부분관측가능 환경의 경로 계획을 위한 효율적인 가치 반복 알고리즘

김영기*, 김해천*, 이재성*
*중앙대학교 소프트웨어학부
e-mail:kimyoungi99@naver.com, code.bug.station@gmail.com
curseor@cau.ac.kr

An Efficient Path Planning Algorithm for Partially Observable Maps Based on Value Iteration Algorithm

Young Ki Kim*, Hae-Cheon Kim*, Jaesung Lee*
*Dept. of Software, Chung-Ang University

요 약

경로 계획은 에이전트가 로봇이 특정 목적지에 도착할 수 있도록 에이전트가 수집한 정보를 바탕으로 경로를 설정하는 작업을 뜻한다. 부분 관측만 가능한 맵인 경우 에이전트 이동마다 새로 수집되는 정보들을 바탕으로 마르코프 의사결정 과정을 사용한 가치 반복 알고리즘이 널리 사용되지만, 제안된 가치 반복 알고리즘 사용 시 매 행동마다 모든 공간의 최적 경로를 계산하기 때문에 시간이 오래 걸리는 문제점이 있다. 이에 본 논문에서는 에이전트가 한 번에 탐색하는 범위가 제한되어 있다는 점에 착안하여 탐색 반경 내에 속하는 공간의 가치 함수 값을 미리 추정하여 효율적으로 최적의 경로를 추정하는 가치 반복 알고리즘을 제안한다.

1. 서론

경로 계획(Path Planning) 문제는 에이전트(Agent)가 로봇이 특정 목적지에 도착할 수 있도록 에이전트가 수집한 정보를 바탕으로 경로를 찾는 문제로, 청소로봇의 경로 [3], 자율주행차량의 주행 경로 [4] 등의 활용되고 있다. 현실의 실제 경로 계획 문제는 에이전트가 맵의 모든 환경 정보를 가지고 있는 경우가 드물기 때문에 부분관측이 가능한 환경에 놓인 에이전트가 직접 탐색하여 정보를 수집하고 최적의 경로를 탐색하는 것을 목표로 한다.

경로 계획 문제의 대다수는 마르코프 의사결정 과정(Markov Decision Process)으로 문제를 모델링하여 해결하는 방식을 취한다. 마르코프 의사결정 과정은 특정 에이전트의 행동과 전이확률, 그리고 보상을 정보로 가진 의사결정 모델로, 이 모델의 목표는 에이전트가 받는 보상을 최대화 하는 것이며, 가치 반복 알고리즘(Value Iteration Algorithm) [1]을 통해 맵의 환경 정보를 바탕으로 최적의 경로를 구하여 문제를 해결해왔다.

하지만, 부분관측만 가능한 맵 위에 있는 경우, 에이전트가 새로운 정보를 수집할 때마다 가치 반복 알고리즘을 매번 실행하여 최적의 경로를 계산해야 하며, 이에 따라 최적 경로 탐색에 매우 오랜 시간이 걸리는 문제점을 가진다. 이에 따라 본 논문은 가치 반복 알고리즘 실행 시 경로의 가치를 계산하는 과정을 변형하여 효율적으로 최적 탐색을 실행하는 알고리즘을 제안하고자 한다.

2. 관련 연구

마르코프 의사결정 과정은 에이전트의 상태와 그에 따른 다음 행위들과 그 확률을 수학적으로 나타낸 과정의 일종으로 $< S, A, P, R >$으로 정의된다. 여기서 S는 상태의 집합, A는 행동의 집합, $P(s'|s,a)$는 상태 s에서 행동 a를 택하여 상태 s'로 전이될 확률, $R(s,a)$ s에서 행동 a를 택할 때 얻게 되는 보상이라 정의한다.

