2001년 한국정보처리학회 춘계 학술발표논문집 제8권 제1호

강화 학습 및 감독 학습 기반의 지능형 판매 에이전트 시스템

이경은, 고세진, 이필규
인하대학교 전자계산공학과
e-mail:g21981354@inhavision.inha.ac.kr

Reinforcement and Supervised Learning Based Intelligent Sales Agent System

Kyung-Eun Lee, Se-Jin Ko, Phill-Kyu Rhee
Dept. of Computer Science, In-Ha University

요 약

인터넷상에서의 대부분의 검색 환경이 그렇듯이, 인터넷 쇼핑몰에서의 검색 환경 역시 고객 중심으로 제공하는 것이 중요하다. 특히, 고객의 행동 패턴 분석을 통해 얻어진 정보는 고객 중심의 검색 환경을 구성하는 데에 가장 중요한 요소라고 할 수 있으며, 또한 시시각각 변화하는 고객의 심리에 따라서 판매 전략도 달라질 수 있어, 이에 대한 여러 방법들이 연구되고 있는 추세이다. 본 논문에서는 고객과 시스템과의 상호작용으로부터 학습을 최대화시키기 위해 강화학습 기반의 플래닝과 학습의 통합 방법을 통하여 실시간적이고 동적인 인터뷰를 구성하는 방법과 이를 통해 얻어진 개인화된 판매전략과 결정 수와의 통합으로 고객이 원하는 적합한 상품을 추천할 수 있는 방법을 제시한다.

1. 서론

인터넷의 발달로 인해 넘쳐나는 정보를 어떻게 적절히 검색할 것인가에 대한 문제는 특히 사용자 인터페이스 측면에서 가장 중요한 이슈이며, 정보 시스템에서의 인터페이스는 점점 사용자 중심적이 되어가고 있는 추세이다. 온라인 상에서의 전자 상거래 발전이 활기를 띄고 있는 가운데, 맞춤 서비스의 중요성도 이런 맥락에서 주목을 받고 있다. 이미 국내 기업의 여러 웹 사이트들도 초기 단계지만, 이러한 맞춤 서비스 제공을 목적으로 개인의 성향을 분석하기 위한 여러 시도들을 하고 있다.

이러한 맞춤 서비스를 위해 가장 기본적으로 필요한 사항은 고객의 취향 및 의도로서, 사용자와 시스템과의 상호작용 속에는 이런 종류의 많은 정보를 내포하고 있다. 특히, 상호작용으로부터 얻은 정보를 통해 사용자의 취향 및 의도의 순간적인 변화를 인지할 수 있고, 보다 나은 검색 환경을 제공하기 위한 중요한 기준으로 사용될 수도 있다. 또한 전자 상거래에 있어서 사용자 관심의 변화에 주목하는 동적인 인터페이스 기능은 사용자의 만족 도를 높이는 데에 크게 기여하는 반면에 고정적인 질문 순서에 입하면서 내리는 의사 결정은 복잡한 문제에서 큰 문제가 될 수 있으며[1], 무엇보다도 비효율적이다. 또한 고객의 취향 및 의도에 대한 개인적인 정보를 얻기 위해 다양한 기계 학습(Machine Learning) 알고리즘들이 적용되고 있는데[3,4], 이들 방법은 사용자와 시스템과의 상호작용으로부터의 학습에는 적절하지 못하며, 결국 사용자의 과거 행동에 국한되어 학습할 수밖에 없는 한계를 가지고 있다[2].

본 논문에서는 고객의 현재 행동으로부터 학습이 가능하도록, 강화학습(Reinforcement Learning) 방법을 적용하여 고객의 특성에 맞게 최적화되고 개인화된 인터뷰를 구성하는 방법을 제안하며, 이를 통해 얻은 개인화 된 판매 전략과 결정수(Decision Tree)와의 통합을 적용한 상품 추천 방법을 제안한다.

2. 플래닝(Planning)과 학습(Learning)의 통합

플래닝은 에이전트의 행동에 대해 환경이 어떻게

반응할 것인가를 예측하기 위해 환경 모델을 필요로 하는 반면, 학습은 환경 모델을 따로 필요로 하지 않는다. 이렇듯, 플래닝과 학습 사이에는 커다란 차이점이 존재하지만, 평가 함수(Value Function)를 계산한다는 점에서 큰 유사점을 지닌다.

