퍼지 및 지능시스템학회 논문지 2004, Vol. 14, No. 2, pp. 142-149

사용자 프로파일에 기초한 유즈넷 뉴스그룹 자동 결정 방법

Automatic Determination of Usenet News Groups from User Profile

김종완* · 조규철* · 김희재* · 김병만**
Jong-Wan Kim*, Kyu-Cheol Cho*, Hee-Jae Kim*, and Byeong Man Kim**

*대구대학교 컴퓨터 · IT공학부, **금오공과대학교 컴퓨터공학부

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. 하지만, 초보자인 경우는 어떤 뉴스그룹이 자신의 관심사와 관련이 있는지를 판단하기가 용이치 않다. 따라서, 본 연구에서는 다양한 뉴스그룹들 중에서 사용자의 취향과 유사한 뉴스그룹들을 코호넨 신경망을 이용하여 추천해주는 방법을 제공한다. 신경망을 학습시키기 위한 뉴스 문서의 키워드들을 선택하기 위해 예제 문서들로부터 후보 용어들을 추출하고 퍼지 추론을 적용하여 대표 용어들을 선택한다. 하지만 신경망의 학습패턴을 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.

Abstract

It is important to retrieve exact information coinciding with user's need from lots of Usenet news and filter desired information quickly. Differently from email system, we must previously register our interesting news group if we want to get the news information. However, it is not easy for a novice to decide which news group is relevant to his or her interests. In this work, we present a service classifying user preferred news groups among various news groups by the use of Kohonen network. We first extract candidate terms from example documents and then choose a number of representative keywords to be used in Kohonen network from them through fuzzy inference. From the observation of training patterns, we could find the sparsity problem that lots of keywords in training patterns are empty. Thus, a new method to train neural network through reduction of unnecessary dimensions by the statistical coefficient of determination is proposed in this paper. Experimental results show that the proposed method is superior to the method using every dimension in terms of cluster overlap defined by using within cluster distance and between cluster distance.

Key words : Usenet news filtering, fuzzy inference, Kohonen network, statistical coefficient of determination, dimensionality reduction.

1. 서 론

유즈넷(Usenet) 뉴스는 ARPA 인터넷에 속한 사용자들 간에 새로운 정보의 분배, 조회 및 송수신할 수 있는 환경을 제공하여 학술연구, 취미, 오락 활동에 도움을 주는 세계적인 정보 교환 시스템이다. 이메일(email)이 특정 송수신자 사이에 정보를 교환하는 것에 비하여 유즈넷은 세계 각국의 누구나 그 정보에 접하여 자신의 의견을 제시하거나 도움을 구할 수 있다. 유즈넷을 이용하는 사람의 숫자는 헤아릴 수 없이 많으며, 국제적인 학술정보망에 가입된 기관의 컴퓨터 사용자는 대부분 이를 사용하는 것으로 보아도 될 정도이다.

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. 하지만, 초보자인 경우는 어떤 뉴스그룹이 자신의 관심사와 관련이 있는지를 판단하기가 용이치 않다. 뉴스그룹 대신에 사용자의 프로파일, 즉 사용자의 관심사를 나타내는 주요 단어들을 사용할 수만 있다면 이러한 문제를 조금이나마 완화시킬 수 있을 것이다. 따라서 본 논문에서는 이러한 작업의 일환으로 사용자의 프로파일로부터 유즈넷 뉴스그룹을 자동으로 결정하는 방법을 제시하고 이의 성능을 평가하고자 한다.

사용자 프로파일과 부합되는 뉴스그룹을 결정하기 위해서는 각 뉴스그룹을 대표하는 용어들을 추출하고 이들을 사용

접수일자 : 2003년 9월 15일
완료일자 : 2004년 3월 24일
이 논문은 2003학년도 대구대학교 학술연구비 지원에 의한 논문임.

자 프로파일과 비교하여 유사성이 높은 뉴스그룹들을 선택하는 일들이 필요하다. 뉴스그룹을 대표하는 용어들을 추출하기 위해, 인터넷에 접속된 뉴스서버들에 접속해서 뉴스를 수집하였다. 그리고 수집된 뉴스들로부터 후보 용어들을 추출하고 퍼지추론을 적용하여 대표 용어들을 선택하였다. 제안 방법의 성능은 대표 용어들을 선택하는 방법에 의해 영향을 크게 받으며, 또한 뉴스그룹에서 대표 용어를 추출하는 문제 자체가 불확실성을 내포하고 있다. 따라서 이러한 문제 해결에 효과적인 퍼지추론을 이용한 대표 용어의 선택 방법[1]을 선택하였다. 사용자 프로파일과 각 그룹의 대표 용어들과의 유사성을 판단하기 위한 방법으로는 정보검색 분야에서 널리 쓰이는 코사인 유사도 방법[2], 신경망을 이용한 방법 및 기타 학습 기법을 이용한 방법들을 쓸 수 있다. 어떤 방법을 사용하느냐에 따라 성능에 차이가 있을 수 있다. 하지만, 본 논문에서 성능 자체에 초점을 두기보다는 제안 방법의 유용성에 더 관심을 두고 있기 때문에 본 저자가 다루기 쉽고 관심을 갖고 있었던 신경망을 이용한 방법을 선택하였다.

