키워드를 이용한 뉴스 필터링 에이전트 시스템

진승훈*, 이승아*, 김종완*, 권영직*
*: 대구대학교 컴퓨터정보공학부
e-mail : GLIDE77@hitel.net

News filtering agent system using keyword

Seung-Hoon Jin*, Seung-A Lee, Jong-Wan Kim*, Young-Jik Kwon*
*Department of Computer and Information Engineering, Taegu University

요약

인터넷의 급성장과 함께 인터넷을 통해 제공되는 서비스 중 사용자들에게 제공되는 뉴스서비스는 사용자가 원하지 않은 뉴스들까지 제공됨으로써 원하는 뉴스만을 골라서 제공받을 수 있는 시스템의 필요성이 증가하고 있다. 본 논문에서는 사용자가 입력하는 키워드를 이용하여 각 뉴스서버에서 제공하는 뉴스 중 사용자의 요구에 적합한 뉴스를 필터링하는 에이전트 시스템을 구현하였다.

1. 서론

1990년대 이후 인터넷이 급속도로 발전하고, 일반 사용자들에게 보급되면서 인터넷을 통해 제공되는 정보의 양도 기하급수적으로 증가하고 있다. 하지만 사용자의 입장에서 보면 아직도 웹상에서 존재하는 많은 자료들 중에서 찾고자 하는 정확한 정보를 빠른 시간 안에 검색할 수 있다거나, 원하는 정보만 필터링 되어져 제공받고 있다고 할 수는 없다. 특히 인터넷 사용자들이 많이 사용하는 기능 중의 하나인 뉴스서비스의 경우 매일 사용자에게 전달되는 많은 뉴스와 스팸 메일을 포함한 광고들 중에서 실제적으로 필요로 하는 뉴스를 검색해 내는 필터링의 기능이 절실히 요구되고 있다.

텍스트로 구성된 문서들 중에서 사용자가 원하는 내용이나 키워드를 포함한 문서만을 필터링해서 제공하는 기능은 이미 오래 전부터 정보검색이라는 분야에서 활발히 연구되어져 왔다. 이러한 정보검색 기능은 이제 인터넷을 통해 제공되는 많은 정보들 중에서 사용자의 요구에 맞는 정확한 문서를 검색해 내는데 까지 그 활용이 확대되고 있다.

본 논문에서는 뉴스 필터링에 대한 사용자 요구를 해결하기 위해 인터넷에 접속된 뉴스서버들에 접속해서 뉴스를 모아오고, 모아온 뉴스들 중에서 사용자가 원하는 키워드에 맞는 뉴스를 걸러낼 수 있는 뉴스 필터링 시스템을 구현하였다. 이 시스템에서는 사용자의 기호를 학습하여 뉴스를 필터링 하였는데 사용자 기호의 학습을 위하여 신경망 기법 중에서 코호넨 네트워크를 이용하였다.

본 논문의 구성을 살펴보면 먼저 2장에서 기존의 관련 연구를 통하여 코호넨 네트워크(Kohonen network)와 사용자의 기호를 이용한 뉴스 필터링 시스템에 대해 살펴본다. 3장에서는 본 논문에서 제안한 사용자 기호를 학습해서 뉴스를 필터링 하도록 구현한 시스템을 설명하고, 4장에서는 시스템을 실제 실험하고, 실험결과를 평가한다. 그리고, 5장에서는 논문의 결론 및 향후 연구 방향을 제시한다.

2. 관련 연구

2.1 신경회로망

인간의 두뇌 작용을 신경세포들 간의 연결 관계로 모델링 한 것을 신경회로망이라 하는데, 신경회로망은 사람의 학습 능력과 마찬가지로 교사(teacher, trainer)가 가르쳐주면서 학습하는 지도학습(supervised learning) 신경회로망과 교사 없이 스스로 학습하는 비지도 학습(unsupervised learning) 신경회로망으로 분류할 수 있다.

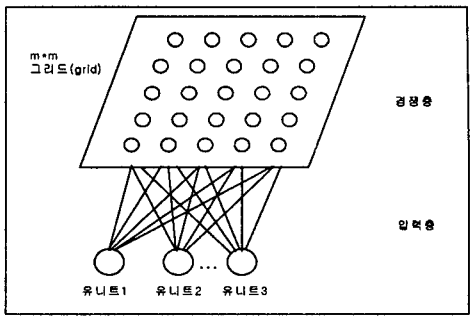
지도 학습은 실험 데이터의 입력벡터와 그에 대응하는 출력 값을 함께 신경회로망에 입력시킨 후 학습시키는 방법이다. 대표적인 알고리즘들로는 델타규칙(delta rule)과 오류역전파(error backpropagation:EBP) 학습 규칙이 있다.

