강화학습을 사용한 개인화된 웹 검색

Personalized web searching with Reinforcement Learning

이승준*, 장병탁**

*, **서울대학교 전기컴퓨터공학부

Seung Joon Yi*, Byoung Tak Zhang**,
*,**School of Computer Science and Engineering, Seoul National University
{sjlee,btzhang}@bi.snu.kr

ABSTRACT

본 논문에서는 사용자의 취향에 맞춰 특정 웹 문서를 탐색하는 개인화된 웹 검색기의 구현을 다룬다. 사용자의 취향은 사용자의 직접적인 평가와 사용자의 검색 과정을 통해 얻어지는 간접적인 평가를 사용한 강화 학습을 사용하여 학습된다. 웹 문서의 검색은 사용자의 취향과 현재 문서와의 관련도를 보상으로 사용한 강화 학습을 통하여 이루어진다.

Keywords : Reinforcement Learning, Web spidering, Document filtering

Ⅰ. 서 론

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[2]. 개인화된 지능적 정보 에이전트는 사용자의 정보 요구와 선호도를 직접, 간접적으로 학습하여 사용자 프로파일을 구축하게 된다.

그리고 웹 검색 에이전트는 복잡하게 서로 연결되어 있는 웹 상에서 특정 목적에 맞는 정보를 탐색하는 일을 수행한다. 서치 엔진 등에서 사용되는 일반적인 웹 검색 에이전트는 가능한 한 광범위한 검색이 목표이기 때문에 비 지향적인 탐색을 사용하지만 특정 주제나 종류의 문서를 검색하는 경우와도 같이 목표를 가지고 탐색을 행하는 경우에는 지향적인 탐색 방법을 사용할 경우 보다 나은 성능을 보인다[4].

본 논문은 동적인 인터넷 환경 하에서 이러한 개인화된 지능적 정보 에이전트와 웹 검색 에이전트의 결합으로 동적인 정보 집합을 개인의 정보 요구와 선호도에 따라 검색하는 통합적 에이전트 시스템을 다룬다.

Ⅱ. 강화 학습

2.1 강화 학습

강화학습은(reinforcement learning) 동적인 환경 하에서 시행착오를 거쳐 환경으로부터 주어지는 보상(reward)을 최대화하기 위한 학습 방법이다[6].

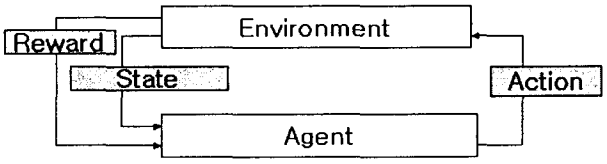

그림 1.강화 학습 framework

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

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

2.2 Q-Learning

Q-Learning [7]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.

$$Q^*(s,a) = E\{r_{t+1} + \gamma \max_{a'} Q^*(s_{t+1}, a') | s_t = s, a_t = a\} \quad (2)$$

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.

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

그림 2. Q-Learning

모든 행동이 무한히 시행되며 학습률 α를 적절히 줄이며 학습시킬 경우 $Q(s,a)$는 모든 s,a에 대해 optimal Q-value 인 $Q^*(s,a)$에 수렴한다는 것이 증명되어 있다[7]. 이 방법은 모델의 정보 없이 행동의 적합성을 나타내는 Q값만을 학습하므로 구현하기 간단하며 실제 여러 문제에 사용되어 좋은 결과를 보이고 있다.

III. 개인화된 문서 여과

정보의 종류와 양이 증가할수록 사용자가 필요로 하는 정보를 찾기까지의 시간과 노력은 오히려 증가하게 된다. 이러한 정보과잉 상태를 해결하기 위해 제시되는 것이 지능적 정보 에이전트 시스템이다. 지능적 정보 에이전트 시스템은 사용자로부터의 직접, 간접적인 정보를 통하여 사용자의 선호도를 학습하여 사용자의 선호도에 맞춰 정보 필터링, 추천 등의 개인화된 정보 제공을 하게 된다.

개인화된 정보 여과 에이전트는 아래와 같이 구성되어 있다. 즉 유저 프로파일을 바탕으로 웹 정보 중 사용자의 선호도와 정보 요구에 맞는 정보를 사용자에게 제공한다. 다시 사용자는 그 정보에 대한 평가를 에이전트에게 제공하고 이를 바탕으로 에이전트는 사용자의 프로파일을 학습해 나가게 된다.

