지능형 에이전트의 환경 적응성 및 확장성

백 혜 정† · 박 영 택††

요 약

로봇이나 가상 캐릭터와 같은 지능형 에이전트가 자율적으로 살아가기 위해서는 주어진 환경을 인식하고, 그에 맞는 최적의 행동을 선택하는 능력을 가지고 있어야 한다. 본 논문은 이러한 지능형 에이전트를 구현하기 위하여, 외부 환경에 적응하면서 최적의 행동을 배우고 선택하는 방법을 연구하였다. 본 논문에서 제안한 방식은 강화 학습을 이용한 행동기반 학습 방법과 기호 학습을 이용한 인지 학습 방법을 통합한 방식으로 다음과 같은 특징을 가진다. 첫째, 강화 학습을 이용하여 환경에 대한 적응성을 학습함으로 지능형 에이전트가 변화하는 환경에 대한 유연성을 가지도록 하였다. 둘째, 귀납적 기계학습과 연관 규칙을 이용하여 규칙을 추출하여 에이전트의 목적에 맞는 환경 요인을 학습함으로 주어진 환경에서 보다 빠르게, 확장된 환경에서 보다 효율적으로 행동을 선택을 하도록 하였다. 셋째, 본 논문은 지능형 에이전트를 구현하는데 있어서 처음부터 모든 상태를 고려하기 보다 상태 탐지기를 이용하여 새로운 상태가 입력될 때마다 상태를 확장시키는 방식을 이용하였다. 이러한 방식은 필요한 상태에 대하여서만 고려함으로 메모리를 획기적으로 축소 할 수 있으며, 새로운 상태를 동적으로 처리 할 수 있어, 환경에 대한 변화에 능동적으로 대처 할 수 있다.

A study on environmental adaptation and expansion of intelligent agent

Hae-Jung Baek† · Young-Tack Park††

ABSTRACT

To live autonomously, intelligent agents such as robots or virtual characters need ability that recognizes given environment, and learns and chooses adaptive actions. So, we propose an action selection/learning mechanism in intelligent agents. The proposed mechanism employs a hybrid system which integrates a behavior-based method using the reinforcement learning and a cognitive-based method using the symbolic learning. The characteristics of our mechanism are as follows. First, because it learns adaptive actions about environment using reinforcement learning, our agents have flexibility about environmental changes. Second, because it learns environmental factors for the agent's goals using inductive machine learning and association rules, the agent learns and selects appropriate actions faster in given surrounding and more efficiently in extended surroundings. Third, in implementing the intelligent agents, we considers only the recognized states which are found by a state detector rather than by all states. Because this method consider only necessary states, we can reduce the space of memory. And because it represents and processes new states dynamically, we can cope with the change of environment spontaneously.

키워드 : 지능형 에이전트(Intelligent Agent), 강화 학습(Reinforcement Learning), 기호 학습(Symbolic Learning), 통합 방식(Hybrid System), 행동 선택/학습(Action Selection/Learning)

1. 서 론

오래 전부터, 차세대 사용자 인터페이스로서 로봇이나 가상 캐릭터와 같은 지능형 에이전트에 대한 연구가 진행되어 왔다[2, 11]. 이러한 지능형 에이전트는 주어진 환경을 인식하고, 필요한 목적을 성취하기 위하여 적절한 행위를 선택하는 능력을 가져야 한다. MIT의 Maes는 이러한 지능형 에이전트를 동적이고 복잡한 환경에서 일련의 목적을 만족 시키는 시스템으로 정의 했으며, 센서를 통해서 환경을 파악하고 실행자를 통해서 환경에 적절한 행동을 수행한다고 하였다[4].

이러한 지능형 에이전트를 구현하는 방법으로는 외부 환경을 인식하고 모델링한 후, 문제 해결을 위한 계획(Planning)을 세우고 이를 수행하는 고전적인 하향식 방법론이 있다. 이러한 방법은 변화하는 환경을 모델링 하기 힘들뿐 아니라 계획을 세우고 이를 수행하는 과정이 길어져 주어진 환경에 대하여 빠르게 대응 할 수 없다는 단점을 가진다. 본 논문은 이러한 문제점을 해결하기 위하여 기존 인공 생명에서의 연구 결과를 사용한다. 인공 생명은 생명체가 나타내는 현상을 컴퓨터나 로봇과 같은 인공 매체상에 재현함으로 생명의 일반적인 특성에 대해 연구하는 학문으로서, 이들은 생

※ 본 논문은 숭실대학교에 의해서 지원되었습니다.
† 준 회 원 : 숭실대학교 대학원 컴퓨터학과
†† 정 회 원 : 숭실대학교 컴퓨터학과 교수
논문접수 : 2003년 9월 15일, 심사완료 : 2003년 10월 31일

명체가 환경과 직접 상호작용하면서 현재 주어진 환경에 맞는 적절한 행동을 수행하면서 결국 목적을 성취하는 상향식 방법을 모델링한 행동 기반 구조를 사용한다.

초기에 행동 기반 구조를 이용한 지능형 에이전트 시스템은 내부상태와 외부 환경을 인식하여 최적의 행동을 선택하는데 시스템에 의해 정의된 행동 구조들을 이용하였다. 이러한 방식은 동적인 외부 환경에 대한 적응성이 떨어지기 때문에, 학습 방법을 이용한 행동 선택 방법이 연구되어져 왔으며, 강화 학습을 이용한 환경、적응에 대한 연구가 많이 이뤄져 왔다. 여기서 강화 학습은 현재 상황에 대한 행동을 수행하고 이에 따라 보상을 받음으로 현재 상황에 가장 적절한 행동을 학습하는 방법으로 환경에 대한 사전 지식 없이 적절한 행동을 학습하는 장점을 가진다. 하지만 강화 학습은 최적의 상태와 행동을 찾아내는 수렴 속도가 느리며, 기억 공간이 많이 필요한 단점을 가진다.

