강화학습 기법을 이용한 최적경로 탐색

구다솔*, 이태경*
*동국대학교 컴퓨터과학과
e-mail: ekthf0201@dongguk.ac.kr

Optimal Path Search using Reinforcement Learning Technique

Da-Sol Gu*, Tae-Kyung Lee*
*Dept. of Computer Science, Dong-guk University

요 약

본 논문에서는 사용자로부터 실시간으로 전송 받은 교통정보 이용하여 강화학습에 의한 최적 경로 탐색을 제안한다. ITS(Intelligent Transportation Systems)를 서비스하기 위한 시스템을 구축하기에는 많은 시간적 비용과 물질적 비용이 소모된다. 이를 보완하기 위해 사용자의 단말기로부터 실시간으로 수집한 교통 정보를 이용하여 강화학습기법을 적용한다. 강화학습의 목표는 환경 내에서의 에이전트가 행동에 대한 보상의 총합을 최대화 하는 것이다. 본 논문에서는 실시간으로 사용자의 단말기로부터 획득한 교통 정보를 이용하여 강화학습기법을 적용하고, 최단경로탐색 알고리즘을 분석하여 비교한다.

1. 서론

오늘날 대부분의 자동차나 스마트폰 등에는 내비게이션이 탑재되어있다. 초창기 내비게이션은 사용자의 현재 위치에서 목적지까지의 최단경로를 탐색하는 단순한 기능에 국한되었다. 최근에는 기존에 제공하던 최단경로 탐색과 교통 정보를 수집하여 교통 상황에 따른 최적 경로를 추천하는 기능을 제공하기도 한다. 대표적인 예로 ITS(Intelligent Transportation Systems) 서비스를 들 수 있다. 지능형 교통 체계(ITS)는 효과적인 교통체계를 실현하기 위해 필요한 기반을 제공하며, 첨단기술을 활용하여 기존의 교통체계를 좀 더 효율적으로 사용하거나 새로운 교통서비스를 제공함으로써 교통문제를 해결하는데 목적을 두고 있다[1][2].

하지만 ITS 서비스를 제공하기 위한 시스템을 구축하는데 물질적으로나 시간적으로 많은 비용이 소모되고, 교통수집 장치가 설치되어있지 않은 도로는 최적 경로 탐색에 취약하다는 단점이 있다. 이를 보완할 방법으로 내비게이션을 이용하는 사용자로부터 실시간으로 교통정보를 수집하여 서비스를 제공하는 방법이다. 사용자로부터 교통정보에 관련된 데이터를 직접 획득함으로써 교통 수집 장치 구축에 드는 비용을 줄일 수 있다. 교통정보를 수집하여 사용자에게 제공하기 위해서는 새로운 정보를 학습하고, 지속적 학습으로 오류를 줄여 합리적으로 사용하기 위해서는 기계학습이 필요하다.

기계학습은 새로운 지식을 반복적으로 학습하여, 학습한 데이터를 효율적으로 사용하는 기술을 말한다. 경험을 통한 환경 적응을 위해서는 반드시 학습이 필요하다.

강화학습 알고리즘은 그 에이전트가 앞으로 누적 될 보상을 최대화 하는 일련의 행동으로 정책을 찾는 방법이다. 즉, 환경을 탐색하는 에이전트가 현재의 상태를 인식하여 어떤 행동을 취하면 에이전트는 정해진 보상을 얻는다[3].

본 논문에서는 강화학습(Reinforcement Learning) 기법을 이용하여 실시간으로 사용자의 단말기로부터 획득한 교통정보로 사용자가 출발지에서 목적지까지의 최적 경로를 추천 해주는 알고리즘을 제안하고, 최단 경로 알고리즘과 분석하여 비교한다.

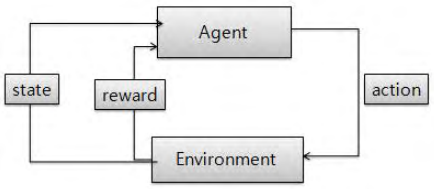
2. 관련 연구

2.1 최단경로탐색 알고리즘

Dijkstra 알고리즘은 최단경로 탐색 알고리즘 중 대표적인 알고리즘 중 하나이다. 이 알고리즘은 그래프 상의 출발점에서 모든 지점까지의 최단 거리를 구하는 알고리즘이다. Dijkstra 알고리즘은 간선의 가중치가 모두 양수일 때 유효하다. 모든 간선의 가중치가 음이 아닐 때, 방향 그래프 G=(V , E)에서 단일 출발점에서 최단 경로를 해결하는 알고리즘이다. 방향 그래프의 V는 정점, E는 간선을 나타내는데, 가까운 정점부터 차례로 모든 정점에 대한 간선들의 가중치를 고려하여 하나의 간선을 선택하고, 이 간선에 연결된 정점을 추가하는 과정을 반복한다. 최단 경로를 찾는 방법은 가중치가 가장 가까운 정점을 선택하는 탐욕적(Greedy) 전략을 사용한다[4][9].