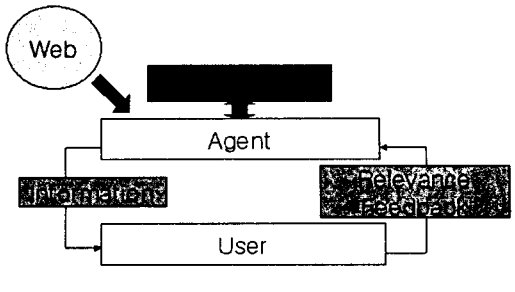

그림 3. 웹 정보 여과 에이전트

3.1 사용자 프로파일

지능적 정보 에이전트의 사용자 프로파일은 사용자의 선호도를 효과적으로 해석하고 개인화된 정보 과정 단계에서 이를 충분히 반영하는데에 사용된다. 정보 검색의 대표적인 모델인 벡터 공간 모델(vector space model)에서는 사용자가 정보 요구를 표현하는 질의문(query)와 문서(document)를 단어들의 벡터 형태로 표현한다. 본 논문에서도 기본적으로 사용자의 프로파일은 이러한 단어들의 벡터 형태로 표현되게 된다. 이러한 프로파일을 수정하여 사용자의 선호도를 잘 반영하게 하는 것이 프로파일 학습의 목적이 된다.

3.2 사용자 평가 피드백

사용자 프로파일은 직접, 간접적으로 지능적 정보 에이전트 시스템에 의해 학습되게 된다. 사용자들이 자신의 추상적인 정보 요구를 단어로 표현하기는 어려우나 실제 예를 제시받고 그 예가 자신의 요구에 적합한지를 판단하기는 상대적으로 쉽다. 따라서 이러한 사용자의 판단(relevance feedback)을 사용하여 사용자의 프로파일을 학습할 수 있다.

직접적인 프로파일 학습은 사용자에게 직접 문서 등을 제시한 후 그에 대한 평가를 요구하는 것이다. 이 방법은 사용자의 명시적인 평가를 받음으로써 효율적 학습이 가능하지만 사용자가 일일이 평가해야 하는 큰 문제점이 있다.

간접적인 프로파일 학습은 사용자의 행동을 관찰하여 간접적인 평가를 스스로 얻는다. 즉 웹 브라우징 시의 브라우징 행동, 북마크 여부, 링크 선택 등의 행동을 기반으로 사용자의 평가를 추론하는 것이다. 사용자의 직접적인 평가 없이 평가를 얻을 수 있으나 개인적이나 시간적 편차의 소지가 크다는 단점이 있다. 직접적 평가와 간접적 평가값간의 관계를 학습하는 방법으로 이러한 문제를 해결할 수 있다[5].

3.3 사용자 프로파일 업데이트

벡터 공간 모델에서 사용자 적합성 평가를 통한 사용자 프로파일 향상은 프로파일 벡터의 가중치 변경을 통하여 적용된다. 이러한 방법들 중 가장 일반적인 것이 Rocchio가 제안한 알고

리즘이다. 이 알고리즘은 아래와 같이 기존의 사용자 프로파일 벡터에 대해 적합하다고 평가된 문서 집합의 벡터를 더하고 부적합한 문서 집합의 벡터를 빼는 방식이다.

$$Q' = Q_0 + \frac{1}{n_1}\alpha\sum R_i - \frac{1}{n_2}\beta\sum S_i \quad (3)$$

Q_0:초기 질의문 벡터

R_i: Q_0에 대해 적합한 문서집합

S_i: Q_0에 대해 부적합한 문서집합

이 방법은 검색 대상 문서 집합이 정적인 경우에 사용되는 배치(batch)알고리즘으로 많은 문서를 필요로 한다. 이 방법을 실시간화 한 방법이 Widrow-Hoff방법으로 다음과 같은 방법으로 유저 프로파일 W와 문서 벡터 X의 차이를 최소화한다.

$$W = W - 2\eta(W \cdot X - y)X \quad (4)$$

y:사용자의 적합성 평가

이러한 사용자 프로파일 업데이트 과정은 정답이 제시되지 않고 적합도만을 가지고 학습한다는 데 있어서 강화 학습 문제라고 할 수 있다. 강화 학습 프레임웍 내에서의 프로파일 수정 방법이 [5]에서 다음과 같이 제시된 바 있다. 적합한 문서 내에 있는 단어의 가중치는 다음과 같이 수정된다.