본 논문은 강화 학습의 환경에 대한 적응성에 대한 장점을 최대한 살리면서 지연되는 수렴속도와 과중한 기억 공간에 대한 문제를 해결하기 위한 방법론을 연구하였다. 본 논문에서 제안하는 방법은 강화 학습과 같은 행동기반 학습 방법과 기호 학습을 이용하여 규칙을 추출하는 인지 학습 방법을 통합한 방식으로 다음과 같은 특징을 가진다. 첫째, 강화 학습을 이용하여 외부 환경에 적응성 연구를 수행하였으며, 이는 지능형 에이전트의 환경 유연성을 가지도록 하였다. 둘째, 귀납적 기계학습과 연관 규칙을 이용하여 경험에서 필요한 규칙을 추출하고 이를 일반화함으로 주어진 환경뿐 아니라 확장된 환경에서도 보다 빠르고, 효율적으로 행동을 선택을 하도록 하였다. 셋째, 본 논문은 지능형 에이전트를 구현하는데 있어서 모든 상태를 고려하기 보다 상태 탐지기를 이용하여 새로운 상태가 입력될 때마다 상태를 확장시키는 방식을 이용하였다. 이러한 방식은 필요한 상태에 대하여서만 고려함으로 메모리를 획기적으로 축소할 수 있으며, 새로운 상태를 동적으로 처리 할 수 있어, 환경에 대한 변화에 능동적으로 대처 할 수 있다.

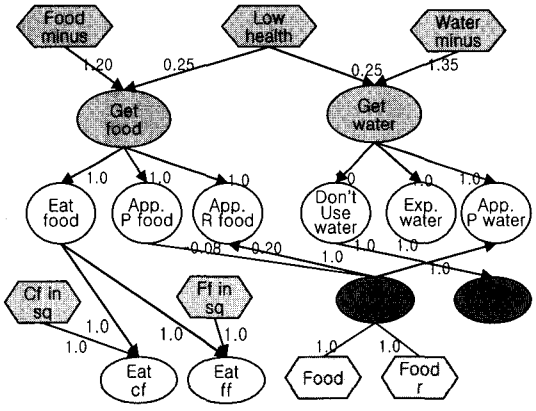
본 논문에서는 주어진 환경에 대한 적응성뿐 아니라 확장된 환경에서의 효율적인 환경 적응성에 대한 학습 효과를 집중적으로 연구하기 위하여 단일 목적에서의 행동 선택 방법으로 연구 범위를 제한하였다. 이는 단일 목적에서의 행동 학습/선택 효율성을 높여, 이를 기반으로 차후의 연구인 다중 목적을 다루는 행동 선택에 대한 전체적인 학습 효율을 높이기 위한 것이다.

2장에서는 지능형 에이전트에 대한 행동 선택 방법에 대한 기존 연구에 대하여 알아보고, 3장에서는 본 논문에서 제안하는 강화 학습과 인지 학습을 통합한 행동 학습/선택 방법에 대하여 설명한다. 마지막으로 4장에서는 시뮬레이션과 실험에 대한 결과를 통하여 제안 시스템의 우수성을 보이도록 한다.

2. 관련 연구

로봇이나 가상 캐릭터의 연구에서 현재 상황을 인식하고 상황에 맞는 적절한 행동을 선택하는 행동 선택 문제에 대한 많은 연구가 진행되어져 왔다. 본 논문에서는 이 중 대표적인 두가지 방법에 대해서 설명하도록 한다. 처음은 Toby Tyrrell의 시스템으로서 초기 시스템 디자인시 행동에 대한 설계를 하고 이를 기반으로 행동을 수행하며, 둘째는 Humphry의 시스템으로서 학습을 통하여 환경 적응성을 가지는 행동을 학습하고 수행한다.

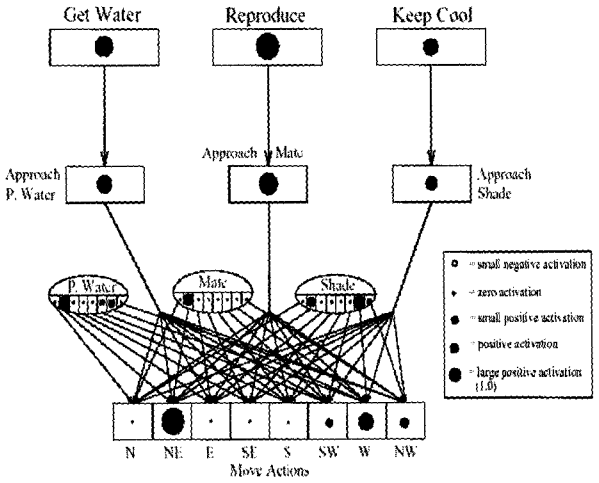
첫째, Tyrrell의 연구는 실제 동물의 생태를 연구하는 생태학의 관점에서 연구 된 것으로, 여러 가지 내적 욕구의 만족도를 최대화 시키는 행동을 선택하도록 한다[8]. 이때, Toby는 (그림 1)에서 처럼 행위에 대한 규칙들을 정의해 놓고 있다. 예를 들어 "Get food"에 대한 행동은 음식을 먹거나, 음식에 가까이 가거나 예전에 음식을 먹었던 위치로 가까이 가는 것으로 분류되고, 음식을 먹는 행위는 씨리얼이나 과일을 먹는 것을 의미하고 가까이 가는 행위는 8방향으로 움직이는 것을 의미한다는 것을 사전에 정의하고 있다. 또한 이러한 규칙에 대하여 각각의 세부적인 가중치를 부여 하고 있어, 이러한 규칙과 가중치를 통해서 현재 상황에 대한 행동 중요치를 계산하여 최적의 행동을 선택하게 된다.