2.2 강화학습 (Reinforcement Learning)

강화학습은 인간이나 동물의 학습을 모방하여 작동 환경에 대한 학습모형을 사용하지 않는 대표적 기계학습 방법이다. 에이전트가 어떠한 행동을 하였을 때, 좋은 피드백을 받을 경우 그 행동을 더 강화하고, 나쁜 피드백을 받을 경우 해당 행동을 하지 않도록 훈련하는 것이다. 강화학습은 액션을 수행하게 되면 환경으로부터 그에 대한 평가 피드백이 돌아오고 다음 상태로 이전하게 된다.



(그림 1) 에이전트와 환경의 상호작용

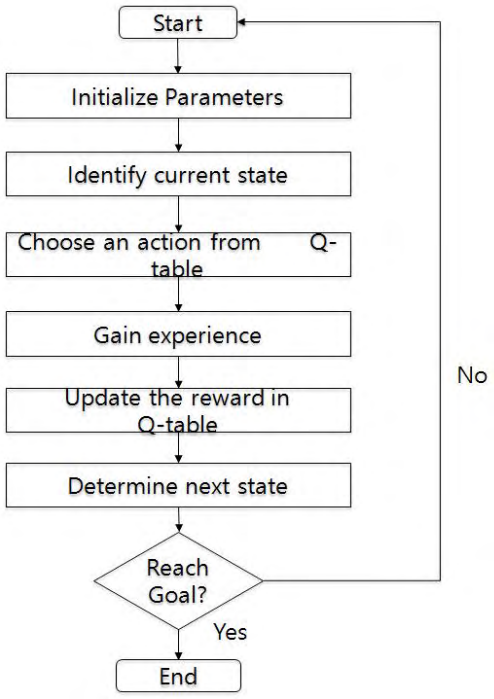
정책(Policy)은 특정 상태가 주어졌을 때 어떤 행동(Action)을 취할 것인지를 정해준다. 강화학습은 상태(State)와 행동(Action), 보상(reward)의 개념을 이용하면 알고리즘을 통해 각 상황의 예측 값을 결정 할 수 있다. 강화학습에서 가장 중요한 요소는 보상 함수를 적절히 설정하는 것이다. 보상 함수를 어떻게 설정 하느냐에 따라 학습의 효율이 달라진다[5].

2.3 Q-learning

Q-learning은 마르코프 의사 결정 과정(MDP)을 기반으로 한 강화학습의 off-policy 기법 중의 하나이다. Q-learning의 중요한 요소는 State(상태)와 Action(행동)이다. Q-learning 알고리즘은 (그림 2)과 같다.

$$s,a) \leftarrow (1-\alpha)*Q(s,a)+\alpha[r+\gamma \max_{,} Q(s',a')] \quad (1)$$

위 식(1)에서 r은 보상, s는 state로 현재의 상태를 나타내며, a는 Action으로 어떤 상태에서 수행 가능한 행동을 나타낸다. 상태와 행동은 R(s,a)과 Q(s,a)로 나타낸다. 여기서 R은 보상(Reward)을 나타내고, R(s,a)은 상태에서 a라는 행동을 실행하였을 때 보상을 나타낸다. Q(s,a)는 상태 s에서 a라는 action을 실행 하였을 때 Q-value값이다. α는 에이전트의 학습 율을 나타낸다. 0≤ α ≤1의 범위를 가지고 학습 속도에 영향을 미치는 파라미터이다. 값의 크기에 따라 학습의 속도에 영향을 미치며, 너무 작은 값이든, 큰 값이든 값의 크기에 따라 학습이 제대로 진행되지 않을 수 있다[6]. Q-learning에서의 행동 선택을 위해서 생성된 Q값 중 가장 큰 값을 갖는 행동을 선택하기 위해 적용 시킬 수 있는 방법이 ε-greedy 방법이다. ε-greedy는 확률은 임의의 행동을 취하도록 해서 다양한 행동집합에 의해 학습이 진행되도록 하는 선택 방법이다[7][8][10].



