반도체디스플레이기술학회지 제21권 제2호(2022년 6월)
Journal of the Semiconductor & Display Technology, Vol. 21, No. 2. June 2022.

물품 출고 시간 최소화를 위한 강화학습 기반 적재창고 내 물품 재배치

김여진[*] · 김근태[*] · 이종환[*†]
[*†]금오공과대학교 산업공학과

Minimize Order Picking Time through Relocation of Products in Warehouse Based on Reinforcement Learning

Yeojin Kim[*], Geuntae Kim[*] and Jonghwan Lee[*†]
[*†]Department of Industrial Engineering, Kumoh National Institute of Technology

ABSTRACT

In order to minimize the picking time when the products are released from the warehouse, they should be located close to the exit when the products are released. Currently, the warehouse determines the loading location based on the order of the requirement of products, that is, the frequency of arrival and departure. Items with lower requirement ranks are loaded away from the exit, and items with higher requirement ranks are loaded closer from the exit. This is a case in which the delivery time is faster than the products located near the exit, even if the products are loaded far from the exit due to the low requirement ranking. In this case, there is a problem in that the transit time increases when the product is released. In order to solve the problem, we use the idle time of the stocker in the warehouse to rearrange the products according to the order of delivery time. Temporal difference learning method using Q_learning control, which is one of reinforcement learning types, was used when relocating items. The results of rearranging the products using the reinforcement learning method were compared and analyzed with the results of the existing method.

Key Words : Reinforcement Learning, Q_learning, TD (Temporal Difference learning), Relocation, Machine Learning

1. 서 론

물류 분야에서 인공지능, 기계학습이 활용됨에 따라 증가하는 물동량을 처리하기 위해 자동화 창고가 도입되고 있다. 본 연구에서의 자동화 창고 시스템은 물품이 소요량 순위 기준으로 위치를 선정해 입고된 후 그 자리에 머물러 있다가 출고시점에 출고되는 형식이다. 현재 시스템은 소요량 순위가 낮아 입고 시 출구로부터 멀리 적재된 물품의 경우 출고 시간이 증가한다는 문제점이 있다. 문제점을 해결하기 위해 창고 내 스토커의 유휴시간을 활용하여 물품의 출고 시간 순위에 따라 출고 시간이 가장 빠른 물품을 출구에 가장 가까운 위치로 재배치를 진행하여 물품 출고시간의 단축여부를 연구한다. 물품들을 재배치할 경우에 기계학습의 한 방법론인 강화학습을 사용하였다. 강화학습의 대표적인 방법론으로는 동적계획법(DP, Dynamic Programming)이 있다. 동적계획법은 모든 환경이 갖춰진 모델이 필요한 방법론이다. 모든 경우의 수를 알아야 사용할 수 있는 모델이므로 현실에 적용하기 어렵다. 따라서 모든 환경을 구축하지 않아도 되는 Q_learning 제어를 이용한 시간차 학습(TD, Temporal Difference Learning) 방법[1]을 사용하였다. 본 연구에서는 소요량 기준으로 적재하는 기존 방식과 강화학습을 이용하

[†]E-mail: shirjei@kumoh.ac.kr

여 재배치를 적용한 방식의 물품 출고 시간을 비교 분석하고자 한다.

2. 이론적 배경

2.1 강화학습

강화학습은 기계학습의 한 분야로 주어진 환경(Environment)에서 에이전트 (Agent)가 보상 (Reward)을 최대화하는 행동 (Action)을 선택하도록 학습해 나가는 것이다. Fig. 1은 강화학습의 진행 과정을 나타낸 것이다.

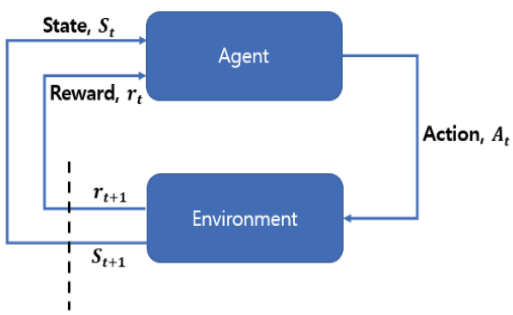

Fig. 1. Reinforcement Learning Concept Diagram.