(그림 1) 설계된 행동 선택 구조

Tyrrell의 연구의 있어서의 핵심은 이러한 정의된 규칙과 가중치를 이용할 때, 단지 한가지 목적만을 고려하는 것이 아니라, 여러 가지 목적을 동시에 만족할 수 있는 최적의 행동을 선택하는 방법에 있다. 즉, Tyrrell는 (그림 2)처럼 여러가지 목적이 있을 때, 이러한 모든 목적을 동시에 고려하여 가장 이득이 되는 행동을 선택하는 Free-Flow Hierarchy 구조를 사용한다는 것이다. 하지만 Tyrrell의 행동 선택은 근본적으로 시스템 설계시 디자이너가 정의한 방식을 그대로 이용한다. 즉, 시스템 초기에 외부 환경 모델링을 정확하게 수행해야 하며, 이로 인해 동적인 외부 환경에

대한 적응성을 수행하기에 적합하지 않은 단점을 가지게 된다.



(그림 2) Tyrrell's Free-Flow Hierarchy

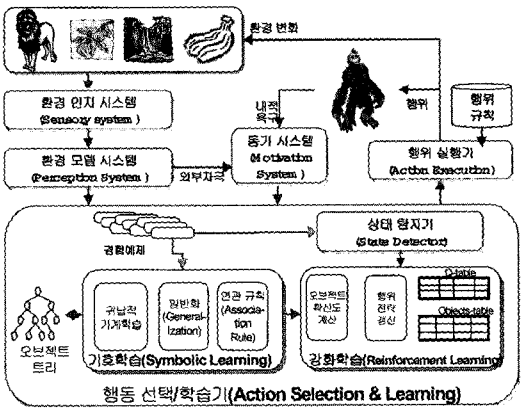
둘째, Humphry는 Tyrrell의 정적인 구조를 개선하기 위하여 학습을 이용한다. 여기서 학습은 외부 환경에 적응하기 위해 필요한 지식을 습득하는 과정으로서 Humphry는 행동에 대한 보상을 통하여 학습을 수행하는 강화 학습을 이용한다[5]. Humphry는 전체 환경에 대하여 하나의 강화 학습을 수행하는 것은 불가능 하다고 판단하여 전체 시스템을 여러 가지 목적에 따라 하위 시스템을 나누고 각각의 시스템에 대하여 강화 학습의 일종은 Q-learning을 이용하여 행동의 패턴을 학습하였다. 이는 강화 학습의 단점인 수렴속도의 지연과 메모리의 부담을 완화하기 위한 방법이었다. 최종적으로 Humphry는 단일 목적에 대하여 Q-Learning을 이용하고 다중 목적에 대하여 W-Learning을 이용하여 환경에 맞는 적절한 행위를 선택 하도록 하였다. 이러한 Humphry 방식은 2단계의 학습을 이용함으로 외부 환경에 대한 적응성 면에서 좋은 효과를 보여주나 기본적으로 강화 학습을 이용함으로, 학습 속도와 메모리의 부담을 가지게 되며, 이로 인하여 환경에 대한 학습이 비효율적일 수 있다.

본 논문에서 제안한 방식은 Tyrrell의 행동 선택에 대한 효율과 Humphry의 외부 환경에 대한 행동 적응성을 모두 고려하였다. 그래서, 본 논문에서 제안한 통합 방식은 Humphry의 방식과 마찬가지로, 외부 환경에 대한 학습을 수행한다. 하지만, Humphry의 방식의 학습 지연과 메모리 사용 부담에 대한 문제를 해결하기 위해서 본 논문은 기호 학습를 이용하여 효과적인 학습을 수행하여 Humphry의 단점을 완화 시켰다. 이는 Tyrrell가 시스템 디자인시 설계했던 규칙들을 에이전트가 환경과 상호작용하면서 동적으로 이끌어 낼 수 있도록 했으며, 이 규칙을 이용하여 기존 강화 학습 보다 빠른 환경 적응을 보일 수 있었다. 이처럼, 우리는 이러한 통합된 학습을 통하여 외부 환경의 빠른 적응성을 얻을 수 있었으며, 확장된 외부 환경에서 효율적인 행동 선택을 수행할 수 있었다.

3. 제안한 행동 선택/학습 알고리즘

본 논문은 지능형 에이전트의 행동 선택의 학습을 통하여 외부 환경에 대한 적응성 및 확장성을 고려하기 위해서 행동 기반 방식뿐 아니라 인지를 이용한 알고리즘을 제안하였다. 이러한 두가지 통합 방식은 행동 수정에 대한 연구를 수행해 온 타 학문에 대한 결과와 일맥 상통한다. 즉, 심리학, 교육학에서 행동 수정을 위한 학습 방법에 대해서 행동과 결과간의 연합에 의한 행동 기반 방식뿐 아니라 정신적 조작에 의한 학습 방법을 이용한다고 정의하고 있다[12]. 이에 본 논문은 행동과 결과간의 연합에 의한 방식으로 강화학습을 이용하고 정신적 조작에 의한 학습방법으로 기호 학습을 이용하여 행동 학습에 대한 효율을 높였다.

다음 그림은 본 논문에서 제안하는 지능형 에이전트의 전체적인 구조이다. 그림에서 보듯이 지능형 에이전트는 크게 외부 환경을 인식하고 필요한 정보를 추출하는 부분, 이를 기반으로 행동을 선택하고 학습하는 부분, 선택된 행동을 수행하는 부분으로 나눌 수 있다.