코호넨 신경망(Kohonen network)은 교사의 지시 없이 뉴스그룹 문서들로부터 추출된 키워드 즉 대표 용어만 가지고 자연스럽게 뉴스그룹간의 연관 관계를 찾을 수 있다는 장점이 있다. 이에 본 연구에서는 코호넨 신경망을 학습 알고리즘으로 채택하였다. 하지만, 신경망의 훈련벡터로 사용되는 패턴을 관찰해 보면, 많은 뉴스그룹에서 선택된 특정한 키워드부분이 비어있는 희소성 문제를 발견할 수 있다. 이러한 희소성 문제를 해결하기 위해, 본 논문에서는 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

본 저자들이 알고 있는 바로는 본 연구처럼 사용자 프로파일을 이용하여 유즈넷 뉴스그룹을 결정하는 연구는 없다. 단, 정보필터링 분야의 유즈넷 뉴스 필터링이 본 연구와 어느 정도 관련성이 있다. 일반적으로 프로파일에 기초한 필터링 시스템에서는 사용자의 관심사가 수동 혹은 자동으로 획득되어 프로파일(profile)로 작성된다. 다음으로, 수집된 웹 문서나 뉴스 기사, 도착한 전자 우편 메시지들은 시스템의 속도와 효율성 등을 고려하여 설계된 문서 표현 방법에 의해 지정된 형태로 저장된다. 마지막으로, 이미 작성된 사용자의 프로파일과 저장된 문서들을 비교 함수를 이용하여 각 문서가 사용자의 정보 요구를 만족할 정도를 산정하며, 이러한 만족 정도를 이용하여 사용자에게 제공할 문서를 결정하게 된다.

사용자가 직접 입력한 단어들을 이용하여 필터링하는 방법으로 대표적인 시스템으로는 SIFT등이 있다[3]. SIFT 시스템은 Yan에 의해 인터넷 뉴스 기사들을 필터링하기 위하여 제작되었으며, 프로파일에 포함될 단어들을 사용자가 선호, 비 선호로 구분하여 표현함으로서 수동으로 사용자 프로파일을 생성되도록 하였다. 또한 사용자가 프로파일을 쉽게 구성할 수 있도록 하기 위하여 상위에 랭크(rank)된 기사들에서 중요 단어들을 추출하여 후보 프로파일로 사용자에게 제공하며, 기사를 읽는 도중에 발견되는 단어들을 선호나 비선호의 리스트에 추가할 수 있도록 하였다. 그러나 이 시스템에서는 프로파일을 작성하기 위하여 사용자에게 매번 추가될 단어들을 선택하게 함으로써 너무 많은 부담을 지우고 있다. 또한 사용자의 부담을 완화하기 위하여 후보 프로파일을 제공하고 있지만 후보 프로파일에 있는 단어들은 기사에서 제목 부분에 나타나는 단어들만을 제공함으로서 사용자가 선택할 수 있는 단어들이 제한될 수밖에 없다.

사용자로부터 정보 요구를 단어로서 명백하게 입력받지 않고 사용자의 행동을 관찰하여 반응을 나타낸 문서들을 이용하여 정보 요구를 추출하는 방법으로 InfoScope[4]가 있다. InfoScope 시스템은 인터넷 뉴스를 필터링하기 위하여 Stevens에 의해 개발된 시스템이다. 이 시스템에서는 사용자가 관심을 가지는 메시지의 간접 증거로서 기사의 읽음, 기사의 저장, 기사의 반송 등과 같은 행동을 관찰하여 규칙을 추론하고, 이를 검증하기 위하여 사용자에게 추론된 규칙을 제공하는 방법을 사용하였다. Morita와 Shinode는 InfoScope 시스템에서 사용된 방법을 개선하여 사용자가 기사를 읽었다는 증거로서 읽는데 소비한 시간을 이용하여 실험을 수행하였다[5]. 이 실험에서 6명의 사용자들을 대상으로 6주 동안 관찰한 결과, 20초 이상 읽은 기사를 사용자의 관심 문서들로 선택할 경우 사용자가 직접 관심 문서를 지정한 경우보다 성능이 더 우수하다는 사실을 발견하였다.

협력기반 필터링은 특정 사용자와 관심 분야가 동일한 집단이 존재하고, 그림, 표 등과 같이 텍스트 형태로 표현할 수 없는 정보들을 필터링 할 경우에 유용한 방법이다. 협력기반 방법의 특징은 각 사용자들이 자신이 살펴본 문서들에 대하여 평가치를 부여한다는 것이며, 이렇게 부여된 평가치는 다른 사용자와의 정보 요구에 관한 유사성과 문서의 점수를 예측하는데 사용된다. 사용자간의 유사성은 기존의 문서에 대한 평가에 의해 계산되며 이러한 사용자간 유사성을 바탕으로 필터링될 문서의 점수를 예측하게 된다. 협력 필터링을 이용한 대표적인 시스템은 Tapestry[6], GroupLens[7] 등이 있다.

