유즈넷 뉴스 그룹 결정 방법을 활용한 성능평가

김종완*, 김희재*, 김병익*
대구대학교 컴퓨터·IT공학부*

Performance Analysis by utilizing a Determination Method of Usenet News Groups

Jong-Wan Kim*, Hee-Jae Kim*, Byung-Ik Kim*
School of Computer and Information Technology, Daegu Univ.*

요 약

많은 양의 유즈넷 뉴스 중에서 사용자가 찾고자 하는 정확한 정보를 빠른 시간 안에 검색하고, 원하는 정보만 필터링 하는 것은 중요하다. 그러나 뉴스 문서는 이메일과 달라서 미리 자신에게 맞는 뉴스그룹을 등록해 주어야만 정보를 얻을 수 있다. 본 연구에서는 다양한 뉴스그룹들 중에서 사용자의 취향과 유사한 뉴스그룹들을 코호넨 신경망을 이용하여 추천해주는 방법을 제시한다. 신경망을 학습시키기 위한 뉴스 문서의 키워드들을 선택하기 위해 예제 문서들로부터 후보 용어들을 추출하고 퍼지 추론을 적용하여 대표 용어들을 선택한다. 하지만 신경망의 학습 패턴을 관찰해 보면, 많은 부분이 비어있는 희소성 문제를 발견할 수 있다. 이에 본 연구에서는 통계적인 결정계수를 도입하여 불필요한 차원을 제거한 후 신경망을 학습시키는 새로운 방법을 제안한다. 제안된 방법은 모든 차원을 활용할 때 보다 클러스터내 거리와 클러스터간 거리의 척도를 이용한 클러스터 중첩도 면에서 우수한 분류 성능을 보여줌을 확인하였다.

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. 하지만 많은 뉴스그룹에서 선택된 특정한 키워드부분이 비어있는 희소성 문제가 발생되어 이 문제를 해결하기 위해, 본 연구에서는 사용자가 제시하는 목표변수(즉 유사한 뉴스그룹)와 관련성이 높은 입력변수(여기서는 선택된 대표 용어)를 선정하여, 이를 기준으로 학습시키는 것이 입력변수의 전체 차원을 함께 학습시키는 것보다 유용할 것이라는 판단 하에 통계적인 방법을 도입하였다.

본 논문은 아래와 같이 구성된다. 2장에서는 관련 연구를 서술하고, 3장에서는 제안된 방법에 관해서 설명한다. 4장에서는 여러 가지 실험결과를 제시하며, 5장에서 결론을 맺는다.

2. 관련 연구

2.1 대표 용어 추출

뉴스그룹의 예제 문서들로부터 뉴스그룹을 가장 잘 대변하는 대표 용어의 선택은 중요하다. 문서 집합에서 대표 용어를 추출하고 이들의 가중치를 부여하는 문제는 기존의 대표적인 선형 분류기인 Rocchio와 Widrow-Hoff 알고리즘들[1]이 학습 문서 집합을 대표

하는 중심 벡터를 구성하는 것과 성격이 동일하다. 이들 알고리즘들은 용어의 가중치 산정시 발생 빈도수(TF)와 역문헌 빈도수(IDF)를 결합하는 방법을 취하고 있지만, 문서내 또는 문서 집합내 용어들간의 관련성을 용어의 가중치 계산에 반영하고 있지는 않다. 따라서 TF가 높은 용어는 높은 가중치를 가지게 되는데 대표 용어로서 실제 중요하지 않는 용어임에도 문서내에 자주 발생만 되면 높은 가중치 값을 부여받을 수 있다는 단점을 지니고 있다[2].

이러한 문제를 해결하기 위해, 특정 용어의 중요도 계산에 사용되는 입력 정보(예: TF, IDF)들은 정량적으로 정확히 해석될 수 없는 부정확하고 불확실한 특성을 내포하고 있으므로, 이러한 불확실성의 문제 해결에 효과적인 퍼지추론을 적용하여 후보 용어들의 가중치를 계산하고 이 값들에 따라 선택 우선 순위를 부여하는 방법도 있다[3]. 이 방법은 소수의 긍정적 학습 문서 집합들에 대해서 실험한 결과 비교적 우수한 결과를 보여주었으므로, 본 연구에서도 이 방법을 채택한다.