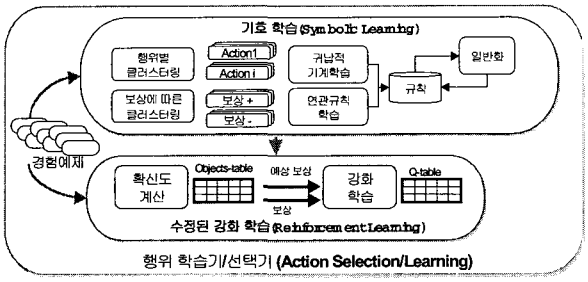


(그림 3) 지능형 에이전트의 시스템 구조

본 논문은 외부환경을 인식하는 부분과 선택한 행동을 수행하는 부분은 간략화하고, 행동을 선택하고 학습하는 행동 선택/학습기부분을 집중적으로 연구하였다. 본 행동 선택/학습기의 특징은 에이전트의 행동 적응에 있어, 강화 학습과 기호 학습을 이용한 통합 방식을 이용한다는 것과, 에이전트를 구현하는데 처음부터 모든 상태를 고려하기 보다 상태 탐지기를 이용하여 새로운 상태가 입력될 때마다 상태를 확장시키는 방식을 이용한다는 것이다. 먼저 본 논문에서 제안한 통합방식의 전체적인 행동 선택 및 학습 구조를 알아보고, 세부적으로 본 논문에서 제안하는 기호 학습과 수정된 강화 학습에 대하여 알아 본 후, 마지막으로 구현 효율을 위해 제안한 상태 탐지기에 대해서 설명하도록 한다.

3.1 행동 선택 및 학습 구조

본 논문에서는 기본적으로 강화 학습을 주된 학습 시스템으로 사용하고 기호 학습을 강화 학습의 가속화를 위한 도구로 활용하였다. 즉, 기존에 강화 학습의 문제점인 학습 속도 지연을 해결하기 위해서 기호 학습의 결과를 강화 학습에 적용하였으며, (그림 4)와 같은 구조를 제안하였다.



(그림 4) 지능형 에이전트의 행위 선택/학습 설계도

그림에서 보듯이, 본 논문에서 제안하는 지능형 에이전트의 행동 학습 및 선택 알고리즘은 2단계로 수행된다. 첫 번째 단계는 기호 학습을 통한 규칙 추출이다. 입력 예제는 행위별 보상별로 클러스터링을 수행한 후 기계학습과 연관 규칙을 이용하여 규칙을 추출한다. 본 논문에서는 이때 추출되는 규칙중에서 어떤 환경요인들이 에이전트의 목적에 맞는지 판단에 관한 규칙을 이용한다. 첫 번째 기호 학습 단계에서 목적은 주로 에이전트의 목적에 맞는 오브젝트를 추출하는데 있다. 둘째 단계는 수정된 강화 학습으로, 환경에 대한 상호 작용을 통하여 강화 학습을 수행하면서, 기호 학습을 통하여 수행된 학습 결과를 이용함으로 보다 빠르게 학습을 수행하도록 하였다. 기존의 강화 학습이 직접 행위에 대한 보상을 받으면서 행위 테이블을 갱신하는데, 본 논문에서는 직접 행위에 대한 보상 이외에 기호 학습 결과를 이용하여 예측할 수 있는 예상 보상값을 이용하여 행위 테이블을 빠르게 갱신한다.

본 논문에서 제안한 2단계 행동 학습 알고리즘은 기존의 강화 학습만을 고려한 학습 알고리즘에 비하여 학습 속도를 향상 시킬수 있으며, 기호 학습만을 고려한 학습 알고리즘에 비하여 환경에 유연성을 가지고 빠르게 행동을 적용할 수 있는 장점을 가진다. 다음절에서 두단계 행동 학습 알고리즘에 대하여 구체적으로 설명하도록 한다.

3.2 기호 학습(Symbolic Learning)

본 논문에서 기호 학습의 역할은 환경에서 에이전트 목적에 필요한 오브젝트를 학습하여, 강화 학습에 적용함으로, 에이전트의 환경 적응 속도를 향상시키는 데 있다. 이처럼, 본 논문에서 기호 학습은 강화 학습에 대한 보조적인 역할로서, 현재 목적에 필요한 오브젝트와 목적에 필요하지 않은 오브젝트에 대한 규칙을 학습 하는데 있다. 지능형 에이전트가 겪는 경험들을 통하여 규칙을 다양하게 추출하기 위하여 귀납적 기계학습, 연관 규칙을 이용하였다. 이때, 귀납적 기계학습은 단편적인 규칙을 추출하고, 연관 규칙을 통하여서는 환경의 공간적인 연관성에 관한 규칙을 추출하고, 일반화를 통하여서는 추출된 규칙과 기반 지식을 통하여 규칙을 일반화하도록 하였다.

지능형 에이전트가 겪는 경험들은 (Xi, Xe, A, R)로 표현한다. 이때 Xi는 지능형 에이전트의 내부적인 상태를 표현하고, Xe는 지능형 에이전트가 볼 수 있는 외부적인 환경을 표현했으며, A는 지능형 에이전트가 수행한 행위, R은 지능형 에이전트가 받은 보상을 표현 하였다. 기본적으로, 이렇게 표현된 경험들은 각 기호 학습의 필요에 따라 입력 예제를 재구성하여 적용된다. 다음은 본 논문에서 제안한 각각의 규칙 추출을 위한 기호 학습 방법에 대해서 자세히 설명하도록 한다.

3.2.1 귀납적 기계학습을 통한 규칙 추출

먼저, 지능형 에이전트의 경험들은 행위별로 클러스터링을 수행하여, 목적에 필요한 행위에 대하여 규칙을 추출한다. 이때 입력 예제들은 (Xi, Xe, A, R)로 표현 됨으로 보상여부에 따라서 행위에 대한 옳고 그름을 판단 할 수 있다. 이를 통해서 목적에 맞는 환경적인 요인과 목적에 맞지 않는 환경적인 요인을 찾을 수 있다. 본 논문에서는 엔트로피 개념을 활용하는 C4.5이라는 귀납적 기계학습 알고리즘을 이용한다. C4.5은 Ross Quinlan의 분류모델(Classification Model)로서, 클러스터를 대상으로 각 클러스터를 대표하는 특성(feature)을 발견하고 분석할 수 있다. 본 논문에서는 보상을 받는 행위에 대한 대표적인 특성인 환경요인을 밝히기 위해 정보 이론(Information Theory)에 근거하는 gain값을 사용하는데 gain값을 구하는 식은 다음과 같다[4, 7].

