퍼지추론과 코호넨 신경망을 사용한 뉴스 필터링 시스템의 분류 능력

김종완*, 조규철*, 김병만**
대구대학교 정보통신공학부*, 금오공대 컴퓨터공학부**
{jwkim, kccho}@webmail.daegu.ac.kr, bmkim@se.kumoh.ac.kr

Classification Performance of News Filtering System by Fuzzy Inference and Kohonen Network

Jong-Wan Kim*, Kyu-Cheol Cho*, Byeong Man Kim**
School of Computer and Information Engineering, Daegu University*
School of Computer Engineering, Kumoh National Institute of Technology**

요 약

많은 양의 유즈넷 뉴스 중에서 찾고자 하는 정확한 정보를 빠른 시간 안에 검색하고, 원하는 정보만 필터링 하는 것은 중요하다. 하지만 뉴스 문서는 이메일과 달라서 미리 자신에게 맞는 뉴스그룹을 등록해 주어야만 정보를 얻을 수 있다. 본 연구에서는 다양한 뉴스그룹들 중에서 사용자와 취향이 가장 유사한 뉴스그룹을 코호넨 신경망을 이용하여 분류하는 서비스를 제공한다. 신경망을 학습시키기 위한 뉴스 문서의 키워드들을 선택하기 위해 예제 문서들로부터 후보 용어들을 추출하고 퍼지 추론을 적용하여 대표 용어들을 선택한다. 뉴스 필터링 시스템의 분류 성능을 평가하기 위하여 유클리드 거리 면에서 비교한 결과, 제안한 방법의 유용성을 확인할 수 있었다.

1. 서론

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 수집하도록 한다. 그리고 수집된 뉴스들의 대표 용어들을 추출하기 위해서 뉴스들로부터 후보 용어들을 추출하고 퍼지추론을 적용하여 대표 용어들을 선택한다. 제안 방법의 성능은 대표 용어들을 선택하는 방법에 의해 영향을 크게 받는다. 따라서 뉴스그룹에서 대표 용어를 추출하는 문제는 불확실성을 내포하고 있으므로 이러한 문제 해결에 효과적인 퍼지추론을 대표 용어의 선택 방법에 적용하였다. 제안된 뉴스 필터링 시스템에서는 추출된 대표 용어로 뉴스 문서를 학습시키기 위해 신경망 기법 가운데 대표적인 비지도 학습 알고리즘인 코호넨 신경망을 이용하였다. 코호넨 신경망은 지속적인 사용자의 피드백을 요구하지 않는 비지도 학습의 한 종류로 주어진 키워드 즉 대표 용어만 가지고 뉴스그룹들을 학습시킬 수 있다는 장점이 있다. 이에 본 연구에서는 코호넨 신경망을 필터링 알고리즘으로 채택하였다.

2. 제안된 뉴스 필터링 성능 향상 방법

2.1 뉴스 필터링 시스템의 기본 구조 및 대표 용어 선택 방법

본 연구에서 제안하는 뉴스 필터링 시스템의 기본 구조는 그림 1과 같다. 그림에서 알 수 있듯이, 사용자가 키워드를 입력하면 뉴스 필터링 시스템이 유즈넷 뉴스 서버(NNTP server)에게 요청을 하고, 유즈넷 뉴스 서버는 이를 다시 인터넷상의 뉴스 서버들에 접속하여 뉴스 문서들을 내려 받는다. 이 뉴스 문서들을 뉴스 필터링 시스템에게 전송하면, 퍼지추론으로 대표 용어를 추출하고 코호넨 신경망으로 학습하여 사용자의 의도와 유사한 뉴스그룹을 제시해 준다.

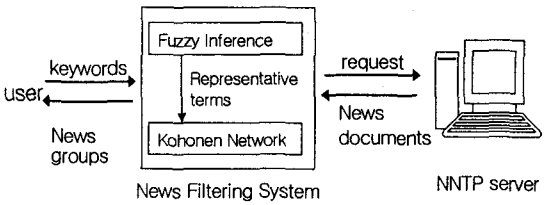


그림 1. 제안된 뉴스 필터링 시스템

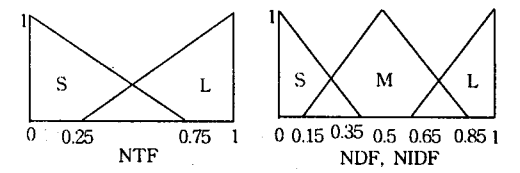
유즈넷 뉴스들을 필터링하기 위해서는 사용자의 관심 내용을 가장 잘 대변하는 대표 용어의 선택이 중요하다. 특정 용어의 중요도 계산에 사용되는 입력 정보들은 정량적으로 정확히 해석될 수 없는 부정확하고 불확실한 특성을 내포하고 있다. 따라서 본 연구에서는 소수의 긍정적 학습 문서 집합들에 대해서 실험한 결과 기존의 대표 용어 추출 방법들보다 비교적

