코호넨 신경망을 사용한 유즈넷 뉴스 필터링[†]

진승훈[0], 김종완, 김병만*
대구대학교 정보통신공학부
*금오공과대학교 컴퓨터공학부
{glide77[0],jwkim}@webmail.taegu.ac.kr, bmkim@se.kumoh.ac.kr

Usenet News Filtering using Kohonen Network

Seung-Hoon Jin[0], Jong-Wan Kim, Byeong Man Kim*
School of Computer and Information Engineering, Daegu University
*School of Computer & Software Engineering, Kumoh National Institute of Technology

요 약

With the proliferation of internet, it is increasingly needed to realize personalized news filtering service reflecting user's interest. In this paper, we implement a filtering agent for personalized news service. In the proposed system, Kohonen network for an unsupervised learning is used to train keywords provided by users and the personalization is achieved by using the trained neural network. After we trained and tested our filtering agent, we could provide users news groups considering their interests.

1. 서 론

1990년대 이후 인터넷이 급속도로 발전하고, 일반 사용자들에게 보급되면서 인터넷을 통해 제공되는 정보의 양도 기하급수적으로 증가하고 있다. 따라서 사용자들은 웹 상에서 존재하는 많은 자료들 중에서 찾고자 하는 정확한 정보를 빠른 시간 안에 검색하고, 원하는 정보만 필터링 되어져서 제공받기를 원하고 있다[2]. 하지만 사용자 입장에서 보면 아직 그러한 서비스가 만족스럽게 제공받고 있지 못하다. 특히 인터넷 사용자들이 많이 사용하는 기능 중의 하나인 뉴스 서비스의 경우 매일매일 사용자에게 전달되는 많은 뉴스와 스팸메일을 포함한 광고들 중에서 실제적으로 필요로 하는 뉴스를 검색해 내는 필터링의 기능이 절실히 요구되고 있다[3].

본 논문에서는 수많은 뉴스서버들에서 제공하는 뉴스들 중 사용자가 원하는 정확한 뉴스만을 필터링 해주는 서비스에 대한 사용자 요구를 해결하기 위해 먼저, 인터넷에 접속된 뉴스서버들에 접속해서 뉴스를 모아오도록 한다. 사용자는 미리 관심있는 분야에 대한 키워드를 입력할 수 있고, 시스템이 각 뉴스서버들에서 모아온 뉴스들 중에서 사용자가 입력한 키워드에 맞는 뉴스를 걸러낼 수 있도록 뉴스 필터링 시스템을 구현하였다. 또한 이 시스템에서는 사용자가 입력한 키워드를 통해 사용자의 기호를 학습하여 뉴스를 필터링하기 위해 신경망 기법 가운데 대표적인 비지도 학습 알고리즘인 코호넨 신경망을 이용하였다. 코호넨 신경망은 지속적인 사용자의 피드백을 요구하지 않는 비지도 학습의 한 종류로 사용자가 입력한 키워드만 가지고 뉴스그룹들을 학습시킬 수 있어서, 프로파일을 이용한 뉴스 그룹의 순위를 부여할 수 있다는 장점이 있다. 이에 본 연구에서는 코호넨 신경망을 필터링 알고리즘으로 채택하였다.

† 본 연구는 한국과학재단 목적기초연구(2000-1-51200-008-2) 지원으로 수행되었음.

2. 시스템 구조와 학습

2.1 시스템의 기본 구조

본 논문에서 구현한 뉴스 필터링 시스템은 자바언어로 구현된 Bigus의 뉴스 필터링 시스템[4]을 참고하여, 사용자 인터페이스를 GUI로 하기 용이한 Swing을 사용하여 자바언어로 구현하였다. 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.kornet.net 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 경우 그 존재하는 뉴스그룹의 경우에는 서버 응답 메시지의 첫 시작이 "211"로 시작한다. 이것을 이용하여 뉴스그룹의 존재 유무를 판단한다.

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. 존재하는 문서일 경우 서버 응답 메시지의 첫 부분이 "223"으로 시작한다. 뉴스그룹의 존재 유무의 판단과 같은 방법으로 문서의 유무도 판단한다.

2.2 학습 방법

제시된 뉴스 필터링 시스템은 코호넨 신경망을 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 신경망에 대한 입력 벡터로 취급해서 학습한다. 코호넨 신경망 학습 알고리즘은 아래와 같이 6단계로 구성된다[1].

[단계 1] 연결강도벡터 W를 초기화한다.
N개의 입력으로부터 M개의 출력 뉴런 사이의 연결강도를 임의로 생성되는 작은 값으로 초기화한다. 이웃반경은 충분히 크게 잡은 후 점차 줄어든다.

[단계 2] 새로운 입력벡터 X를 제시한다.

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

$$dj = \sum_{i=0}^{N-1} [X_i(t) - W_{ij}(t)]^2 \quad (1)$$

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

$$j^* = \min_j dj, \quad j \in \text{출력뉴런} \quad (2)$$

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

$$Wij(t+1) = Wij(t) + \alpha \cdot (Xi(t) - Wij(t)) \quad (3)$$

$$\alpha = \alpha_0 \cdot (1/epoch) \quad (4)$$

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. 본 연구에서 α값은 초기값 α_0로 0.9를 사용하였다.