Tapestry 시스템은 협력 필터링에 기초한 최초의 시스템으로, 개인의 전자 메일, 인터넷 뉴스 기사들을 필터링하기 위하여 개발되었다. 이 시스템에서는 문서에 대한 평가치를 선호, 비선호와 같이 이진(binary) 값으로 부여하였으며, 문서가 선택될 수 있는 조건과, 사용자로부터 조건들에 부여된 값으로 구성된 프로파일들을 이용하여 문서에 대한 점수를 예측하였다. 그러나 이 시스템은 선택된 문서들이 사용자가 부여한 조건 값에 의해 너무 많은 영향을 받고 있으며, 사용자가 부여하는 조건 값은 동일한 조건일 경우에도 개인에 따라서 많은 차이를 나타낸다는 문제점 있다. GroupLens는 인터넷 뉴스를 필터링하기 위하여 Minnesota대학에서 개발한 시스템으로 Tapestry시스템과는 달리 문서에 대한 평가치를 부여하기 위하여 5등급의 정수를 사용하였다.

지금까지 인터넷 뉴스 필터링을 중심으로 관련연구들을 소개하였다. 하지만, 이러한 방법들과 본 제안방법을 직접적으로 비교하기가 곤란하다. 왜냐하면 본 논문에서 제안한 방법은 필터링 방법이라기보다는 새로운 형태의 뉴스 사용자 인터페이스로 볼 수 있기 때문이다. 또한, 제안 방법은 기존의 뉴스 필터링 방법에 적용되어 비교 대상의 문서를 줄이는데 사용될 수도 있다. 즉, 모든 기사에 대해서 처리할 필요없이 프로파일과 관련된 뉴스그룹들에 속하는 기사들에 대해서만 처리를 하면 되기 때문에 속도 측면에서 잇점을 가질 수 있다.

3. 제안된 뉴스그룹 자동 결정 방법

본 논문에서 제안하는 내용은 뉴스그룹 대신에 사용자 프로파일을 이용하는 뉴스리더(news reader)의 핵심부분, 즉, 사용자 프로파일과 관련이 있는 뉴스그룹을 자동으로 결정하는 방법에 관한 것이다. 본 방법이 실제 뉴스 시스템에 적용될 경우의 기본 구조는 그림 1과 같다. 새로운 형태의 뉴스리더는 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 뉴스리더가 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.

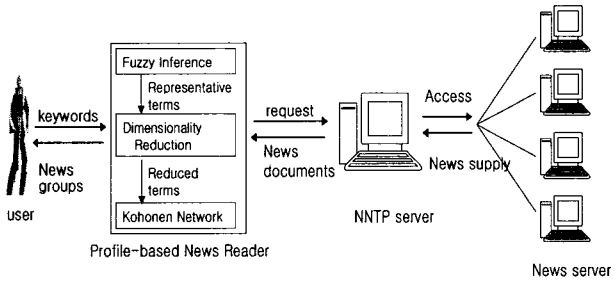


그림 1. 제안된 뉴스 시스템 구조
Fig. 1. Architecture of the proposed news system

3.1 대표 용어 선택 방법

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[8]이 학습 문서 집합을 대표하는 중심 벡터를 구성하는 것과 성격이 동일하다. 이들 알고리즘들은 용어의 가중치 산정 시 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 문서 내에 자주 발생만 되면 높은 가중치 값을 부여받을 수 있다는 단점을 지니고 있다[9].

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[1]. 이 방법은 소수의 긍정적 학습 문서 집합들에 대해서 실험한 결과 비교적 우수한 결과를 보여주었으므로, 본 연구에서도 이 방법을 채택하였다. 그 방법을 설명하면 아래와 같다.

퍼지추론을 이용한 대표 용어 중요도를 계산하기 위해 뉴스 문서들은 불용어 처리 과정을 거치고, Porter stemmer를 사용하여 영문 단어를 추출하고 대표적인 한글 형태소 분석기[10]를 사용하여 한글 명사를 추출하여 이를 후보 용어들의 집합으로 변형하며, 이 집합으로부터 각각의 용어들의 TF(Term Frequency), DF(Document Frequency), IDF(Inverse Document Frequency) 정보가 구해진다. 이들 정보들이 퍼지추론을 위한 퍼지시스템의 입력으로 이용되는데, 자세히 설명하면 아래와 같다.

TF(Term Frequency)

각 용어의 발생 빈도수는 퍼지 계산에 사용되어지기 위해 정규화(NTF) 되어야 하며 아래의 식 (1)을 사용하였다.

$$NTF_i = \frac{\frac{TF_i}{DF_i}}{\underset{j}{Max}\left[\frac{TF_j}{DF_j}\right]} \quad (1)$$

TF_i : 예제 문서 집합에서 i번째 단어의 발생 빈도수
DF_i : 예제 문서 집합에서 i번째 단어를 포함하는 문서의 수

DF(Document Frequency)

각 용어의 예제 문서 집합 내에서의 문서 발생 빈도수를 나타내며 TF와 마찬가지로 아래의 식 (2)를 사용하여 정규화(NDF) 하였다.

