논문 2008-45CI-3-5

적응적인 학습을 위한 텍스트 마이닝 기술

(Text Mining Techniques for Adaptable Learning)

김 천 식*, 정 명 희*, 홍 유 식**

(Cheonshik Kim, Myunghee Jung, and You-Sik Hong)

요 약

지금까지 이러닝 시스템을 통해서 학습 능력을 향상시키는 기술이 많이 나와 있다. 대부분의 이러닝 시스템에서 학습자들은 강의 자료와 학습문제를 통해서 학습을 한다. 그러나, 때로는 학습자간의 자료공유나 토론을 통해서 학습능력과 학습 의욕을 향상시킬 수 있다. 이 경우에 일반적으로 게시판을 통해서 학습 자료를 공유하거나 MSN과 같은 메신저를 사용하여 학습자들끼리 토론 및 자료를 공유한다. 하지만, 이와 같은 형태의 학습 공유 유형은 학습 자료가 주제별로 분류되어 있지 않기 때문에 학습자가 관련 자료를 검색하는 일이 쉽지 않다. 그 결과 학습에 크게 도움이 되지 않는다. 대부분의 텍스트 마이닝 기술은 문서데이터의 집합으로부터 요약 데이터를 추출하거나 유사한 문서의 집합을 분류하는 기술이다. 따라서, 본 논문에서 학습자가 학습능력을 향상시킬 수 있도록 이러닝 시스템에 텍스트 마이닝 기술을 적용하여 효과적으로 이러닝 자료를 분류하여 학습자에게 도움이 되는 시스템을 구현하고 평가하였다.

Abstract

Until now, there are many technologies to improve studying ability using e-learning system. In most of e-learning system, learners are studying through the lecture materials and studying problems. The studying ability and intention, however, can be improved through the shared materials and discussion. In this case, learning materials are shared by the learners' discussion and shared materials through the board Internet and MSN. Such data was not classified by learners; it was not easy for the learners to search related valuable information. Therefore, it was not helping to learning. The technologies of most text mining extract summary data from the collection of document or classify into similar document from the complex document. In this paper, we implemented e-learning system for learners to improve learning abilities and especially, applied text mining technology to classify learning material for helping learners.

Keywords : e-Learning, Text Mining, Clustering, Classification.

Ⅰ. 서 론

첨단 정보통신(IT) 기술을 교육에 활용하는 이러닝은 지난 10여 년간 추진해 온 정보통신기술(ICT) 활용 교육과 최근의 엠(m·mobile)-러닝, 유(u·ubiquitous)-러닝을 포괄하는 용어로서 정보통신 기술과 교육을 접합하는 새로운 교육 형태이다. 교육의 차세대 패러다임으로 등장하고 있는 이러닝은 동시성·다양성·개별성·효율성 등의 특징을 가지고 있으며, 교실에 한정돼 있던 교육 장소를 사회 모든 곳으로 확장시켜 학습자 중심의 교육을 가능하게 한다[1~4]. 특히 이러닝은 빠르게 변화하는 사회에서 필요한 의사소통 능력, 상호 협력 방법, 문제해결 능력, 고차원적 사고 능력 등을 습득하는 데 주효한 교육혁신 수단이라 할 수 있다. 프리챌 포털은 최근에 '마이티(myT:Make Your Tomorrow)'를 오픈했다. 이러닝 네트워크 포털을 지향하는 마이티는 동영상 강의 뿐 아니라 강사들의 홈페이지를 블로그 형태로 구성해 수강생들과 쌍방향 학습 교류가 쉽도록 배려했다. 개별 평가 시스템을 도입, 학생의 수준 진단 및 평가와 이를 기초로 한 문제풀이 및 학습 전략 제시 등을 통해 효율적인 학습 관리가 가능하다. 나아가 업그레이드 작

* 정회원, 안양대학교 디지털미디어공학전공
(Major in Digital Media Engineering, Anyang University)
** 정회원, 상지대학교 컴퓨터공학과
(Dept. of Computer Science, Sangji University)
※ 이 논문은 2007년 정부(교육인적자원부)의 재원으로 한국 학술 진흥 재단의 지원을 받아 수행된 연구임 (KRF-2007-D00306 - I00563)
접수일자: 2008년4월26일, 수정완료일: 2008년5월6일

업을 통해 회원 개인 역시 자신의 강의실, 블로그형 개인 계정 등을 통해 강사들과 1대1 맞춤형 학습 커뮤니케이션이 기능하도록 하고 있다. 이러닝 시스템에서는 일반적으로 온라인 강의 자료와 학습 평가 시스템으로 구성되어 있다. 온라인에서 학습자는 정기적으로 강의 자료를 청취하고 학습문제를 평가받고 평가의 결과에 피드백 등의 방법으로 학습 능력을 향상시킬 수 있다. 이와 같은 형태는 대부분의 이러닝 시스템에서 공통적으로 볼 수 있다. 그러나 학습자가 강의 자료나 강의 평가만으로 부족한 점이 있을 수 있다. 수많은 의문과 이와 관련한 답들이 궁금할 수 있다. 이는 마치 블로그나 네이버의 지식인을 결합한 것과 같다. 그러나 특별이 이러닝 시스템에서는 과목별 관심사나 의문점을 해결하는 지식정보를 축적할 수 있다면 누구든지 이와 관련한 의문점을 해소 할 수 있을 것이다. 즉, 스스로 자기가 모르를 정보를 지식이 저장된 저장소에서 찾아서 스스로 학습할 수 있는 적응적인 학습을 하도록 하는 것이 본 논문의 목표이다.