2.2 학습에서 차원 감축의 효과

학습 데이터의 차원은 여러 방식으로 특징지을 수 있다. 데이터집합은 다양한 패턴을 가지며, 각 패턴은 여러 속성과 클래스 라벨을 갖는다. 그러나 패턴의 차원 즉 속성 중에는 분류에 기여하는 속성도 있지만, 그렇지 않은 속성도 존재한다. 따라서 훈련 집합으로부터 하나의 개념을 유도하는데 걸리는 시간과 새로운 패턴의 클래스를 결정하는 것은 사용된 학습 알고리즘과 함께 제시된 속성의 개수, 즉 사용된 차원의 개수도 중요한 역할을 차지한다.

학습 작업에서 어느 속성이 클래스를 예측하는데 기여하는지 결정하는 것은 기계학습의 중심 문제이다. 과거에는 영역 전문가들이 학습 문제에 기여하는 것으로 예측되는 속성을 선택하였다. 하지만 배경 지식이 부족한 문제에서는 그러한 속성들을 자동으로 식별하는 작업이 요구된다. 대표적인 방법들로 속성들 모두에 걸쳐서 평균 유사성 척도를 계산하는 근사 이웃(nearest neighbor) 알고리즘들이 제안되었다[4]. 하지만 단순한 근사 이웃 알고리즘들은 모든 속성들을 동일한 가중치로 판단하므로 속성들 사이의 패턴 분류 기여도를 적절하게 산정하지 못하였다. 이를 해결하기 위해 여러 가지 가중치를 부여하는 방식도 제안되었다[5]. 이 방법은 일종의 주성분 분석(PCA: Principal Component Analysis) 기법으로서, 낮은 특성값(singular value)을 갖는 차원을 삭제하는 Singular Value Decomposition(SVD) 방법을 채택하고 있다. 주성분 분석 기법은 원 변수들(y_i)의 선형 결합으로 이루어지는 주성분 예측변수들($\hat{y}_i$)를 구해서, 이 변환된 변수들을 패턴분류에 사용하는 것이므로 어떤 입력 성분이 패턴 분류에 기여하는 지 알 수는 없다. 하지만 본 연구에서는 특정 성분이 패턴 분류 학습에 필요한 것인지 결정하는 것이 중요하므로, 이러한 방법보다는 패턴들간의 분류 기여도를 결정해주는데 유용한 통계학의 결정계수를 활용하려고 한다.

3. 제안된 뉴스 그룹 결정 방법

3.1 뉴스 필터링 시스템의 기본 구조 및 대표 용어 선택 방법

본 연구에서 제안하는 뉴스 필터링 시스템의 기본 구조는 그림 1과 같다. 사실 뉴스 필터링 작업은 학습 단계와 테스트 단계로 나뉘어 진다. 먼저, 학습 단계에서는 사용자가 특정한 유즈넷 뉴스서버(NNTP server)를 지정하면, 이 뉴스서버에 접속하여 뉴스 문서들을 내려 받는다. 그리고 각 뉴스그룹에 속한 문서들에 퍼지추론을 적용시켜 대표 용어를 추출하고 결정계수 기법으로 차원을 감축시킨 후 코호넨 신경망으로 학습한다. 테스트 단계에서는 뉴스리더는 사용자 키워드 프로파일을 읽어 이를 코호넨 신경망에 제시하고 그 결과로 사용자의 의도와 가장 유사한 뉴스그룹 목록을 얻고 이를 바탕으로 기존 뉴스 프로토콜에 따라 뉴스 기사들을 내려 받게 된다.

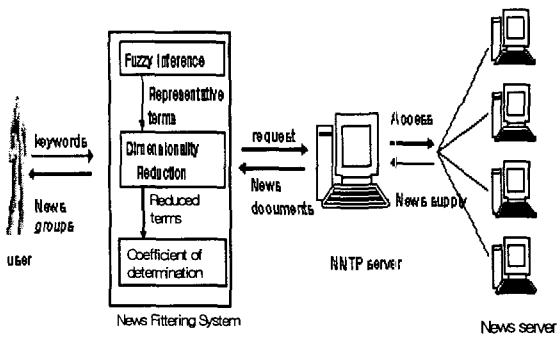


그림 1. 제안된 뉴스 필터링 시스템

유즈넷 뉴스들을 필터링하기 위해서는 사용자의 관심 내용을 가장 잘 대변하는 대표 용어의 선택이 중

요하다. 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[3]을 사용하여 후보 용어들의 가중치를 계산하고 이 값들에 따라 선택 우선 순위를 부여하였다. 그 방법을 설명하면 아래와 같다.