(그림 2) Q-learning algorithm flow chart

2.4 교통 정보 수집

본 논문에서는 사용자의 단말기로부터 최적경로 탐색을 위한 지정된 노드를 통과한 구간의 평균주행속도를 교통관제서버에서 전송받고, 이 데이터를 활용하여 최적경로를 탐색한다. 하지만 도로 주행 특성상 과속 차량과 저속 및 정차 차량을 구분해야하고, 단말기 센서의 오작동 또한 고려하여야 한다. 신뢰성 있는 데이터를 확보하기 위해 사용자들의 평균주행속도의 누적평균을 기준으로 오차범위를 설정한다. 단말기로부터 전송받은 평균주행속도가 누적평균을 기준으로 한 오차범위를 벗어날 경우 이 데이터는 신뢰성이 낮은 데이터로 인식하여 제외시킨다.

2.5 보상 규칙

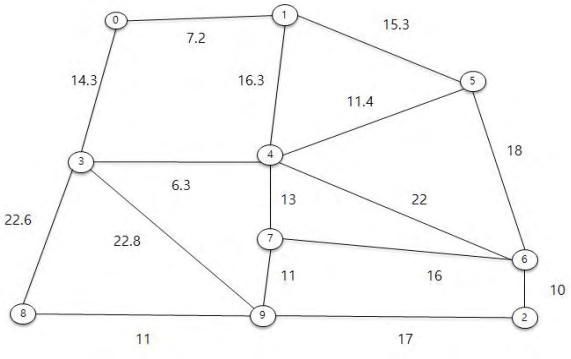
교통 정보 수집을 바탕으로 사용자로부터 받은 정보를 이용하기 위하여 본 논문에서는 Q-learning 알고리즘에서 보상 값의 변화를 주고자 보상 규칙을 변경하였다.

$$\frac{am}{rs} * rb \quad (2)$$

위 식(2)은 강화학습에서 Q-learning을 위한 보상 값의 변화를 위한 식이다. 식(2)에서 rb는 기준 보상, rs는 제한속도를 나타내며, am은 누적 평균을 나타낸다. 그리고 강화학습 기법은 에이전트에 대한 벌점을 부가하는데, 현재 이동한 거리에 벌점을 가하여 에이전트의 이동 거리가 증가 할수록 에이전트가 획득할 수 있는 보상 값이 감소하게 된다.

3. 알고리즘 분석 및 비교

3.1 Dijkstra 알고리즘 분석



(그림 3) 교통 네트워크

(그림 3)은 Dijkstra 알고리즘과 강화학습을 적용한 경로 탐색을 비교하기 위해 구성한 교통 네트워크이다. (그림 3)에 제시된 노드 간의 거리는 노드와 노드 사이의 거리를 나타내기 위해 각 노드에 좌표 값을 부여하여 각 노드간의 직선거리를 나타내었다.

<표 1> 교통 네트워크에 대한 노드 간의 거리 정보

번호	노드 간 거리 (노드 번호)					좌표(x ,y)
0	7.2(1)	14.3(3)				(8,36)
1	7.2(0)	16.3(4)	15.3(5)			(14,40)
2	10.0(6)	17.0(9)				(28,0)
3	14.3(0)	6.3(4)	22.6(8)	22.8(9)		(5,22)
4	16.3(1)	6.3(3)	11.4(5)	22.0(6)	13.0(7)	(11,24)
5	15.3(1)	18.0(6)				(22,27)
6	10.0(2)	18.0(5)	16.0(7)			(28,10)
7	13.0(4)	16.0(6)	11.0(9)			(12,11)
8	22.6(3)	11.0(9)				(0,0)
9	17.0(2)	22.8(3)	11.0(7)	11.0(8)		(11,0)

<표 1>은 (그림 3)의 노드와 노드 사이의 직선거리를 나타낸다. 교통 네트워크에 제시 된 거리를 계산하기 위해 각 노드에 좌표 값을 부여하여 각 노드간의 직선거리를 나타내었다.

본 논문에서는 실험 상황을 위하여 노드 1에서 노드 2까지의 최단거리와 최적경로를 실험한다. 노드 1에서 노드 2까지의 최단경로는 다음과 같다.

