강화학습을 이용한 웹 정보 검색

정 태 진[0]　　장 병 탁
서 울 대 학 교　전 기 컴 퓨 터 공 학 부
{tjjeong, btzhang}@bi.snu.ac.kr

Web Information Search Using Reinforcement Learning

Tae-Jin Jeong[0]　　Byoung-Tak Zhang
School of Computer Science and Engineering,
Seoul National University

요 약

현재 인터넷상에서 제공되고 있는 대부분의 서치엔진들은 정보소스에 접근해서 이를 가져오는 웹 로봇(webbot)이라고 불리우는 에이전트를 이용한다. 그런데 이런 웹 로봇들이 웹 문서를 검색하는 방법은 극히 단순하다. 물론 많은 정보를 가지고 오는 것에 초점이 맞추어져 있어서 정확도를 중시하지 않는 것에도 한 원인이 있다. 범용 검색엔진과는 달리 검색하는 영역을 축소하여 특정 주제에 관련된 정보만을 더 정확히 찾아주는 검색엔진의 필요성이 증가하고 있다. 이에 본 논문에서는 강화 학습 방법을 이용하여 웹 상에 존재하는 정보 중에서 특정 주제의 웹 페이지를 보다 더 정확히 찾는 방법을 제시한다. 강화 학습은 웹 상의 하이퍼링크를 따라가는 문제에 있어서 미래에 이로움을 주는 행동의 효용성을 측정하는데 있어서 이점을 보인다. 강화 학습을 이용하여 제시된 방법을 통한 실험에서는 일반적인 방법보다 더 적은 링크를 따라가고도 더 정확한 결과를 보였다.

1. 서 론

인터넷상의 정보의 증가로 인해 사용자는 예전보다 더 많은 정보에 접근할 수 있는 반면 다른 한 편으로는 급속하게 늘어난 수많은 정보 중에서 자신에게 유용한 정보를 찾는데 더 많은 시간을 투자해야만 한다. 그리므로 컴퓨터가 사람을 대신하여 웹으로부터 양질의 정보를 찾아낸다면 사용자에게 매우 유용한 도구가 될 것이다. 이를 위해서는 웹을 포함한 인터넷에 존재하는 대규모의 전자 문서로부터 원하는 정보를 정확히 찾고 그 텍스트의 내용을 분석하여 사용자가 요구하는 정확한 정보를 제공할 수 있는 정보 분류 및 검색 기술의 개발이 필수적이다. 정보 분류 및 검색 기술을 향상시키는 방법들이 여러 가지지만 그 중에서도 사용자 관심도(user interest)를 반영하여 검색의 효율을 높이고 개인화된 검색 서비스를 제시할 수 있는 방법이 먼저 고찰되어져야 한다. 본 논문에서는 강화학습 알고리즘을 이용하여 웹 문서를 검색하는데 있어서 더욱 효과적으로 검색할 수 있는 방법을 제시하고 이를 실험을 통해서 알아보겠다. 먼저 2장에서는 강화 학습 알고리즘을 설명하고 3장에서는 이 알고리즘이 웹 문서를 검색하는데 있어서 어떻게 적용되고 있는지를 살펴보고 4장에서는 이를 실험을 통하여 일반적인 검색방법보다 효율이 큰 것을 제시하고 끝으로 향후 연구에 대하여 토론하겠다.

2. 강화 학습 과정

기계학습을 크게 분류하면 명시적으로 학습 목표가 주어지는 분류(classification), 예측(prediction)등과 같은 감독학습(supervised learning)과 명확하게 학습의 목표가 주어지지 않는 군집화(clustering)와 같은 무감독 학습(unsupervised learning)으로 나누어 볼 수 있는데, 강화학습은 그 중간적인 특성을 띠고 있다. 강화 학습의 특성은 주어진 환경과의 상호 작용에 의한 학습(Learning from interaction with environment), 지연되는 보상(delayed reward), 그리고 시도와 오류(trial and error)로 기술할 수 있다[2].이러한 특성을 학습 과정으로 표현하면 다음과 같다.

① 시각 t에 학습자가 선택 가능한 행동 집합과 학습자의 행동에 따른 가설공간의 가설이 주어진다.
② 행동 집합에서 이제까지 학습한 지식을 토대로 가장 좋은 결과를 유도할 만한 행동을 선택한다.
③ 특정 행동 선택에 따른 가설을 선택하고 이에 따라 가설공간을 탐색한다.
④ 환경으로부터 행동에 대한 평가를 받는다.(evaluative feedback).
⑤ 학습한 지식과 평가간의 차이를 고려하며 시각이 증가한다.