퍼지추론을 이용한 대표 용어 중요도를 계산하기 위해 뉴스들은 불용어를 처리하고, Porter stemmer[6]와 국민대 강승식교수가 만든 한글 형태소분석기[7]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 TFi(예제 문서 집합에서 i번째 단어의 발생 빈도수)를 DFi(예제 문서 집합에서 i번째 단어를 포함하는 문서의 수)로 나누어 계산하며, NDF(Nomalized Document Frequency)는 DFi를 TD(예제 문서의 수)로 나누어 구하며, NIDF(Normalized Inverse Document Frequency)는 IDFi(i번째 단어의 역문헌 빈도수)를 역문헌 빈도수 최대값으로 나누어 계산한다. 자세한 내용은 [3]을 참조한다.

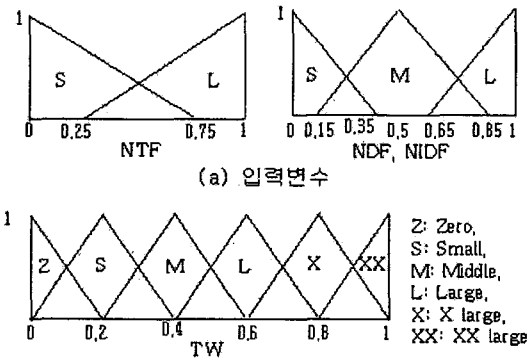


그림 2. 퍼지 입출력변수

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 나 누었다.

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 출력변수 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)[8]으로 비퍼지화(defuzzification)한 값이 해당 용어의 중요도 값으로 결정되어진다.

표 1. 퍼지 추론규칙

NTF = S

NIDF	S	M	L
S	Z	S	M
M	S	L	X
L	S	X	XX

NTF = L

NIDF	S	M	L
S	Z	Z	S
M	Z	M	L
L	S	L	X

3.2 결정계수를 이용한 학습패턴의 차원 축소

학습패턴의 차원 축소에 있어 우리가 사용하고자 하는 통계학의 결정계수(coefficient of determination) R^2는 n개의 입력변수들과 목표변수와의 상관관계로 회귀선에 의하여 설명되는 편차

가 기여하는 비율을 의미하므로, 일반적으로 추정된 회귀모형의 제곱합은 아래의 식 (1)와 같이 분해된다. 즉 회귀선에 의한 총편차(SST: Total Sum of Squares)는 회귀선에 의하여 설명되지 않는 편차(SSE: Sum of Squares due to residual errors)와 회귀선에 의해 설명되는 편차(SSR: Sum of Squares due to regression)로 정의된다[9].

$$\sum(y_i-\bar{y})^2=\sum(y_i-\hat{y}_i)^2+\sum(\hat{y}_i-\bar{y})^2 \quad (1)$$

여기서 y_i는 i번째 개체의 실제값이고, $\hat{y}_i$는 i번째 개체의 예측값이고, $\bar{y}$는 변수 y_i의 평균값이다. 식 (2)의 r^2은 표본결정계수(sample coefficient of determination)의 정의로서, 총편차(*SST*)를 설명하는데 있어서 회귀식에 의하여 설명되는 편차(*SSR*)가 기여하는 비율을 나타낸다.

$$r^2=\frac{SSR}{SST}=1-\frac{SSE}{SST} \quad (2)$$

만약 모든 관찰치들 y_i가 회귀선상에 위치한다면 $y_i=\hat{y}_i$가 되므로, $SSE=0$이며 $r^2=1$이다. 이와 반대로, 입력변수 x와 목표변수 y사이에 회귀관계가 전혀 없어서 추정된 회귀선의 기울기가 0 이면 $\hat{y}_i=\bar{y}$가 성립되며, 이 경우에는 r^2의 값이 0 이 된다. 즉 r^2의 값이 0 에 가까운 값을 가지는 회귀선은 쓸모가 없는 회귀선이므로 회귀분석의 의미가 없으며, r^2의 값이 큰 값을 가질수록 회귀선의 유용성이 높아진다. 따라서 뉴스그룹을 분류하는 대표용어 중에서 목표변수인 뉴스그룹 클래스에 대한 기여도가 낮은 즉 결정계수값이 적은 변수(즉 용어)는 제거하는 것이 패턴의 분류율을 높이는데 기여할 수 있다.