$$NDF_i = \frac{\frac{DF_i}{TD}}{\underset{j}{Max}\left[\frac{DF_j}{TD}\right]} \quad (2)$$

TD : 예제 문서의 수
DF_i : 예제 문서 집합에서 i번째 단어를 포함하는 문서의 수

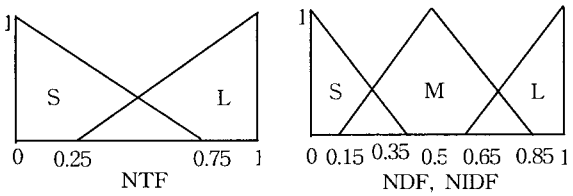
IDF(Inverse Document Frequency)

각 용어의 전체 예제 문서 집합 내에서의 역문헌 빈도수를 나타내며 아래의 식 (3)을 사용하여 정규화(NIDF) 하였다.

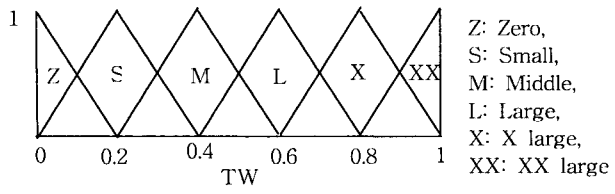
$$NIDF_i = \frac{IDF_i}{\underset{j}{Max}\left[IDF_j\right]} \quad (3)$$

IDF_i : i번째 단어의 역문헌 빈도수

그림 2는 퍼지 추론을 위하여 사용된 입출력 변수들의 멤버쉽 함수를 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 멤버쉽 함수 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

Fig. 2. Fuzzy input/output variables

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

표 1. 퍼지 추론규칙

Table 1. Fuzzy inference rules

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

NTF = S

NIDF / NDF	S	M	L
S	Z	Z	S
M	Z	M	L
L	S	L	X

NTF = L

3.2 결정계수를 이용한 학습패턴의 차원 축소

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[12]. 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[13]. 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 사용하였다.

학습패턴의 차원 축소에 있어 우리가 사용하고자 하는 통계학의 결정계수(coefficient of determination) R^2는 각 입력변수들과 목표변수간의 상관 관계를 나타내는 값으로 0에서 1사이에 있으며, 입력변수 x와 목표변수 y사이에 높은 상관관계가 있을수록 1에 가까워진다[14]. 목표변수(y)의 변화는 하나의 입력변수만으로 충분히 설명되는 경우는 거의 없다. 따라서 적절한 입력변수를 여러 개 잘 선택하여 이들의 함수로 목표변수를 설명한다면 보다 정확한 예측을 할 수 있을 것이다. 결정계수는 n개의 입력변수들과 목표변수와의 상관관계로 회귀선에 의하여 설명되는 편차[1]가 기여하는 비율을 의미하므로, 일반적으로 추정된 회귀모형의 제곱합은 아래의 식 (4)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[15].

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 \qquad (4)$$

여기서 y_i는 i번째 개체의 실제값이고, $\hat{y}_i$는 i번째 개체의 예측값이고, $\bar{y}$는 변수 y_i의 평균값이다.

식 (5)의 r^2은 표본결정계수(sample coefficient of determination)의 정의로서, 총편차(SST)를 설명하는데 있어서 회귀식에 의하여 설명되는 편차(SSR)가 기여하는 비율을 나타낸다.

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (5)$$

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 0이면 $\hat{y}_i = \bar{y}$가 성립되며, 이 경우에는 r^2의 값이 0이 된다. 즉 r^2의 값이 0에 가까운 값을 가지는 회귀선은 쓸모가 없는 회귀선이므로 회귀분석의 의미가 없으며, r^2의 값이 큰 값을 가질수록 회귀선의 유용성이 높아진다. 따라서 뉴스그룹을 분류하는 대표 용어 중에서 목표변수인 뉴스그룹 클래스에 대한 기여도가 낮은 즉 결정계수값이 적은 변수(즉 용어)는 제거하는 것이 패턴의 분류율을 높이는데 기여할 수 있다.

이와 같이 회귀분석을 이용하여 목표변수에 영향을 미치는 입력변수를 찾을 수 있는데, 가능한 모든 후보변수들을 입력변수로 사용하여 예측모형을 만드는 경우에는, 데이터를 계속 수집하고 관리하는데 많은 노력과 비용이 필요할 뿐만 아니라 일부 회귀계수 추정치의 분산과 예측값의 분산이 커지게 되어 이를 신뢰할 수 없게 된다. 따라서 불필요한 변수들이 들어 있는 완전모형보다는 필요한 변수들만 있는 축소모형이 보다 바람직한 회귀모형이라 할 수 있다[14]. 이러한

1) 편차는 관측값이 평균값에 비하여 흩어진 정도를 의미한다.

축소모형을 구축하기 위해 본 연구에서는 모든 후보 입력변수인 퍼지추론으로 구해진 용어들을 대상으로 뉴스그룹을 식별하는 입력변수들을 선택하기 위하여 결정계수를 계산한다.

