강화학습 기반 사용자 프로파일 학습

김영란[0] 한현구
한국외국어대학교 컴퓨터 및 정보통신공학과

Learning User Profile with Reinforcement Learning

Younglan Kim[0] Hyungoo Han
School of Computer & Information Communinations Engineering,
Hankuk University of Foreign Studies

요 약

정보검색 태스크에서 사용자 모델링의 목적은 관련정보 검색을 용이하게 해주기 위하여 사용자의 관심도 또는 필요정보의 모델을 학습하는 것으로 시간적인 속성(temporal characteristics)을 가지며 관심이동을 적절하게 반영하여야 한다. 강화학습은 정답이 주어지지 않고 사용자의 평가만이 수치적으로 주어지는 환경에서 평가를 최대화 한다는 목표를 가지므로 사용자 프로파일 학습에 적용할 수 있다. 본 논문에서는 사용자가 문서에 대해 행하는 일련의 행위를 평가값으로 하여 사용자가 선호하는 용어를 추출한 후, 사용자 프로파일을 강화학습 알고리즘으로 학습하는 방법을 제안한다. 사용자의 선호도에 적응하는 능력을 유지하기 위하여 . 지역 최대값들을 피할 수 있고, 가장 좋은 장기간 최적정책에 수렴하는 R-Learning을 적용한다. R-learning은 할인된 보상값의 최적화보다 평균 보상값을 최적화하기 때문에 장기적인 사용자 모델링에 적합하다는 것을 제시한다

1. 서론

사용자에게 맞춤 정보를 제공하는 개인 웹 에이전트는 사용자 프로파일을 최적의 상태로 유지하여야 한다. 사용자 프로파일은 사용자들의 행동패턴(action pattern) 또는 선호도(preference)[1]를 학습하여 유지할 수 있으며, 최적의 상태로 유지하기 위해서는 사용자의 관심이동(concept drift)을 적절하게 반영할 수 있어야 한다.

선호도를 학습하는 방법에는 명확한 목적 진술이나 연관성 평가를 요구하는 명시적인 방법[2]과 사용자가 인식할 수 없는 상태에서 사용자 행위를 기록하여 이들 행위에 따라 추론하는 묵시적인 방법[3]이 있다. 선호도 학습은 정답이 주어지지 않고 사용자의 평가만이 수치적으로 주어지는 환경에서 평가를 최대화하는 것이 목표이므로 강화학습을 적용할 수 있다.

강화 학습은 한 에이전트가 자신이 놓여진 환경으로부터의 보상을 최대화할 수 있는 최적의 행동전략을 학습하는 것이다. 강화학습 기법 중에서 Q-learning은 할인상수 γ 가 고정된 값으로 반영되므로 최단기간의 보상에서는 부-최적화 정책(sub optimal policy)에 수렴할 수 있다. 그러나 R-learning[4]은 이런 지역 최대값들을 피할 수 있고, 장기간동안 가장 좋은 최적정책에 수렴한다.

본 논문에서는 사용자의 문서에 대한 관심도에서 사용자가 선호하는 용어를 추출하여, 이를 강화학습 알고리즘으로 사용자 프로파일을 학습하는 방법을 제안한다. 사용자의 선호도에 대한 적응력을 유지하기 위하여 R-Learning을 적용함으로써 할인된 보상값을 최적화하는 것보다 평균 보상값을 최적화하는 R-learning이 장기적인 사용자모델링에 적합하다는 것을 제시한다.

2. 관련연구

2.1 사용자 모델링

사용자 모델링을 실세계에 적용하는데 제약이 되는 중대한 쟁점은 현재 4가지이다[1]. 첫째, 실세계의 사용자 모델링작업에서 정밀도를 갖기 위해 학습 알고리즘은 대량의 데이터 집합을 필요로 한다. 둘째, 관심도를 표현한 레이블된 데이터를 필요로 한다. 셋째, 계산복잡도이다. 마지막으로, 관심이동이다. 정보검색 태스크에서 사용자 모델링의 목적은 관련정보 검색을 용이하게 해주기 위하여 사용자의 관심도 또는 필요정보의 모델을 학습하는 것으로 시간적인 속성을 갖는다. 최근에 수집된 사용자 데이터는 이전의 시점에서 오는 데이터 보다 좀더 정밀한 현재의 지식, 선호도를 반영한다고 가정할 수 있다. 그러나 최근 데이터에 모델을 제한하는 것은 지나치게 특정 모델로 유도될 수 있다.