가치 반복 알고리즘은 마르코프 의사결정 과정을 해결하는 동적 계획(Dynamic Programming) 알고리즘으로, 벨만 방정식을 사용한 계산의 반복을 통해 각 상태의 가치를 나타내는 가치 함수(Value Function)를 구하는 알고리즘이다 [1]. 가치 반복 알고리즘은 각 행동에 대한 가치를 계산의 반복을 통해 계산하여 최적의 가치 함수(Optimal Value Function)를 찾을 수 있어 널리 사용되었으나, 상태가 매우 많은 마르코프 의사결정 과정이나 보상이나 전이확률(Transition Probability)이 명확하지 않을 경우 최적 가치의 근사치를 찾기 위해 여러 번 계산을 반복해야 하는 문제를 가진다.

이를 해결하기 위해 가치 반복 알고리즘의 반복 횟수를 줄이고도 최적 가치 함수에 근사할 수 있는 방식이 연구 되었다. 이 방식을 근사 가치 반복 알고리즘(Approximate Value Iteration Algorithm)이라 하며, 대표적으로 가치 함수의 일부분만 구한 뒤 지도 학습 알고리즘을 사용하여 가장 오차가 적은 함수를 맞추는 피팅(fitting) 알고리즘 [2] 등이 연구되었다.

하지만, 이러한 알고리즘의 변형을 통해 계산의 적은 반복으로도 좋은 결과를 얻어낼 수 있지만, 상태들이 부분 관측 가능한 경우 마르코프 의사결정 과정의 보상이 바뀌기 때문에 결국 여러 번 계산해야 하는 문제를 가진다. 이에 따라 본 논문은 가치 반복 알고리즘의 부하를 효율적으로 줄일 수 있는 방식을 제안하고자 한다.

3. 제안하는 알고리즘

3.1. 문제 정의

본 논문에서는 에이전트란 $N \times M$ 크기의 맵 위에서 움직이는 객체로, 맵의 정보를 수집하는 객체로 정의한다. 맵은 격자로 나뉘며 벽과 에이전트가 위치할 수 있는 곳으로 구별한다. 맵의 각 지역은 마르코프 의사결정 과정 $< S, A, P, R >$에서 상태로 표시하며, $|S| = m$, A={위, 아래, 오른쪽, 왼쪽}, P는 에이전트가 택한 방향으로 전이할 확률로 정의한다. 에이전트는 w만큼 거리의 주변 상태들만 관측 할 수 있으며, 에이전트는 한번 움직일 때 마다 주변 상태들을 관측하고 기억한다. (그림 1) 은 w가 1일 때의 에이전트가 탐색 가능한 공간을 그림으로 나태난 것이다.

Unknown	Unknown	Unknown	Unknown	Unknown
Unknown	Searched	Searched	Searched	Unknown
Unknown	Searched	Agent	Searched	Unknown
Unknown	Searched	Searched	Searched	Unknown
Unknown	Unknown	Unknown	Unknown	Unknown

(그림 1) $w = 1$ 경우 에이전트의 탐색 범위

부분 관측으로만 상태 수집이 가능한 경우, 목표 상태를 찾을 때까지 상태들을 돌아다니면서 탐색을 방법이 보편적으로 사용되었다. 이에 따라 탐색을 유도하고자 탐색하지 않은 지역에는 x의 보상, 목표지점에 도달하면 y의 보상을 설정하였다. (그림 2)는 위의 예시를 보여준 것이다.

Agent	WALL Searched	y Target	x Unknown
0 Searched	WALL Searched	0 Searched	x Unknown
0 Searched	0 Searched	0 Searched	x Unknown
x Unknown	x Unknown	x Unknown	x Unknown

(그림 2) 맵의 예시

2. 일반적인 가치 반복 알고리즘

일반적인 가치 반복 알고리즘은 다음과 같다. s'는 s의 다음 상태, $V(s)$는 가치함수, $Q(s,a)$는 큐 함수를 나타낸다. $Q(s,a)$는 상태 s에서 행동 a를 택하였을 때의 가치 즉, 미래의 보상의 기대치라 할 때의 의사코드를 표현하면 <표 1>과 같다.