먼저 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 114개의 클래스 변수를 목표변수로, 목표변수들과 관련성이 있는 후보용어들을 입력변수로 두고, 모든 후보용어들에 대하여 목표 클래스 변수와의 결정계수를 계산한다. 계산된 결정계수값중에서 미리 지정한 임계치 이하인 가장 낮은 결정계수값을 갖는 후보용어를 하나씩 제거하는 후진소거(backward elimination) 변수선택법을 수행한다[14]. 남아있는 모든 후보용어들에 대하여, 계산되는 결정계수값이 모두 지정한 임계치를 초과할 때까지 후진소거 변수선택법을 반복하여 수행한다.

4.2절의 실험 결과에 보면, 126개 뉴스그룹의 경우에는 결정계수의 임계치를 기준으로 임계치 이하인 후보용어들을 제거하는 방법을 사용하여 전체 용어의 약 20 내지 40% 정도가 제거되는 효과를 얻을 수 있었다.

3.3 코호넨 신경망을 이용한 뉴스그룹 분류

본 논문에서는 뉴스그룹을 학습하여 키워드 중심으로 분류하기 위하여 신경망 중 목표값 없이 학습 데이터만을 단순히 신경망의 입력으로 사용하여, 신경망이 스스로 연결가중치들을 학습시키는 비지도 학습 신경망의 일종인 코호넨 신경망을 사용하였다. 코호넨 신경망을 사용한 이유는 교사의 지시 없이 뉴스그룹 문서들로부터 자연스럽게 연관 관계를 분류할 수 있기 때문이다. 대표적인 비지도 학습 신경망인 코호넨 신경망의 학습방법은 먼저 각 뉴런이 연결강도벡터와 입력벡터의 거리가 얼마나 가까운가를 계산한다. 그리고 각 뉴런들은 학습할 수 있는 특권을 차지하려고 서로 경쟁하게 되는데 거리가 가장 가까운 뉴런이 승리하게 된다. 이 승자뉴런의 연결강도벡터는 입력벡터에 가장 가까운 것으로 이 뉴런만이 출력신호를 보낼 수 있는 유일한 승자뉴런이 되고, 이 뉴런과 인접한 이웃 뉴런들만이 제시된 입력벡터에 대한 학습이 허용된다.

학습 알고리즘은 아래와 같은 일반적인 코호넨 학습규칙을 따랐다[16]. 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 연결강도를 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 정수이다. α는 0과 1사이의 값을 가지는 이득항인데, 학습회수인 epoch가 증가함에 따라 점차 작아진다. 본 연구에서 α의 초기값 α_0는 0.9를 사용하였다.

[단계 6] 단계 2로 가서 반복한다.

4. 실험 및 분석

4.1 실험 데이터 수집 및 학습 방법

본 논문에서 제안된 방법은 사용자 인터페이스로 Swing을 사용하여 자바언어로 구현되었다[17]. 먼저, 훈련 데이터를 수집하려고 자바의 java.net.Socket 클래스를 이용하여 유즈넷 뉴스서버인 news.kornet.net에 접속한 후, NNTP 프로토콜을 통해서 뉴스그룹을 선택하고 각 뉴스그룹에서 뉴스문서를 내려 받았다. 이때 이미 삭제되었거나 옮겨진 뉴스그룹과 10개 이하의 문서를 가지고 있는 뉴스그룹은 제외시켰다.

실험은 126개의 뉴스그룹을 대상으로 하였으며, 퍼지추론으로 대표 용어를 추출하는 경우에 뉴스그룹당 10개의 문서를 임의로 선택하는 경우와 20개의 문서를 임의로 추출한 경우를 실험하였다. 실험을 두 가지 경우로 수행한 이유는 추출된 용어의 개수가 차원 축소 실험에 얼마나 영향을 미치는지 확인하기 위한 것이다. 출력뉴런의 크기는 5*5로 정하였으며, 훈련은 각각 1000회 실시하였다. 훈련 데이터는 각 뉴스그룹에서 퍼지추론으로 추출하고 결정계수를 적용해서 일부 연관도가 낮은 성분을 제거한 용어들을 데이터베이스에 저장해 놓고, 각 뉴스그룹의 문서에서 용어들을 분석한다. 본 논문에서 추출된 대표 용어는 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 정규화(normalization)를 수행한다[17]. 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.

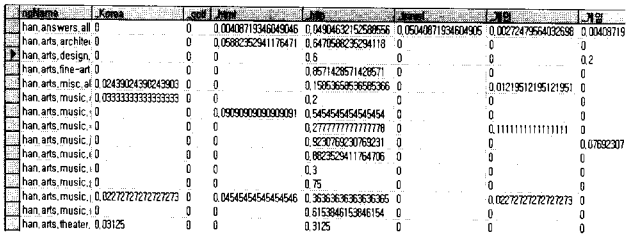


그림 3. 126개의 뉴스그룹의 정규화된 입력벡터
Fig. 3. Normalized input vectors for 126 news groups

학습이 끝난 후 각 뉴스그룹의 코호넨 신경망의 출력층 위치와 연결강도 벡터를 그림 4와 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. 뉴스그룹의 일부 출력층 학습 결과
Fig. 4. Some trained neurons of news groups