그러나 이를 위해서 Q&A 게시판을 만들어서 학습자에게 이용하도록 하였지만 학습자들이 학습에 대한 만족감을 느끼지 못했다. 그와 같은 이유는 학습자들이 학습정보를 교환하려고 할 때 대부분 중복되는 질문이 있을 수 있다. 이 경우 중복데이터가 만들어지는데 만일 같은 유형의 질문을 한곳에 모을 수 있다면 학습자가 자기가 관심을 갖고 있던 정보를 쉽게 찾을 수 있기 때문에 보다 효율적일 것이다. 이를 해결하기 위해서 우리는 텍스트 마이닝 기술을 활용하여 관련 있는 데이터를 한곳에 모아서 학습자가 쉽게 정보를 활용할 수 있도록 할 것이다.

따라서 본 논문에서는 지식정보를 체계적으로 관리할 수 있는 시스템을 구현하고자 한다. 이 시스템은 지식을 생성하고 지식을 분류하고 지식을 검색하는 일련의 과정을 보다 편리하고 정확하게 관리하는 시스템을 구현하고자 한다.

Ⅱ. 이러닝을 위한 데이터마이닝

1. 텍스트 마이닝

텍스트 마이닝은 정보마이닝 기술의 한 부분이다. 그리고 정보마이닝 기술은 지식경영(KM)의 한 부분이다. 이 경우 지식은 집합적인 전문성, 경험, 노하우 그리고 조직의 지혜를 말한다. 지식은 단순한 데이터 및 정보 그 이상이다. 이것은 문맥, 의사결정 프로세스에 도움을 주는 사실을 포함한다. 비즈니스 세계에서 지식은 전통적인 데이터베이스에서 발견되는 구조화된 데이터뿐만 아니라 워드 문서, 메모 그리고 편지, 전자우편, 뉴스, 웹 페이지 등과 같은 다양한 구조화 되지 않은 형태로 나타날 수 있다[5~6].

지식 관리를 가능하게 하는 주요기술로서의 텍스트 마이닝은 정보들 사이의 관계를 밝힌다는 점에서 데이터마이닝과 유사하다. 그러나 다음과 같은 점에서 차이점을 갖는다. 실제 데이터마이닝은 통계적인 애플리케이션이고, 이전에 확인되지 않은 연결성이나 상호 연관성을 밝히기 위한 기계적인 학습 알고리즘이다. 현재까지 데이터마이닝은 고객의 행동을 해석하고, 예측모델을 세우고자 하는 조직에게 귀중한 직관을 제공한다. 그러나 데이터마이닝은 대부분 구조화된 수치적 데이터를 가지고 작업하는데, 이 데이터는 데이터웨어하우스 또는 데이터 마트 중심부에 저장되어 있다.

데이터마이닝과는 달리 텍스트 마이닝은 텍스트 문서와 같이 구조화 되지 않은 데이터를 대상으로 작업한다. 특히 온라인 텍스트 마이닝은 인터넷상의 구조화되지 않은 데이터를 탐색하고 그것으로부터 어떤 의미를 유도하는 프로세스를 의미한다. 텍스트 마이닝은 데이터 파일에 통계적인 모델을 적용하는 것 이상의 개념이다. 사실 텍스트 마이닝은 텍스트의 집합에서의 관계를 밝히고 이러한 관계를 찾아내는 지식 작업자의 창조성을 이용하여 새로운 지식을 발견하는 것이다. 많은 텍스트 마이닝 알고리즘은 지식작업자의 머리에서 존재하는 아이디어와 논리를 보완함으로써 새로운 지식 발견을 지원한다.

2. 텍스트 분석

텍스트 분석의 역사는 인터넷 탐색보다 도 더 오래되었다. 실제 과학자들은 수십 년 동안 컴퓨터가 자연어를 이해하도록 하기위해 노력해왔다. 텍스트 분석은 그러한 노력들의 집합인 것이다. 많은 연구가들이 자연언어 문서들로부터 의미를 유도하고 이들이 나타내는 문제를 탐구해 왔지만, 텍스트 마이닝에 대한 많은 기준들과 기술적인 접근 방법들이 아직 존재하지 않기 때문에 논쟁의 여지가 없는 실정이다. 자동화된 텍스트 분석을 이용하여 적용될 수 있는 분야는 다음과 같다.

■우편관리 : 텍스트 분석이 많이 사용되는 곳이 바로 메시지 라우팅(routing)인 데, 컴퓨터는 누가 그것을 취급해야 하는지 결정하기 위해 메시지를 읽는다. 텍스트 마이닝의 다른 응용은 메시지의 성질을 통계적으로 분

석하는 것이다.

■문서 관리 : 텍스트 마이닝은 저장고에 있는 수천 만 개의 문서를 취급하는 것을 돕는다. 이는 문서를 저장고에 넣을 때 문서의 의미와 관련한 다른 문서들을 찾아내는 것으로 언제라도 관련 문서의 위치를 알 수 있는 상세한 인덱스를 만들 수 있다.