Billsus와 Pazzani[5]는 사용자의 흥미 변화에 적응할 수 있도록 관찰 시간별로 적응 속도의 차이를 두는 단기 모델과 장기 모델을 적용한 NewsDude라는 지능 에이전트를 기술하였다. Nearest-Neighbor 분류 알고리즘을 사용하여 뉴스에서의 최근 관측을 사용자 관심에 대한 단기간 모델로 만든다. 만약에 단기간 모델이 충분한 확신이 있는 예측을 만들 수 없으면, 분류는 장기간 동안에 수집된 관측을 기반으로 하는 나이브 베이스 분류기에게 위임된다. 이 구조는 장기간 후에 수집된 데이터의 잠재적인 이익을 희생하지 않고, 시스템이 급격하게 변하는 관심도에 적응할 수 있게 한다.

Widmer 와 Kubat[6]은 숨겨진 정황과 관심의 이동을 가진 환경에서는 효과적인 학습을 위해서 명확히 나타나지 않는 정황 변화를 탐지해 낼 수 있고 변화를 빠르게 복원하고, 새로운 정황에 대해서 기존 관심의 조절을 가능케 하며, 다시 나타나게 될 관심을 위해서 과거 경험을 사용할 수 있는 알고리즘을 필요로 한다고 기술하였다.

2.2 사용자 행위분석을 이용한 프로파일 학습

Syskill & Webert[3]는 웹을 브라우징하는 동안 사용자가 아이콘(thumbs up 또는 thumbs down)을 클릭하게 하여

*본 연구는 2002학년도 교내학술연구비 지원에 의하여 연구되었음.

관심도를 명시적으로 식별한다. 사용자 평가가 세션마다 따로 유지되고 MDL(minimum description length)원리를 이용하여 사용자 프로파일을 학습한다. 이에 적합한 문서를 분류하기 위해 나이브 베이스 분류기를 사용한다.

Letizia[7]는 사용자가 다양한 action을 수행한 것을 가지고 (예: 프린트 또는 북마크에 저장) 사용자가 웹 페이지에 관심이 있는지 추론한다. 사용자 행위를 추적하는 것으로 검색된 문서의 현재 위치에서 링크에 공존하거나 또는 자율적으로 탐색하여 관심도가 있는 사항을 예상하도록 한다. 에이전트는 Best-First Search로 구성된 브라우징 전략을 자동으로 조작할 수 있게 한다. 이 탐색전략은 사용자 행위로부터 나온 관심도를 휴리스틱 기법으로 추론하여 확장한다.

2.3 강화학습 기법

강화학습은 마코프 의사결정 문제(MDP: Markov decision Problem) 프레임 워크를 기반으로 한다. MDP는 상태집합 S, 행동집합 A, 그리고 상태(또는 행동)에 사상되는 보상함수로 구성된다. 학습의 목표는 장기간 환경으로부터 받는 보상의 기대값을 최대화하는 정책 $\pi : S \rightarrow A$ 를 결정하는 것이다[8].

그러므로 강화학습은 입력상태와 출력행동의 쌍으로 명확한 훈련 예들이 제공되는 감독학습과는 다르다. 특히, 시간차 (TD: Temporal-Difference)학습은 학습할 환경에 대한 명확한 모델이 있어야 하는 DP(Dynamic Programming)과 online 또는 시뮬레이션된 환경과 경험으로 학습을 하는 MC(Monte Carlo)방식의 절충적인 형태로 강화학습기법에서 가장 많이 사용되는 형태이다. 강화학습 기법에는 Q-learning, TD (λ) , 그리고 R-learning 등이 있다. Q-learning과 R-learning은 비모델 기반의 강화학습(model free reinforcement learning)으로 사전에 환경에 대한 별다른 모델을 설정하거나 학습할 필요가 없으며, 다양한 상태와 행동들을 충분히 자주 경험할 수 있으면 최적의 행동전략에 도달할 수 있어 다양한 응용분야에 적용되고 있다.

