# 강화학습을 사용한 연관성 피드백

# Relative Feedback with Reinforcement Learning

이승준*, 장병탁

{sjlee*,btzhang}@bi.snu.kr

서울대학교 전기컴퓨터공학부

Seung Joon Yi*, Byoung Tak Zhang,
School of Computer Science and Engineering, Seoul National University

**요약**

본 논문은 웹 문서 여과시 사용자 모델링을 위해 사용되는 연관성 피드백 방법을 강화 학습 프레임웍에서 분석하고 강화학습 기반의 새로운 연관성 피드백 알고리즘을 제안한다. 제안된 방법은 강화 학습 프레임웍상에서 기존의 방법을 일반화한 것으로 기존의 연관성 피드백 방법이 현재의 프로파일만을 상태로 사용하는 데 비해 과거 history부터 얻는 추가 정보를 사용하는 방법이다.

**Keywords** : Reinforcement Learning, Relevance Feedback, Web browsing

## Ⅰ. 서 론

인터넷의 발달로 인해 이용 가능한 전자 정보량은 기하급수적으로 증가하고 있다. 정보의 양이 방대해질수록 사용자가 각각의 정보들을 일일이 확인해 보기는 보다 힘들어지게 되고, 따라서 사용자를 대신하여 이런 정보들 중 사용자의 선호도에 맞는 정보를 선별해 주는 자동화된 시스템이 보다 요구되고 있다[1].

개인화된 지능적 정보 에이전트(Personalized intelligent information agent) 란 월드 와이드 웹과 같은 방대한 정보 집합 속에서 사용자의 정보 요구 혹은 관심, 선호도에 대한 관련 정보를 제시함으로써 사용자를 돕도록 의도되어진 지능적인 시스템이라고 정의할 수 있다. 개인화된 지능적 정보 에이전트는 사용자의 정보 요구와 선호도를 직접, 간접적으로 학습하여 사용자 프로파일을 구축하게 된다[2].

이러한 선호도를 학습하기 위해 널리 이용되는 방법이 사용자의 평가로부터 사용자의 선호도를 학습하는 연관성 피드백이다.

이러한 연관성 피드백 방법은 정답이 주어지지 않고 사용자의 평가만이 수치적으로 주어지는 환경에서 평가를 최대화한다는 데에서 강화학습의 문제로 볼 수 있다. 본 논문은 연관성 피드백 알고리즘을 강화학습 관점에서 분석하고 강화학습 알고리즘을 적용한 확장된 연관성 피드백 알고리즘을 제안한다.

## Ⅱ. 관련 연구

### 2.1 연관성 피드백

연관성 피드백 방법은 정보 추출 분야에서 오랜 기간 연구되어 왔고 피드백이 없을 경우에 비해 많은 성능 향상을 가져오는 것이 알려져 왔다. 일반적인 연관성 피드백 모델에서는 각 문서들에 사용자에 의한 적합도 평가값이 주어지게 되고, 문서와 그에 해당하는 적합도 평가값으로부터 보다 평가값이 높은 문서들을 얻을 수 있도록 질의문(query)이 수정되게 된다[3].

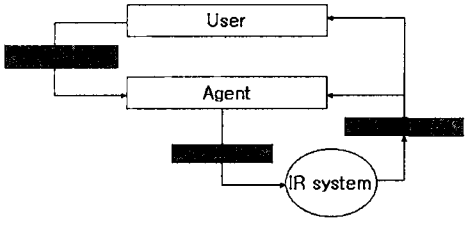


그림 1. 연관성 피드백 framework

가장 대표적인 연관성 피드백 방법은 Rocchio의 알고리즘이다. 질의문은 다음과 같이 수정되게 된다.

$$Q' = Q_0 + \frac{1}{n_1}\alpha\sum_{i=1}^{n_1} R_i - \frac{1}{n_2}\beta\sum_{i=1}^{n_2} S_i \quad (1)$$

이 방법은 전체 문서를 한꺼번에 처리하는 배치 알고리즘으로써 이를 실시간화한 방법으로는 Widrow-Hoff 알고리즘이 알려져 있다. WH 알고리즘은 다음과 같이 질의문을 수정한다.

$$\Delta W = -2\eta(WX - y)X \quad (2)$$

### 2.2 강화학습

강화학습은(reinforcement learning) 동적인 환경 하에서 시행착오를 거쳐 환경으로부터 주어지는 보상(reward)을 최대화하기 위한 학습 방법이다[4].