4.2 학습 성능평가 및 결정계수 도입 효과분석

학습 성능을 평가하기 위해서 본 논문에서는 코호넨 신경망이 사용자가 의도하는 대로 뉴스그룹을 클러스터링 해주는지를 확인하였다. 실험을 위해서 사용자가 입력한 키워드를 이용하여 테스트용 입력벡터를 생성한다. 사용자가 입력한 키워드와 미리 입력되어 있는 키워드와의 거리를 계산하기 위하여 사용자가 입력하지 않은 키워드의 값을 0으로 하여 입력벡터의 차원을 일치시켰다. 사용자가 입력한 키워드의 가중치는 학습할 때 사용한 입력벡터에서 실제로 해당 키워드가 나타난 뉴스그룹을 대상으로 평균한 값을 사용하였다. 표 2는 사용자가 입력한 키워드 프로파일을 보여준다. 예를 들면, 표 2의 키워드 프로파일 중 'html'의 경우를 그림 3에 보이는 입력벡터를 대상으로 계산해 보면, 값이 0인 부분을 제외한 나머지 값들인 0.0040, 0.0588, 0.0909, 0.0454 등을 평균한 값이 'html' 키워드의 가중치로 결정된다.

표 2. 테스트에 사용된 사용자 정보 사례
Table 2. User information used for testing

userid	passwd	name	keywords
kc	****	홍길동	html, http, 서버, 시스템

테스트용 입력벡터가 결정되면 코호넨 신경망에 제시하여 가장 가까운 출력뉴런을 선정하고, 이 뉴런에 속하는 뉴스그룹들을 사용자에게 제시한다. 그림 5는 사용자(kc)가 자신의 ID를 입력한 후의 결과 화면으로, 사용자가 입력한 키워드와 미리 학습된 정보를 이용하여 가장 가까운 뉴스그룹을 보여준다. 그림 5에서는 출력뉴런 (4,1)이 승자뉴런으로 선정되었다.

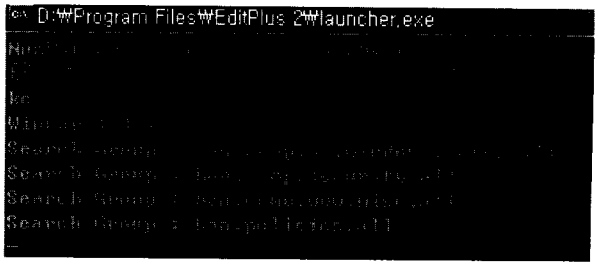


그림 5. 사용자(kc)에게 추천한 뉴스 그룹
Fig. 5. Recommended news groups for user(kc)

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

$$Dw_j = \frac{1}{|C_j|} \sum_{i \in C_j} \sqrt{[X_i - W_j]^2} \qquad (10)$$

여기서 X_i는 클러스터 j에 속하는 i번째 학습패턴을, W_j는 j번째 출력뉴런의 연결강도벡터 즉, j번째 클러스터의 중심벡터를 의미하고, C_j는 j번째 클러스터에 속하는 패턴들의 집합을, $|C_j|$는 j번째 클러스터에 속하는 패턴들의 수를 나타낸다. 따라서 이들간의 거리인 Dw_j는 j번째 클러스터의 중심벡터와 학습패턴간의 거리를 의미하며, 이를 모든 출력뉴런들의 합으로 정의하여 전체 클러스터의 수로 나눈 아래의 식은 클러스터내 거리(Dw)를 나타낸다.

$$Dw = \frac{1}{k} \sum_{j=1}^{k} Dw_j \qquad (11)$$

여기서 k는 출력뉴런들의 수, 즉 클러스터의 개수를 나타낸다.

클러스터간 거리 Db는 식 (12)와 같이 각 클러스터별로 j번째 클러스터 자신을 제외한 다른 클러스터들과의 거리(Db_j)를 계산하고 이를 평균한 식 (13)으로 정의된다.

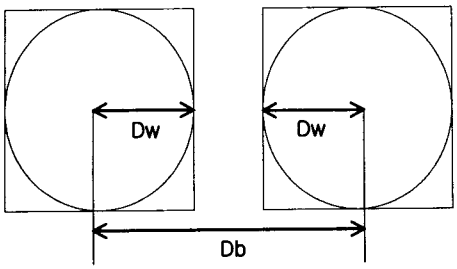
$$Db_j = \sum_{m=1, m \neq j}^{k} \sqrt{[W_j - W_m]^2} \qquad (12)$$

여기서 W_j와 W_m은 각각 j번째 뉴런과 m번째 뉴런의 연결가중치를 나타내므로, 이들간의 거리는 클러스터들 사이의 거리를 의미한다.