Q-Learning은 상태 s에 대한 행동 a의 가치함수 Q(s,a)를 기반으로, 각 단계마다 학습자는 최대 가치함수 값을 가진 행동 a를 선택한다. Q-Learning은 미래 보상값들을 할인하기 때문에 장기간동안 보상값들을 유지하는 행동보다는 단기간에 끝나는 행동을 선호한다. 그러나, 학습자는 지역 최대값에 수렴하지 않는 것을 보장하기 위해 때때로 부-최적화 행동을 탐험할 필요가 있다. Q-Learning은 매 번의 행동 후에 Q(s,a) 값은 다음 규칙으로 갱신한다.

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left[r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)\right] \quad (1)$$

r 은 행동 a 의 수행에 대한 보상(reward)이다. γ 는 Q 가치함수 값의 적합도를 보장하는 할인상수(discount rate: $0 \le \gamma \le 1$) 이다. 이것은 먼저 행한 행동에 대한 보상값은 높은 값으로 정하고, 같은 행동에 대해서도 시간이 흐름에 따라 그 보상값을 감소시키는 것이다. 즉, 할인상수가 γ 라면 어떤 행동이 n번째 단계에 행해졌을 때의 보상값은 초기 보상값에 γ^n 을 곱한 값이 된다. α 는 학습률(learning rate)이다.

Schwartz[4]는 단계마다 평균 보상을 최대화하는 R-learning을 제안했다. 상태 s에서 행동 a를 선택하면 가치 함수 R(s,a)로 추론한다. 각 상황에서 학습자는 가끔 탐험의(exploratory) 부-최적화 행동을 선택하는 경우를 제외하고는 최대 R(s,a) 값을 가진 행동을 선택한다. R(s,a)값은 각 행동 후에 다음 학습 규칙으로 조정된다.

$$R(s_t,a_t) \leftarrow R(s_t,a_t) + \alpha\{r - \rho + \max_{a'} R(s_{t+1},a_{t+1}) - R(s_t,a_t)\} \quad (2)$$

Q-learning 과 차이점은 즉시 보상값 r 에서 평균보상값 ρ 를 감산하고, 다음 행동-가치 $R(s_{t+1},a_{t+1})$ (모든 행동 a에 대한 최대값)을 할인하지 않는다는 것이다. 평균 보상값 ρ 는 다음과 같이 평가된다.

$$\rho_t \leftarrow \rho_t + \beta(r + R(s_{t+1},a_{t+1}) - R(s_t,a_t) - \rho_t) \quad (3)$$

3. 사용자 프로파일 학습

개인 웹 에이전트의 사용자 모델링 설계는 다음과 같은 사항을 만족하여야 한다[9]. 첫째, 별개의 주제에서 복수의 선호도를 표현할 수 있는 사용자 프로파일 이어야 한다. 둘째, 변화된 사용자의 선호도에 신속하고 합리적으로 적응하는 융통성이 충분히 있어야 한다.

이와 같은 사항을 만족시키기 위하여, 사용자 프로파일을 주제별로 동적으로 생성하며 사용자의 관심도가 일정 기간 내에 정기적인 출현인가 아닌가를 구별하고 단계마다 평균 보상값을 최대화하여 학습한다.

3.1 연관성 평가

기존의 사용자 프로파일과 검색된 문서간의 유사도가 높다는 것은 사용자가 그 주제에 관해 관심도를 정기적으로 나타낸다는 것으로 간주한다.

사용자의 프로파일과 문서간의 유사도를 측정하기 위해 웹 문서를 TF 벡터로 변환하고, 용어 빈도 순위가 높은 n개로 특징 벡터를 구성한다. 유사도 정량화는 두 벡터간의 cosine 유사도 측정법을 사용한다.

