	$Q_t(s,a_N)$
$t+1$	$Q_{t+1}(s,a_1)$	⋯	$Q_{t+1}(s,a_m)$	⋯	$Q_{t+1}(s,a_N)$

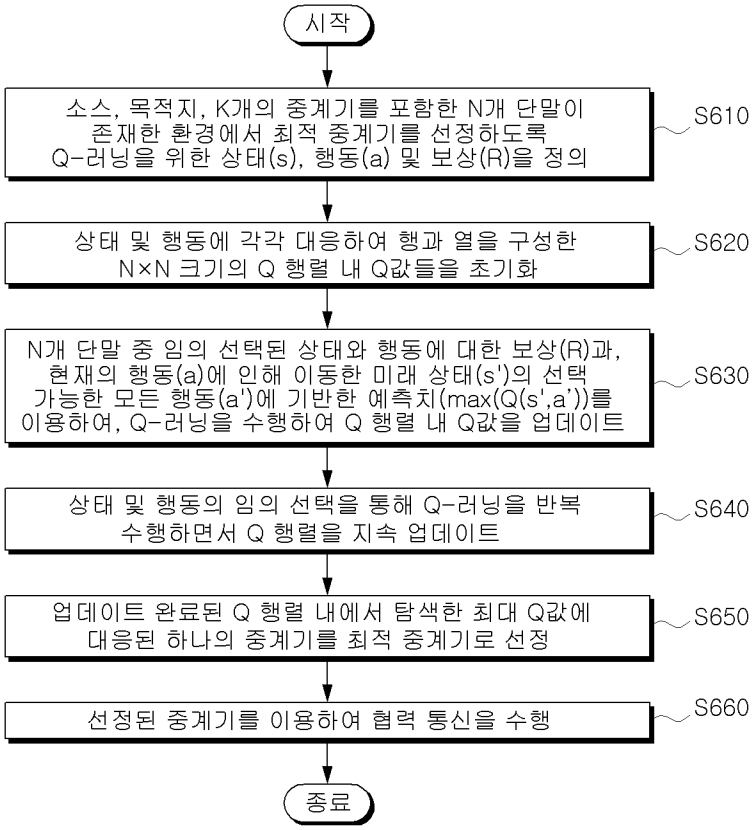
도면4


시작
환경 설정 및 현재 상태를 확인
S410
ε-greedy 알고리즘에 따라
상태 및 행동 업데이트
S420
행동 및
보상 저장
행동에 따른 보상 관찰
S430
학습 종료
S440
No
Yes
최적 중계기 선정
S450
끝