$$w_k \leftarrow w_k + \beta r \quad , \text{ if } \ k \in D_i \quad (5)$$

r:문서의 적합도 평가치

IV.웹 검색

4.1 강화 학습을 사용한 웹 검색

웹 검색 문제는 문제의 특성상 강화 학습과 밀접한 관련이 있다. 웹 검색 에이전트는 여러 웹 페이지를 링크를 따라 이동하며 목적하는 문서를 찾는다. 즉 각 웹 페이지는 상태(state), 웹 페이지 내부의 링크는 행동(action)에 각각 대응되고 목적하는 웹 페이지에 도달했을 경우 보상(reward)를 받는 것에 대응되게 된다. 즉 보상을 얻기 위해서는 현재의 보상만이 아니라 미래의 보상을 고려해야 한다는 점과 올바른 답이 주어지는 것이 아니라 보상 형태로 목표가 정해진다는 점에서 웹 검색 문제는 강화 학습 문제에 적합하다[4].

실제 웹 문서 검색에 강화 학습을 적용할 때에는 주제에 맞는 문서는 즉치 보상(immediate reward) 에 대응되게 된다. 웹 검색 에이전트는 주소만 알면 어디든지 바로 가 볼 수 있기 때문에 현재의 '위치'가 상태나 행동에 영향을 끼치지 않는다. 따라서 상태(state)는 현재까지 방문한 모든 웹 페이지의 집합으로 놓을 수 있고, 행동은 이제까지 방문한 모든 웹 페이지로부터 수집한 링크들의 합집합이 된다. 에이전트는 미래의 보상이 최대화되는 링크를 우선적으로 검색하게 된다.

4.2 사용자 프로파일을 사용한 보상 책정

검색 엔진 등에서 쓰이는 일반적인 검색 엔진의 경우는 가능한 한 많은 문서를 수집하는 것 자체가 목적이기 때문에 비 지향적인 검색을 사용한다. 반면 강화 학습을 웹 검색에 사용할 경우 목적하는 문서를 찾았을 경우 적합한 보상을 주어야 한다. 사전에 정해 놓은 문서를 찾는 경우에는 그 문서들에 대해 고정된 보상을 책정하는 경우를 쓸 수 있으나 동적인 환경에서의 실제 검색 문제에서는 그것이 불가능하다. 따라서 적당한 기준에 맞춰 동적으로 보상을 책정해 줄 필요성이 있다.

웹 검색 에이전트의 목적이 사용자의 정보 요구와 선호도에 맞는 문서의 검색이기 때문에 사용자 프로파일을 바탕으로 보상을 책정하는 것이 자연스럽게 된다. 즉 현재 문서가 얼마나 사용자 프로파일에 유사한지의 값을 바로 보상으로 줄 수가 있다. 현재 문서가 프로파일과 얼마나 유사한지를 측정하는데는 다양한 방법이 사용 가능하다. 그 중의 하나로 다음 cosine measure를 사용할 수 있다.

$$Cosine(u, f) = \frac{\sum_{i=1}^{t} u_i f_i}{\sqrt{\sum_i^t u_i^2 \sum_i^t f_i^2}} \quad (6)$$

4.3 상태 및 행동 근사

앞서 설정한 바에 의하면 상태와 행동은 매우 범위가 크고 동적으로 변하게 된다. 이러한 상태 및 행동에 대한 데이터를 저장하는 데는 현실적으로 힘든 저장 공간이 요구되게 된다. 또한 강화 학습이 제대로 이루어지려면 모든 상태들을 충분한 횟수만큼 방문해야 한다. 즉 실제로 구할 수 있는 데이터로는 충분한 학습이 힘들다. 따라서 위의 설정을 근사화할 필요가 있다.

우선 상태가 현재까지 방문한 웹 페이지와 무관하다고 가정한다. 과거에 어떠한 웹 페이지를 방문하였든 간에 특정 링크의 가치는 변하지 않는다는 가정으로 이 경우 상태는 하나로 줄어들고 가능한 행동은 모든 각각의 링크들이 된다. 상태가 하나이므로 각각의 모든 개별적인 링크들에 대해 각 링크의 적합도(Q-value)를 저장하면 된다. 행동들의 Q값을 모두 테이블에 저장하려면 큰 저장공간이 요구되기 때문에 저장 공간을 절약하고 일반화 성능을 향상시키기 위해서 일반적으로 Q값 저장은 함수 근사기(Function approximator)를 사용한다. 웹 검색

에서는 링크가 위치하고 있는 문서의 내용 정보, 링크를 구성하는 문장 및 주위 문장의 내용 정보 등으로부터 링크의 적합도를 근사하여 저장하여 사용할 수 있다[4].