$$V^{\pi}(s) = E_{\pi}\{ R_t | s_t = s\}$$

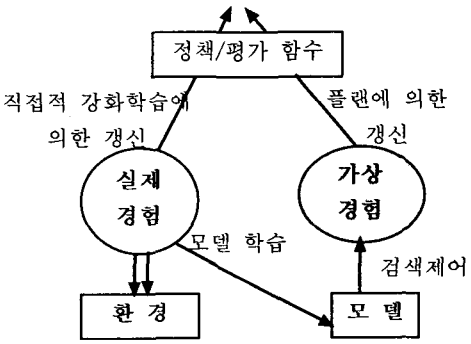
$$= E_{\pi}\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\} \quad [식\ 1]$$

$$Q^{\pi}(s,a) = E_{\pi}\{ R_t | s_t = s,\ a_t = a\}$$

$$= E_{\pi}\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s,\ a_t = a\} \quad [식\ 2]$$

$V^{\pi}(s)$는 특정 정책 하에서의 상태 s의 값으로서, 상태 s에서 시작하여 π를 따를 때의 기대되는 반환 값으로, 정책 π에 대한 상태-평가 함수(State-Value Function)이다. E_{π}는 학습자가 정책 π를 따를 때 주어진 기대되는 평가 값이다.

$Q^{\pi}(s,a)$는 특정 정책 하에서 상태 s안에서 취해질 행동 a의 값으로서, 상태 s에서 시작하고 정책 π 하에서 행동 a를 취했을 때의 기대되는 평가 값이며, 평가 함수 V^{π}, Q^{π} 는 학습자의 경험으로부터 평가된다.



[그림 1] Dyna-Q 구조

플래닝을 통해서 환경 모델을 실제 환경에 보다 가깝도록 향상시키고, 강화 학습 방법을 사용하여 평가 함수와 정책을 직접적으로 향상시킨다. 전자의 경우를 모델학습(Model-Learning)이라 하고, 후자의 경우를 직접적인 강화학습(Direct Reinforcement Learning)이라고 한다. 경험, 모델, 평가 값, 정책 사이에서의 가능한 관계를 Dyna-Q 구조[2]로서 도식화하면 [그림 1]과 같다.

[그림 2]에서 모델(s, a)는 상태-행동 쌍 s, a에 대한 모델의 내용으로서, 예측된 후속 상태와 보상 값이며, (d) 단계에서는 직접적 강화학습이, (e) 단계에서는 모델 학습, (f) 단계에서는 플래닝이 이루어진다. 이 과정에서 행동과 모델 학습, 직접적 강화학습 과정들 안에서는 계산이 거의 필요가 없으며, 단지 각각이 시간의 한 단위를 소비한다고 가정하며, 각 단계에서 남는 시간은 계산이 집중적으로 필요한 플래닝 과정에 쓰여진다.

모든 s∈S와 a∈A(s)에 대해서 Q(s,a)와 모델(s,a)를 초기화 한다.
다음을 반복해서 실행:
(a) s ← 현재 상태(nonterminal state)
(b) a ← ϵ - greedy (s,Q)
(c) 행동 a 실행 ; 결과 상태 s, 보상값 r을 관측한다.
(d) $Q(s,a) \leftarrow Q(s,a)+\alpha[r+\gamma \max_{a'} Q(s',a')-Q(s,a)]$
(e) 모델(s,a) ← s', r (결정적 환경이라고 가정)
(f) N 번 반복:
s ← 먼저 관측된 임의의 상태
a ← s 안에서 이전에 취해진 임의의 행동
s', r ← 모델(s,a)
$Q(s,a) \leftarrow Q(s,a)+\alpha[r+\gamma \max_{a'} Q(s',a')-Q(s,a)]$

[그림 2] Dyna-Q 알고리즘

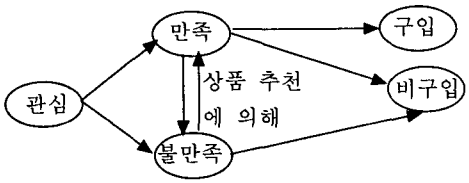
[그림 2]의 알고리즘에서 볼 수 있듯이, 학습은 실제 경험상에서, 플래닝의 경우엔 가상적인 경험상에서 작용하면서 Dyna-Q 에서는 학습과 플래닝이 정확히 같은 알고리즘에 의해서 이루어진다.

3. 시스템 구현

3.1 상호작용의 정의

환경과 학습자 사이의 상호작용을 정의하기 위해서는 상태(states), 행동(actions), 보상(rewards)의 형식적 구조가 필요하다. 일반적으로 행동은 우리가 어떻게 하는 것을 배우길 원하는 일종의 의사 결정일 수 있고, 상태는 행동을 결정하는데 유용한 것을 알 수 있는 어떤 것이라고 할 수 있다.