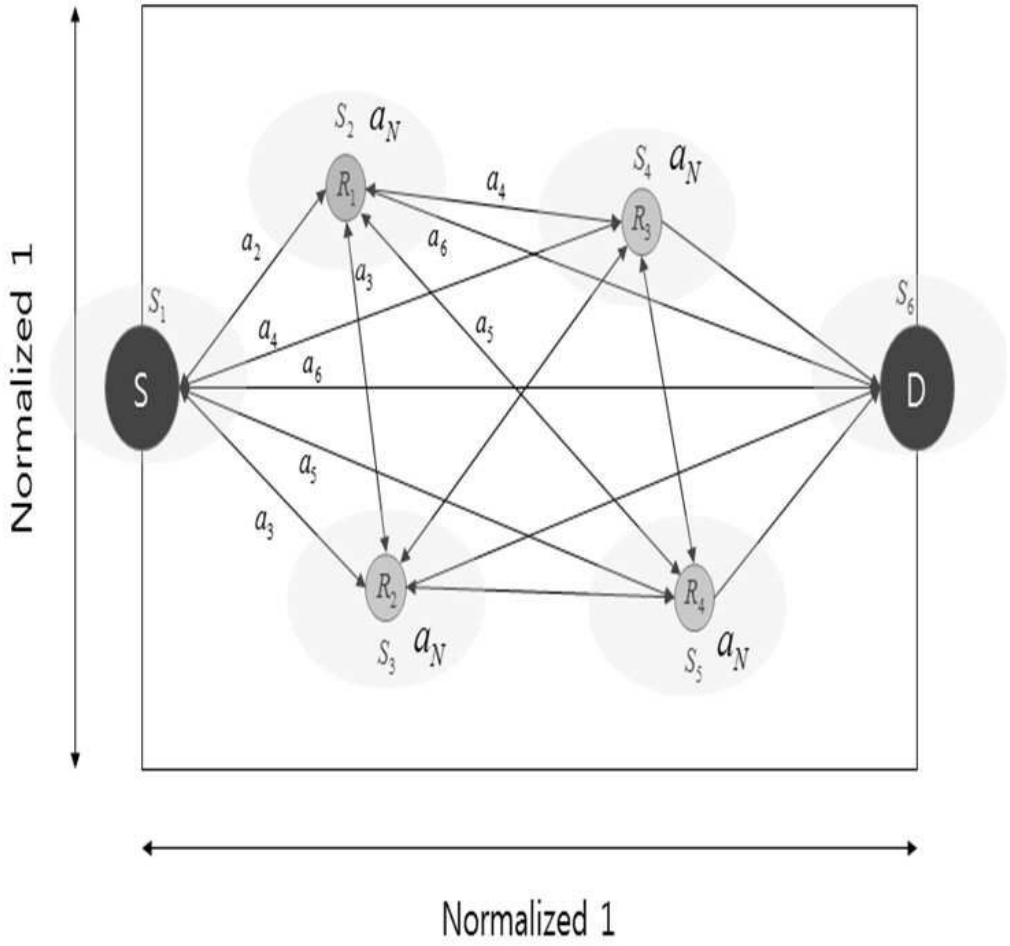
도면5


100
설정부
110
초기화부
120
학습부
130
결정부
140

도면6


시작
소스, 목적지, K개의 중계기를 포함한 N개 단말이 존재한 환경에서 최적 중계기를 선정하도록 Q-러닝을 위한 상태(s), 행동(a) 및 보상(R)을 정의
S610
상태 및 행동에 각각 대응하여 행과 열을 구성한 N×N 크기의 Q 행렬 내 Q값들을 초기화
S620
N개 단말 중 임의 선택된 상태와 행동에 대한 보상(R)과, 현재의 행동(a)에 인해 이동한 미래 상태(s')의 선택 가능한 모든 행동(a')에 기반한 예측치(max(Q(s',a'))를 이용하여, Q-러닝을 수행하여 Q 행렬 내 Q값을 업데이트
S630
상태 및 행동의 임의 선택을 통해 Q-러닝을 반복 수행하면서 Q 행렬을 지속 업데이트
S640
업데이트 완료된 Q 행렬 내에서 탐색한 최대 Q값에 대응된 하나의 중계기를 최적 중계기로 선정
S650
선정된 중계기를 이용하여 협력 통신을 수행
S660
종료

도면7


Normalized 1
Normalized 1
S
D
R_1
R_2
R_3
R_4
S_1
S_2 a_N
S_3 a_N
S_4 a_N
S_5 a_N
S_6
a_2
a_3
a_4
a_5
a_6

도면8

$$R=\begin{bmatrix} 0 & \frac{SNR_{S_{12}}}{d\rho} & \frac{SNR_{S_{13}}}{d\rho} & \frac{SNR_{S_{14}}}{d\rho} & \frac{SNR_{S_{15}}}{d\rho} & 0 \\ \frac{SNR_{S_{12}}}{d\rho} & 0 & \frac{SNR_{S_{23}}}{d\rho} & \frac{SNR_{S_{24}}}{d\rho} & \frac{SNR_{S_{25}}}{d\rho} & \frac{SNR_{S_{26}}}{d\rho} \\ \frac{SNR_{S_{13}}}{d\rho} & \frac{SNR_{S_{23}}}{d\rho} & 0 & \frac{SNR_{S_{34}}}{d\rho} & \frac{SNR_{S_{35}}}{d\rho} & \frac{SNR_{S_{36}}}{d\rho} \\ \frac{SNR_{S_{14}}}{d\rho} & \frac{SNR_{S_{24}}}{d\rho} & \frac{SNR_{S_{34}}}{d\rho} & 0 & \frac{SNR_{S_{45}}}{d\rho} & \frac{SNR_{S_{46}}}{d\rho} \\ \frac{SNR_{S_{15}}}{d\rho} & \frac{SNR_{S_{25}}}{d\rho} & \frac{SNR_{S_{35}}}{d\rho} & \frac{SNR_{S_{45}}}{d\rho} & 0 & \frac{SNR_{S_{56}}}{d\rho} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\Rightarrow R=\begin{bmatrix} 0 & \frac{2}{0.8\times 2} & \frac{1.5}{0.7\times 2} & \frac{2.5}{0.4\times 2} & \frac{1}{0.3\times 2} & 0 \\ \frac{2}{0.8\times 2} & 0 & \frac{1}{0.2\times 2} & \frac{0.5}{0.2\times 2} & \frac{0.7}{0.2\times 2} & \frac{4}{0.2\times 2}+100 \\ \frac{1.5}{0.7\times 2} & \frac{1}{0.2\times 2} & 0 & \frac{0.3}{0.2\times 2} & \frac{1}{0.2\times 2} & \frac{3.5}{0.3\times 2}+100 \\ \frac{2.5}{0.4\times 2} & \frac{0.5}{0.2\times 2} & \frac{0.3}{0.2\times 2} & 0 & \frac{0.1}{0.2\times 2} & \frac{5}{0.6\times 2}+100 \\ \frac{1}{0.3\times 2} & \frac{0.7}{0.2\times 2} & \frac{1}{0.2\times 2} & \frac{0.1}{0.2\times 2} & 0 & \frac{4.5}{0.7\times 2}+100 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

도면9

s \ a	a_1	a_2	a_3	a_4	a_5	a_6
s_1	0	0	0	0	0	0
s_2	0	0	0	0	0	0
s_3	0	0	0	0	0	0
s_4	0	0	0	0	0	0
s_5	0	0	0	0	0	0
s_6	0	0	0	0	0	0

$Q=$

도면10

$$Q=\begin{bmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix}$$

$$Q=\begin{bmatrix}82.0227&91.1364&87.5649&88.0682&85.9632&0\\83.1591&90.0000&88.8636&86.3636&86.0390&\boxed{100}\\82.9968&92.2727&86.5909&85.9091&86.7208&96.2121\\84.8636&91.1364&87.2727&85.2273&84.6753&94.6970\\83.5379&91.5909&88.8636&85.4545&84.4481&93.8312\\0&0&0&0&0&0\end{bmatrix}$$

$$Q=\begin{bmatrix}0&0&0&0&1.3333&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix}$$

$$Q=\begin{bmatrix}0&26.1613&0&95.8560&44.9728&0\\0&0&66.5324&0&45.0816&0\\79.9982&0&32.6684&0&45.3728&84.6667\\0&0&0.6000&0&43.5728&103.3333\\70.3497&34.3611&41.2379&60.2400&0&0\\0&0&0&0&0&0\end{bmatrix}$$

도면11