결정계수를 이용해서 뉴스그룹 데이터를 분류하려면 목표변수가 있어야 한다. 따라서 본 연구에서는 목표변수인 뉴스그룹의 클래스는 뉴스그룹 도메인 이름을 기준으로 지정하였다. 예를 들면, NNTP 서버인 news.kornet.net에 있는 126개의 뉴스그룹을 영역기준으로 분류하였다. 즉 han.answers.all을 클래스 1, han.arts.architecture.all을 클래스 2, 나머지도 이런 식으로 분류하였더니 모두 114개의 그룹이 나왔다. 이 클래스 변수를 임시로 목표변수로 보고, 뉴스그룹에 결정계수를 적용하여 계산한 결과 중에서 관련정도가 아주 낮은 속성을 제거한다. 4.2절의 실험 결과에 보면, 126개 뉴스그룹의 경우에는 결정계수 임계치를 기준으로 전체 차원의 약 20% 내지 30% 정도가 제거되는 효과를 얻을 수 있었다.

4. 실험 및 분석

4.1 실험 데이터 수집 및 학습 방법

본 논문에서 제안된 방법은 자바 언어로 구현되었다. 먼저, 훈련 데이터를 수집하려고 자바의 java.net.Socket 클래스를 이용하여 유즈넷 뉴스서버인 news.kornet.net에 접속한 후, NNTP 프로토콜을 통해서 뉴스그룹을 선택하고 각 뉴스그룹에서 뉴스 문서를 내려 받았다.

실험은 126개의 뉴스그룹을 대상으로 하였으며, 퍼지추론으로 대표 용어를 추출하는 경우에 뉴스그룹당 20개의 문서를 임의로 추출한 경우를 실험하였다. 출력뉴런의 크기는 5*5로 정하였으며, 훈련은 1,000회 실시하였다. 훈련 데이터는 각 뉴스그룹에서 퍼지추론으로 추출하고 결정계수를 적용해서 일부 연관도가 낮은 성분을 제거한 용어들을 데이터베이스에 저장해 놓고, 각 뉴스그룹의 문서에서 용어들을 분석한다. 본 논문에서는 126개 뉴스 그룹에서 28개의 단어를 추출하여 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 정규화(normalization)를 수행하였다. 정규화는 (각 단어의 빈도수)/(뉴스그룹에서 각 단어들이 나타난 총 빈도수)으로 계산하여 각 단어들이 뉴스그룹내에서 나타나는 비율로 계산한다. 그림 3은 실험에 사용된 정규화된 입력벡터를 나타낸다.

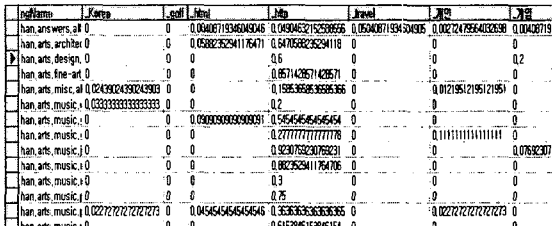


그림3. 126개의 뉴스그룹의 정규화된 입력벡터

4.2 학습 성능평가 및 결정계수 도입 효과분석

학습 성능을 평가하기 위해서 본 논문에서는 코호넨 신경망이 사용자가 의도하는 대로 뉴스그룹을 클러스터링 해주는 지를 확인하였다. 실험을 위해서 사용자가 입력한 키워드를 이용하여 테스트용 입력벡터를 생성한다. 사용자가 입력한 키워드와 미리 입력되어있는 키워드와의 거리를 계산하기 위하여 사용자가 입력하지 않은 키워드의 값을 0으로 하여 입력벡터의 차원을 일치시켰다. 일례로, 사용자(kc)가 입력한 키워드 프로파일이 아래와 같다고 하면, 4개의 성분을 제외한 나머지 요소들은 0으로 채워진다.