<표 1> 가치 반복 알고리즘의 의사코드

알고리즘 1: 일반적인 가치 반복 알고리즘

1. Initialize $V(s), R(s,a)$
2. while $V(s)$ *is not converged* do
3. for all $s \in S$ do
4. for all $a \in A$ do
5. $Q(s,a) \leftarrow R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a)\, V(s')$
6. end for
7. $V(s) \leftarrow \max_a Q(s,a)$
8. end for
9. end while

맵의 모든 정보를 알고 있는 경우, 맵의 정보가 변할 일이 없기 때문에 위의 가치 반복 알고리즘은 단 한 번만 실행하여 최적의 경로를 계산한다. 하지만, 부분관측 가능한 맵 위에서 에이전트가 새로운 공간을 관측하는 경우에는 맵의 정보가 바뀌기 때문에 여러 번의 계산을 다시 거쳐 최적의 경로를 찾아야 한다. 그러므로 목표지점에 도달하기까지 알고리즘 실행 횟수가 많아져 매우 비효율적이라는 문제를 가진다.

3. 제안하는 알고리즘

이에 본 연구에서는 에이전트는 정보 습득 시 새로 정보를 습득한 지역의 가치 함수 값을 미리 계산하여 반영하여 가치 함수 값을 빠르게 수렴시킬 수 있도록 알고리즘을 변형시켰다. <표 2>은 본 연구에서 제안하는 가치 반복 알고리즘의 의사코드를 나타낸 것이다.

<표 2> 제안하는 알고리즘의 의사코드

알고리즘 2: 제안하는 가치 반복 알고리즘

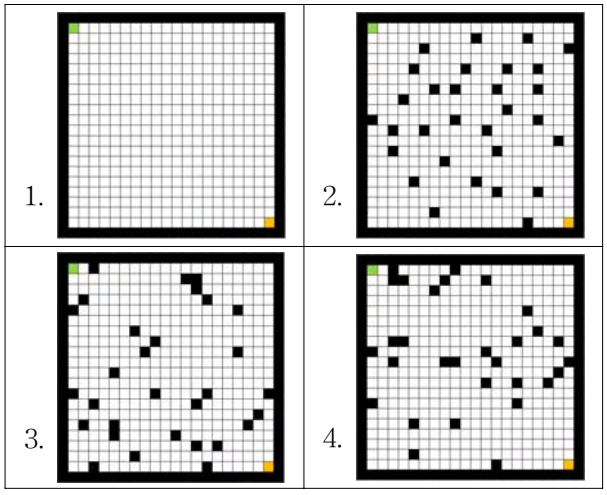
1. Initialize $R(s,a)$
2. for all $s \in S$ do
3. for all $a \in A$ do
4. if $R_p(s,a) = x$ and $R(s,a) \neq x$ then
5. $V(s) \leftarrow V(s) - x$
6. end for
7. end for
8. for k time do
9. for all $s \in S$ do
10. for all $a \in A$ do
11. $Q(s,a) \leftarrow R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a)\, V(s')$
12. end for
13. $V(s) \leftarrow \max_a Q(s,a)$
14. end for
15. end for

여기서 $R_p(s,a)$는 에이전트의 움직임으로 인해 마르코프 의사결정 과정이 변하기 전의 보상 값을, k는 수렴을 위한 과정 반복계산 수치를 뜻한다.

본 연구에서 제안하는 가치 반복 알고리즘은 최종적인 가치 값에 정량적인 규칙을 도입하여, 만약 새로 탐색된 부분에 목표 지점이 없으면 그 탐색에 대하여 x만큼의 보상을 감산하여 멀리 떨어진 큐 함수의 값에 미리 반영하여 가치 함수 값의 빠르게 수렴하도록 유도했다. 그리고 가치함수 계산횟수를 k번으로 설정하여 가치함수의 시간 복잡도를 감소시켜 계산 속도에 효율성을 증가시켰다. 최종적으로 위 규칙을 가치반복 알고리즘에 반영하여 에이전트의 탐색을 효율적으로 성공하도록 마무리한다.