경로 1 : 1->4->6->2
경로 2 : 1->4->7->6->2
경로 3 : 1->4->7->9->2
경로 4 : 1->4->5->6->2
경로 5 : 1->5->6->2

경로 1의 주행거리 : 16.3 + 22 + 18 = 56.3
경로 2의 주행거리 : 16.3 + 13 + 16 + 10 = 55.3
경로 3의 주행거리 : 16.3 + 13 + 11 + 17 = 57.3
경로 4의 주행거리 : 16.3 + 11.4 + 18 + 10 = 55.7
경로 5의 주행거리 : 15.3 + 18 + 10 = 43.3

<표 2> 최단거리 결과

최단경로 탐색	1-> 5-> 6-> 2
이동거리	15.3 + 18.0 + 10 = 43.3

Dijkstra 알고리즘은 1->5->6->2 의 최단 경로가 나타난다. 위 경로 5의 주행 거리가 최단거리 주행 경로임을 확인 할 수 있다.

3.2 Q-learning 알고리즘 분석

<표 1>의 교통 네트워크의 거리 정보로 강화학습을 시켰을 때의 경로는 다음과 같다.

경로 1 : 1->4->6->2,
경로 2 : 1->4->7->6->2
경로 3 : 1->4->7->9->2
경로 4 : 1->4->5->6->2
경로 5 : 1->5->6->2

위 의 경로의 결과로 보았을 때, 최단경로와의 차이가 없다. 이에 본 논문에서는 강화학습과 Dijkstra 알고리즘의 비교를 위해 제안한 알고리즘의 제한속도와 평균주행속도를 나타내어 강화학습을 하려고 한다. <표 3>은 노드 간 평균주행속도 정보이다.

본 논문에서는 강화학습의 효율성을 실험하기 위해 <표 1>의 노드 간의 거리에 실시간으로 전송 받은 평균주행속도와 노드 간에 제한속도를 부여받아 강화학습 한 결과를 최단경로와 비교한다.

<표 3> 노드 간 평균주행속도 정보

번호	노드 간 평균주행속도(km) (노드 번호)					좌표(x ,y)
0	30(1)	50(3)				(8,36)
1	30(0)	80(4)	10(5)			(14,40)
2	60(6)	10(9)				(28,0)
3	50(0)	20(4)	60(8)	10(9)		(5,22)
4	80(1)	20(3)	20(5)	30(6)	30(7)	(11,24)
5	10(1)	50(6)				(22,27)
6	60(2)	50(5)	60(7)			(28,10)
7	30(4)	60(6)	30(9)			(12,11)
8	60(3)	20(9)				(0,0)
9	10(2)	10(3)	30(7)	20(8)		(11,0)

<표 3>은 실시간 교통 상황을 적용한 교통관제 서버에서 전송받은 노드 간의 평균주행속도를 나타낸다. 사용자로부터 수집한 정보를 기반으로 한 노드간의 제한속도는 60Km로 일정한 값을 부여한다. 부여한 제한속도 60Km와 <표 3>의 값으로 본 논문에서 제안한 위 식(2)로 계산 한

결과는 <표 4>에 나타난다. 노드 간 강화 학습의 값은 클수록 보상이 큰 것으로 최적의 경로에 적합하다는 것을 의미한다.

<표 4> 노드 간 강화학습 결과 정보

번호	노드 간 강화학습 결과 값(노드 번호)					좌표(x ,y)
0	50(1)	83(3)				(8,36)
1	50(0)	133(4)	16(5)			(14,40)
2	100(6)	16(9)				(28,0)
3	83(0)	33(4)	100(8)	16(9)		(5,22)
4	133(1)	33(3)	33(5)	50(6)	50(7)	(11,24)
5	16(1)	83(6)				(22,27)
6	100(2)	83(5)	100(7)			(28,10)
7	50(4)	100(6)	50(9)			(12,11)
8	100(3)	33(9)				(0,0)
9	16(2)	16(3)	50(7)	33(8)		(11,0)

<표 4>의 정보로 강화학습 한 노드1에서 노드2까지의 최적 경로는 다음과 같다.