우수한 것으로 알려진 방법 [1]을 사용하여 후보 용어들의 가중치를 계산하고 이 값들에 따라 선택 우선순위를 부여하였다. 그 방법을 설명하면 아래와 같다.

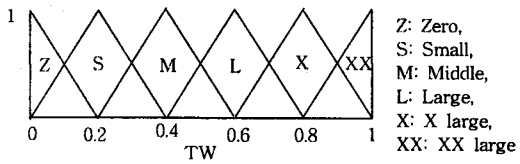
퍼지추론을 이용한 대표 용어 중요도를 계산하기 위해 뉴스들은 불용어를 처리하고, Porter stemmer[2]를 사용한 스테밍 과정에 의해 후보 용어들의 집합으로 변형되며, 이 집합으로부터 각각의 용어들의 TF(Term Frequency), DF(Document Frequency), IDF(Inverse Document Frequency) 정보가 구해진다. 이들 정보들을 정규화하여 퍼지추론을 위한 퍼지시스템의 입력으로 이용한다. 정규화 과정을 간단히 설명하면, NTF(Nomalized Term Frequency)는 TF_i(예제 문서 집합에서 i번째 단어의 발생 빈도수)를 DF_i(예제 문서 집합에서 i번째 단어를 포함하는 문서의 수)로 나누어 계산하며, NDF(Nomalized Document Frequency)는 DF_i를 TD(예제 문서의 수)로 나누어 구하며, NIDF(Normalized Inverse Document Frequency)는 IDF_i(i번째 단어의 역문헌 빈도수)를 역문헌 빈도수 최대값으로 나누어 계산한다.

그림 2는 퍼지 추론을 위하여 사용된 입출력 변수들의 멤버쉽함수를 나타내고 있다. 용어별로 구해진 NTF, NDF, NIDF 값들을 퍼지추론에 적합한 형태로 퍼지화 시켜야 한다. 본 논문에서는 그림 2와 같은 삼각형 형태의 퍼지 수를 사용하였다. 그림 2(a)에서 NTF 입력변수 값은 S(Small)과 L(Large)로 2개의 멤버쉽함수 부분으로 나누었고,

NDF와 NIDF들은 S(Small), M(Middle), L(Large)로 하였다. 그림 2(b)에서 중요도를 나타내는 퍼지 출력변수인 TW(Term Weight)는 6개의 멤버쉽함수 부분으로 나누었다.



(a) 입력변수



(b) 출력변수

그림 2. 퍼지 입출력변수

표 1은 NTF 퍼지 입력값의 소속 정도에 따라 두 부분으로 나누어 규칙들을 표현하고 있다. 작성 과정의 예를 살펴보면, NTF가 S, NDF가 L, 그리고 NIDF가 S일 경우, 해당 용어가 대부분의 예제 문서들에 등장함으로 인해 관련성을 높게 평가 할 수 있지만 NTF와 NIDF가 낮은 값을 취함으로 관련 정도는 S(낮음)으로 설정하였다. 이와 같은 과정으로 다른 모든 규칙들의 후건부를 설정하였다.

NTF, NDF, NIDF 퍼지 입력값을 위의 결과로 생성된 18개의 추론 규칙별로 이들의 전건부의 소속 함수에 적용시킨다. 각각의 소속 정도가 구해지면 이들 중에서 최소값을 취한다. 그 결과 규칙별로 하나씩의 퍼지 값이 생성되며 이 퍼지 값들을 퍼지 출력 변수 TW에 따라 6개의 그룹으로 분류하고 그룹별로 해당 그룹에 속한 퍼지 값들 중 최대값을 취하여 총 6개의 퍼지 값들을 생성한다. 최종적으로 이들 6개의 퍼지 값들을 무게중심법(center of gravity)으로 비퍼지화(defuzzification)한 값이 해당 용어의 중요도 값으로 결정되어진다.