본 논문에서 제안하는 시스템에서는 상품에 대한 고객의 심리 상태로서 관심, 만족, 불만족, 구입, 비구입의 총 5개의 상태를 가정한다. 웹 상에서 이루어지는 고객의 행동에 따라서 각각의 상태가 정의되며, 각 상태간의 가능한 전이는 [그림 3]과 같다.



[그림 3] 고객의 심리 상태 전이도

고객이 인터뷰를 통해 간단한 질의를 마치고 쇼핑을 희망하면, 일단 소비자의 심리 상태는 관심에 해당한다. 그리고, 관련된 상품의 이미지를 보고, 특정 상품에 관련한 보다 자세한 정보를 요구하게 되면, 그 상품에 어느 정도 만족을 느끼는 것으로 가정한다. 반면에 상품에 대한 이미지를 그냥 보기만 하고, 적극적인 정보 요구가 더 이상 없을 때, 또는 상품에 대한 이미지들을 쇼핑 한 후, 다시 인터뷰를 요청한 경우는 이전의 인터뷰에 대한 불만족 상태로 간주한다. 그 상태에서 로그아웃하고 나가게 되면, 비구입의 상태로 끝나게 된다고 가정하며, 궁극적인 목표 상태는 구입 상태이다.

상호 작용의 또 하나의 요소로서, 인터뷰의 항목을 판매 에이전트가 고객에 대해 상품을 판매할 목적으로 하는 행동으로 가정한다. 따라서 상품 속성의 개수가 행동의 가지 수를 결정 짓게 된다.

강화 학습에서 궁극적인 목적은 마지막으로 얻게 되는 보상 값의 전체 합을 최대화시키는 것이며, 보상 값은 최종적인 목적을 성취했을 때만 양의 보상 값을 부여하고 하위 목적을 성취했을 때는 부여하지 않는다.

3.2 실험 데이터

실험에 적용한 도메인은 의류에 관한 데이터들로서, 의류에 대한 속성들을 기준으로 인터뷰 항목을 지정하였다. 20명의 사용자를 대상으로 실험하였고, 의류 이미지는 각 의류 회사의 홈페이지 및 잡지 site에서 수집되었다.

3.3 시나리오

고객이 로그인 하게되면, 고객에 대한 인식 과정으로서, 고객의 프로파일이 로드 되고 이를 기반으로 판매 에이전트의 판매 전략이 세팅된다. 이는 특정 고객에 대한 판매 에이전트의 경험으로서 적용되며, 고객과의 상호 작용 동안에도 학습되어진다. 처음 들어오는 고객에 대해서는 고객의 나이, 성별, 직업과 같은 정적인 특성을 기반으로 판매 에이전트가 가지고 있는 지식에 의해 인터뷰 내용이 구성된다.

인터뷰는 단계적으로 이루어지며, 실시간적인 플래닝을 기반으로 단계마다 개인화 된 동적인 인터뷰 내용이 구성된다.

사용자의 구입에 대한 상품 정보는 결정 수(Decision Tree)에 의해서 학습되며, 결과적으로 판매 에이전트가 학습한 고객의 구매 패턴은 고객이 구입을 결정할 때, 중요시 여기는 상품의 속성들을 학습하게 된다. 따라서, 상품 추천 시, 이러한 속성들을 기준으로 하게 되고, 특히 결정 수에서 고객에 의해 구입된 과거의 상품들 상에서의 특정 속성 값에 대한 정보를 얻어 이를 추천 시 적용한다.

상품 추천을 제안하는 시기는 고객의 심리가 "불만족" 상태일 때, 인터뷰의 맨 끝 단계일 때와 보다 자세한 검색 조건을 요구할 때이다. 보다 자세한 검색 조건을 요구하는 형식은 인터뷰를 통해 상품 이미지를 보다가 다시 인터뷰를 요청하는 것이 이에 해당한다.

3.4 평가 방법

본 시스템에서 제안하는 인터뷰 방식의 적응성 및 성능을 평가하기 위해서 PARADISE 평가 구조[5]를 적용하였고, 별도로 추천에 대한 평가 방법은 예측 값과 실제 사용자 평가 값 사이의 차이를 표시하는 MAE(Mean Absolute Error)방식을 적용하였다. PARADISE는 어떠한 평가 방법이 시스템의 전체적인 성능을 가장 잘 예측할 수 있는 가를 User Satisfaction과 같은 의미 있는 외부적 기준과 연관된 성능을 가정함으로써 이해할 수 있도록 해준다. 본 실험에서는 다음의 3가지 요소를 중심으로 평가 방법을 구성하였다.