강화학습의 대표적인 방법론으로 동적계획법(DP, Dynamic Programming), 몬테카를로 방법 (MC, Monte Carlo Method), 시간차 학습 방법 (TD, Temporal Difference learning)이 있다. 강화학습은 시간에 따라 상태, 행동, 보상을 순차적으로 처리한다. 이러한 강화학습을 적용할 수 있는 문제는 순차적으로 행동을 결정해야하는 문제를 수학적으로 정의한 마르코프 결정 과정 (MDP, Markov Decision Process)으로 표현된 문제이다[2]. MDP는 상태, 행동, 보상함수, 감가율, 상태 변환 확률로 구성되어 있다. 가치 함수를 통해 에이전트가 최적의 행동을 선택하도록 한다. 벨만 방정식을 사용해 에이전트가 현재 상태에서 취하는 행동에 대한 기대값을 나타낸 Q함수를 업데이트해 나간다. Q함수에 관한 식은 (1)과 같다.

$$Q_\pi(\mathrm{s},\mathrm{a}) = E_\pi[R_{t+1} + \gamma Q_\pi(S_{t+1}, A_{t+1}) \,|\, S_t = s, A_t = a] \quad (1)$$

벨만 방정식은 특정 정책을 따랐을 때 가치함수 사이의 관계식을 의미한다. 벨만 방정식을 Q함수를 활용해 나타내면 (2)와 같다.

$$Q(\mathrm{s},\mathrm{a}) = E[R_{t+1} + \gamma max_{a'} Q(S_{t+1}, a') \,|\, S_t = s, A_t = a] \quad (2)$$

2.2 Q_learning

강화학습의 대표적인 방법론인 DP는 MDP에 대한 완전한 정보를 필요로 하기 때문에 현실 문제에 적용하기 힘들다는 단점이 있다. 현실문제 적용이 어려운 단점을 보완한 것이 MC 방법론이다. MC 방법론은 MDP에 대한 정보가 없을 때 처음부터 끝까지 에피소드를 진행하여 최적의 정책을 찾아가는 방법이다. MC 방법은 MDP 정보가 필요 없다는 장점이 있지만 에피소드를 처음부터 끝까지 진행해야 하기 때문에 시간이 오래 걸린다는 단점이 있다. DP와 MC방법론의 장점을 결합한 방법론은 TD 알고리즘이다. TD방법론은 단계마다 바로바로 가치함수를 업데이트한다. TD방법 중 하나인 Q_learning은 현재 행동하는 정책과는 독립적으로 학습을 진행한다[3]. 즉 에이전트는 행동하는 정책을 따라 계속 탐험해 나가고 행동하는 것과는 별개로 목표 정책을 두어 학습을 진행한다. 이를 Off-Policy라고 한다. Q_learning을 통해 Q함수를 업데이트하는 과정은 식 (3)과 같다. 식 (3)을 보면 실제로 다음 상태에서 행동을 해보는 것이 아닌 다음 상태에서 가장 큰 Q함수 값에 해당하는 행동을 선택해 업데이트를 하는 방식이다.

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma max_{a'} Q(S_{t+1}, a') - Q(S_t, A_t)) \quad (3)$$

Q_learning에서는 Q함수를 업데이트할 때 벨만 최적 방정식을 사용한다. 벨만 최적 방정식은 식 (4)와 같다.

$$Q^*(s,a) = E[R_{t+1} + \gamma max_{a'} Q^*(S_{t+1}, a') \,|\, S_t = s, A_t = a] \quad (4)$$

Q_learning은 행동 정책과 업데이트 시 사용하는 정책을 다르게 설정함으로써 SARSA의 문제점이었던 에이전트가 갇혀버리는 현상을 해결하였다[4].