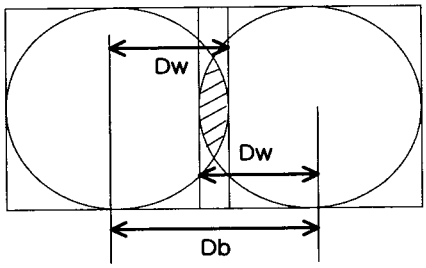
$$Db = \frac{1}{k} \sum_{j=1}^{k} Db_j \qquad (13)$$

일반적으로 좋은 패턴 분류기는 클러스터내 거리는 줄이면서 클러스터간 거리는 늘리는 것이므로 이를 제안된 차원 감축 방법의 성능을 평가하는데 활용한다[18]. 표 3에는 126개 뉴스그룹의 10개 문서를 대상으로 25개의 용어를 추출하여 사용한 기준 방법과 결정계수의 임계치를 0.01과 0.02로 변경하여 20개 및 16개의 용어를 사용하여 20%와 36%씩 용어 수를 줄인 제안된 방법을 비교하였다. 실험 결과를 분석해 보면, 임계치를 높일수록 클러스터내 거리와 함께 클러스터간 거리도 줄어서 제안된 방법의 유용성이 없는 것처럼 보인다. 하지만 제안된 방법과 기준 방법에 대하여 클러스터내 거리(Dw)와 클러스터간 거리(Db)를 함께 고려해서 이들의 중첩도(overlap)를 묘사한 아래의 그림 6을 살펴보면, 10개의 문서를 대상으로 한 실험에서 제안된 방법의 임계치와 상관

없이 Dw를 반경으로 하는 원이나 정사각형으로 클러스터를 묘사할 때 두 클러스터내 거리합인 $2*Dw$ 보다 클러스터간 거리 Db가 훨씬 큰 case 1에 해당하므로, 제안된 방법의 클러스터들이 중첩되지 않으면서도 Dw가 작아서 보다 밀집된 (compact) 클러스터를 형성해준다고 말할 수 있다.



case 1: 클러스터 분리($2Dw \leq Db$)



case 2: 클러스터 중첩($2Dw > Db$)

그림 6. 클러스터 중첩도

Fig. 6. Cluster Overlap

표 4의 20개 문서를 대상으로 실험한 경우에는, 예상한 바와 같이 용어의 수가 3개 늘어났으며 0.01과 0.02의 결정계수 임계치를 사용할 때, 각각 32%와 43%의 용어 수를 줄일 수 있었다. 특히 표 3의 경우와 다르게 표 4의 경우에는, 두 클러스터내 거리합인 $2*Dw$ 보다 클러스터간 거리 Db가 작은 그림 6의 case 2에 해당하므로 클러스터들간에 중첩이 있음을 알 수 있다. 그래서 클러스터 중첩도 계산이 어려운 원 대신에 정사각형으로 간주하고 최대한 원의 면적과 유사하도록 중첩 사각형 면적의 절반을 계산하는 식 (14)에 따라 클러스터간의 중첩도를 계산한 결과, 제안된 방법들이 임계치에 상관없이 클러스터간 중첩도가 50% 이상 개선됨을 알 수 있었다. 중첩도 계산 과정을 예로 설명하면, Dw=0.4이고 Db=0.61인 경우에, 중첩도=(2*0.4 - 0.61)*0.4=0.076 이 된다. 나머지 중첩도는 같은 방식으로 계산한 것이다.

$$\text{중첩도} = (2Dw - Db) * Dw \tag{14}$$

표 3과 표 4의 향상률 계산 과정은 기준 방법과 제안된 방법의 차이를 기준 방법과 비교한 백분율로 계산한 것이다. 예를 들어, 표 3의 기준 방법의 Dw=0.1에서 임계치=0.01을 사용한 제안된 방법의 향상률=(0.1-0.08)/0.1*100=20% 가 된다. 나머지도 같은 방식으로 계산하여 얻은 결과이다.

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. 하지만 여전히 용어 수 감소에 따른 신경망 학습 속도 개선 효과는 유효하기 때문에 제안된 방법은 패턴 분류율 향상과 신경망의 학습 속도를 개선시키는 일석이조의 효과는 기대할 수 있다.

표 3. 126개 뉴스그룹의 10개 문서를 대상으로 실험

Table 3. Experiments with 10 documents per news group

사용된 용어 수	기준방법 (25개 용어)	임계치=0.01 (20개 용어)	향상률 (%)	임계치=0.02 (16개 용어)	향상률 (%)
Dw	0.1	0.08	20.0	0.07	30.0
Db	0.49	0.36	-26.5	0.35	-28.6
중첩도	0	0	0	0	0

표 4. 126개 뉴스그룹의 20개 문서를 대상으로 실험

Table 4. Experiments with 20 documents per news group

사용된 용어 수	기준방법 (28개 용어)	임계치=0.01 (19개 용어)	향상률 (%)	임계치=0.02 (16개 용어)	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	0.0324	57.4

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 퍼지 추론을 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, 각 뉴스그룹 내에 있는 문서들의 개수에 따른 편차를 줄이기 위하여 정규화를 수행하였다. 셋째, 학습에 불필요한 중복된 속성들을 제거하기 위하여 통계학의 결정계수를 활용하여 패턴 분류율을 향상시켰다. 넷째, 제안된 방법을 패턴 분류율면에서 성능을 평가하기 위하여, 클러스터내 거리 및 클러스터간 거리의 척도 면에서 비교하였다. 특히 클러스터내 거리합이 클러스터간 거리 보다 커지는 클러스터 중첩의 정도를 정의하고, 이를 기준으로 제안된 방법의 우수성을 확인하였다. 마지막으로 불필요한 성분 제거에 따른 신경망의 학습 속도 개선 효과를 얻을 수 있었다.