$P_k = (w_{1k}, w_{2k}, \Lambda, w_{nk})$ 프로파일 주제 k의 벡터

$d_j = (w_{1j}, w_{2j}, \Lambda, w_{nj})$ 웹 문서 j의 벡터

$$sim(d_j, p_k) = \frac{\overrightarrow{d_j} \bullet \overrightarrow{p_k}}{|\overrightarrow{d_j}| \bullet |\overrightarrow{p_k}|} = \frac{\sum_{i=1}^{n} w_{i,j} \times w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{n} w_{i,k}^2}} \quad (4)$$

3.2 선호도 학습

선호도를 학습하는 방법에는 사용자에게서 연관성 평가를 직접 받는 명시적인 방법과 사용자 행위를 기록하여 이들 행위에 따라 추론하는 묵시적인 방법이 있다. 선호도 학습은 정답이 주어지지 않고 사용자의 평가만이 수치적으로 주어지는 환경에서 평가를 최대화 한다는 목표를 가지므로 강화학습을 적용할 수 있다.

강화학습은 시각 t 에 상태 s_t 에서 학습자의 정책에 따라 행동 a_t 를 행하고, 자신이 위치한 환경으로부터 스칼라 값의 보상값 r_{t+1} 을 받고 상태는 s_{t+1} 로 전이된다. 여기서 정책은 주어진 시각에 학습자가 취할 수 있는 행동집합에서 특정 행동을 선택하는 기준이다. 강화학습의 목표는 장기간 환경으로부터 받는 기대 보상값을 최대화 시키는 것이다[8].

본 논문에서 학습 에이전트의 상태 s_t 는 사용자 프로파일에 있는 용어들이며, 행동 a_t 는 웹 문서에 대한 사용자의 행위 정보 분석에 의해 용어를 재구성하는 것이다. 보상값 r_{t+1} 는 사용자가 문서 d_j 에 보인 사용자의 행위를 관심도로 관측하여 정의한다.

학습 에이전트의 목표는 사용자가 관심을 보인 문서에서 가중치를 부여하여 용어를 선별한 후 최적의 프로파일을 유지하는 것이다. 사용자가 관심을 보인 문서에서 행동-가치 함수(action-value function)가 용어추출정책에 얼마나 유용한가를 평가한다.

본 논문에서 제안한 사용자의 선호도를 학습하는 과정에 대한 의사코드는 그림1과 같다.

1. Preprocess the jth web document
2. Get the initial profile p
3. Initialize ρ and $R(s,a)$, for all s,a
4. Repeat until to finish user behavior
   (1) $s \leftarrow$ current state
   (2) Take action a in s (ε-greedy policy)
   (3) Extract user preference from user behavior
   (4) Take reward value
   (5) $R(s_t,a_t) \leftarrow R(s_t,a_t) + \alpha[r - \rho + \max_{a'} R(s_{t+1},a_{t+1}) - R(s_t,a_t)]$
   (6) if $R(s,a) = \max_a R(s,a)$, then $\rho_t \leftarrow \rho_t + \beta(r + R(s_{t+1},a_{t+1}) - R(s_t,a_t) - \rho_t)$
   (7) Update *term_weight*
5. Update user profile
6. $j \leftarrow j+1$
7. Goto step 1

그림 1 Learning Pseudo Code

사용자 프로파일은 주제별로 용어벡터로 구성한다. 가장 관련이 있다고 추론되는 프로파일을 찾은 후, 연관성 평가값(식 4)이 임계치보다 작으면 새로운 주제로 프로파일을 생성하고 사용자 질의어로 용어를 구성한다. 학습과정에서 프로파일 용어의 재구성은 ε-greedy 정책을 사용하여 선택한다. 프로파일의 n개의 용어들은 사용자가 관심을 보인 문서의 용어벡터와 프로파일 벡터를 병합한 용어들에서 가중치가 높은 $n-\varepsilon$개와 무작위로 선택한 ε개로 구성한다. 학습 과정에서 문서에 대한 가치함수 $R(s,a)$의 비율이 높으면 ε의 비율을 감소시킴으로써 최적의 정책을 찾을 수 있다.