■자동화된 도움 데스크 : 어떤 회사들은 고객의 질의에 대답하기 위해 텍스트 마이닝을 사용한다. 고객의 편지들과 전자 우편들은 텍스트 마이닝 애플리케이션에 의해 처리된다. 애플리케이션은 고객이 무엇을 원하는지 알 수 있다면, 그들에게 적당한 정보를 자동적으로 보낸다.

2.1 문서 분류 기술

Yang (1999)는 SVM (Support Vector Machine), 신경망 (NN : Neural Networks), k-NN (Nearest Neighbors), LLSF (Linear Least-squares Fit), 베이지언(Bayesian) 분류를 포함하는 분류 기술을 평가하였다[7].

(1) 결정트리(Decision Tree)

결정트리는 정보이론에 기초한 연역적 유도 학습 방법으로 가장 많이 사용되어지는 것 중 하나로써 1949년 Shannon과 Weaver에 의해 처음으로 소개 되었다. 이 분류기법을 기반으로 한 영국의 Timberlake사에서 만든 CART(Classification And Regression Trees)시스템은 결정트리 알고리즘을 이용해서 만든 범주 데이터 및 연속 데이터의 분류시스템의 한 예이다.

결정트리는 기계학습 분야에서 널리 사용되는 규칙-표현 방법으로 객체를 분류하는 규칙들이 트리의 형태로 나타난다. 모든 예제에 대해서 그 예제가 속하는 범주가 결정되고 속성을 의미하는 각각의 노드는 속성 값에 따라서 서로 다른 링크로 분리되므로 트리의 뿌리에서부터 단말노드까지의 경로는 하나의 규칙으로 변환 가능하다. 특히 문서 범주화에서 결정트리를 이용하는 경우에는 특징이 노드로 사용되며 특징의 값이 가지로 사용된다.

(2) 베이지안 분류(Bayesian Classification)

Carnegie Mellon 대학에서 개발한 Rainbow 문서 자동 분류시스템은 나이브 베이지안 이론을 이용하여 개발 되었다. 나이브 베이지안 이론은 학습문서에 나타난 어휘들이 특정범주의 문서에 나타날 확률을 계산하여 새로운 문서의 범주를 예측하는 방법으로 문장에 속해 있는 용어들과 범주와의 결합 확률값(joint probability)을 사용하며, 특징들 사이의 독립을 가정하여 베이지안 확률로 입력문서에 대한 범주의 확률을 계산하는 방법이다.

(3) SVM (Support Vector Machine)

SVM는 이원 패턴 인식문제를 해결하기 위해 제안된 분류기법으로 두개의 범주를 구성하는 데이터들을 가장 잘 분리해 낼 수 있는 결정면을 찾아내는 모형이다. 특히 통계적 학습이론(statistical learning theory)에 기반을 둔 SVM은 기존의 통계적 이론에서 이용되는 경험적 리스크 최소화 원칙(ERM : Empirical Risk Minimization)이 아닌 구조적 리스크 최소화 원칙(SRM : Structural Risk Minimization) 을 이용하여 일반화 오류를 줄이기 때문에 패턴 인식 등에서 우수한 성능을 보여주고 있다.

(4) 최근접 이웃 (k-NN : Nearest Neighbors)

k-NN은 40년 동안 패턴 인식(Pattern Recognition)에서 연구가 되어온 잘 알려진 통계적 접근 방법으로써 문서 범주화 연구가 시작된 초기부터 적용되어 왔으며 가장 좋은 성능을 내는 분류기법 중에 하나이다. k-NN은 학습문서 중에서 새로 입력된 문서와의 유사도가 가장 높은 k개의 문서를 추출하여 각 문서에 미리 할당된 범주를 기반으로 입력문서의 범주를 결정하는 방법으로 학습과정에서는 학습문서들에서 특징을 추출하여 특징 벡터로 표현하는 작업만을 수행한다.

2.2 데이터 마이닝 도구

1) 마이닝 도구 : IMT

Intelligent Miner for Text(IMT)는 소프트웨어 개발 도구이다. IMT를 가지고 개발한 애플리케이션은 편지나 웹페이지 그리고 온라인 뉴스 서비스 같은 텍스 소스로부터 정보를 얻을 수 있다. IMT는 텍스트로부터 패턴을 뽑아내주거나 주어진 화제로부터 문서를 조직화하고, 주어진 주제에 일치하는 서류를 조사할 수 있는 능력을 제공한다[8~9]. IMT는 문서 분석도구들과 향상된 검색 엔진을 가지며 결과를 나타내는데 있어서 기능성과 능력을 향상시켰다. 웹 도구들은 텍스트 도출 능력을 향상시킬 수 있는 모든 요소들을 제공한다. IMT는 또한 사용자의 필요에 의해서 변경되어질 수 있는 몇몇 애플이케이션을 가진다. IMT 군집 분석은 그룹 내에서 문서의 조합을 나누는데 있어 전적으로 자동화 되어진

과정이다. 각각의 그룹 내에서 문서들은 서로 비슷한 면을 갖는다. 문서의 목차가 군집분석의 기초로 사용되어질 때 다른 그룹들은 그 수집된 것에서 토론되어진 다른 주제들이나 테마에 일치한다. 그러므로 군집분석은 수집된 것이 포함하고 있는 것을 알아내는 한 방법이다.

