인공생명 기법을 이용한 정보검색 에이전트의 원형

Prototype of Information Retrieval Agents Using Artificial Life Technologies

김학균, 조성배
연세대학교 컴퓨터과학과

Hak-Gyoon Kim, Sung-Bae Cho
Computer Science Department, Yonsei University

요 약

인터넷의 웹은 여러 곳에 분산되어 있을 뿐만 아니라 끊임없이 동적으로 변화하는 특성이 있기 때문에, 보통의 인덱스를 통한 정보검색 방법에는 한계가 있다. 이러한 웹의 특성을 적절히 살리면서 원하는 정보를 신속하게 검색하기 위하여, 본 논문에서는 여러 개의 에이전트가 인공생명 기법에 의해 조직되어 정보를 검색하는 온라인 에이전트를 소개한다. 이것은 각각의 에이전트에 의하여 검색된 문서가 얼마나 질의에 가까운가에 따라서 해당 에이전트와 유사한 것을 복제하거나 제거한다. 사용자의 질의에 적절한 문서를 제공하는 에이전트들만이 살아 남아 문서를 가져오도록 함으로써, 불필요한 문서를 검색하지 않게 되어 단위 시간에 원하는 문서를 많이 얻어올 수 있는 장점이 있다. 실제 웹 환경에서 실험한 결과 종래의 폭우선 검색이나 랜덤검색에 비하여 좋은 결과를 내는 것을 볼 수 있었다.

1. 서 론

빠른 속도로 성장하는 웹의 특성으로는 문서의 위치에 따라 상이한 시간이 걸리고, 문서의 형식, 스타일, 내용이 여러 종류이고, 문서가 동적으로 추가되거나, 지워지거나 수정 될 수 있다는 것 등이 있다. 이러한 웹 환경에서 사용자가 원하는 자료를 얻기 위해 검색을 할 경우 많은 시간이 요구된다. 이러한 일을 자동적으로 에이전트가 해 줄 경우 시간 및 노력이 경감 될 것이다. 이에 따라 웹에서 정보를 자동으로 검색해주는 여러가지 검색 에이전트들이 발표되고 있다.

보통의 검색 에이전트들은 자동으로 색인을 해주는 로봇 에이전트를 두어, 사용자의 질의가 들어올 때 만들어진 색인을 이용하여 결과를 출력한다. 그러나 이러한 인덱스 기반의 정보 검색 에이전트에는 한계가 있는데, 요약해 보면 다음과 같다.

첫째, 웹의 문서가 바뀌거나, 추가되거나, 삭제되는데 대처하기가 어렵다. 웹의 문서가 변화되면 그 내용을 다시 색인하여야 하지만 검색 에이전트는 그 문서가 변화되었는지를 알 수 없다. 이러한 이유 때문에 빠르게 성장하고 있는 웹의 환경에 대처하기 힘들게 된다. 둘째, 검색 에이전트들은 모든 문서들을 탐색함으로 인하여 네트웍에 큰 부하를 주게 된다. 즉, 문서 순회 알고리즘으로 폭우선 탐색이나 깊이우선 탐색과 같은 것을 채택하고 있는데 필요 유무에 관계없이 문서들을 탐색하기 때문에 네트웍에 큰 부담을 주고 있다. 셋째, 단일 에이전트를 기반으로 하고 있기 때문에 동시에 여러 지역에 있는 문서들을 얻어올 수 없다. 여러 개의 에이전트를 기반으로 문서를 탐색 할 경우 더 빠른 시간에 여러 지역에 있는 문서를 가지고 올 수 있는 장점을 얻어낼 수 있다.

이러한 종래 검색 엔진의 단점을 극복하고 개선시킬 수 있는 알고리즘의 필요성이 절실하여, 웹의 특성을 살리면서 더 정확한 정보를 얻어오기 위한 여러가지 방법들이 제안되었다 [1, 4, 5, 6]. 사용자를 위해서 스스로 판단하는 자생적인 에이전트나 반인공지능적인 에이전트는 많은 사람들에 의해서 온라인상에서 나날이 커져 가는 사용 가능한 정보를 관리하기 위