4. 실험 결과

본 논문에서 제안한 알고리즘이 얼마나 효율적인지 확인하기 위해 실험을 진행하였다. 실험은 20×20 크기의 맵 4개에서 진행하였다. (그림 3)은 맵 4개의 형태를 보인 것이다.



(그림 3) 실험에 사용된 맵 4개의 그림

1번 맵은 방해물을 두지 않았고 맵 2, 3, 4는 각각 30개의 방해물을 무작위로 설정하였으며 전이 확률은 에이전트가 선택한 행동의 방향으로 86%의 확률로 이동하게 하였으며 행동의 방향을 기준으로 각각 7%의 확률로 왼쪽, 오른쪽으로 이동하게 설정하였다. 가치함수 계산횟수 k는 10으로 설정했다. <표 3>과 <표 4>는 각 알고리즘 당 20번 실험한 결과의 이동횟수와 실행시간의 평균과 표준편차를 기록한 것이다.

<표 3> 이동 횟수 실험 결과 (단위, 횟수)

	Value Iteration	Proposed
Map1	159.7 ± 69.26	126.5 ± 68.31
Map2	148.6 ± 81.86	143.7 ± 50.91
Map3	167.5 ± 57.55	170.5 ± 69.10
Map4	168.9 ± 76.95	181.2 ± 67.79

<표 4> 실행 시간 실험 결과 (단위, 초)

	Value Iteration	Proposed
Map1	355.5 ± 270.95	7.6 ± 6.08
Map2	286.0 ± 315.62	8.1 ± 4.68
Map3	334.4 ± 228.52	11.0 ± 6.76
Map4	324.3 ± 328.19	14.5 ± 8.27

실험 결과, 이동 횟수의 경우 기존 가치 반복 알고리즘과 본 논문에서 제안한 알고리즘과 큰 차이가 없었다. 이동횟수 기준에서 Map1, 2에서는 더 나은 성능을 보였으며, Map3, Map4의 경우 이동횟수의 효율성이 저하되었는데, 이는 충분한 가치함수 계산이 부족했던 것으로 추정된다. 하지만 실행시간에서는 큰 차이가 있음을 확인할 수 있어, 본 논문의 제안 알고리즘이. 위 실험을 통해 가치함수 값이 빨리 수렴하도록 가치 함수의 값에 특정 규칙을 반영하는 것이 가치 함수 값을 효율적으로 찾아내고, 이를 통해 적절한 경로를 효율적으로 찾아낼 수 있음을 확인했다.

5. 결론

본 연구는 부분 관측 가능한 맵에서 매 움직임마다 모든 맵의 가치 함수 값을 다시 계산하는 가치 반복 알고리즘의 구조에서 가치 함수 값을 미리 반영한 후 가치 반복 알고리즘을 시행하여 효율적인 방식을 찾는 것을 목표로 진행하였다. 그리고 실험 결과, 본 논문에서 제안한 방식이 결과성능을 비슷하게 유지함과 동시에 시간적으로 효율적임을 확인할 수 있었다. 추후 연구를 통해 맵마다 가치함수의 수렴주기를 추측하여 시간적인 효율을 추구함과 동시에 이동횟수상의 효율도 추구하는 연구를 진행할 수 있을 것이다.

사사

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음 (20170001000031001)

참고문헌

[1] Sutton, Richard S. and Andrew G. Barto. "Reinforcement learning: An introduction," 2011.

[2] Mann, Timothy, and Shie Mannor. "Scaling up approximate value iteration with options: Better policies with fewer iterations," International Conference on Machine Learning, Beijing, China, pp. 127-135, Jun. 2014.

[3] Yongwoo Shin, Seungyoon Choi and Taechoong Chung, "Development of Intellectual Online Complete Coverage Planning Algorithm for Cleaning Robot," Journal of KISS : Software and Applications, Vol. 40, No. 10, pp. 592-599, Oct. 2013.

[4] Sungwook Jo, kyungil Lim, Yoonsub Kim and Jungha Kim, "The Safety Path planning for Generating Overtaking Path of Unmanned Ground Vehicle," 한국자동차공학회 춘계학술대회, pp. 545-548, May 2015.