2) 마이닝 도구 : Clementine

SPSS의 클레멘타인(Clementine)은 데이터마이닝의 모든 과정을 지원하는 소프트웨어 솔루션으로 인공신경망 기법, 의사결정 나무분석, 로지스틱 회귀분석, 요인분석 및 2단계 군집분석 등 방대하고 폭 넓은 모델링 알고리즘을 제공한다. 비즈니스 전문지식을 데이터에 접목시켜 예측 모델을 빠르게 개발하고 이것을 비즈니스 현장에 적용시켜 의사결정의 효율성을 향상시키는데 많은 도움을 준다. 또 효율적인 마이닝 프로세스를 위하여 CRISP-DM(Cross Industry Standard Process for Data Mining)이라는 세계적인 마이닝 방법론을 전폭적으로 적용하고 있다.

Ⅲ. 이러닝 데이터를 위한 마이닝 시스템 설계

3.1 전체 시스템 구조

(그림 1)은 이러닝 시스템에서 생성되는 다량의 데이터 중에서 학습에 도움이 되는 질문 및 답변과 관련한 데이터를 텍스트 마이닝 기법으로 분류하기 위한 구조이다. 이 구조에서는 우리가 이러닝 시스템에서 흔히 볼 수 있는 게시판 형태의 데이터가 내용별로 분류가 되어 있지 않은 경우가 많으므로 이를 자동화 하여 학습 향상에 도움을 주고자 한다. (그림 1)을 단계별로 설명하면 다음과 같다.

■질문 항목(Question Item) : 학습자가 교수자의 강의 내용에 대한 질문을 올리면 교수자 또는 또 다른 학습자가 이 질문에 대해서 답변을 달아서 많은 학습자가 지식을 공유할 수 있도록 하는 것이 목적이다. 이와 같은 질문과 답변은 적절히 구룹화되어 있지 않기 때문에 학습자는 검색을 통해서 연관된 내용을 쉽게 검색할 수 있도록 하는 것이 목적이다.

■전처리(Pre-Processing) : 마이닝을 위해서 토큰을 분리하고, 불용어를 제거하고, stemming 처리를 수행한다.

■모델 만들기 (Model Building) : 문서를 분석하기 전에 분석하기 좋도록 하기 위해서 각 문서에 대해서

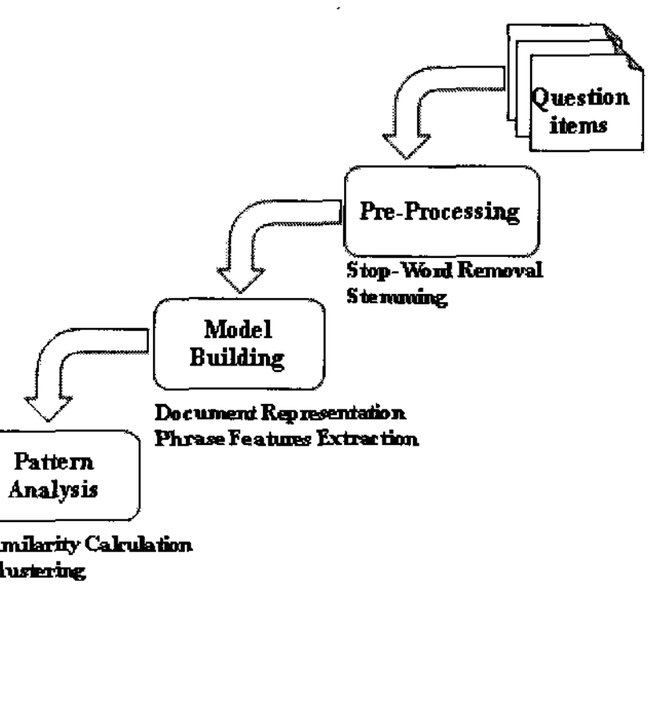


그림 1. 텍스트 마이닝 구조
Fig. 1. Structure of text mining.

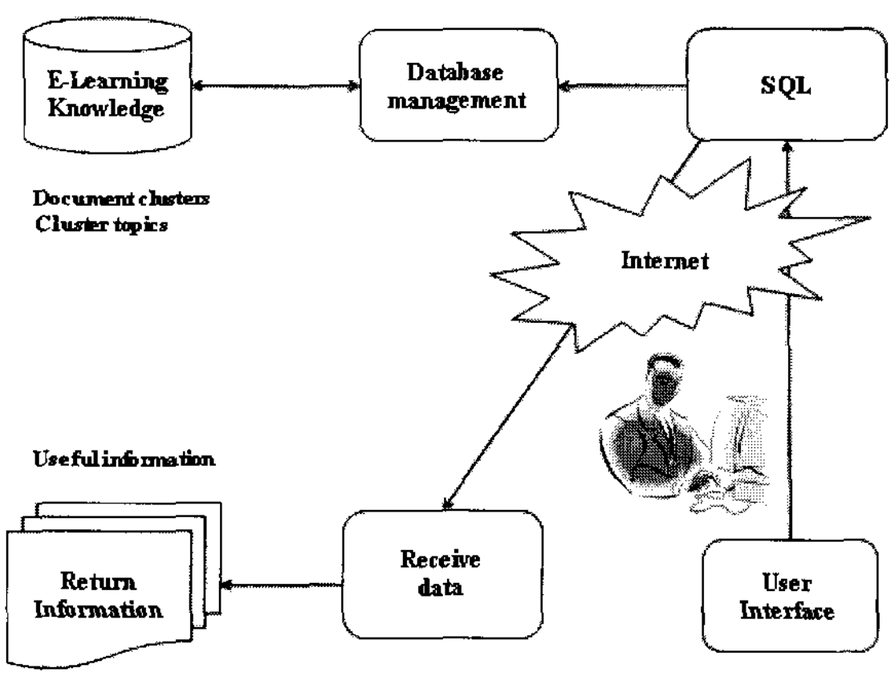


그림 2. 정보검색 과정
Fig. 2. Information retrieval process.