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \times Entropy(S)$$

$$Entropy(S_v) = -P\oplus Log_2 P\oplus - P\ominus Log_2 P\ominus,$$

$$P\oplus = \frac{\text{보상을받은수}}{\text{전체}},\quad P\ominus = \frac{\text{보상을받지못한수}}{\text{전체}}$$

여기서 S는 전체 집합이며, A는 속성을 나타내는 것으로 여기서는 환경요인을 의미한다. Sv는 속성의 값을 나타내며, P+는 보상을 받은 예제집합을 P-는 보상을 받지 않은 예제 집합을 말한다. 이러한, gain값을 이용한 속성 추출로 현재 목적에 필요한 오브젝트와 목적에 필요하지 않은 오브젝트를 학습할 수 있다.

3.2.2 연관규칙을 통한 규칙 추출

일반적으로, 공간적 상관관계와 시간적인 일련의 상관관계를 고려하기 위해서 연관 규칙을 이용한다. 공간적 상관관계는 A가 있으면 항상 B가 있다라는 환경내의 오브젝트

간의 관계를 말하는 것이고, 시간적인 상관 관계는 A에 도달할 때 항상 수행되는 일련의 행동이나 관찰되는 일련의 오브젝트들간의 관계를 말하는 것이다. 이러한 상관 관계를 파악하는데, 지지도와 신뢰도라는 척도로 그 타당성이 판단된다. 지지도란 전체 입력에 대해서 연관된 항목 집합을 가진 입력이 차지하는 비율을 의미하고, 신뢰도는 조건부 항목 집합에 대해 규칙에 포함되는 모든 항목 집합이 차지하는 비율을 의미한다. 본 논문에서는 공간적 상관 관계를 위해서 연관 규칙을 이용하였으며, 이를 구현하기 위해 Apriori 알고리즘을 이용하였다. Agrawal et al.이 제안한 Apriori 알고리즘은 비교적 빠른 수행 속도를 가지며 다음과 같은 두단계의 연관 규칙 생성 과정을 거친다[9]. 첫째, 최소 지지도 이상을 갖는 빈발 항목 집합을 발견하는 단계이다. 이때, 빈발 항목 집합(Lk)은 이전 단계(k-1)의 빈발 항목 집합에서 K개의 가능한 항목 집합을 생성하여 현 단계의 빈발 항목 집합의 부분 집합이 아닌 경우를 제거하여 후부 항목 집합으로 한다. 이때, 생성한 후부 집합에서 최소 지지도 이상을 가지는 집합을 빈발 항목 집합으로 한다. 둘째, 발견한 빈발 항목 집합의 모든 부분집합을 생성하여 최소 신뢰도 이상인 규칙을 발견하는 단계이다.

본 논문에서는 복잡도를 줄이기 위하여 보상을 가지는 입력 예제에 한하여 연관 규칙을 적용하도록 하여 목적에 맞는 환경요인에 대한 공간적 상관관계를 파악하여 지능형 에이전트의 행동 선택에 이용하였다. 예를 들어 "Hungry"를 느꼈을 때 "Wheat"가 먹는것임을 인식하고, "Wheat가 보통 Water옆에 있다"는 공간적 상관관계를 파악했다면, 지능형 에이전트가 Water를 인지 했을 때 Wheat가 주변에 있다는 것을 인식하여 행동을 취하도록 하였다. 이러한 공간적 상관 관계를 통하여 지능형 에이전트가 환경에서 보다 유연하게 생활 할 수 있도록 하였다.

3.3 수정된 강화 학습

본 논문에서 제안하는 통합 방식은 지능형 에이전트에 의해서 경험을 통하여 매번 강화 학습을 수행하면서, 주기적으로 기호 학습을 이용하여 규칙을 추출 하는 방식이다. 이때, 추출되는 규칙은 강화 학습을 위한 상태 테이블에 영향을 주도록 하였다. 본 논문에서 제안한 수정된 강화 학습에 대하여 설명하기 앞서 기본적인 강화 학습에 대해서 간단히 설명하도록 하겠다.

본 논문에서 사용한 강화 학습의 기본 알고리즘은 Q-learning을 사용한다. Q-learning은 문제의 상태 및 행동공간을 Q-table로 만들고 그 상태에 대한 행동의 적합도를 Q-value로 가지며 행동에 따른 결과로 이 값을 갱신함으로서 학습을 하는 방법이다. 이때, 값을 갱신 하는 최적의 행위 전략 Π(St)는 행위 함수로부터 결정 된다[10].

$$\Pi(S_t) = \arg\max Q(S_t, A)$$

즉, 현재 생태(St)에서 가장 값이 높은 행위(A)를 선택하는 것이다. 이때, Q-learning의 행위 함수는 지능형 에이전트의 계속되는 경험을 이용하여 다음과 같이 갱신된다[10].

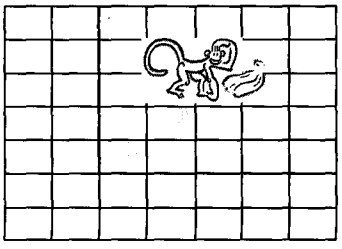
$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Max Q(S_t, A_t) - Q(S_t, A_t)]$$

여기서 St는 현재 상태를, At는 현재 상태에서 취한 행위를, St+1는 현재 상태에서 행위를 취했을 때 일어나는 다음 상태를, R은 직접적인 보상을 말한다.