3. 실 험

3.1 적재 창고 환경

본 연구에서 사용한 적재창고의 환경은 가로 3, 높이 3의 9개칸 2개로 총 18칸을 가지며 2개 사이에 스토커가 이동하면서 물품을 운반한다. 입구와 출구의 위치는 동일하고 창고 내 물품 이동에 사용되는 스토커 (Stocker)는 하나이다. 적재창고 구조는 Fig. 2와 같다. 스토커는 3개의 모터 (상승, 주행, 포크)를 가지며 각 모터 (가속, 감속, 주행)의 속도는 무게에 영향을 받고, 총 18칸의 적재창고에서 스토커의 최대 이동 시간은 1회 약 30초이다. 적재 창고의 입/출구로부터 각 위치별까지의 스토커 도달 시간 순위는 Fig. 3과 같다. 1열과 2열의 위치별 순위는 동일하다.

1열	(1,3,1)	(1,3,2)	(1,3,3)
	(1,2,1)	(1,2,2)	(1,2,3)
입/출구	(1,1,1)	(1,1,2)	(1,1,3)

2열	(2,3,1)	(2,3,2)	(2,3,3)
	(2,2,1)	(2,2,2)	(2,2,3)
	(2,1,1)	(2,1,2)	(2,1,3)

Fig. 2. Warehouse Structure.

1열	6	6	6
	4	4	5
입/출구	1	2	3

Fig. 3. Ranking by Location Considering Weight.

3.2 데이터

데이터는 해양수산부에서 제공하는 2021년 1월부터 2022년 3월 9일까지의 각 항구별 선박 입출항 현황을 사용하였다. 적재 창고 내 물품이 입고, 출고되는 것이 선박이 항구에 입항, 출항되는 것과 유사하기 때문에 해당 데이터를 사용하였다. 하지만 선박이 평균적으로 항구에 정박되어 있는 시간이 창고 내 물품이 적재되어 있는 평균 시간보다 길기 때문에 정박된 시간이 8시간 이내인 선박들만 데이터로 사용하였으며, 창고의 칸 개수가 18개이기 때문에 적재되어 있는 물품의 개수가 18개를 넘지 않도록 전처리를 진행하였다. 전체적인 물동량이 높은 창고로 가정하기 위해 적재되는 물품의 종류는 9가지로 설정하였다. 즉 총 9개의 항구로부터 데이터를 추출하였다.

입고 시 적재 위치를 선정함에 있어 필요한 데이터는 Table 1과 같다.

Table 1. Required Data for Product Input

구분	변수
입력	예측 소요량
	입고 시간
	물품 S/N
출력	적재 위치

물품 재배치를 위해 필요한 데이터는 Table 2와 같다.

Table 2. Required Data for Products Relocation

구분	변수
입력	물품의 현재 적재 위치
	물품의 출고 시간 순위
	물품 S/N
출력	물품의 재배치 위치

3.3 실험 방법

스토커의 유휴 시간 발생 시 각 물품별 출고시간을 고려하여 출고시간이 가장 빠른 물품을 출구와 가장 가까운 위치로 재배치한다. 물품의 출고시간에 맞게 해당 물품만 출구 가까이에 있으면 되므로 전체 물품을 재배치하는 것이 아니라 해당하는 물품에 대해서만 재배치를 진행한다. 만약 출구와 가장 가까운 위치에 다른 물품이 적재되어 있다면 빈 공간에 물품을 옮긴 후 해당 자리에 출고시간이 가장 짧은 물품을 적재한다. 과정을 나타내면 Fig. 4와 같다. Fig. 4의 숫자는 물품의 출고 순위를 의미한다.

[단계 1] 재배치가 필요한 물품 (1번 물품)을 선정한다.
[단계 2] 출구와 가까운 자리에 다른 물품 (7번 물품)이 있다면 해당 물품을 빈공간으로 옮긴다.
[단계 3]1번 물품을 출구와 가장 가까운 자리로 재배치한다.
[단계 4]7번 물품을 다시 1번 물품이 있던 자리로 옮긴다.