해 필요한 인간과 컴퓨터의 상호 작용을 줄일 수 있는 방법으로 나타났다[8]. 기계학습 기술은 효과적인 정보 에이전트를 생산하는데 제안되어졌다. 예를 들어 미리 주어진 정보에 따라 검색을 수행하고 학습능력을 배양시키는데 기초를 두어 사용자에게 결과를 제공하는 에이전트를 말하는 것이다[7]. 가중치 키워드 벡터 표현이나 수행능력의 피드백과 같은 기술은 정보 검색과 정보 필터링에 적용 되어 지고 있다[5, 6]. 유전자 알고리즘과 인공생명에서 제기된 방법에서 집단을 기초로 한 진화와 개인을 기초로 한 학습능력이 동시에 배양되는 다중 에이전트 정보 시스템을 가져오게 했다

본 논문에서도 이러한 웹의 특징을 살리면서 기존 검색 에이전트의 단점을 극복하기 위한 알고리즘을 제안한다. 이것은 국지적인 결정을 할 수 있는 여러 개의 분산된 에이전트들의 집합을 기초로 한 알고리즘이다. 이 알고리즘은 각각의 에이전트에 의하여 검색된 문서가 얼마나 질의에 가까운가에 따라 해당 에이전트의 주변에 또 다른 에이전트를 생성하거나 제거한다. 또한 학습능력과 사용자의 의해 제공되는 적합도의 피드백 정보를 이용하여 에이전트가 경험을 가지고 그들의 행동을 적응시켜 나갈 수 있도록 한다.

2. 알고리즘

사용한 알고리즘은 그림 1과 같다. 이것은 여러 개의 에이전트를 기초로 하고 있으며, 각각의 에이전트들은 자신만의 고유한 유전자형과 에너지를 가지고 생성, 소멸을 반복해 나간다. 에너지의 임계값은 상수이기 때문에 에이전트의 재생성 여부는 다른 에이전트와 관계없으며, 자신만의 국지적인 탐색을 할 수 있다. 사용자는 초기에 키워드들(Q)과 시작하는 문서의 위치를 제공한다.

```
초기 북마크된 URL에 대해 n 개의 에이전트 생성;
각 에이전트의 유전자형 및 에너지 초기화;
while (살아있는 에이전트가 수 > 0)  {
      검색대상 키워드들 입력;
      임의의 에이전트 A 선택;
      에이전트 A가 검색할 다음 링크 선택;
      해당 링크의 문서 D_A 수집;
      D_A 의 적절성에 따라 에이전트의 에너지 E_A 조정;
      If  (E_A > 임계값)
            A를 돌연변이 시켜 새로운 에이전트 A' 생성;
            E_A = E_A' = E_A / 2;
      Else if  (E_A < 0 )
            에이전트 A 삭제;
}
```

그림 1 알고리즘

2.1 초기화

각각 에이전트를 초기의 분산된 링크들로 초기화 시키고, 각각의 에너지를 임계값 $\varepsilon/2$로 초기화 시킨다. 에이전트들은 자신만의 유전자형을 갖는데, 이것은 자신의 존재 유무와 재생산 시에 중요한 인자로 작용한다. 에이전트의 유전자형은 β, γ, 에너지, 상태정보로 구성되는데 그 역할을 살펴보면 다음과 같다.

β는 선택되어질 링크에 대해서 현재의 문서가 가지고 있는 정보들을 얼마나 믿을 수 있는지를 나타내는 지표이고, γ는 사용자의 검색능력 판단을 얼마나 믿을 수 있는 지를 나타내는 지표이다. 에너지는 에이전트의 검색 능력을 나타내는 지표로서 에이전트 존재 유무를 결정하는 중요한 인자이다. 상태 정보는 에이전트가 살아있는지 소멸되었는지를 나타낸다.

2.2 검색 대상 링크선택

현재의 문서에서 정보를 가지고 올 다음 문서의 하이퍼 링크를 선택하기 위하여 각 링크가 주어진 질의어와 얼마나 가까운지를 계산한 후 적절성을 평가한다. 이것은 질의어와 가

까이에 위치한 링크는 그만큼 질의어에 관계 있는 문서일 확률이 높을 것이라고 기대되기 때문이다. D_A에 있는 각각의 링크(l)에 대하여 적절성(λl)을 표현하면,

$\lambda l = \sum match(k,Q)/dist(k,l)$ 과 같다

k는 D_A에 있는 키워드를 나타내며, $match(k, Q)$는 k가 질의어에 속하면 1, 아닐 경우에는 0을 가지며, $dist(k, l)$는 키워드 k와 현재의 링크 l 사이에 있는 링크의 개수를 나타낸다. 이 값들을 합이 1인 분포로 정규화하면, 확률 분포 $Pr[l] = Exp(\beta\lambda l) / \sum \beta\lambda l'$ 과 같다. 그리고 에이전트는 확률 분포에 따라 링크를 선택한다. β값이 클 경우에는 적절성(λl)이 큰 링크들이 분포가 커지므로 선택되어질 확률이 높게 된다. 그러나 β값이 작아질 경우에는 링크를 랜덤하게 선택한다.