[단계 6] 단계 2로 가서 반복한다.

3. 구현 및 실험

먼저, 훈련 데이터(training data)를 모으기 위하여 자바의 Socket Class를 이용하여 NNTP Server (news.kornet.net)에 접속한 후, 각 뉴스그룹에서 뉴스문서를 내려 받았다. 이때 이미 삭제되었거나 옮겨진 뉴스그룹과 10개 이하의 문서를 가지고 있는 뉴스그룹은 제외시켰다.

실험 결과 131개의 뉴스그룹을 검색하여 조건에 맞는 71개의 뉴스그룹을 훈련데이터로 사용하였으며, 출력뉴런의 크기는 5*5이고, 훈련은 1000회 실시하였다. 훈련 데이터는 각 뉴스그룹에서 임의로 단어들을 선택하여 데이터베이스에 저장해 놓고, 각 뉴스 그룹의 문서에서 단어들을 분석하여 입력된 단어들의 개수를 알아낸다. 본 논문에서는 총 31개의 단어를 임의로 선출하여 사용하였으며, 각 뉴스그룹의 문서의 수와는 상관없이 단어의 개수만을 파악했을 경우 문서의 수가 많은 뉴스그룹에서는 대체적으로 단어의 빈도수가 많다. 예를 들어, "han.comp.os.linux.networking" 뉴스그룹의 경우 문서의 수가 1448개인 반면, "han.answers" 뉴스그룹은 24개의 문서만 데이터베이스에 저장되어 있다. 이런 편차를 줄이기 위하여 본 논문에서는 정규화(normalization)를 수행한다.

정규화는 (각 단어의 개수)/(뉴스그룹에서 각 단어들이 나타난 총합)으로 계산하여 각 단어들이 뉴스그룹 내에서 나타나는 비율로 한다. 예를 들어, "han.answers"에서 각 단어들이 나타난 총합이 416번이며, "메일"이란 단어는 284번 나타났다. 이 경우에 "han.answers"에서 "메일"이라는 키워드의 비율은 "284/416 = 0.682" 가 된다. 나머지 단어들도 마찬가지 방식으로 계산한 결과가 그림 1과 같다.

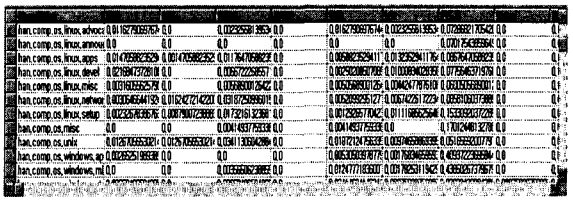

그림 1 정규화된 입력벡터

학습이 끝난 후 각 뉴스그룹의 코호넨 신경망의 출력층 위치와 연결강도 값을 그림 2와 같이 데이터베이스에 저장한다. 그림 2는 학습에 사용된 뉴스그룹들이 학습이 완료된 후 2차원 출력층에 배열된 예의 일부를 보여준다. 그림에서 알 수 있듯이, "리눅스"와 관련된 뉴스그룹들이 코호넨 신경망의 (0,4) 출력 뉴런에 모여 있음을 확인할 수 있다.

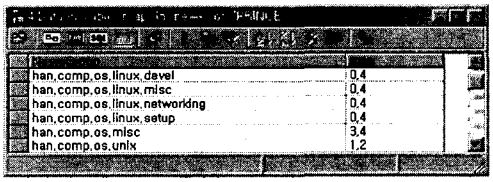

그림 2 각 뉴스 그룹의 위치정보

표 1은 사용자가 입력한 키워드 프로파일을 나타낸 것이다. 사용자가 입력한 키워드를 이용하여 테스트용 입력벡터를 생성한다. 사용자가 입력한 키워드와 미리 입력되어있는 키워드와의 거리를 계산하기 위하여 사용자가 입력하지 않은 키워드의 값을 0으로 하여 차원을 일치시켰다. 사용자가 입력한 키워드는 각 뉴스 그룹에서 출현한 비율의 평균값을 사용하였다.

표 2 사용자 정보

glide77	****	진승훈	자바, 삼바, 리눅스, 오피스, information

테스트용 입력벡터가 계산되면 코호넨 신경망에 제시하여 가장 가까운 출력뉴런을 선정하고, 이 뉴런에 속하는 뉴스그룹들을 사용자에게 제시한다. 그림 3은 사용자(glide77)가 자신의 ID를 입력한 후의 결과 화면으로, 사용자가 입력한 키워드와 미리 학습된 정보를 이용하여 가장 가까운 뉴스그룹을 보여준다. 그림 3에서는 출력뉴런 (4,2)가 승자뉴런으로 선정되었다. 본 논문에서는 선정된 승자뉴런과 관련된 뉴스그룹들의 순위(ranking)를 부여하여 사용자에게 순위 순으로 제시한다. 순위는 식 (5)과 같이 (모든 키워드들의 빈도수 비율의 합 / 키워드가 포함된 수)를 이용하여 계산하였다.

$$ranking(k) = \frac{\sum_{i=1}^{p} f_i(k)}{d} \quad \text{for all } k \qquad (5)$$

p는 프로파일에 등록된 키워드 수, k는 키워드, d는 키워드가 해당 뉴스그룹에 나타난 경우의 수를 나타낸다.