프로파일의 각 용어의 가중치는 사용자의 관심도를 학습한 결과를 반영한다. 처음 생성되는 용어벡터의 가중치는 tf 값으로 하며, 기간이 경과함에 따라 다음과 같이 적용한다.

$$term\_weight = TF(w) \times periodic\_weight(w) + R(s,a) \quad (5)$$

TF(w)는 문서 D에서 용어 w의 빈도수이고, periodic_weight (w)는 그림2의 알고리즘으로 산출한다.

프로파일의 어떤 주제에 관해 일정 기간내에 정기적인 관심도를 표명하고 있는지는 다음과 같은 주기에 대한 가중치 알고리즘을 사용한다. Periodic_cnt는 프로파일에 용어가 나타나는 횟수이며, 브라우징하는 문서내에 같은 용어가 존재하면 증가한다. β는 periodic_cnt를 반영하는 비율로 실험에 의해 설정한다.

```
If (일정기간내 용어에 대한 관심도 표명)
        periodic_cnt++
Else
        periodic_cnt--
periodic_weight □ periodic_cnt^β
```

그림 2 Periodic Weight Algorithm

관심도 평가는 웹 문서에 대한 사용자의 행위를 관측하여 의미있는 정보를 분석하여 보상값으로 받는다. 행위정보는 문서의 저장(SD), 인쇄(PD), 즐겨찾기에 추가(BD), 스크롤 이동(SM), 문서를 읽는 시간(TD)으로 구성한다. SD,PD,BD중에서 어느 것이라도 사용한다면 강한 관심도를 나타낸다고 하고 SM과 TD는 노이즈를 고려하고 중요도를 낮게 한다.

4. 결론 및 향후 연구과제

본 논문에서는 사용자의 문서에 대한 관심도에서 사용자가 선호하는 용어를 추출하여, 이를 할인된 보상값을 최적화하는 것보다 평균 보상값을 최적화하는 R-learning 강화학습 알고리즘으로 사용자 프로파일을 학습하는 방법을 제안하였다.

현재까지의 연구결과에 성능향상을 위해 보완해야 할 점은 사용자의 행위 관련 메시지에서 노이즈를 필터링하는 것과 현재상태에서 과거를 얼마나 추정하여 용어를 선택할 것인가를 결정하는 연구가 진행되어야 한다.

5. 참고문헌

[1] Geoffrey I. Webb, Michael J. Pazzani, Daniel Billsus, "Machine Learning for User Modeling," User Modeling and User-Adapted Interaction, 11, pp.19-29, 2001.

[2] Thorsten Joachims, Dayne Freitag, Tom Mitchell, "WebWatcher: A tour guide for the World Wide Web," In Proc, of the 15th Int. Joint Conference on Artificial Intelligence, 1997.

[3] Michael Pazzani, Daniel Billsus, Jack Muramatsu, "Syskill & Webert: Identifying interesting Web sites," In Proc. Of the 13th Annual National Conference on Artificial Intelligence, pp. 54-61, 1996.

[4] A. Schwartz, "A reinforcement learning method for maximizing undiscounted rewards," In proc. Of the 10th International Conference on Machine Learning, pp.298-305, 1993.

[5] D. Billsus, M.H. Pazzani, "A Hybrid User Model for News Story Classification, " In Proc. Of the Seventh International Conference on User Modeling, Springer-Verlag, 99-108, 1999.

[6] Gerhard Widmer, Miroslav Kubat, " Learning in the Presence of Concept Drift and Hidden Contexts," 1996

[7] Henry Lieberman, "Letizia: An Agent That Assists Web Browsing," IJCAI-95, pp.475-480, 1995.

[8] Richard S. Sutton, Andrew G. Barto, An Introduction Reinforcement Learning, MIT Press, 1998.

[9] Belkin, N. "User modeling in Information Retrieval: Tutorial Overhead," 6th International Conference on User Modeling, 1997, http://www.scils,rutgers.edu/~belkin/um97oh/ .

[10] Sridhar Mahadevan, " Average Reward Reinforcement Lerning: Foundations, Algorithms, and Empirical Results," Machine Learning, 22, pp. 159-196, 1996.