비지도 학습 신경회로망은 출력층의 목표값을 필요로 하지 않으므로 미리 결정된 해(解)와의 비교가 필요하지 않다. 대신 학습 알고리즘은 비슷한 입력패턴들이 같은 출력뉴런으로 학습되도록 연결가중치들을 수정하여 준다. 따라서 학습과정은 학습 데이터의 통계적 성질을 추출하고, 유사한 벡터들을 같은 클래스로 분류하여 준다. 대표적인 알고리즘들로는 코호넨으로 대표되는 경쟁학습(competitive learning) 알고리즘과 Grossberg로 대표되는 ART(Adaptive Resonance Theory) 모델이 있다.

2.2 코호넨 네트워크

코호넨 네트워크에서 학습방법은 먼저 각 뉴런이 연결강도(weight) 벡터와 입력(input)벡터의 거리가 얼마나 가까운가를 계산한다. 그리고 각 뉴런들은 학습할 수 있는 특권을 부여받으려고 서로 경쟁하게 되는데 거리가 가장 가까운 뉴런이 승리하게 된다. 이 승자뉴런의 연결강도 벡터는 입력벡터에 가장 가까운 것으로 이 뉴런만이 출력신호를 보낼 수 있는 유일한 뉴런이 되고, 이 뉴런과 인접한 이웃 뉴런들만이 제시된 입력벡터에 대한 학습이 허용된다.

코호넨 네트워크의 학습원칙은 '승자 독점(winner take all)'으로, 승자뉴런을 결정하고 난 후에 코호넨의 학습규칙에 따라 뉴런의 연결강도를 조정해야 한다.



(그림 1) 코호넨 네트워크

본 논문에서는 사용자의 기호를 학습하여 원하는 정보를 검색결과로 제시하기 위하여 신경회로망 중 목표값 없이 학습 데이터만을 단순히 신경회로망의 입력으로 사용하여, 신경회로망이 스스로 연결가중치들을 학습시키는 비지도 학습 회로망의 한 종류인 코호넨 네트워크를 사용하였다.

2.3 필터링 에이전트

정보시스템이 발전하고 사용자에게 제공되는 정보의 양이 증가하면서 사용자가 필요로 하는 정확한 정보를 빨리 검색하여 제공할 수 있는 시스템에 대한 요구가 발생하게 되었고, 정보검색과 필터링 시스템 등의 형태로 발전하게 되었다. 또한 이러한 시스템은 인터넷의 사용이 확산되고, 사용자의 시스템에 대한 의존도가 높아지면서 사용자의 요구를 파악하여 그 요구에 대한 작업을 사용자 대신 수행할 수 있는 에이전트 시스템으로 발전하였다.

현재 인터넷 상에서 운영되고 있는 필터링과 관련된 에이전트는 여러 가지 형태가 있으며, 필터링하는 정보의 종류에 따라 웹 문서 필터링 에이전트, 상용뉴스 필터링에이전트, Usenet 뉴스 필터링 에이전트로 구분한다.

상용뉴스 필터링 에이전트는 인터넷을 통해 제공되는 상업용 뉴스를 사용자에게 제공하는 시스템으로 사용자가 미리 입력한 프로파일에 따라 그에 적합한 새로 추가되는 뉴스를 필터링하여 제공한다. 대표적으로 NewsHound,, Personal Journal,

PointCast Network 등이 있다.

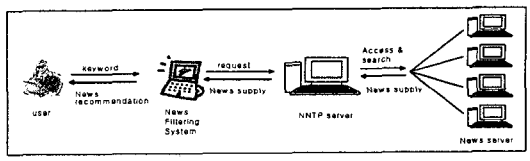
상용뉴스 서비스와 구분되는 Usenet 뉴스 서비스는 주제별로 구성되며 수많은 사용자들이 자유로이 내용을 게시하고, 무료로 뉴스 서비스를 제공받을 수 있다. 그러나 상용뉴스 서비스에 비해 사용자는 원하지 않는 뉴스까지 모두 제공받게 되므로 Usenet 뉴스 서비스에서의 필터링 기능이 더욱 필요하다고 할 수 있다. 이러한 Usenet 뉴스 서비스에서 사용자의 기호에 맞는 뉴스를 필터링 해주는 에이전트는 SIFT, InfoScan, BORGES, NewsClip, MAXIMS, Mailagent 등이 있다.