- 작업 성공 : 인터뷰 및 추천의 전체적인 작업 성공 여부
- 인터뷰의 효과성: 시스템 반응(플래닝에 걸린 시간+ 추천에 걸린 시간), 사용자 반응(인터뷰에 대한 만족, 불만족도)
- 인터뷰의 질: 인터뷰 단계, 추천의 질

$$UserSatisfaction = \sum_{i=1}^{n} w_i * N(measure_i)$$

[식 3] UserSatisfaction

강화학습을 통해 얻은 각 개인별 판매 전략과 결정 수와의 통합으로 추천을 한 경우와 결정 수만을

통해 추천을 한 경우의 비교를 MAE 방식을 적용하여 평가하였다.

$$E=\frac{\sum|P-v|}{n}$$

P: 사용자 상품 선호도 예측 값
v: 사용자 실제 평가 값
[식 4] Mean Absolute Error

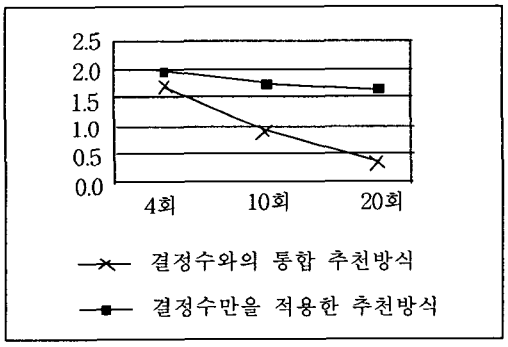
3.5 결과

본 시스템에서 적용한 성능 함수는 각각에 가중치가 부여된 작업 성공 여부, 인터뷰의 질과 효과성의 조합을 사용하여 가장 잘 예견되어질 수 있음을 보이고 있다. 특히, 작업 목적을 달성하기 위한 보다 많은 성공의 경험, 보다 짧은 인터뷰가 성능 향상에 기여함을 보였다.

$$UserSatisfaction = .37N(\text{인터뷰단계}) + .65N(\text{작업성공}) - .12N(\text{시스템반응})$$

[식 5] UserSatisfaction을 적용한 결과

또한 MAE를 이용한 추천 방식의 성능 평가에 있어서, Dyna-Q 알고리즘과 결정 수를 함께 통합된 추천 방식이 결정 수만을 적용한 추천 방식보다 훨씬 높은 성능을 나타내고 있다. 그래프에서 살펴보면, MAE 값이 훨씬 적게 나오고 있음을 확인할 수 있다.



[그림 4] MAE 방식을 적용한 추천 방식의 비교

3.6 결론

본 논문에서는 고객과 시스템 사이의 상호작용으로부터 목표 지향적인 학습을 위해 개인화된 동적 인터뷰 구성 방식과 추천 방식을 강화 학습 방법을 적용하여 제안하였다. 이는 강화 학습 방법이 Feature Selection에서도 좋은 결과를 보여주고 있음을 보여주는 예로서, 결정 수와의 통합에 있어서 그 성능을 보다 향상시켜주는 결과를 보였다. 이는 실제 상품 매장에서의 판매원이 고객을 상대로 여러 단계의 인터뷰를 통해 고객이 원하는 상품을 보다 쉽게 찾을 수 있도록 상품의 범위를 축소시킴과 동시에, 자신의 경험을 통해 고객의 요구에 맞는 상품을 추천하는 판매 전략을 시도하는 실제 모델을 적용시킨 예로서, 순간마다 고객에 대해서 반응하면서, 고객에 적합한 판매전략을 세우는 내부 메커니즘과 동일하다. 이러한 강화학습 기반의 인터뷰 방식과 추천 방식은 다양한 기계 학습 알고리즘과의 통합으로 보다 상호 작용적인 사용자 모델링 기술에 활발히 적용되어질 수 있으며, 사용자의 감성을 고려한 지능형 인터페이스 에이전트의 설계에 있어서도 활발한 연구가 진행되고 있다.

참고문헌

[1] Ivo Vollrath, Wolfgang Wilke, and Ralph Bergmann, "Case-based Reasoning Support for Online Catalog Sales", Journal of IEEE Internet Computing , pp47-54, July August 1998.

[2] Richard S. Sutton and Andrew G. Barto, "Reinforcement Learning", The MIT Press, 1998.

[3] Henry Lieberman, "Letizia: An Agent That Assists Web Browsing", Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, pp1704-1710, Montreal Canada, 1995.

[4] Armstrong, R., Freitag, D., Joachims, T., and Mitchell, T., "Web-Watcher: A learning apprentice for the World Wide Web", AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, March, 1995.

[5] Walker, M., Litman, D., Kamm, C., and Abella, A., "PARADISE: A general framework for evaluating spoken dialogue agents", Proc. 35th Annual Meeting of the Association for Computational Linguistics and 8th Conf. of the European Chapter of the Association for Computational Linguistics, pp.271-280, 1997.