표 1. 퍼지 추론규칙

NDF \ NIDF	S	M	L
S	Z	S	M
M	S	L	X
L	S	X	XX

NTF = S

NDF \ NIDF	S	M	L
S	Z	Z	S
M	Z	M	L
L	S	L	X

NTF = L

2.2 코호넨 신경망을 이용한 뉴스그룹을 분류

본 연구에서 제안된 뉴스 필터링 시스템은 뉴스그룹 문서들을 분류하는데 코호넨 신경망을 이용하여 학습하였다. 코호넨 신경망을 사용한 이유는 교사의 지시 없이 뉴스그룹 문서들로부터 자연스럽게 연관관계를 분류할 수 있기 때문이다. 학습 알고리즘은 아래와 같은 일반적인 코호넨 학습규칙을 따랐다[3]. 제안된 시스템에서는 뉴스문서에 대해서 각 키워드들이 몇 번 나타나는지를 코호넨 신경망에 대한 입력벡터로 취급해서 학습한다.

[단계 1] 연결강도벡터 W를 초기화한다.
[단계 2] 새로운 입력벡터 X를 제시한다.
[단계 3] 입력벡터와 모든 뉴런들 간의 거리를 계산한다. 입력과 출력 뉴런 j 사이의 거리 d_j는 다음과 같이 계산한다.

$$d_j = \sum_{i=0}^{N-1} [X_i(t) - W_{ij}(t)]^2 \quad (6)$$

[단계 4] 최소거리에 있는 출력 뉴런을 승자뉴런으로 선택한다. 최소거리 d_j 인 출력뉴런 j^*를 선택한다.

$$j^* = \min_j d_j, \quad j \in \text{출력뉴런} \quad (7)$$

[단계 5] 승자뉴런 j^*와 그 이웃들의 연결강도를 재조정한다. 뉴런 j^*와 그 이웃 반경내의 뉴런들의 연

결강도를 다음 식에 의해 재조정한다.

$$W_{ij}(t+1) = W_{ij}(t) + \alpha \cdot [X_i(t) - W_{ij}(t)] \quad (8)$$

$$\alpha = \alpha_0 \cdot (1/epoch) \quad (9)$$

여기에서 j는 j^*와 이의 이웃반경내의 뉴런이고, i는 0에서 N-1까지의 정수값이다. α는 0과 1사이의 값을 가지는 이득항인데, 학습회수인 $epoch$가 증가함에 따라 점차 작아진다. 본 연구에서 α의 초기값 α_0는 0.9를 사용하였다.

[단계 6] 단계 2로 가서 반복한다.

3. 실험 및 분석

3.1 실험 데이터 수집 및 학습 방법

본 논문에서 구현한 뉴스 필터링 시스템은 사용자 인터페이스로 Swing을 사용하여 자바언어로 구현하였다[4]. 먼저, 훈련 데이터를 수집하려고 자바의 java.net.Socket 클래스를 이용하여 유즈넷 뉴스 서버인 news.kornet.net에 접속한 후, NNTP 프로토콜을 통해서 뉴스그룹을 선택하고 각 뉴스그룹에서 뉴스 문서를 내려 받았다. 이때 이미 삭제되었거나 옮겨진 뉴스그룹과 10개 이하의 문서를 가지고 있는 뉴스그룹은 제외시켰다.

실험은 126개의 뉴스그룹을 대상으로 하였으며, 퍼지추론으로 대표 용어를 추출하는 경우에 그룹당 10개의 문서를 임의로 선택하는 경우와 20개의 문서를 임의로 추출한 경우를 실험하였다. 실험을 두 가지 경우로 수행한 이유는 추출된 용어의 개수가 분류 성능 실험에 얼마나 영향을 미치는지 확인하기 위한 것이다. 출력뉴런의 크기는 5*5로 정하였으며, 훈련은 각각 1000회 실시하였다. 훈련 데이터는 각 뉴스그룹에서 퍼지추론으로 용어들을 추출하고 이를 데이터베이스에 저장한 후, 각 뉴스그룹의 문서에서 용어들을 신경망으로 분석한다.