User(kc) : html, http, 서버, 시스템

테스트용 입력벡터가 결정되면 코호넨 신경망에 제시하여 가장 가까운 출력뉴런을 선정하고, 이 뉴런에 속하는 뉴스그룹들을 사용자에게 제시한다. 그림 4는 사용자(kc)가 자신의 ID를 입력한 후의 결과 화면으로, 사용자가 입력한 키워드와 미리 학습된 정보를 이용하여 가장 가까운 뉴스그룹을 보여준다. 그림 4에서는 출력뉴런 (4,1)이 승자뉴런으로 선정되었다.

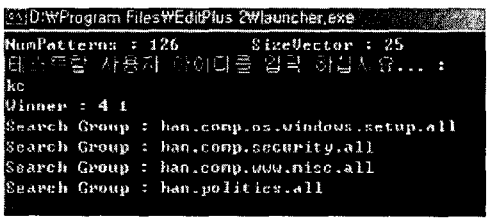


그림4. 사용자(kc)에게 추천한 뉴스 그룹

차원을 축소하기 위한 결정계수의 효과를 살펴보기 위하여, 차원을 강소시켰을 때와 그렇지 않을 때의 클러스터내 거리(Dw) 및 클러스터간 거리(Db)를 아래와 같이 정의하고 계산하였다.

$$Dw_j = \frac{1}{|C_j|}\sum_{i \in C_j}\sqrt{[X_i - W_j]^2} \qquad (3)$$

여기서 Xi는 클러스터 j에 속하는 i번째 학습패턴을, Wj는 j번째 출력뉴런의 연결강도벡터 즉, j번째 클러스터의 중심벡터를 의미하고, Cj는 j번째 클러스터에 속하는 패턴들의 집합을, |Cj|는 j번째 클러스터에 속하는 패턴들의 수를 나타낸다. 따라서 이들 간의 거리인 Dwj는 j번째 클러스터의 중심벡터와 학습패턴간의 거리를 의미하며, 이를 모든 출력뉴런들의 합으로 정의하여 전체 클러스터의 수로 나눈 아래의 식은 클러스터내 거리(Dw)를 나타낸다.

$$Dw = \frac{1}{k}\sum_{j=1}^{k} Dw_j \qquad (4)$$

여기서 k는 출력뉴런들의 수, 즉 클러스터의 개수를 나타낸다.

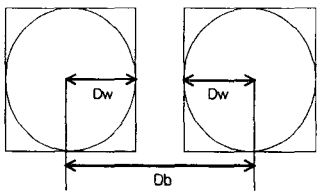
클러스터간 거리 Db는 식 (3)과 같이 각 클러스터별로 j번째 클러스터 자신을 제외한 다른 클러스터들과의 거리(Dbj)를 계산하고 이를 평균한 식 (4)로 정의된다.

$$Db_j = \sum_{m=1, \neq j}^{k}\sqrt{[W_j - W_m]^2} \qquad (5)$$

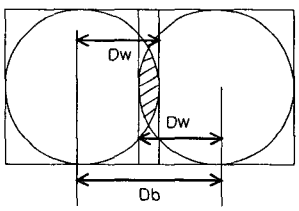
여기서 Wj와 Wm은 각각 j번째 뉴런과 m번째 뉴런의 연결가중치를 나타내므로, 이들 간의 거리는 클러스터들 사이의 거리를 의미한다.

$$D_b = \frac{1}{k}\sum_{j=1}^{k} Db_j \qquad (6)$$

일반적으로 좋은 패턴 분류기는 클러스터내 거리는 줄이면서 클러스터간 거리는 늘리는 것이므로 이를 제안된 차원감축 방법의 성능을 평가하는데 활용한다[10].



case 1: 클러스터 분리 (Db ≥ 2Dw)



case 2: 클러스터 중첩 (Db < 2Dw)

그림 5. 클러스터 중첩도

표 2의 20개 문서를 대상으로 실험한 경우에는, 용어의 수가 28개이며, 0.01과 0.02의 결정계수 임계치를 사용할 때, 각각 32%와 43%의 용어 수를 줄일 수 있었다. 특히, 두 클러스터내 거리합인 2*Dw 보다

클러스터간 거리 Db가 작은 그림 5의 case 2에 해당하므로 클러스터들 간에 중첩이 있음을 알 수 있다. 그래서 클러스터 중첩도 계산이 어려운 원 대신에 정사각형으로 간주하고, 최대한 원의 면적과 유사하도록 중첩 사각형 면적의 절반을 계산하는 식 (7)에 따라 클러스터간의 중첩도를 계산한 결과, 제안된 방법들이 임계치에 상관없이 클러스터간 중첩도가 50% 이상 개선됨을 알 수 있었다. 중첩도 계산 과정을 예로 설명하면, Dw=0.40이고 Db=0.61인 경우에, 중첩도 =(2*0.4 - 0.61)*0.4=0.076 이 된다. 나머지 중첩도는 같은 방식으로 계산한 것이다.