3. 시스템 구조

3.1 시스템의 기본 구조

본 논문에서 구현한 뉴스 필터링 시스템은 기본적으로 자바언어로 구현하였으며 사용자 인터페이스를 GUI로 하기 위해 Swing 1.1 버전을 사용하였다.(그림 2) 또한 java.net.Socket class를 사용해서 NNTP Server에 접속하였고, NNTP Protocol을 통해서 뉴스그룹을 선택하고, 뉴스문서의 목록 및 내용을 조회할 수 있도록 하였다. 유즈넷 접속과 뉴스그룹, 뉴스문서 조회에 대한 기능을 NewsHost class에 구현하였는데 유즈넷은 news.netple.com 같은 도메인으로 접속할 수 있는 서버가 있고, 각 서버마다 여러 개의 그룹이 있다. 그러나 존재하지 않는 뉴스그룹이 상당히 많기 때문에 이 프로그램을 사용하여 뉴스서버에서 각 뉴스그룹에 접속할 경우에 존재하지 않는 뉴스그룹의 경우에는 Exception이 발생한다. 이런 예외상황이 발생하면 무시하고 다음 뉴스그룹으로 넘어간다.

뉴스서버에서 뉴스를 읽어올 때 먼저 뉴스의 시작번호와 끝번호를 읽어온 후 그 시작번호부터 끝번호까지 뉴스를 읽어오도록 명령어를 실행한다. 이때 처음에 읽어왔던 시작번호와 끝번호의 정보와는 달리 뉴스가 그만큼 존재하지 않는 경우가 종종 있다. 이 경우 서버에서 보낸 코드를 이용하여 처리한다. 예를 들어, 올바른 뉴스문서의 번호를 서버에 요청하였을 경우 "223" 이란 코드를 포함한 값이 리턴된다.



(그림 2) 뉴스 필터링 시스템 구조

3.2 학습 방법

뉴스 필터링 시스템 신경망 기법 중 코호넨 네트워크를 이용하여 사용자의 기호를 학습하게 하였다. 먼저, 사용자는 자신이 원하는 뉴스에 포함될 키워드를 입력할 수 있고, 시스템은 각 뉴스문서에 대해서 각 키워드들이 몇 번 나타나는지를 코호넨 네트워크에 대한 입력 벡터로 취급해서 학습한다.

코호넨 신경회로망 학습 알고리즘은 아래와 같이 6단계로 구성된다.

[단계 1] 연결강도를 초기화한다.

N개의 입력으로부터 M개의 출력 뉴런 사이의 연결강도를 작은값의 임의의 수로 초기화한다. 이웃 반경은 충분히 크게 잡은 후 점차 줄어든다.

[단계 2] 새로운 입력벡터를 제시한다.

[단계 3] 입력벡터와 모든 뉴런들 간의 거리를 계산한다. 입력과 출력 뉴런 j 사이의 거리 dj는 다음과 같이 계산한다.

$$d_j = \sum_{i=0}^{n-1} (X_i(t) - w_{ij}(t))^2$$

[단계 4] 최소거리에 있는 출력 뉴런을 승자 뉴런으로 선택한다. 최소거리 dj 인 출력뉴런 j*를 선택한다.

[단계 5] 승자 뉴런 j*와 그 이웃들의 연결강도를 재조정한다. 뉴런 j*와 그 이웃 반경내의 뉴런들의 연결강도를 다음 식에 의해 재조정한다.

$$Wij(t+1) = Wij(t) + \alpha(x_i(t) - Wij(t))$$

여기에서 j는 j*와 j*의 이웃반경내의 뉴런이고 i는 0에서 N- 1까지의 정수값이다. α는 0과 1사이의 값을 가지는 이득항(gain term)인데 시간이 경과함에 따라 점차 작아진다. α = α * (1/iteration) 의 값을 이용 초기값 0.9에서 0.05까지 줄어들면서 계산한다.

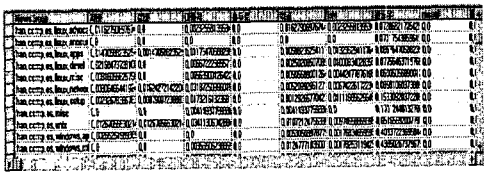
[단계 6] 단계 2로 가서 반복한다.