본 논문에서 추출된 단어는 126개의 그룹에서는 용어 추출에 사용된 문서의 수에 따라 25개와 28개의 단어를 사용하였다. 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 정규화를 수행한다[4]. 정규화는 (각 단어의 빈도수)/(뉴스그룹에서 각 단어들이 나타난 총 빈도수)으로 계산하여 각 단어들이 뉴스그룹 내에서 나타나는 비율로 한다. 예를 들어, "han.answers"에서 각 단어들이 나타난 총 빈도수가 416이며, "메일"이란 단어는 284번 나타났다. 이 경우에 "han.answers"에서 "메일"이라는 단어의 비율은 "284/416 = 0.682" 가 된다. 나머지 단어들도 마찬가지 방식으로 계산한 결과가 그림 3에 나타난다.

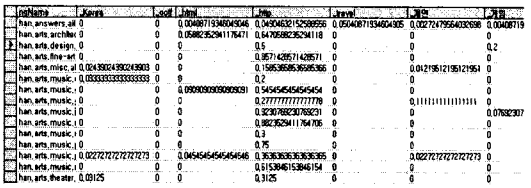


그림 3. 126개의 뉴스그룹의 정규화된 입력벡터

학습이 끝난 후 각 뉴스그룹의 코호넨 신경망의 출력층 위치와 연결강도 벡터를 그림 3과 같이 데이터베이스에 저장한다. 그림 4는 학습에 사용된 뉴스그룹들이 학습이 완료된 후 2차원 출력층에 배열된 예의 일부를 보여준다. 그림에서 알 수 있듯이, 126개의 뉴스그룹 중에서 코호넨 신경망의 (4,1) 출력 뉴런에 모여 있는 뉴스그룹들은 유사한 그룹과 상관정도가 낮은 그룹이 함께 분류되는 현상을 발견하게 된다. 이것은 분류기 자체의 성능도 일부 있지만, 그보다는 학습에 사용된 대표 용어들이 혼재하고 있다는 사실에 보다 기인한다.

NewsGroup	map
han.comp.os.windows.setup.all	4,1
han.comp.security.all	4,1
han.comp.www.misc.all	4,1
han.politics.all	4,1
han.comp.lang.c.all	4,2

그림 4. 뉴스그룹에 대해 학습된 출력층 일부

3.2 학습 성능평가

학습 성능을 평가하기 위해서 본 연구에서는 코호넨 신경망과 일반적인 k-nearest neighbor 방법의 유클리드 거리를 비교하였다.

사용자가 입력한 키워드를 이용하여 테스트용 입력벡터(U)를 생성한다. 사용자가 입력한 키워드와 미리 입력되어있는 키워드와의 거리를 계산하기 위하여 사용자가 입력하지 않은 키워드의 값을 0으로 하여 입력벡터의 차원을 일치시켰다. 사용자가 입력한 키워드는 각 뉴스 그룹에서 출현한 비율의 평균값을 사용하였다.

실험 방법은 이 테스트벡터 U를 코호넨 신경망의 입력벡터로 사용하여 승자뉴런을 선정하고 그들간의 유클리드 거리(Euclidean distance)와 코사인 유사도(Cosine Similarity)를 계산한다[5]. 유클리드 거리는 아래의 식 (10)과 같이 계산된다. 즉 사용자가 제시한 키워드 벡터와 학습된 벡터간의 유클리드 오차를 구해준다.

$$L(W, U) = \sqrt{\sum_i (W_i - U_i)^2} \quad (10)$$

여기서 W는 코호넨 신경망 또는 k-nearest neighbor 방법으로 학습한 벡터를 나타내고, U는 사용자가 필터링 시스템에 테스트시 제시한 키워드 벡터를 나타낸다.

코사인 유사도는 아래의 식 (11)과 같이 계산된다. 식(11)에서 보듯이, 코사인 유사도는 학습된 벡터와 테스트 벡터간의 매칭정도를 비교하는 척도이다.

$$\sigma(W, U) = \frac{\sum_i (W_i \times U_i)}{\sqrt{\sum_i (W_i)^2} \times \sqrt{\sum_i (U_i)^2}} \qquad (11)$$

한편 k-nearest neighbor 방법의 k를 결정하기 위해, 신경망에 포함된 클러스터의 개수를 사용하였다. 예를 들면, 테스트벡터 Ui를 코호넨 신경망에 입력시켜서 재생한 결과가 출력뉴런 (2,1)로 결정된 경우, 테스트벡터와 출력뉴런의 연결강도벡터간의 오차를 계산한다. 또한 이 (2,1) 뉴런이 4개의 뉴스그룹을 대표하고 있다고 판정된 경우에 k=4로 보고, k-nearest neighbor 방법의 상위 4개 뉴스그룹의 벡터와 Ui 벡터간의 유클리드 오차를 계산 후 이를 평균한 값과 비교한다. 표 2의 사용자를 대상으로 유클리드 거리를 계산한 결과는 표 3과 표 4에 주어진다.