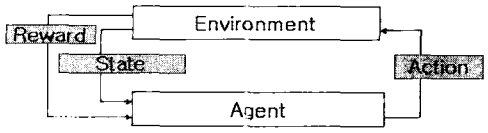


그림 2. 강화 학습 framework

학습의 주체인 에이전트(agent)는 환경의 상태(state)를 관측하고 과거의 경험을 바탕으로 행동(action)을 선택하면 그에 따른 보상(reward)을 환경으로부터 받게 된다. 강화학습의 목표는 장래까지 고려한 보상, 즉 다음과 같은 값을 최대화할 수 있는 행동을 학습하는 것이 된다.

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} \ldots = \sum_{k=0}^{\infty}\gamma^k r_{t+k+1} \quad (3)$$

### 2.3 Q-Learning

Q-Learning [5]은 현존 강화 학습 방법들 중 대표적으로 쓰이는 방법으로써 시간 변화에 따른 적합도 차이를 학습에 이용하는 TD-Learning의 한 종류이다. Q-Learning에서는 아래에 정의된 optimal Q-value $Q^*(s,a)$를 직접 학습한다. 이 값은 상태 s에서 행동 a를 취한 후 최적으로 행동했을 경우의 보상의 총합을 나타낸다.

$$Q^*(s,a) = E\{r_{t+1} + \gamma \max_{a'} Q^*(s_{t+1}, a') | s_t = s, a_t = a\} \quad (4)$$

Q-Learning의 한 step은 다음과 같이 이루어진다.

1. 현재 상태를 s라 하자.
2. 행동 a를 선택한다.
3. a를 행해서 받은 보상을 r,
다음 상태를 t라 하면
4. Q(s,a)를 다음과 같이 수정한다.

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a'} Q(t,a')) \quad (5)$$

그림 3. Q-Learning

모든 행동이 무한히 시행되며 학습률 $\alpha$를 적절히 줄이며 학습시킬 경우 $Q(s,a)$는 모든 s,a에 대해 optimal Q-value인 $Q^*(s,a)$에 수렴한다는 것이 증명되어 있다[5]. 이 방법은 모델의 정보 없이 행동의 적합성을 나타내는 Q값만을 학습하므로 구현하기 간단하며 실제 여러 문제에 사용되어 좋은 결과를 보이고 있다.

## III. 강화학습으로서의 연관성 피드백

### 3.1 일반적인 연관성 피드백

그림 1과 그림 2를 비교해 보면 연관성 피드백을 강화 학습의 테두리에 넣을 수 있음을 알 수 있다. 다음과 같이 강화 학습의 각 요소를 대응시킴으로써 일반적인 임의의 연관성 피드백 방식을 구현할 수 있다. 행동은 질의문 자체 (혹은 수정) 이 되고 그에 따른 보상은 문서에 대한 사용자의 평가, 일반적인 연관성 피드백 방법은 구체적으로 다음과 같은 대응이 가능하다. Action은 질의문 자체 (혹은 그 수정)이 되고 Reward는 문서에 대한 사용자의 피드백, 그리고 State는 현재까지의 모든 history로 대응 가능하다. 이 프레임웍에서는 임의의 연관성 피드백 방식을 모두 포괄이 가능하나 직접적인 구현은 state수가 지수적으로 증가하기 때문에 불가능하다.

### 3.2 Rocchio algorithm

로키오 방법에서는 전체 데이터를 사용하여 batch 방식으로 질의문을 구한다. 이 경우 history의 순서가 고려되지 않기 때문에 문서가 고정이 될 경우 history는 고정되고 state는 하나로 줄어들게 된다. 즉 하나의 state에서 한 번의 action을 행해 한 번 학습하게 되기 때문에 여러번의 시행을 통해 학습하는 강화학습 프레임웍으로 확장이 무의미하다.

### 3.3 Widrow-Hoff algorithm

온라인 학습 방법인 Widrow-Hoff 방식에서는 기존의 history는 사용하지 않고 현재의 데이터만 가지고 질의문을 고쳐 나가게 된다. 이

방법에서는 기존 history 정보 전체 대신 기존의 질의문 벡터 하나만을 상태의 근사값으로 사용하게 된다. 즉 부분적으로 관측가능한(Partially Observable)한 상황이다[6]. 일반적인 하나의 벡터만을 사용하기 때문에 간단하게 구현 및 실행이 가능하지만 이제까지의 history 정보를 사용하지 않기 때문에 일반적인 RL 하에서보다 이론적으로 성능이 높을 수 없다.