2.3 문서수집 및 적절성 평가

문서내에서 선택되어진 링크와 연결된 문서를 수집한다. 수집된 문서의 적절성은 문서내에 있는 매치되는 키워드의 개수를 전체 키워드의 개수로 나누어서 계산된다. 이렇게 계산되어진 문서의 적절성에 따라 에이전트의 에너지가 증가한다. 또한 에이전트에게 문서를 수집하는데 소요되는 네트웍 자원의 사용은 에너지의 감소를 가져온다. 여기서 선택된 문서가 이전에 이미 방문 되어진 것일 경우 에너지의 증가를 기대할 수 없으나 자주 방문 되어지는 것으로 보아 관련이 있을 확률이 높다는 것을 감안하여 보너스 에너지를 부여 할 수도 있다. 에이전트는 문서의 적절성을 전체 키워드에서 질의어에 속하는 단어가 차지하는 비율로 계산한다.

2.4 에이전트의 생성 및 제거

에이전트의 존재 유무는 에이전트의 에너지와 임계값 ε의 비교를 통해 결정되어진다. 결국 에이전트는 새로운 에이전트를 생성하거나 에너지의 부족으로 사라질 수도 있다. 재생산되어질 경우 새로운 에이전트는 부모와 같은 URL 시작점을 가지며, 진화를 가능키 위해 돌연변이를 일으킨다. 이러한 진화론적인 방법으로 수행능력이 좋은 에이전트들로 집단을 편향 시킬 수 있는 장점을 가지고 있다. 여러 번의 스텝을 통해 축적된 에너지를 기초로 하기 때문에, 단기간에 잘못된 방향으로 진행하는 것을 무마시키고 다시 새로운 방향으로 진행할 수 있다. 에이전트의 재생성시에 새로운 에이전트의 유전자형은 β와 γ의 변이를 통해 결정되어 진다. β는 문서의 적절성을 나타내므로 β은 문서에 따라 계속 증가해 나갈 수 있다.

이 알고리즘의 결과는 에이전트의 집단에 의해 방문 되어진 문서의 흐름이 된다. 초기치 에너지가 부족할 경우에는 에이전트의 집단이 어떠한 적절한 자료를 얻기 전에 사라질 수도 있다. 따라서 에너지의 초기치 배분은 에이전트 집단의 효율성에 지대한 영향을 미치게 된다.

3. 실험 결과

제안한 방법의 유용성을 보이기 위해 랜덤 검색과 폭우선 검색을 비교 대상으로 실험하여 보았다. 폭우선 검색 방식은 하나의 문서안에 있는 모든 링크를 검색하는 에이전트 순회 방식이다. 이 방식은 한 문서가 어떠한 주제에 대해서 밀집해 있을 경우에는 좋은 성능을 발휘할 수 있으나 그 문서가 관련이 없을 경우에는 필요 없이 모든 문서를 검색하는 큰 단점을 가지고 있다. 또한 모든 문서를 소모적으로 순회하기 때문에 네트웍에 많은 부하를 줄 수 있다.

그림 2와 그림 3은 웹에서 테스트 한 결과로써 초기의 URL은 Yahoo에 있는 각각 다른 주제로 된 디렉토리를 주었다. 아래의 결과에서 볼 수 있듯이 단위 시간에 살아있는 에이전트의 개수는 현재 탐색하고 있는 에이전트를 나타내며, 증가추세로 가고 있는 것은 그만큼 관련 문서를 탐색하고 있는 에이전트의 수가 증가하는 것을 의미한다. 그러나 폭우선 검색 방식은 시간이 지날수록 관련 문서를 탐색하는 에이전트의 수는 증가하지만 모든 문서를 탐색하므로 단위 시간에 살아있는 에이전트의 수는 인공생명 기법을 이용한 에이전트의 수보다 적은 수치를 나타낸다. 축적된 에너지가 많으면 많을수록 에이전트가 탐색할 수 있는 기회가 많아질 뿐만 아니라 그만큼 에이전트 집단이 관련 문서들 쪽으로 나아가고 있는 것을

의미한다.