그래프 형태의 구문 구조로 나타낸다.

■패턴 분석(Pattern Analysis) : 문서간의 유사도를 계산하고, 가까운 문서에 대해서 클러스터링 과정을 수행하면서 문서 그룹화 처리를 수행한다.

■이러닝 지식(E-Learning Knowledge) : 텍스트 마이닝의 최종 결과물로서 많은 텍스트 데이터 중에서 유용한 텍스트를 추출한 결과 지식이다.

(그림 2)는 정보지식(Q&A)를 마이닝 알고리즘으로 분류한 다음 주제별로 분류된 정보를 활용하여 학습자에게 필요한 정보를 검색하는 과정을 그림으로 나타낸 것이다.

3.2 마이닝 알고리즘

질문의 항목과 내용은 순수한 텍스트 데이터로서 데이터가 효율적으로 분류되고 분석되지 못하여 이러닝의 지식으로 사용하기 적합하지 못하다. 이를 해결하기 위해서 본 절에서는 마이닝 알고리즘을 사용하여 이를 해

결하고자 제안한 알고리즘으로 텍스트 데이터를 분류하여 학습에 도움이 되도록 할 것이다.

1) 벡터 공간 모델(Vector Space Model)

벡터 공간모델은 정보 필터링, 문서내의 정보검색, 색인과 유사도를 계산하기 위한 수학 모델로, 다차원 선형공간에서 벡터(vector)정보를 이용해서 자연어를 포함한 문서의 중요도를 분석하기 위한 방법을 제시한다.

문서는 색인단어의 벡터로 나타낼 수 있고, 문서의 유사도는 벡터에 위치하는 단어들 간의 거리로 계산해 낼 수 있다[10]. 문서 x와 질의 y사이의 벡터 유사도 측정은 두 벡터 x와 y사이의 상관도로 구할 수 있으며, 이는 두 벡터간 사이의 각의 코사인 값으로 계산 될 수 있다. 이는 수식 (1)과 같이 나타낼 수 있으며, 이를 코사인 유사도(cosine coefficient similarity)라고 한다. $sim(x,y)$ 값은 0과 1사이의 값이 된다. 따라서, 벡터 공간 모델은 문서가 질의와 관련 여부만을 예측하기보다는 질의와의 유사도 값에 따라 순위를 매기기 때문에, 일정한 유사도 값 이상의 문서를 검색 하기위하여 $sim(x,y)$ 값에 임계값을 둘 수도 있다[12].

$$sim(x,y) = \frac{\sum_{i=1}^{K} x_i y_i}{\sqrt{\sum_{i=1}^{K} x_i^2 \cdot \sum_{i=1}^{K} y_i^2}} \quad (1)$$

한 개의 문서벡터에서 단어(Term)의 가중치는 다양한 방법으로 결정될 수 있다. 공통적인 방법은 $tf \times idf$ 방법을 사용하는 것이다. 이것은 단어의 가중치가 두 가지 요인에 의해서 결정된다. 문서 i에 단어 j가 얼마나 자주 나타나는 가 (단어발생 빈도 $tf_{i,j}$) 와 전체 문서에서 단어 j가 얼마나 자주 나타나는 가(문서 발생빈도 df_j)로 결정 된다. 정확하게, 문서 i에서 단어 j의 가중치는 수식 (2) 와 같다.

$$w_{i,j} = tf_{i,j} \times idf_j = tf_{i,j} \times \log N/df_j \quad (2)$$

수식 (2)에서 N은 문서집합에서 문서의 개수이고 idf는 역 문서 발생을 나타낸 것이다. 이 수식은 문서 집합에서 단어의 개수가 적은 것에 높은 가중치를 부여하는 방법이다. 일단 단어의 가중치가 결정되면, 질의와 문서 벡터사이의 유사성을 측정 하는 순위 함수가 필요하다. 수식(1)은 이의 유사도를 계산하는 수식이다. 정확하게 문서 D_j 와 질의 Q사이의 유사도는 다음 수식 (3)과 같이 정의 한다.

$$sim(Q,D) = \frac{\sum_{j=1}^{V} w_{Q,j} \times w_{i,j}}{\sqrt{\sum_{j=1}^{V} w_{Q,j}^2 \times \sum_{j=1}^{V} w_{i,j}^2}} \quad (3)$$

수식 (3)에서 $w_{Q,j}$는 질의에서 단어 j의 가중치이고 $w_{i,j}$은 $tf_{Q,j} \times idf_j$로 정의 한다. 수식에서 분모는 문서의 점수에서 문서의 길이의 영향을 제거하는 표준화 요소이다.