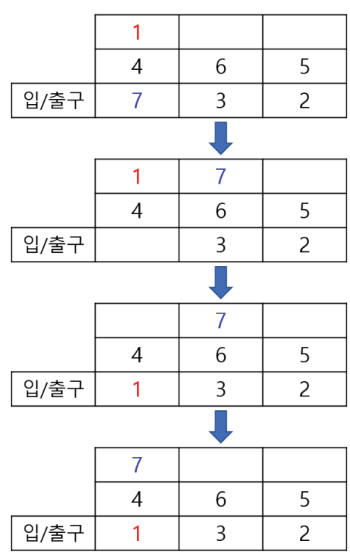

Fig. 4. Relocation Process.

Fig. 4와 같은 단계에 강화학습을 적용하기 위해서는 창고 내부 상태 그대로가 아닌 재배치 진행을 위한 창고 상태를 Fig. 5와 같이 따로 생성해야 한다. 재배치가 필요한 물품의 S/N는 그대로 두고 나머지 S/N와 빈칸은 모두 1로 설정하였다. 그림에서 10은 빈칸을 알려주기 위해 임의로 창고 내 빈칸 중 하나를 선정하여 설정하였다. 빈칸으로 설정한 칸 (10)이 action을 취하면서 학습을 진행한다. Fig. 6은 목표 상태를 나타낸 것이다. 초기 상태가 목표 상태와 같아지면 episode를 종료한다.

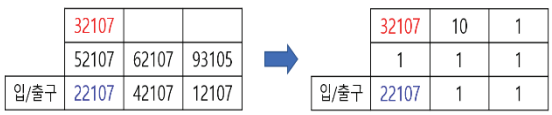

Fig. 5. Initial State Creation.

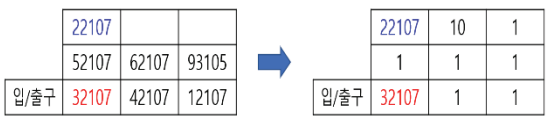

Fig. 6. Goal State Creation.

재배치 진행 시 강화학습 방법론 중 하나인 Q_learning 제어를 이용한 TD방법을 사용하였다. 목표 상태에 도달하기까지 최대한 적은 수의 행동으로 도달하기 위해 이동 거리에 따른 보상 방식을 추가해주었다[5]. MDP 구성은 다음과 같이 설정하였다.

1. 상태 (State) : 현재 물품이 적재 되어있는 창고 내부 상태
2. 행동 (Action) : 창고 내에서 자리를 이동해야 하는 물품 위치와 빈 공간 중 하나를 선택하여 action으로 할당하였다.
3. 보상 (Reward)
   (1) 창고 내부 상태가 목표 상태에 도달하였을 때 +1점을 부여하였다.
   (2) 3차원 환경이기 때문에 유클리디언 거리[6]를 이용하여 각 물품의 현재 위치로부터 목표 위치까지의 거리를 합산한 후 (-0.1)을 곱한 것을 Reward값으로 부여하였다. 거리 공식은 (5)와 같다.
4. 감가율 (γ) : 0.9
5. 상태변환확률 : 1.0
6. 학습률(α) : 0.3

$$d_{ij} = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2} \quad (5)$$

4. 실험 결과

에피소드는 50회 진행하였다. Reward per Episode그래프는 Fig. 7과 같다

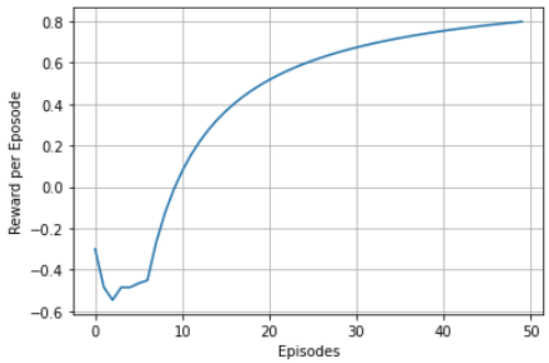

Fig. 7. Reward per Episode.