대부분의 실제 응용은 동적인 환경과 각 학습 시스템과의 상호 작용이 고려되어야 하는 특성을 가지고 있는데

강화 학습은 그 특성상 이러한 응용에 적합하다. 웹 상에 존재하는 웹 문서의 특성도 이러한 동적인 상태 즉, 문서가 있던 사이트가 없어지거나 관련 링크가 변경된 경우와 사용자의 정보욕구의 변화에 따른 사용자 프로파일의 변화 등의 동적인 상황이 많이 존재한다. 따라서 웹 문서를 수집하고 여과하는데 강화학습을 이용할 수 있다. 기본적인 강화학습 모델에 대하여 살펴보면 다음과 같다. 학습자가 주어진 환경과 상호 작용을 할 때 상태(state), 행동(action), 보상(reward)이라는 세 가지 기본 틀을 이용한다. 환경은 주로 상태로 표현되며 학습자는 적절한 정책에 따라 행동을 취하게 된다. 이 때, 환경은 학습자에게 행동에 대한 보상을 주게 된다. 아래 그림은 강화 학습 에이전트가 t시각에 행동 a_t를 취하면 행동에 대한 보상 r_t가 환경으로부터 주어진다. 그리고 행동에 의해서 상태 s_t가 s_{t+1}로 변화된다. 강화 학습에서의 환경 모델은 마코프 속성을 만족하는 MDP(Markov Decision Process)이다. 마코프 속성이란 시각 t+1의 환경에서의 반응은 오직 시각 t에서의 상태와 행동만에 의존하는 속성을 말한다[3].

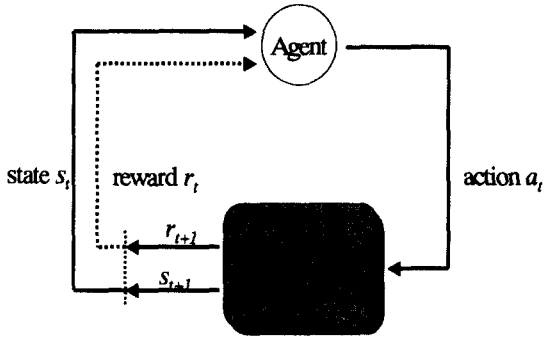

그림 1: 학습자와 환경간의 상호작용

보상 함수는 학습자의 행동에 대한 환경으로부터의 반응으로 보통 스칼라(scalar)값으로 주어진다. 강화 학습에 주어지는 보상은 감독 학습에서처럼 지시적인(instructive) 특성을 갖는 것이 아니라 평가적인(evaluative) 특성을 갖으며 아래 식으로 표현한다.

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

위 식에서 γ 는 할인 상수(discount factor)로 미래에 받게 될 보상이 현재의 상태의 가치나 상태-행동의 가치에 반영되는 정도를 조절한다. 즉 γ 의 값이 1에 가까울수록 t시각 이후에 받게 될 보상을 할인하지 않고 반영하게 된다. 가치함수는 현재 상태의 가치를 평가하여 다음 상태의 가치를 추정하여 학습하는 규칙이다. 일반적인 가치함수는 아래식과 같이 표현한다. 구체적으로는 상태에 대한 가치함수 아래 식 또는 상태-행동 쌍에 대한 가치함수 식으로 나뉘어지는데 이는 t+1 이후에 시간에 대한 기대값으로 표현된다.

$$V(s) = V(s) + \alpha[\, V(s') - V(s) \,]$$

$$V^{\pi}(s) = E_{\pi}[\, \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \,]$$

$$Q^{\pi}(s,a) = E_{\pi}[\, \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a]$$

π는 학습자의 행동 선택에 기준이 되는 정책을 말하며 각 상태의 가치 또는 어떤 상태에서 어떤 행동을 선택했을 때의 가치를 계산하는 것을 무한히 반복하게 되면 결국 최종 목적에 수렴하게 된다.