경로 1의 학습 결과 값 : 133 + 50 + 100 = 283
경로 2의 학습 결과 값 : 133 + 50 + 100 + 100 = 383
경로 3의 학습 결과 값 : 133 + 50 + 50 + 16 = 249
경로 4의 학습 결과 값 : 133 + 33 + 83 + 100 = 349
경로 5의 학습 결과 값 : 16 + 83 + 100 = 199

경로 1의 주행시간 : 0.2 + 0.73 + 0.17 = 1.1
경로 2의 주행시간 : 0.2 + 0.43 + 0.27 + 0.17 = 1.07
경로 3의 주행시간 : 0.2 + 0.43 +0.37 + 1.7 = 2.7
경로 4의 주행시간 : 0.2 + 0.57 + 0.36 + 0.17 = 1.3
경로 5의 주행시간 : 1.53 + 0.36 + 0.17 = 2.06

<표 5> 최적경로 결과

최단경로 탐색	1-> 4 -> 7 -> 6 -> 2
학습 결과 값	133 + 50 + 100 + 100 = 383

강화학습은 노드와 노드간의 거리에 따라 최단경로를 추적하는 Dijkstra 알고리즘과는 달리 강화학습은 평균 주행속도와 제한속도를 기반으로 보상을 주어 최적의 경로를 찾는다. 위 경로의 학습 결과를 보면 최적경로로 나타난 경로 2는 많은 학습 결과 값이 나왔고, 최단 경로의 적합한 경로였던 경로 5는 학습 결과 값이 적게 나왔다. 또한, 주행시간도 최적 경로에 비해 많은 시간이 걸림을 알 수 있다.

<표 6> Dijkstra 알고리즘과 제안한 Q-learning 알고리즘 비교

노드 1 에서 노드 2까지	Dijkstra	Q-learning
경로	1->5->6->2	1->4->7->6->2
강화 학습 결과	199	383
주행 거리	43.3	55.3
주행 시간	2.06	1.07

<표 6>과 같이 동일한 출발지와 목적지의 최단경로를 비교하였을 때, Q-learning 알고리즘을 적용한 추천경로의 주행거리는 Dijkstra 알고리즘의 주행거리보다 증가하였다. 그러나 Q-learning 알고리즘을 적용한 추천경로가 가장 빠른 시간 내에 도달할 수 있는 최적경로를 추천한다.

4. 결론

본 논문에서는 실시간 교통 정보를 이용하여 최단경로 알고리즘과 강화학습기법을 적용한 알고리즘을 분석하여 비교하였다. 최단 경로 알고리즘보다 제안한 강화학습기법이 동적인 교통상황에 맞는 최적의 경로를 찾는데 도움을 준다는 것이 검증되었다. 교통상황에 따라 달라질 수 있는 변수에 대해 최단경로를 찾는 것 보다 강화학습을 이용한 최적경로를 찾는 것이 주행시간 단축에 더 우수하다.

향후 과제로는 사용자로부터 교통정보를 수집하기 위한 실시간 교통관제 시스템을 구축하고, 이 시스템에 강화학습기법을 적용하여 실시간 교통 정보를 이용한 최적경로 시스템을 개발하여 사용자에게 체계적으로 서비스할 수 있는 방안을 연구할 것이다.

참고문헌

[1] 정상호, 김선형, 스마트한 지능형 교통체계 구축을 위한 ITS 관련 법령 정비에 관한 연구, 한국정보통신학회, 2012
[2] 최진섭, 조영태, 정인범, 트래픽 패턴 학습 기반의 다중 교차로 교통 신호 제어, 한국정보과학회, 2014
[3] 정환묵, “소프트 컴퓨팅”, 내하출판사, 2008
[4] 김선경, 김진상, 임재걸, 임태수, “컴퓨터 알고리즘”, 도서출판 한산, 2012
[5] 장정, 승지훈, 김태영, 정길도, Q 학습을 이용한 교통제어 시스템, 한국산학기술학회, 2011
[6] 곽한주, 박귀태, 강화학습과 유전자 알고리즘을 이용한 이동로봇의 최적경로 탐색, 정보 및 제어 학술대회, 2010
[7] 장시영, 공성학, 서일홍, 오상록, Domain Knowledge를 이용한 강화학습, ICCAS, 2001
[8] 정희석, 이종수, 강화학습을 이용한 주행경로 최적화 알고리즘 개발, 한국지능시스템학회 춘계학술대회, 2003
[9] Richard Sutton, Andrea Barto, "Reinforcement Learning : An Introduction", BrandfordBook, 1998
[10] Yit Kwong Chin, Wei Yeang Kow, Wei Leong Khong, Min Keng Tan, Kenneth Tze Kin Teo, Q-Learning Traffic Signal Optimization within Multiple Intersections Traffic Network, UKSimm-AMss, 2012