예를 들어, han.comp.lang.java의 경우 ((0.4904 + 0 + 0.0300 + 0 + 0.0015) / 3) = 0.17401이 된다. glide77 사용자의 경우 표 2와 같이 순위가 계산되어 그림 3과 같은 결과를 제시한다.

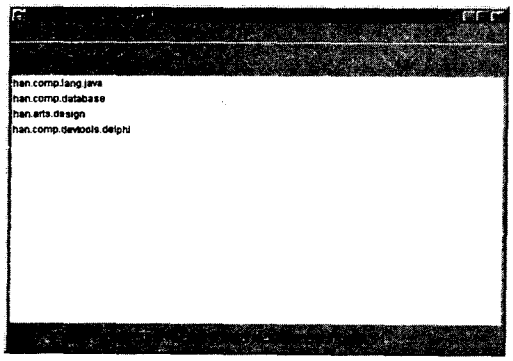

그림 3 결과 화면 (glide77 사용자)

표 3 순위 계산 결과

han.comp.lang.java	0.17401
han.comp.database	0.04869
han.arts.design	0.02580
han.comp.devtools.delphi	0.018390
han.comp.devtools.vb	0.00868055555
han.arts.music.gugak	1.0E-99

표 2에서 "han.arts.music.gugak"의 경우 "1.0E-99"의 아주 작은 값이 계산되었다. 코호넨 신경망에서 훈련을 통하여 "han.arts.music.gugak" 뉴스그룹이 표 2에 있는 뉴스그룹과 같은 뉴런에 분류되었으나 사용자가 입력한 키워드가 "han.arts.music.gugak" 뉴스그룹에서 하나도 존재하지 않았으므로 최하위 값이 선정되었다. 본 연구에서는 임계치를 사용하여 임계치 보다 낮은 뉴스그룹은 관련정도가 적다고 판단하여 제외하였다. 현재는 임계치로 0.01을 사용하여, 순위 계산값이 작은 2개의 뉴스그룹을 제거시켰다. 최종 순위 부여 결과가 그림 3과 같이 계산되어 사용자에게 보여진다.

4. 결론 및 향후 과제

본 연구에서는 사용자가 관심 있는 키워드와 관련 있는 뉴스 그룹을 사용자에게 추천하는 방식으로 유즈넷 뉴스 필터링 시스템을 구현하였다. 학습 방법으로 임의로 선정된 키워드들의 클러스터링에 용이한 코호넨 신경망을 사용하였다.

본 연구의 특징을 다음과 같이 정리할 수 있다. 첫째, 각 뉴스그룹들의 문서의 개수가 서로 달라 비슷한 내용을 지닌 뉴스그룹의 경우라도 문서의 개수가 많은 곳과 적은 곳의 경우 서로간의 단어 빈도수 차이가 많이 나서 거리가 멀어지게되어 비슷한 뉴스그룹으로 분류할 수 없게 된다. 이러한 편차를 줄이기 위하여 정규화를 하였다. 둘째, 테스트시에 입력벡터의 차원을 일치시키기 위하여, 사용자가 입력한 키워드의 경우, 키워드가 나타난 뉴스그룹들의 빈도수의 평균을 구하여 사용하였다. 셋째, 선택된 뉴스그룹을 사용자에게 순위 순으로 제시하여 어떠한 뉴스그룹이 사용자가 입력한 키워드와 가장 유사한 값을 지니고 있는지를 파악할 수 있으며 순위를 부여하지 않고 제시하는 것보다 불필요한 검색을 줄일 수 있다. 넷째, 비슷한 뉴스그룹으로 분류는 되었으나 사용자가 입력한 키워드와 관계가 적은 뉴스그룹들은 사용자에게 제시할 필요가 없으므로 임계치를 사용하여 제거하였다.

본 시스템에서는 테스트를 위하여 입력벡터 키워드를 임의로 선정하였는데, 키워드 추출 방법에 대한 연구가 추가되어야 보다 의미 있는 필터링 기능을 수행할 수 있다. 현재 이 분야에 관한 연구를 수행하고 있다. 향후에는 각 뉴스 그룹의 뉴스 문서를 학습한 후 새롭게 갱신되는 뉴스 문서를 새로운 입력벡터로 사용하여 사용자에게 적당한 문서인지를 파악하여 제공하는 시스템도 추가할 필요가 있다.

참고문헌

[1] 김대수, 신경망 이론과 응용, 하이테크 정보, 1992.
[2] 최중민, "인터넷 정보가공을 위한 에이전트 연구동향," 정보처리학회지, 4권 5호, pp 101-109, 1997.
[3] Point CAst Network http://www.pointcast.com/.
[4] Joseph P. Bigus, Jennifer Bigus, Costructing intelligent agents with JAVA, Wiley, 1998.