향후에는 학습된 뉴스그룹의 문서가 사용자가 원하는 뉴스그룹인지를 피드백 받아서 유용성 여부를 판별해야 한다. 또한 입력벡터의 차원이 보다 큰 복잡한 문제에 적용시켜서 제안된 결정계수를 이용한 차원 감소 효과의 유용성을 확장할 필요도 있다.

참고문헌

[1] Byeong Man Kim, Ju Youn Kim and Jongwan Kim, "Query Term Expansion and Reweighting using Term Co-Occurrence Similarity and Fuzzy Inference,"

Proc. of IFSA/NAFIPS, pp.715-720, 2001.
[2] G. Salton and M. McGill, Introduction to Modern Information Retrieval, New York, McGraw Hill, 1983.
[3] Tak W.Yan and Hector Garcia-Molina, "Distributed selective dissemination of information," Proceedings of the Third International Conference on Parallel and Distributed Information Systems, pp.89-98, IEEE Computer Society, September 1994.
[4] Curt Stevens, "Automating the creation of information filters," Communications of the ACM, Vol.35, No.12, pp.48, 1992.
[5] Masahiro Morita and Toichi Shinoda, "Information filtering based on user behavior analysis and best match text retrieval," Proceedings of the Seventeenth Annual International ACM-SIGIR Conference, pp.272-281, Springer-Verlag, July 1994.
[6] Douglas B. Terry, "A tour through tapestry," In Proceedings of the ACM Conference on Organizational Computing Systems(COOCS), pp.21-30, November 1993.
[7] Paul Resnick, Neophytos Iacovou, etc., "GroupLens: An open architecture for collaborative filtering of netnews," Proceedings of the Conference on Computer Supported Cooperative Work, pp.175-186, ACM, October 1994.
[8] 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.
[9] 김주연, 김병만, 박혁로, "용어 분포 유사도를 이용한 질의 용어 확장 및 가중치 재산정," 한국정보과학회 논문지(B), Vol.27, No.1, pp.90-100, 2000.
[10] 한국어 형태소 분석기와 한국어 분석 모듈 (HAM: Hangul Analysis Module), http://nlp.kookmin.ac.kr/."
[11] 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.
[12] 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.
[13] 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.
[14] 강현철, 한상태, 최종후, 김은석, 김미경, SAS Enterprise Miner 4.0을 이용한 데이터마이닝-방법론 및 활용, 자유아카데미, 2001.
[15] 박성현, "회귀분석", 민영사, 1992.
[16] 김대수, 신경망 이론과 응용, 하이테크 정보, 1992.
[17] 진승훈, 김종완, 이승아, 김영순, 김병만, "코호넨 신경망을 사용한 유즈넷 뉴스 필터링 에이전트 구현", 한국산업정보학회논문지, Vol.7, No.5, pp.21-28, 2002.
[18] R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, 1973.

저 자 소 개

김종완 (Jong Wan Kim)
1987년 : 서울대학교 컴퓨터공학과 (공학사)
1989년 : 서울대학교 컴퓨터공학과 (공학석사)
1994년 : 서울대학교 컴퓨터공학과 졸업 (공학박사)
1995년-현재 : 대구대학교 컴퓨터IT공학부 부교수
1999년-2000년 : 미국 U. of Massachusetts 방문교수

관심분야 : 지능형 에이전트, 퍼지시스템, 인공지능, 정보검색
Phone : 053-850-6575
Fax : 053-850-6589
E-mail : jwkim@daegu.ac.kr

조규철 (Kyu Cheol Cho)
2002년 : 대구대학교 멀티미디어공학과 (학사)
2002년-현재 : 대구대학교 컴퓨터정보공학과 석사과정
관심분야 : 인공지능, 퍼지시스템
Phone : 053-850-4419
E-mail : kccho97@webmail.daegu.ac.kr

김희재 (Hee Jae Kim)
1992년 : 대구가톨릭대학교 통계학과 (학사)
1994년 : 대구가톨릭대학교 전산통계학과 (이학석사)
2002년~현재 : 대구한의대학교 멀티미디어학부 초빙교원
2003년~현재 : 대구대학교 컴퓨터정보공학과 박사과정

관심분야 : 퍼지시스템, 인공지능, 데이타마이닝, 멀티미디어
Phone : 053-819-1466
E-mail : kimhj@daegu.ac.kr

김병만 (Byeong Man Kim)
1987년 : 서울대학교 컴퓨터공학과(학사)
1989년 : 한국과학기술원 전산학과(석사)
1992년 : 한국과학기술원 전산학과(박사)
1992~현재 : 금오공과대학교 컴퓨터공학부 교수

관심분야 : 인공지능, 정보검색, 프로그램 테스팅 및 검증
Phone : 054-467-4277
Fax : 054-467-4473
E-mail : bmkim@se.kumoh.ac.kr