$$\text{중첩도} = (2Dw - Db) * Dw \qquad (7)$$

표 2. 126 뉴스그룹의 20 문서 대상 실험결과

사용된 용어수	기준 방법	임계치 0.01	향상률 (%)	임계치 0.02	향상률 (%)
Dw	0.4	0.35	12.5	0.36	10.0
Db	0.61	0.62	1.6	0.63	3.3
중첩도	0.076	0.028	63.2	0.032	57.4

위의 실험 결과를 통해서 우리는 용어 수를 지나치게 많이 줄이는 것이 반드시 좋지는 않다는 사실과 제거되는 용어를 선정하는 방법이 중요하다는 것을 확인할 수 있었다. 이러한 결과는 기본적으로 제안된 방법이 입력 차원이 보다 많은 문제에 효과적이라는 사실을 입증하여 준다.

5. 결론 및 향후 과제

본 논문에서는 사용자 프로파일에 기반한 뉴스 리더의 주요 부분인 프로파일-뉴스그룹 맵핑 방법을 제안하고 이의 성능을 분석하여 보았다. 뉴스그룹의 문서를 대상으로 퍼지추론을 수행하여 뉴스문서를 대표하는 용어를 추출하였고, 결정계수를 도입하여 패턴 분류 기여도가 낮은 차원을 감축시켰으며, 선정된 용어를 클러스터링하기 적합한 코호넨 신경망으로 학습시켰다.

본 연구에서는 첫째, 퍼지추론을 통한 뉴스문서로부터 대표 용어들을 추출하여 보다 정확도를 높였다. 둘째, 학습에 불필요한 중복된 속성들을 제거하기 위하여 통계학의 결정계수를 활용하여 패턴 분류율을 향상시켰다. 셋째, 제안된 방법을 패턴 분류율 면에서 성능을 평가하기 위하여, 클러스터내 거리 및 클러스터간 거리의 척도 면에서 비교하였다. 특히 클러스터내 거리합이 클러스터간 거리보다 커지는 클러스터 중첩의 정도를 정의하고, 이를 기준으로 제안된 방법의 우수성을 확인하였다.

향후에는 입력벡터의 차원이 보다 큰 복잡한 문제에 적용시켜서 제안된 결정계수를 이용한 차원 감소 효과의 유용성을 확장할 필요가 있다. 또한 클러스터 중첩도외에 새로운 성능평가척도를 개발하여 성능을 비교할 필요도 있다.

참 고 문 헌

[1] David D. Lewis, Robert E. Schapire and James P. Callan and Ron Papka, "Training algorithms for linear text classifiler", Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval, 1996.

[2] 김주연, 김병만, 박혁로, "용어 분포 유사도를 이용한 질의 용어 확장 및 가중치 재산정," 한국정보과학회논문지(B), Vol.27, No.1, pp.90-100, 2000.

[3] Kim, B. M., Li, Q., and Kim, J. W., "Extraction of User Preferences from a Few Positive Documents," Proceedings of The Sixth International Workshop on Information Retrieval with Asian Languages, pp. 124-131, 2003.

[4] D.W. Aha, "Tolerating Noisy, Irrelevant and Novel Attributes in Instance-Based Learning Algorithms," International Journal of Man-Machine Studies, Vol.36 pp.267-287, 1992.

[5] Terry R. Payne and Peter Edwards, "Dimensionality Reduction through Sub-Space Mapping for Nearest Neighbor Algorithms," European Conference on Machine Learning, pp.331-343, 2000.

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

[7] 한국어 형태소 분석기와 한국어 분석 모듈 (HAM: Hangul Analysis Module), http://nlp.kookmin.ac.kr/."

[8] C.C. Lee, "Fuzzy logic in control systems: Fuzzy logic controller-part I," IEEE Trans. Syst. Man, Cybern., Vol.20, No.2, pp.408-418, 1990.

[9] 박성현, "회귀분석", 민영사, 1992

[10] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, 1973.