이처럼 강화 학습은 계속 되는 경험을 통하여 행위 값을 수정하는 Q-table을 유지하고, 이를 기반으로 행위를 선택하게 된다. 하지만 Q-table을 이용하여 지능형 에이전트의 적절한 행위를 학습하는데는 많은 시간이 걸리게 된다. 그래서 본 논문은 기호 학습을 통하여 습득한 규칙을 이용하여 Q-table을 갱신하도록 설계하였다.

이를 위해서 본 논문은 Q-table을 다음과 같이 표현하도록 하였다.

(그림 5)는 Hunger에 대한 Q-table로서, 현재 상태는 지능형 에이전트의 시야에 들어오는 지역전부를 표현 했으며, Action 종류는 Eat와 8방향으로의 Move로 이루어져 있다. 기존의 강화 학습은 Q-table만을 이용하는데, 본 논문은 오브젝트 테이블을 이용하여 환경 요인인 오브젝트들에 대한 학습을 수행하도록 하였다. 오브젝트 테이블은 각 오브젝트가 각 목적에 얼마나 적합한지에 대한 확신도를 가지고 있다. 예를 들면, Hunger라는 문제를 해결해야하는 시스템에서 Wheat가 얼마나 적합한지에 대한 값을 나타내는 것이다. 이때, 오브젝트 테이블의 확신도는 기호 학습을 통하여



환경 오브젝트 테이블

오브젝트	확신도
바나나	1
파인애플	0.75
물	0.5

강화 학습을 위한 행동테이블

환경(State)	Move North	Move East	Move South	Move West	Move NE	Move ES	Move SW	Move MN	eat
****F ***** ***** ***** *****	0	0	0	0	0	0	0	0	0
***** ***F* ***** ***** *****	0	0	0	0	0.5	0	0	0	0
***** ***** *F** ***** *****	0	0	0	0	0	0	0	0	5

상태(State) 행동(Action)

(그림 5) 강화 학습 테이블

계산하게 된다. 이는 강화 학습은 주로 현재 상태에 대한 행동값을 계산하는 것으로, 직접적으로 오브젝트에 대한 학습을 수행하는데는 적당하지 않기 때문이다. 그래서 본 논문에서는 강화 학습 결과와 경험들을 이용하여 기호 학습을 수행하여 오브젝트에 대한 학습을 수행하였다.

본 논문에서 오브젝트의 확신도 계산은 다음과 같은 방법으로 수행하였다. 첫째, 귀납적 기계학습을 통하여 위치 A에 있는 바나나가 먹는것이고, 위치 B에 있는 바나나가 먹는것이다는 경험을 했을 때, 바나나는 먹는것이라는 것을 추출할 수있고, 경험을 통한 확신도를 계산할 수 있다. 이를 오브젝트 테이블에 반영하는 것이다. 둘째, 바나나가 먹는것이고, 망고가 먹는 것이라는 것을 알고, 이들이 비슷한 속성을 가졌다는 것을 파악하고 이들을 한카테고리로 묶고 이 카테고리가 먹는것이다는 것을 인식하게 되며 이에 대한 확신도를 가지게 된다. 이때, 이 카테고리에 속할 확률이 높은 오브젝트가 발견되었을 때, 이 오브젝트는 카테고리의 확신도를 상속 받게 된다. 셋째, 연관 학습을 통하여 오브젝트가 직접적으로 목적에 부합하지 않아도 직접적인 목적과 연관된 오브젝트라고 판단되면 간접적인 확신도를 가지게 하였다. 즉 바나나가 먹는것이고, 바나나 옆에 항상 물이 있다면, 물에 대해서 먹는것에 대한 간접적인 확신도를 가지게 함으로 지능형 에이전트가 행위를 환경에 적응하도록 유도 하였다.

이때, Q-Learning에서 행위함수는 다음과 같이 수정된 보상함수(Func(Rt+1))를 가지게 된다.

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[Func(R_{t+1}) + \gamma MaxQ(S_t, A_t) - Q(S_t, A_t)]$$

$$Func(R_t) \leftarrow \text{if } Rt+1 = 0 \text{ return Pred(Rt+1)} \text{ else return Rt+1 ;}$$

Q-table 내에 보상값이 있다면 그것을 이용하고 보상값이 없으면 오브젝트 테이블을 이용하여 목적에 대한 확신도를 보고 예상 보상값을 이용한다.

다음은 본 논문에서 제안한 강화 학습 알고리즘이다.

1. Q-테이블 및 각 파라미터(a, r)들을 알맞게 초기화 한다.
2. 다음 내용을 반복한다.
   1) Q-테이블로부터 행위를 결정한다.
      이때 임의의 비율로 랜덤 행위를 만들어 주어야 한다.
   2) 행위에 대한 다음 상태와 보상을 얻는다.
      이때 오브젝트 테이블의 확신도를 이용하여 예상 보상값을 계산한다.
   3) 현재 상태의 행위값을 갱신 한다.
   4) 행위 책략을 갱신한다.
   5) 주기적으로 기호 학습을 수행하고 결과를 오브젝트 테이블에 반영한다.
      • 오브젝트 테이블의 확신도 값과 보상값을 갱신한다.

이처럼 본 논문에서 제안하는 지능형 에이전트의 2단계 행동 학습 및 선택 알고리즘은 기호 학습을 통하여 오브젝트를 학습하고, 학습한 내용을 Q-table에 적용함으로 학습의 속도를 향상시켜, 외부 환경의 빠른 적응성과 확장성을 가지도록 하였다.