2) $k-NN$ 알고리즘

문서간의 유사성은 문서간의 거리를 의미한다. 이와 같이 거리를 계산한 후 새로운 문서가 들어오면 이 문서와 가장 가까운 문서가 어떤 특성이 있는지를 판단하고 이 특성에 맞는 문서가 어떤 특성이 있는지를 판단하고 이 특성에 맞는 문서로 분류하도록 하기 위한 알고리즘이 최근접 이웃 클러스터링 (Nearest Neighbor Clustering) 알고리즘이다. 이 알고리즘을 이용하여 본 논문에서는 이러닝의 텍스트 데이터를 문서의 유사성을 판단하여 문서를 분류하고자 한다. 일반적으로 문서분류 문제에서는 다음과 같은 $k-NN$ 방법을 이용하여 문서 분류기를 구현한다[11].

$$f(c_i,x) = \sum_{d \in kNN} sum(x,d_j) \cdot f(c_i,d_j) - b_i \quad (4)$$

수식(4)에서 x는 입력 문서, d는 학습 문서집합에 속하는 임의의 문서, 함수 $f(c_i,x) \rightarrow \{0,1\}$는 문서 x가 분류 c_i에 속하는지의 여부를 나타내는 멤버쉽 함수

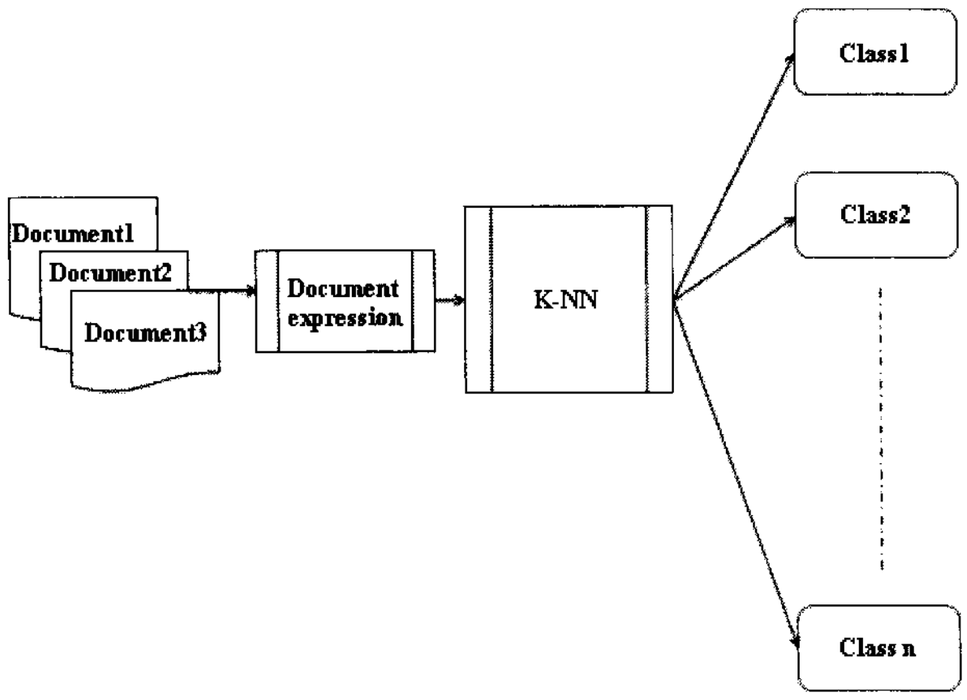


그림 3. K-NN을 이용한 문서 분류 절차
Fig. 3. Document categorization using $k-NN$.

이고, 함수 sim은 두 문서간의 유사도를 나타낸다.

(그림 3)는 수식 (4)알고리즘을 도식화 하여 나타낸 것이다.

Ⅳ. 응용 시스템 구현 및 평가

본 논문에서는 C프로그래밍 언어의 수업을 청취하는 학생에게 온라인 강의를 개설하였다. 온라인 강의를 위해서 교수자의 강의내용을 동영상으로 제공하고 각 단원에 대해서는 단원별 평가를 할 수 있도록 문제를 제공하였다. (그림 4)는 학습자가 1학기 동안 C언어 수업을 온라인으로 듣고 평가를 받을 수 있도록 만든 실험용 이러닝 웹사이트이다. 또한, 이 웹사이트에서 Q&A의 목적은 교수자와 학습자간의 정보공유 및 학습자간의 정보 공유를 목적으로 한다. Q&A 게시판을 통해서 1350건의 Q&A 자료를 수집하였다. 수집된 학습 자료는 주제별로 분류되어 있지 않기 때문에 학습자가 학습하는데 참고용 자료로서 도움이 되지 못하고 있다. 이러한 문제점을 해결하기 위해서 우리는 Q&A 데이터를 텍스트 마이닝 기술로 주제별로 분류하였다. (표 1)은 시험 결과를 나타낸 것이다.