### 3.4 Simplified RL algorithm

본 논문에서 제시하는 방법으로 history를 사용하면서도 일반적인 RL 방법을 적용시 상태의 수가 너무 커지는 것을 보완하기 위하여 상태를 근사하여 사용하는 방법이다.

일반적인 강화 학습에서는 Q-function은 다음과 같이 s와 a의 함수이다.

$$Q(s,a)=f(s,a) \quad (6)$$

만일 Q를 profile과 query의 similarity로 정의하면

$$Q(s,a)=WX(\ W,X\text{는 normal vector}) \quad (7)$$

(5)에 대입하고 W가 근사적으로 같다고 하면

$$\Delta WX=-\alpha WX+\alpha r+\alpha\gamma \quad (8)$$

X를 곱하면

$$\Delta W=-\alpha(WX-r-\gamma)X \quad (9)$$

즉 Widrow-Hoff 알고리즘은 일반적인 강화 학습 방법 중의 특별한 경우, 즉 Q(s,a)=WX인 경우에 해당하게 된다.

보다 복잡한 Q-function을 사용할 경우 history를 더 잘 고려할 수 있어 성능 향상을 기대할 수 있다. Web browsing 시 초기 profile이 학습되지 않은 상태에서는 하나하나의 연관성 피드백이 profile 형성에 큰 영향을 미칠 수 있지만 많은 data가 얻어진 이후에는 상대적으로 영향을 덜 미치는 것을 고려하기 위해서는 Q-function의 분산을 학습 과정에 따라 줄이는 방법을 사용할 수 있다. 본 논문에서 제안하는 Q-function은

$$Q(s,a)=(WX)^{f(s)} \quad (10)$$

로 $f(s)$는 history 증가에 따라 증가하게 된다. 이 경우 Q value의 차이는 similarity가 같을 경우 history 증가에 따라 점차 줄어들게 된다.

근사된 Profile update 식은 위 식을 (5)에 대입하여 각 항을 근사하면

$$\Delta W=-\frac{\alpha}{f(s)}(WX-r) \quad (11)$$

이 되어 학습율이 단조 감소하는 WH와 같은 형태가 된다.

document들의 상호 similarity등을 고려한 더 복잡한 Q-function의 사용도 가능하다. 그럴 경우 위 식들처럼 profile update 식을 만들어 내기가 어렵다. 하지만 그럴 경우에도 일반적인 Q-function 수정 룰인(5)식을 사용하여 학습이 가능하다. 단 Q-function을 파라미터를 사용하여 근사할 경우 (5)식 사용시 최적의 값으로 수렴하지 않을 수 있다[7].

## V. 결론 및 향후과제

본 논문에서는 기존의 웹 브라우징에서 사용되던 연관성 피드백 방법을 강화 학습의 프레임웍 하에서 분석해 보았다. 일반적인 연관성 피드백에 적용 할 수 있는 일반적 강화 학습 방법을 제안하였고 기존의 방법들이 강화 학습 체계에서의 다른 Q함수들로 대응됨을 보였다.

실험은 온라인 실험시 재현성이 떨어지고 시간이 오래 걸리는 문제가 있기 때문에 TREC 데이터를 사용한 오프라인 시뮬레이션으로 계획하였다. 여기 제안된 Q-function 외에도 문서 데이터의 특징을 살릴 수 있는 여러 가지 Q-function들간의 성능 비교도 행할 예정이다.

**감사의 글** : 본 연구는 교육부 BK21 프로그램에 의해 지원 받았습니다.

## VI. 참고문헌

[1] Maes,P. *"Agents that Reduce Work and Information Overload"* Communications of the ACM, July 1994, vol. 37(7), 31-40, 146.

[2] Seo,Y.,Zhang,B.,*"Personalized Web Document Filtering Using Reinforcement Learning"*, Applied Artificial Intelligence, vol. 15, 2001.

[3] Leski, A. "Relevance and Reinforcement in Interactive Browsing", International Conference on Information and Knowledge Management, pp. 119-126, 2000.

[4] Sutton,R.S. and Barto,A.G. *Reinforcement Learning: An Introduction*, MIT Press, 1998

[5] Watkins,C.J. and Dayan,P. *Q-Learning.* Machine Learning, 8(3):279-292, 1992.

[6] Kaelbling,L.P. , Littman,M.L. and Cassandra,A.R.,*"Planning and Acting in Partially Observable Stochastic Domains,"* Artificial Intelligence, Vol. 101, 1998

[7] Boyan,J.A., and Moore,A.W. *"Generalization in reinforcement learning: Safely approximating the value function."* ,Advances in neural Information Processing Systems, volume 7. 1995.