3.4. 상태 탐지기

본 논문은 지능형 에이전트에서 수정된 강화 학습을 위한 학습 테이블인 Q-table을 구현하는데 있어서 처음부터 모든 상태를 고려하기 보다 상태 탐지기를 이용하여 새로운 상태가 입력될 때마다 상태를 확장시키는 방식을 이용하였다. 이는 에이전트의 상태를 이루는 환경적인 요소가 증가 할수록 이를 모두 표현하는 상태 테이블은 기하급수적으로 증가하기 때문이다. 모든 환경적인 요소를 고려한 상태 테이블을 만드는 것은 메모리 부담뿐 아니라 필요한 상태를 찾는데 시간적 부담 된다. 더욱이 에이전트가 주어진 환경에서 학습을 수행하는 상태들은 극히 일부이다. 그래서 본 논문에서는 초기에 모든 상태를 고려하는 것이 아니라 상태 탐지기를 이용하여 새로운 상태가 입력될 때마다 상태 테이블을 확장시키는 방식을 이용한다. 이러한 방식은 현재 학습에 필요한 상태에 대하여서만 고려함으로 메모리를 획기적으로 축소 할 수 있다. 또한 새로운 상태를 동적으로 처리 할 수 있어, 환경에 대한 변화에 능동적으로 대처 할 수 있다

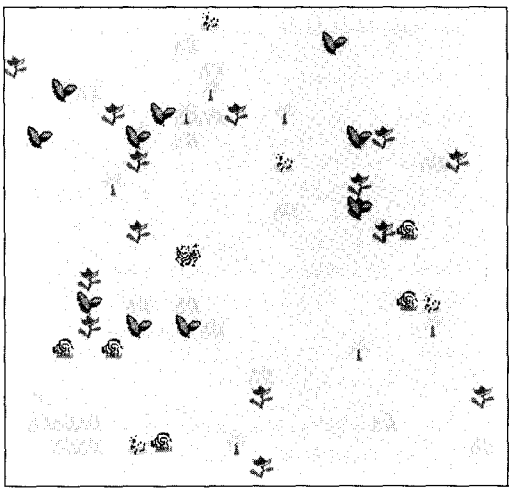
4. 시뮬레이션 및 실험

본 논문에서 제안한 행동 선택 학습 알고리즘의 타당성과 성능을 검증하기 위해서 AI 평원 시뮬레이션에서 실험하도록 한다. 이때, 가상 캐릭터인 지능형 에이전트는 Hunger라는 단일 목적을 가지고 실험을 하도록 한다. 제안한 알고리즘에 대한 평가는 학습 속도와 메모리 사용량 측면을 보고, 외부 환경을 확장하였을 때 학습 속도를 이용하도록 한다. 비교 알고리즘으로는 일반 Q-learning 방법을 이용하도록 한다.

다음은 지능형 에이전트를 실험하기 위한 AI 평원 환경을 보여주고 있다. 이 환경은 Jackson Pauls가 생성한 아프리카 평원을 확장한 것이다[3]. 이 환경에는 물과 늪이 있고 주식으로 <wheat, rice, potato, fish>가 있고, 나무로 <oak, pine>이 있고, 꽃으로 <rose, azalea, lily>가 있다. 이 환경은 기본적으로 격자(grid world)세계로 이루어 졌다. 이 세계에서 각각의 환경들은 파라메터를 조정하여 임의적으로 생성하게 하였다. (그림 6)은 자바로 구현한 40×40의 AI 세계를 보여주고 있다.

이때, 이 세계를 살아가는 지능형 에이전트는 내부적인 욕구로 <hunger>를 가지고 있으며, 시계 범위는 grid에서 2로 제한하였고, 매시간 마다 hunger의 level이 1씩 올라가

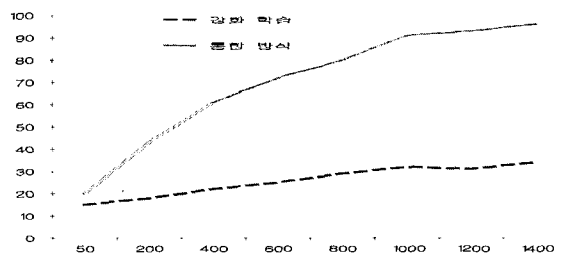
며, 20이 되면 죽게 설정 하였다. 그리고, 지능형 에이전트가 할 수 있는 행동은 먹는것과 이동하는것(8방향)으로 제한 하였고, 지능형 에이전트가 먹을수 있는 것을 먹었을 때 먹은 양에 따라 보상을 받도록 하였다. 이러한 상황에서 지능형 에이전트는 내부적인 욕구와 외부적인 환경을 인식하여 죽지 않도록 행동을 선택하는 방법을 학습하게 되며, 죽었을때는 기존의 지능형 에이전트의 학습 내용을 그대로 잇도록 하였다.



(그림 6) AI 평원

지능형 에이전트가 학습을 하는것은 먹을 것이 무엇인지 알고, 먹을 것이 보였을 때 어떤 식으로 행동을 할 것인지, 그리고 새로운 환경이 있을 때 효율적으로 대처 하는 방법이다. 본 절에서는 지능형 에이전트에 대한 행동 선택을 학습 할때, 학습 속도 면이나 메모리 사용 측면으로 나누어 실험 하도록 한다.

다음은 AI 평원에서 지능형 에이전트의 시간에 따른 평균 수명률을 보여주고 있다. 점선은 일반 Q-learning을 수행한 것이고 실선은 추론된 규칙과 Q-learning을 이용한 통합 방식을 수행한 것이다.



(그림 7) 학습 속도 실험 평가

이 실험을 통하여 기존의 Q-learning에 비하여 추론된 규칙와 연관 규칙를 이용한 Q-learning이 빠르게 생명을 안정화 시켜감을 알 수 있었다. 즉, 본 논문에서 제안한 방식이 에이전트에 대한 학습을 보다 빠르게 수행 시켜 외부 환경에 빠르게 적응 하고 있음을 확인 할 수 있었다.