이러한 함수 근사기는 단어들과 해당 Q값을 대응시켜야 한다. 일반적인 함수 근사(Regression)방법들이 사용 가능하지만 Q값을 이산화할 경우 함수 근사기는 문서 분류기와 동등해지게 되며 문서 정보 추출용으로 고안된 방법들을 바로 적용 가능하다. [4] 에서는 Naive Bayesian Classifier를 사용하였으나 일반적으로 보다 나은 성능을 보이는 타 문서 분류기들을 사용할 경우 보다 나은 성능을 기대할 수 있을 것이다.

V. 통합된 웹 문서 검색 여과 시스템

실제 사용자가 사용하는 시스템은 아래 그림과 같이 개인화된 문서 여과 에이전트와 웹 검색 에이전트, 유저 인터페이스 에이전트로 구성되게 된다.

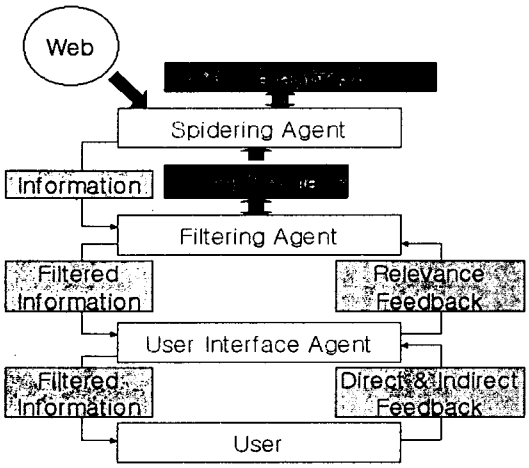

그림 4. 개인화된 웹 검색 및 여과 시스템

유저 인터페이스 에이전트는 여과된 문서에 대한 사용자의 직접적 간접적 평가를 받아 하나의 Feedback 값으로 통합하는 역할을 한다. 이러한 Feedback 값은 다시 정보 여과 에이전트에 입력으로 들어가 사용자의 프로파일을 갱신시키는 데에 이용되게 된다. 이 사용자 프로파일은 검색 에이전트의 강화 학습시의 보상을 책정하는 데에 사용되게 되며, 검색 에이전트는 높은 보상을 받는 문서를 검색해 오게 된다.

사용자의 프로파일과 유사할수록 높은 보상을 받으므로 높은 보상을 받은 문서는 바로 사용자의 프로파일과 유사한 문서가 된다. 이 문서는 다시 사용자에게 보여지게 되어 평가를 받게 된다. 결과적으로 사용자의 프로파일이 사용자의 정보 요구와 취향을 따라 학습되게 되며 검색 에이전트는 이에 따라 웹을 지향적으로 검색하게 된다.

VI. 결론 및 연구 방향

본 논문에서는 사용자의 성향을 학습하는 개인화된 웹 문서 여과 시스템을 사용하여 사용자의 성향을 사용하는 능동적 웹 문서 검색을 수행하는 복합적인 시스템을 제시하였다. 기존의 수동적인 웹 문서 여과 시스템은 많은 양의 문서들 중 사용자의 취향에 맞는 문서를 선택해 주지만 동적인 환경에서 새로 나타나는 정보들을 능동적으로 검색하지는 못한다. 한편 기존의 지향적 웹 검색 시스템은 개인화된 시스템이 아니라 사전에 정해 놓은 특정한 정보들만 검색하는 한계를 가진다. 본 논문에서 제시하는 복합적 시스템은 사용자의 프로파일을 학습한 뒤 그 프로파일을 사용하여 다시 사용자의 기호에 맞는 문서를 적극적으로 검색하는 개인화된 능동적 검색 시스템으로 동적인 환경에서 능동적으로 개인화된 정보 요구에 응할 수 있는 장점을 가진다. 자료가 동적으로 생성되며 개인화 소지가 충분한 논문, 학회 검색 등에의 구체적 적용이 현재 연구되고 있다.

감사의 글 : 본 연구는 BK21-IT 프로그램에 의해 지원 받았습니다.

VII. 참고문헌

[1] Joachims,T. Freitag,D. and Mitchell,T., *"WebWatcher: A Tour Guide for the World Wide Web"*,Proceedings of IJCAI97, 1997.

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

[3] Menczer,F.,*"ARACHNID:Adaptive Retrieval Agents Choosing Heuristic Neighborhoods for Information Discovery"* ,Proceedings of ICML97 , 1997.

[4] Rennie,J. and McCallum,A. *"Efficient Web Spidering with Reinforcement Learning"*

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

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

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