분류의 결과 정확도(precision), 재현율(recall)은 (표 1)과 같다. (그림 5)은 학습자가 학습에 도움을 줄 수 있는 Q&A를 보인 것이다. 재현율과 정확도은 정보검색(IR)에서 중요한 성능 측정 기준으로 사용하는 지표이다. 정확도는 검색 결과 중에 실제로 '관계되는' 문서가 몇 개인가를 의미한다. 즉, 결과의 '정확도'를 의미한다. 재현율은 검색어와 관계되는 문서 전체 중에 몇 개를 찾아내느냐이다. 정확도는 보통 상위 몇 위까지 중 관계되는 문서가 몇 개인가 형태로 평가한다. 웹에는 많은 문서가 있다. 우리가 검색을 할 때, 어떤 특정한 키워드를 준다. 그렇다면 전체 웹에는 그 키워드와 관련 된 문서(=A라고 한다)와 관련 없는 문서가 있을 것이다. 특정 단어의 키워드로 검색을 하면, 검색 엔진은 키워드와 관계된 문서를 찾아 줄 것이다. 검색 엔진이 찾아온 문서들의 집합을 B라고 하자. 이 B에는 일반적으로 관련이 있는 문서와 관련 없는 문서가 섞여 있을 것이다. 그러면 정확도와 재현율은 다음과 같이 정의가 된다.

$$\text{정확도}(precision) = |A \cap B| / |B|$$
$$\text{재현율}(recall) = |A \cap B| / |A|$$

즉, 웹 검색의 경우 재현율이 극도로 높고 (대개 검색결과가 수만 페이지 이상이기 때문에) 사용자들은 상위 몇 개만을 보려하기 때문에 정확도가 매우 높아야만

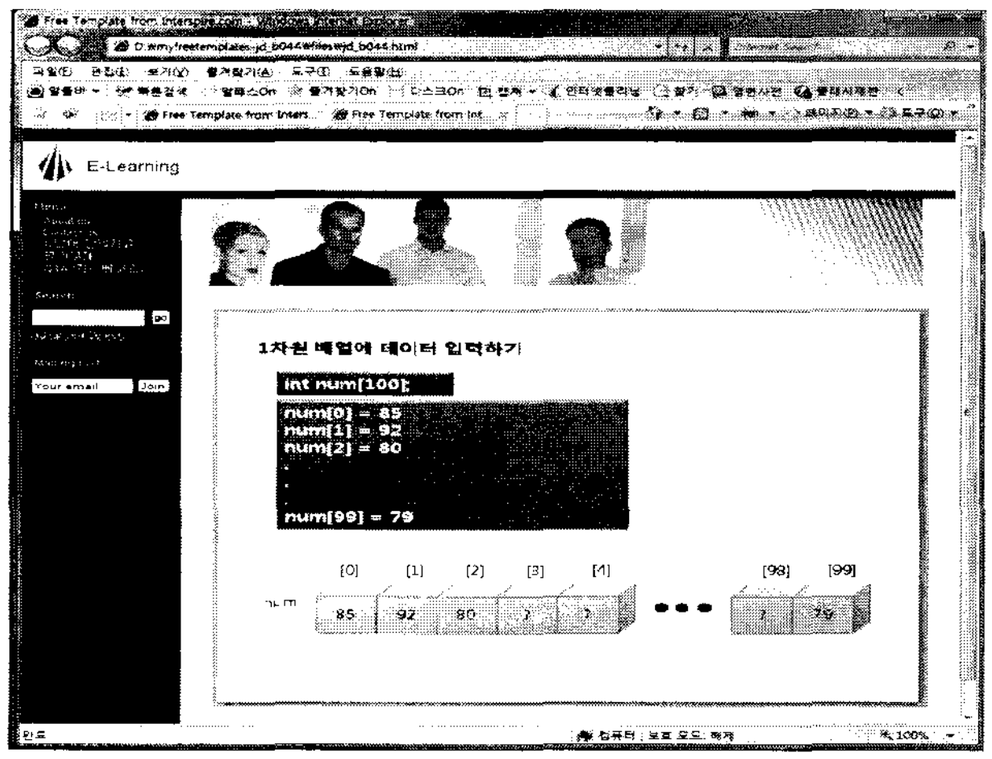


그림 4. 이러닝 시스템
Fig. 4. E-Learning System.