본 논문에서 행동 선택 알고리즘을 구현하는데 있어서 초기에 모든 상태를 고려하는 것이 아니라 새로운 상태가 습득되면 이를 유동적으로 상태 테이블에 부가하는 방식을 이용한다. 이러한 방식은 불필요한 상태를 생략하고 필요한 상태만을 고려함으로 메모리를 줄일 수 있으며, 변화하는 환경에 동적으로 대처할 수 있었다. 본 시뮬레이션 상황에서 모든 상태를 고려한 경우 강화 학습을 위한 테이블 사이즈는 적어도 $25^{11}\times11$이고, 유동적인 상태 탐지기를 이용하고, 10000번째 이후에 테이블 사이즈는 1600×11이었다. 이러한 실험 결과를 통해 보듯이 상태 탐지기를 이용한 방식은 필요한 상태에 대하여서만 고려함으로 메모리를 획기적으로 축소 할 수 있었다.

5. 결론 및 향후 연구

인간은 자신의 욕구를 느끼고 주어진 환경을 인식하여 살아가기 위한 최선의 행동을 끊임없이 선택한다. 이러한 인간 행동 선택에 대한 메카니즘은 로봇과 같은 인공지능 생명체에 대한 행동 선택 문제에 유용하게 사용 할 수 있다. 본 연구는 인간의 행동 선택에 대한 메카니즘을 인공지능 측면에서 연구하여 효율적인 지능형 에이전트를 만들고자 하였다.

본 논문에서 제안하는 방법은 강화 학습과 같은 행동기반 학습 방법과 귀납적 기계학습을 이용하여 규칙을 추출하고 이를 일반화하는 인지 학습 방법을 통합한 방식이다. 이 통합 방식은 심리학, 철학, 교육학에서 행동 수정에 대한 학습은 행동과 결과간의 연합에 의한 행동 기반 방식뿐 아니라 정신적 조작에 의한 학습 방법을 이용한다는 결과에 근거한 것이다. 이에 본 논문은 행동과 결과간의 연합에 의한 방식으로 강화학습을 이용하여 외부 환경에 대한 유연한 학습을 수행 할 수 있었으며, 정신적 조작에 의한 학습 방법으로 기호 학습을 이용하여 강화 학습의 단점인 학습의 지연과 과중한 메모리 사용부분을 완화 시킬 수 있었다. 이를 통해 외부 환경에 대한 빠른 적응성을 보였으며, 확장된 환경에서 이전의 학습 결과를 효과적으로 이용 할 수 있었다. 더불어 본 논문은 지능형 에이전트를 구현하는데 있어 필요한 상태만을 고려하는 방식을 이용함으로 메모리 공간을 줄일 수 있었으며, 변화하는 환경을 동적으로 표현 할 수 있었다. 본 논문에서 제안한 방식은 AI 평원에서의 시뮬레이션 실험을 통해 우수성을 확인 할 수 있었다.

본 논문에서는 지능형 에이전트에서 외부 환경의 적응성과 확장성을 위한 학습에 대한 집중적인 연구를 위해 단일 목적에서의 학습으로 제한하였다. 차후 연구에서는 기본적

으로 통합 방식의 학습을 수행하는 다중 목적들간의 학습에 대한 연구를 수행하고자 한다. 현재, 여러 가지 목적을 동시에 만족시키는 행동을 학습하는 방법에 대한 연구가 진행되고 있다. 이와 더불어, 감정을 이용한 행동 선택과 행동 선택에 따른 감정의 변화에 대한 연구를 수행하고자 한다. 이와 같은 연구는 지능형 에이전트가 인간과 같은 학습을 수행하며, 감정을 표현함으로 차세대의 인간 친화적인 인터페이스로서의 역할을 수행 할 수 있을 것이라고 예상된다.

참 고 문 헌

[1] Brooks, Rodney A., "A robust layered control system for a mobile robot," IEEE Journal of Robotics and Automation, Vol.2, pp.14-23, 1986.

[2] Carlos Gershenson, "Philosophical Ideas on the Simulation of Social Behaviour," Journal of Artificial Societies and Social Simulation Vol.5, No.3, 2002.

[3] Jackson Pauls, Pigs and People, 4th year report, 2001.

[4] J. R. Quinlan, *C4.5 Programs for Machine Learning*, San Mateo, CA : Morgan, Kaufaman, 1993.

[5] Mark Humphrys, "Action selection methods using reinforcement learning," University of Cambridge, 1997.

[6] Pattie Maes, "Modeling adaptive autonomous agents," Artificial Life Journal, Vol.1, No.1-2, pp.135-162, 1994.

[7] T. Mitchell, Machine Learning, McGraw Hill, 1997.

[8] T. Tyrrell, "Computational Mechanism Action Selection," Ph.D. Thesis, University of Edinburg, 1993.

[9] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," In *Proceedings of the 20th VLDB Conference*, Santiago, Chile, Sept., 1994.

[10] R. Sutton, A. Barto, Reinforcement Learning, MIT Press, 1997.

[11] S. Wermter and R. Sun, Hybrid Neural Systems. Springer-Verlag, Heidelberg, 2000.

[12] 문선모, 인간 학습교육적 적용, 양서원, 2001.

백 혜 정

e-mail : hjbaek@multi.soongsil.ac.kr

1995년 숭실대학교 컴퓨터학과(학사)

1998년 숭실대학교 대학원 컴퓨터학과 (공학석사)

1999년~2003년 숭실대학교 대학원 컴퓨터학과 박사

관심분야 : 인공지능, 에이전트, 전문가 시스템

박 영 택

e-mail : park@computing.soongsil.ac.kr

1978년 서울대학교 전자공학과(학사)

1980년 KAIST 전산학(석사)

1992년 Univ. of Illinois at Urbana-Champaign(박사)

1981년~현재 숭실대학교 컴퓨터학과 교수

관심분야 : 인공지능, 에이전트, 전문가 시스템