4. 실험 및 평가

훈련 데이터(training data)를 모으기 위하여 자바의 Socket Class를 이용하여 NNTP Server에 접속한 후, 각 뉴스그룹에서 뉴스문서를 내려 받는다. 이때 이미 삭제되었거나 옮겨진 뉴스그룹과 10개 이하의 문서를 가지고 있는 뉴스그룹은 제외시킨다.

131개의 뉴스그룹을 검색하여 조건에 맞는 71개의 뉴스그룹을 훈련데이터로 사용하였으며, 출력뉴런의 크기는 5*5 이고 훈련은 1000회 하였다.

훈련 데이터를 파싱하여 미리 입력된 단어들의 개수를 알아내고, 정규화(normalization)하기 위하여 각 뉴스 그룹별로 단어들의 비율을 계산 하여 사용한다.



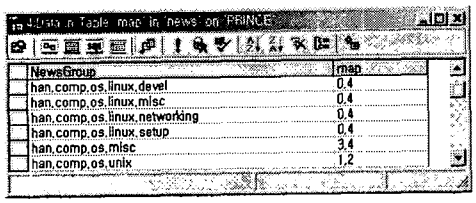
(그림 3) 정규화된 입력벡터

(그림 3)에서는 본 시스템에서 접속된 각 뉴스그룹 별로 불러온 각 뉴스문서에서 출현한 키워드의 비율을 계산한 결과를 보여주고 있다.

사용자가 입력한 키워드를 이용하여 새로운· 입력벡터를 생성한다.

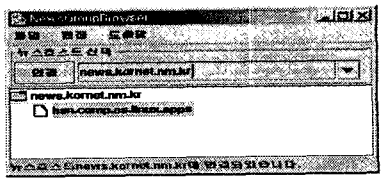
사용자가 입력한 키워드와 미리 입력되어있는 키워드와의 거리 계산을 위하여 사용자가 입력하지 않은 키워드의 값을 0으로 하여 차원을 맞춘다.

사용자가 입력한 키워드는 각 뉴스 그룹에서 출현한 비율의 평균값을 사용하였다.



(그림 4) 각 뉴스 그룹의 위치정보

(그림 4)는 훈련에 사용된 뉴스그룹들이 학습후 2차원 출력층에 배열된 예를 일부 보여준다.



(그림 5) 뉴스그룹 추천

(그림 5)는 사용자가 사용자 ID를 입력한 후의 초기화면이다. 사용자가 입력한 키워드와 미리 학습된 정보를 이용하여 가장 가까운 뉴스그룹을 보여주게 된다.

5. 결론 및 향후 연구 방향

본 시스템에서는 뉴스 그룹을 사용자에게 추천하는 방식을 사용하였다. 사용자는 자신이 입력한 키워드에 적합한 뉴스그룹들을 볼 수 있어 불필요한 검색을 줄일 수 있다. 이와 비슷하게 각 뉴스 그룹의 뉴스 문서를 학습 한 후 새롭게 갱신되는 뉴스문서를 새로운 입력벡터로 사용하여 사용자에게 적당한 문서인지를 파악하여 제공하는 시스템을 추가할 필요가 있다. 테스트를 위하여 입력벡터 키워드를 임의로 선정하였다. 어떤 키워드를 어떻게 추출할 것인가에 관한 연구도 병행하여야 한다.

참고문헌

[1] 이승원, 류제, 우성규, 한광록, 지능형 통합 에이전트의 구현, 2000년 한국정보처리학회 추계 학술발표논문집 제7권 제2호, pp.1437-1440, 2000.

[2] 최중민, 인터넷 정보가공을 위한 에이전트 연구동향, 정보처리학회지 4권 5호, pp 101-109, 1997.

[3] 김대수, 신경망 이론과 응용, 하이테크 정보.

[4] Alexandros Moukas, "Amalthaea: Information Discovery and Filtering using a Multiagent Evolving Ecosystem", The proceedings of the Conference on Practical Applications of Agents and Multiagent Technology, London, April, 1996.

[5] 한선미, 우진운, 지능형 에이전트를 이용한 개인화된 유·무선 뉴스 검색 시스템, 정보처리학회지, 제8-B권 6호, pp.609-616.

[6] Joseph P. Bigus, Jennifer Bigus, Costructing intelligent agents with JAVA, Wiley, 1998.