Table 3은 강화학습을 이용하여 재배치한 경우의 평균 출고시간과 소요량 기준으로 적재한 기존 방식의 평균 출고시간을 나타낸 것이다.

Table 3. Average Delivery Time According to Loading Method

	기존 방식	재배치 진행
평균 출고시간 (s)	20.51	18.83

두 방법의 결과를 비교해본 결과 재배치를 진행한 경우 진행하지 않은 경우보다 평균 출고시간이 약 8.19% 감소한 것으로 나타났다.

5. 결론 및 향후 과제

본 연구에서는 물품의 출고 시간을 최소화하기 위해서 적재 창고 내 스토커의 유휴시간을 활용해 물품을 재배치하는 방식을 제안하였다. 재배치 과정에서 모든 물품에 대해 재배치를 진행하는 것이 아닌 출고시간이 가장 빠른 물품에 대해서만 진행하였다. 그 이유는 실제 창고에서 스토커가 1회 움직이는데 30초 내외의 시간이 필요한데 모든 물품에 대해 재배치를 진행할 경우 스토커가 여러 번 움직이게 되므로 스토커의 이동 시간이 증가하여 이는 비효율적이기 때문이다. 또한 해당 시점에서 출고시간 순위대로 물품을 재배치한다 해도 새로운 물품이 입

고되는 순간 물품들의 출고순위가 바뀔 수 있기 때문에 출고시간이 가장 빠른 물품에 대해서만 재배치를 진행하였다. 재배치 과정에서 적재 창고 내 환경을 바로 사용하지 않고 재배치가 필요한 물품에 대해서만 별도로 초기 상태와 목표 상태를 생성하여 강화학습을 진행하였다.

본 연구에서 진행한 강화학습 Q_learning 제어의 TD 방법을 이용한 재배치는 기존 적재 방식보다 물품의 평균 출고시간이 8.19% 감소하였고, 이는 강화학습을 적재창고에 적용하여 전체 출고시간을 감소시킴과 동시에 재배치의 효율성을 높여 실제 현장에 적용가능성을 제시한 것에 의의가 있다.

본 연구에서는 학습 속도를 고려하여 빈공간과 이동이 필요한 물품에 대해서만 action으로 선택하였다. 추후 연구에서는 모든 위치에 대해서 action으로 고려하여 연구를 진행할 필요가 있다. 또한 적재창고 환경이 18칸인 경우에 대해서만 연구를 진행하였고 적재창고 규모에 따른 출고시간에 대해 비교하지 않아 추후 적재창고 규모를 확장하여 연구를 진행한 후 결과를 비교할 필요가 있다.

감사의 글

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 202101350001, 딥러닝 기반 스마트 자동물류 적재창고 기술개발).

참고문헌

1. Watkins, C.J.C.H., Dayan, P., “Q-learning”, Machine Learning, vol. 8, pp. 279-292, 1992.
2. Howard, Ronald, “Dynamic programming and Markov processes”, Massachusetts Institute of Technology Press, 1960.
3. Richard S. Sutton, Andrew G. Barto, “Reinforcement Learning: An Introduction Second edition, in progress”, Massachusetts Institute of Technology Press, 2015.
4. Lee W.W., Yang H.R., Kim G.W., Lee Y.M., Lee U.R., “Reinforcement Learning with Python and Keras”, wikibooks, 2020.
5. Moon S.U., Jung D.E., Kim J.H., Cho Y.W., “Comparison of Sliding puzzle agent learning performance through Monte Carlo method and Temporal difference learning (SARSA control, Q-learning control) method”, Journal of the institute of Electronics and Information Engineers, pp. 709-712, 2021.
6. Kim S.W., Chung K.S., “The Robust Estimation with Spatial Economics Models using 3-Dimension Weight Matrix considering the Height of the House”, Housing Studies Review, vol. 18, pp. 73~92, 2010.

접수일: 2022년 6월 7일, 심사일: 2022년 6월 21일,
게재확정일: 2022년 6월 22일