3. 강화 학습을 이용한 웹 정보 검색 과정

위에서 설명한 기본적인 강화 학습을 웹 정보를 검색하는 문제에 적용하면 다음과 같다. 우선, 여러 치즈조각이 놓여져 있는 미로를 찾아가는 쥐를 생각해 보면 쉽게 이해가 된다. 주어진 미로의 한 위치에서 쥐는 치즈를 찾기 위해서 다음 위치로 움직이는 행동을 하고 치즈가 있는 위치를 찾아내면 이에 대한 보상으로 치즈를 먹을 수가 있다. 쥐가 우연히 치즈가 있는 위치로 한번에 가게 되면 즉각적인 보상을 받지만 결국에는 여러 개의 치즈를 최단 시간 안에 다 먹기 위해서는 미래의 보상을 고려하면서 최적의 행동을 선택해야한다. 이와 유사하게 웹 상에서 사용자가 원하는 웹 문서를 찾는 문제도 생각해 볼 수 있다. 한 웹 페이지가 주어져 있고 이 페이지에 여러 하이퍼링크가 있을 때, 에이전트는 사용자가 관심 있어 하는 문서를 찾기 위하여 관련 링크를 따라 가면서 정보를 검색하게 된다. 이때 각 웹 페이지들은 상태(state)로 정의하고, 하이퍼링크를 따라가는 것을 행동(action)으로 정의할 수 있다. 각 상태 s에서 에이전트는 그에 따른 보상값(reward) R(s)를 받고 이 상태에서 하이퍼링크를 따라가는 행동(action)을 취하면 이 행동이 에이전트의 최종 목적(찾고자 하는 웹페이지)에 있어서 얼마나 유용한가를 가치함수 Q(s,a)를 통해서 평가 할 수 있다. 이 웹 탐색의 목적은 사용자가 원하는 정보가 있는 웹 페이지를 찾는 것이 된다. 가치함수 Q(s,a)는 한 웹 페이지에서 사용자가 관심도 나타내는 하이퍼링크를 따라가는 행동에 대한 할인된(discounted) 보상값의 합이 된다. 다시 말하면 웹 페이지를 나타나는 단어들과, 하이퍼링크의 앵커 텍스트상의 단어들을 *TFIDF* 벡터 표현으로 나타내어서 이를 사용자 프로파일상의 단어들과의 유사도를 조사하고 이를 보상값으로 정의하여 가치함수 Q(s,a)값을 구했다. 즉 웹 페이지들을 전처리 과정을 통하여 *TFIDF* 벡터로 표현하고 이 웹 페이지들의 하이퍼링크에 나오는 단어들과 이 하이퍼링크를 따라갔을 때 나오는 페이지의 타이틀이나 헤더의 단어를 이용하여 보상함수 R(s)을 구하고 이를 다음의 식을 통하여 Q값을 학습하였다.

$p = (w_{1,p}, w_{2,p}, \cdots, w_{t,p})$ 사용자 프로파일 벡터

$d_j = (w_{1,j}, w_{2,j}, \cdots, w_{t,j})$ 문서나 링크에 대한 벡터

$$\cos(d_j, p) = \frac{d_j \cdot p}{|d_j| \cdot |p|} = \frac{\sum_{i}^{t} w_{i,j} \cdot w_{i,p}}{\sqrt{\sum_{i}^{t} (w_{i,j})^2} \cdot \sqrt{\sum_{i}^{t} (w_{i,p})^2}}$$

$$Q(s_t,a)=\sum_{k=0}^{\infty}\gamma^k R(s_{t+1+k})$$

학습된 결과를 가지고 에이전트가 웹 페이지의 하이퍼링크를 따라갈 때 사용자의 관심도에 나타난 단어들의 Q 값이 가장 큰 하이퍼링크를 따라 가게 됨으로써 너비우선(breadth-first search)방식의 검색 보다 더 효율적인 검색을 할 수 있다.

4. 실험 및 결과

실험에 사용된 데이터는 학회 논문제출을 요청하는 홈페이지(call-for-paper)와 학회 안내 홈페이지(conference homepage)를 대상으로 총 100개의 문서를 수집하였고, 이에 관련된 하이퍼링크는 약 500개를 만들었다. 이 문서들의 내용은 인공지능에 관련된 해외 학회에 대한 정보와 논문 제출 안내를 담고 있다. 수집된 웹 문서에 존재하는 하이퍼링크를 살펴보면 파일을 가지고 있는 곳을 연결해주는 하이퍼링크와 끊어진 링크와 시스템 관리를 처리하는데 사용되는 링크들이 존재하는데 정확한 실험을 위해서 이를 제외하였다. 실험조건상의 사용자 프로파일 색인어수는 10개로 하였고, 사용자 관심도의 변화는 없는 것으로 가정하였다. 실험 결과는 다음 그림과 같다. 에이전트 관련 학회를 검색의 목적으로 하였고 에이전트 관련 학회를 찾기 위해 따라간 하이퍼링크의 퍼센트와 발견한 에이전트 관련 학회 문서의 퍼센트로 하여 비교하였다.