표 2 126개의 뉴스그룹에 사용자 정보

userid	keywords
test1	html, http, 뉴스, 서버, 시스템
test2	Korea, travel, 메시지, 버섯, 친구
test3	게임, 시스템, 종류, 친구, 힘

표 3. 비교 실험 (10개의 문서를 대상)

test vector	Kohonen		k-nearest neighbor	
	Euclidean distance	Cosine Similarity	Euclidean distance	Cosine Similarity
test1	0.727399	0.276260	1.913935	0.682263
test2	1.086579	0.179919	2.061568	0.414824
test3	0.459657	0.147152	2.070134	0.455182

표 4. 비교 실험 (20개의 문서를 대상)

test vector	Kohonen		k-nearest neighbor	
	Euclidean distance	Cosine Similarity	Euclidean distance	Cosine Similarity
test1	0.848920	0.200820	1.947692	0.658038
test2	0.000883	0.188982	2.008213	0.448537
test3	0.997058	0.197627	1.994017	0.585153

결과를 살펴보면 코호넨 신경망을 통한 유클리드 거리가 k-nearest neighbor 방법보다 더 나은 결과를 보였으나, 코사인 유사도면에서는 k-nearest neighbor가 더 나은 결과를 보였다. 이러한 결과에 대해서는 좀더 분석할 필요가 있다.

4. 결론 및 향후 과제

본 연구에서는 사용자가 관심 있는 키워드 즉 대표 용어와 관련 있는 뉴스그룹을 사용자에게 추천하는 방식으로 유즈넷 뉴스 필터링 시스템을 구현하였다. 뉴스그룹의 문서를 대상으로 퍼지추론을 수행하여 뉴스문서를 대표하는 용어를 추출하였으며, 추출된 단어를 클러스터링하기 적합한 코호넨 신경망으로 학습시켰다.

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 멀어지게 되어 비슷한 뉴스그룹으로 분류할 수 없게 된다. 이러한 편차를 줄이기 위하여 정규화를 하였다. 둘째, 사용자가 직접 키워드를 제시하는 대신에 퍼지추론을 통한 뉴스문서로부터 대표 용어들을 추출하여 보다 의미 있는 필터링 기능을 수행하였다. 셋째, 제안된 방법을 패턴 분류율면에서 성능을 평가하기 위하여, 코호넨 신경망을 이용한 방법과 기존의 k nearest neighbor 방법을 유클리드 거리면에서 비교하여 우수성을 확인하였다.

향후에는 뉴스 문서로부터 보다 의미있는 키워드를 추출하도록 퍼지추론을 적용한 방법을 보완할 필요가 있다. 한 가지 방법으로 사용자들로부터 직접 관심있는 키워드를 입력하도록 하고 이를 기준으로 신경망을 학습시킨 후 테스트해 보는 것도 고려해 볼만하다. 또한 적용된 문제 보다 입력벡터의 특징 공간이 보다 큰 복잡한 문제에 적용시킬 수 있도록 제안된 방법을 확장할 필요도 있다.

참고문헌

[1] 노순억, 김병만, 허남철, "퍼지추론을 이용한 소수 문서의 대표 키워드 추출," 한국퍼지 및 지능시스템학회 논문지, Vol.11, No.9, pp.837-843, 2001.

[2] W. B. Frakes and R. Baeza-Yates, Information Retrieval: Data Structure and Algorithms, Prentice-Hall, 1992.

[3] 김대수, 신경망 이론과 응용, 하이테크 정보, 1992.

[4] 진승훈, 김종완, 이승아, 김영순, 김병만, "코호넨 신경망을 사용한 유즈넷 뉴스 필터링 에이젼트 구현", 산업정보학회논문지, Vol.7, No.5, pp.21-28, 2002.

[5] 류근호, 이제환, 정보저장 및 검색, 시그마프레스, 2000.