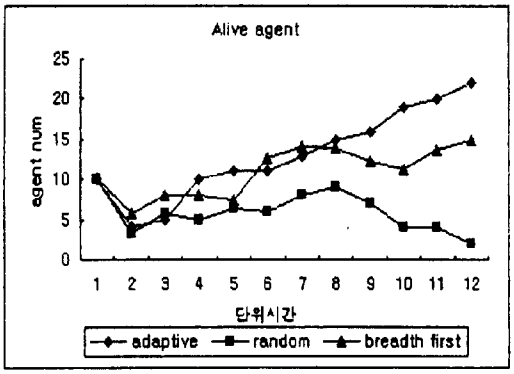

그림 2 단위 시간에 존재하는 에이전트

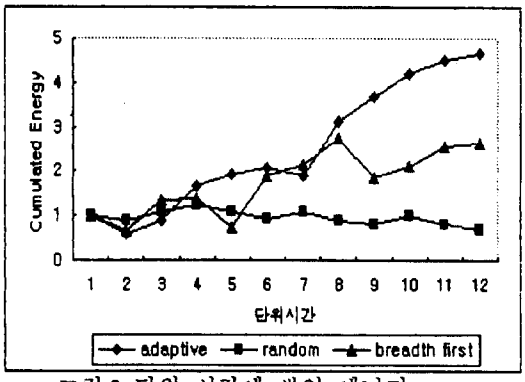

그림 3 단위 시간에 쌓인 에너지

4. 결론

검색 에이전트를 설계할 때 여러가지 측면을 고려해야 한다. 우선, 단어의 빈도수와 같은 확률적 특성이나 네트웍의 구조나 프로토콜과 같은 물리적인 특성을 살펴볼 수 있다. 또한 하이퍼링크로 만들어진 문서들의 관계에 부여된 의미 구조는 또 다른 중요한 측면으로 볼 수 있다. 이렇게 하이퍼링크에 부여된 의미는 서로 공동의 주제에 관련된 문서들의 관계로 볼 수 있다. 이러한 것을 의미적 위상이라고 하는데 본 논문의 알고리즘은 이러한 의미적 위상을 이용한 것이다.

이러한 특성을 이용한 여러 개의 에이전트를 기반으로 한 인공생명 에이전트는 기존의 검색 에이전트의 한계를 극복하고 있다. 인덱스를 기반의 중앙 집으로 하지 않고, 탐색 공간을 줄임으로써 서버에 로드를 줄이고, 문서의 변화에 대처해나가는 것을 볼 수 있다. 이 알고리즘의 단점으로는 것은 문서들이 어떠한 의미구조를 가지고 있지 않을 경우에는 좋지 못한 결과를 가지고 올 수 있다는 점이다.

참고 문헌

[1] F. Menczer, "ARACHNID: Adaptive Retrieval Agents Choosing Heuristic Neighborhoods for Information Discovery," *Machine Learning: Proceedings of the 14th International Conference (ICML97), San Francisco,* 1997.

[2] F. Menczer, R.K. Belew and W. Willuhn, "Artificial life applied to adaptive information agents," *the AAAI Symposium on Information Gathering from Distributed, Heterogeneous Databases,* 1995 .

[3] F. Menczer and W. Willuhn and R.K. Belew, "An Endogenous Fitness Paradigm for Adaptive Information Agents," *CIKM Workshop on Intelligent Information Agents,* 1994.

[4] A. M. Amalthaea, "Information discovery and filtering using a multiagent evolving ecosystem," *In Proc. Conf. Practical Applications of Intelligent Agent Technology,* 1996.

[5] M. Balabanovic and Y. Shoham, "Learning information retrieval agents: Experiments with automated web browsing," *AAAI SSS Info. Gathering from Heterogeneous, Distrib. Envst,* 1995.

[6] R. Armstrong, D. Freitag, T. joachims and T. Mitchell, "Webwatcher: A learing apprentice for the world wide web," *AAAI SSS Info. Gathering from Heterogeneous, Distrib. Envts,* 1995.

[7] P. Maes, "Agents that reduce work and information overload," *Comm. Of the ACM,* 37(7):31-40. 1994.