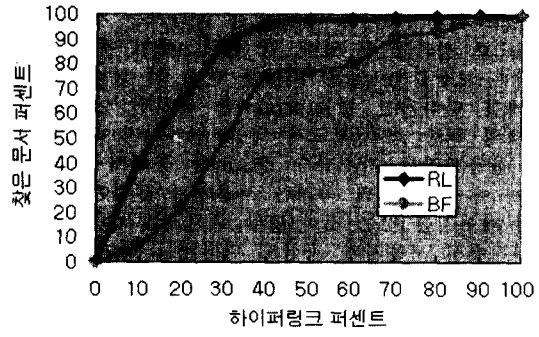

그림 2 실험 결과 1

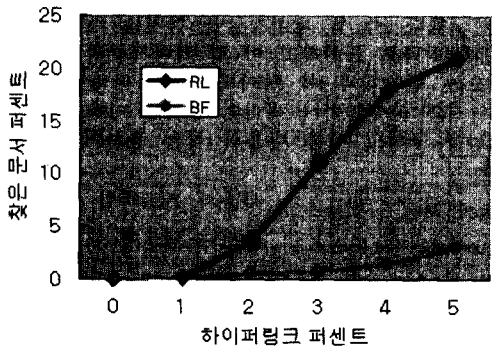

그림 3 실험 결과 2

그림2는 RL(강화학습)과 BF(너비우선)을 적용하여 비교하였는데 x축은 문서를 찾기 위해서 에이전트가 따라간 하이퍼링크의 퍼센트 비율이고 y축은 에이전트가 발견한 문서에 대한 퍼센트를 나타낸다. 그림을 보면 알 수 있듯이 강화학습을 이용한 RL이 너비우선 검색인 BF보다 거의 3배에 가까운 성능 효율을 보이고 있다. 최종적으로 많은 하이퍼링크를 따라가면 RL이나 BF가 원하는 문서를 거의 다 찾아 주기는 한다. 하지만 초기 분포(하이퍼링크의 30%이내)에서는 강화학습을 이용한 검색이 적은 하이퍼링크를 가지고 더 정확하게 원하는 정보를 찾아주는 것을 알 수 있다. 그림 3은 그림 2의 초기 분포를 자세히 보이고 있다. 다른 검색 조건에서도 이와 유사한 분포를 보인다.

5. 결론

지금까지 살펴본 바에 따르면 범용 검색엔진에서 사용하고 있는 검색 알고리즘인 너비우선 방식의 검색보다 강화학습 알고리즘을 이용하여 사용자의 관심도에 따른 검색방식이 더 좋은 성능향상을 보였다. 이는 찾고자 하는 영역을 줄이는 대신, 보다 더 정확하고 빠른 검색을 필요로 하는 특정 주제어 관련 검색엔진에 있어서 크게 도움을 줄 수 있다. 앞으로의 연구에 있어서는 이번 논문에서 제시된 방법을 바탕으로 온라인 상의 하나의 시스템에서 새로운 학회 관련 사이트가 새로 생겨났을 때 이 사이트에 대한 강화학습을 이용한 효율적인 검색을 통하여 사용자에게 추천해 주고 이를 통하여 사용자의 관심도를 학습하는데 중점을 두고 검색된 문서를 여과하는 방법을 생각해 볼 수 있다.

감사의 글: 본 연구는 BK21-IT 프로그램에 의해서 일부 지원 되었슴.

참고문헌

[1] Andrew McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Building domain-specific search engines with machine learning techniques. In *AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*, 1999.

[2] Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 1998.

[3] Tom M. Mitchell, *Machine Learning*, McGraw-Hill Com. Inc., 1997.

[4] G. Salton, *Automatic Text Processing*, Addison-Wesley, 1989.

[5] T. Joachims, D. Freitag, and T. Mitchell. WebWatcher: A tour guide for the WWW. In *Proceedings of IJCAI-97.*

[6] Junhoo Cho, Hector Garcia-Molina, and Lawrence Page. Efficient crawling through URL ordering. In *Computer Networks and ISDN System*, volume 30, 1998.

[7] Personalized Web-Document Filtering Using Reinforcement Learning, Zhang, B.-T. and Seo, Y.-W., *Applied Artificial Intelligence*, vol. 15, 2001.