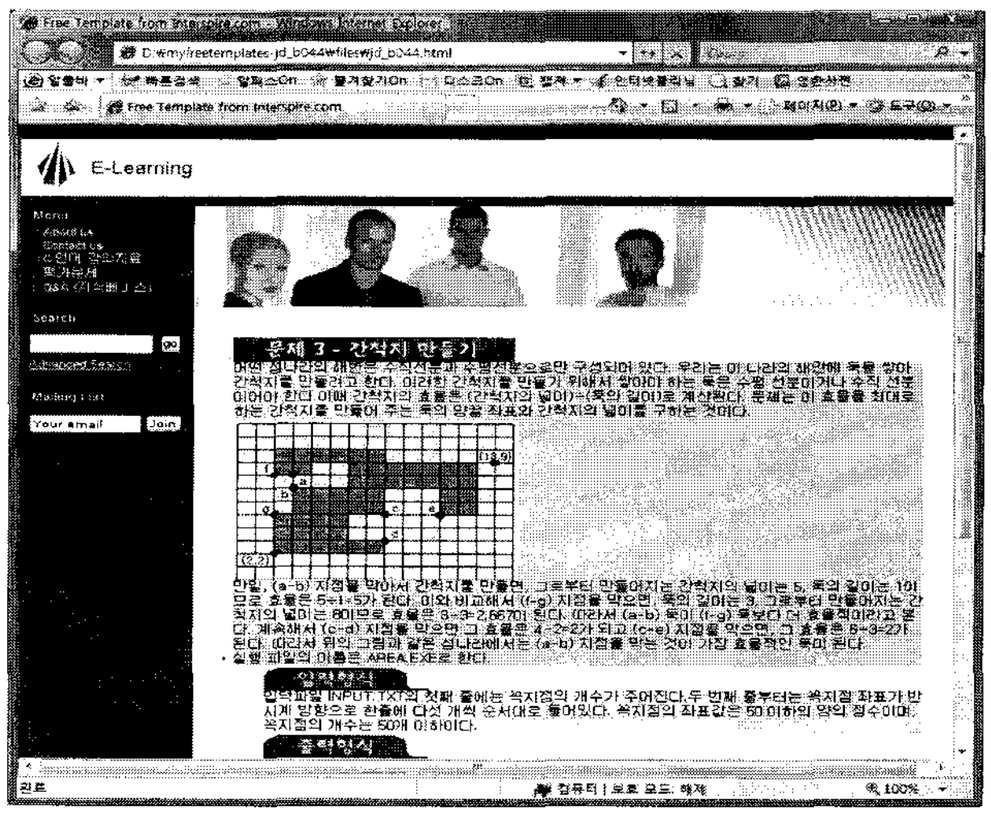


그림 5. 이러닝 시스템 (Q&A)
Fig. 5. E-Learning System (Q&A).

표 1. K-NN 알고리즘의 결과
Table 1. Result of K-NN algorithm.

분류	Precision (%)	Recall (%)
데이터와 연산	100	100
루프(ex, for, while ...)	87	89
판단문(ex, if, if-else ..)	100	97
함수(function)	100	79
배열	100	58
스트링(string)	100	82
포인터(pointer)	90	36
구조체(structure)	80	52
파일	60	47
기타	80	96

한다는 것을 의미한다. 본 논문에서는 재현율 보다는 정확도가 더욱 중요할 것이다.

(그림 6)는 학습자들이 마이닝을 통해서 분류된 Q&A자료를 충분히 활용할 수 있는 실험집단과 Q&A 자료의 접근을 제안한 통제 실험집단으로 구분하여 C 언어 시험을 평가한 결과를 도식적으로 나타낸 것이다.

(그림 6)과 (그림 7)에서 x축은 테스트 항목을 나타낸다. y축은 학생들의 성적을 나타낸다. 이때 통제 실험집단은 본 논문에서 제안한 알고리즘에 의한 Q&A를 제공하지 않은 집단이고 실험집단은 충분히 자료를 제공한 집단이다. 이들에게 1학기동한 충분한 자료를 경험하게 하여 시험문제를 통해서 확인한 결과 문항별로 성적에 있어서 약간의 차이가 나타남을 알 수 있다.

(그림 6)은 1학기 2중간고사 성적을 차트로 나타낸 것이다. (그림 7)은 기말 고사를 나타낸 것이다. 1, 2학기 성적을 차트로 나타낸 도표를 통해서 실험집단이 다소 성적이 좋게 나타났다.

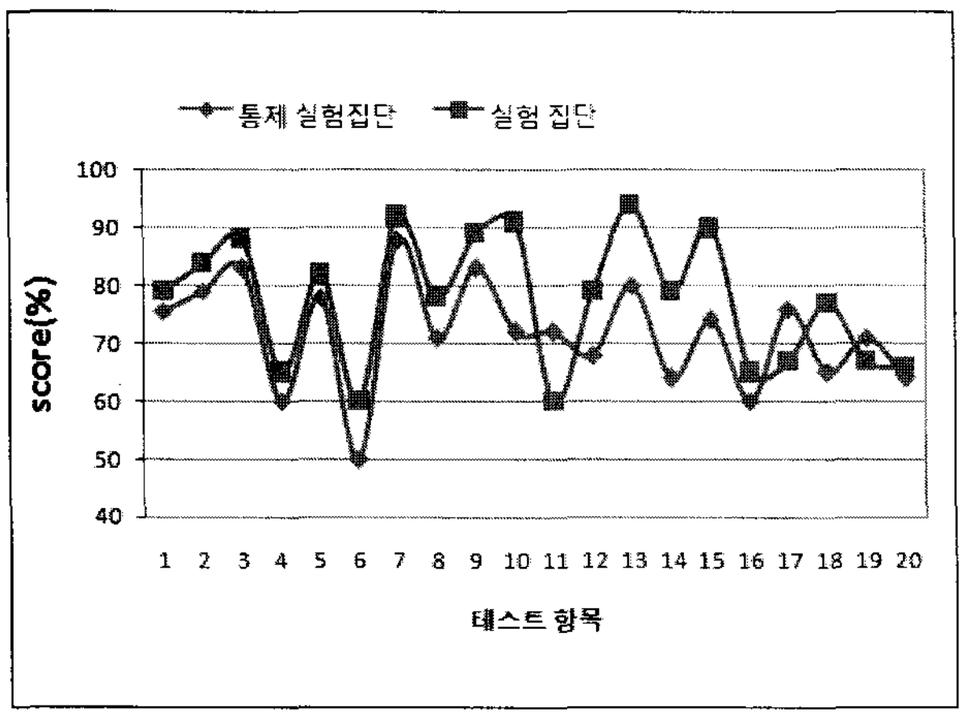


그림 6. 실험집단과 통제 실험집단의 비교
Fig. 6. Comparison of control group and experiment group.

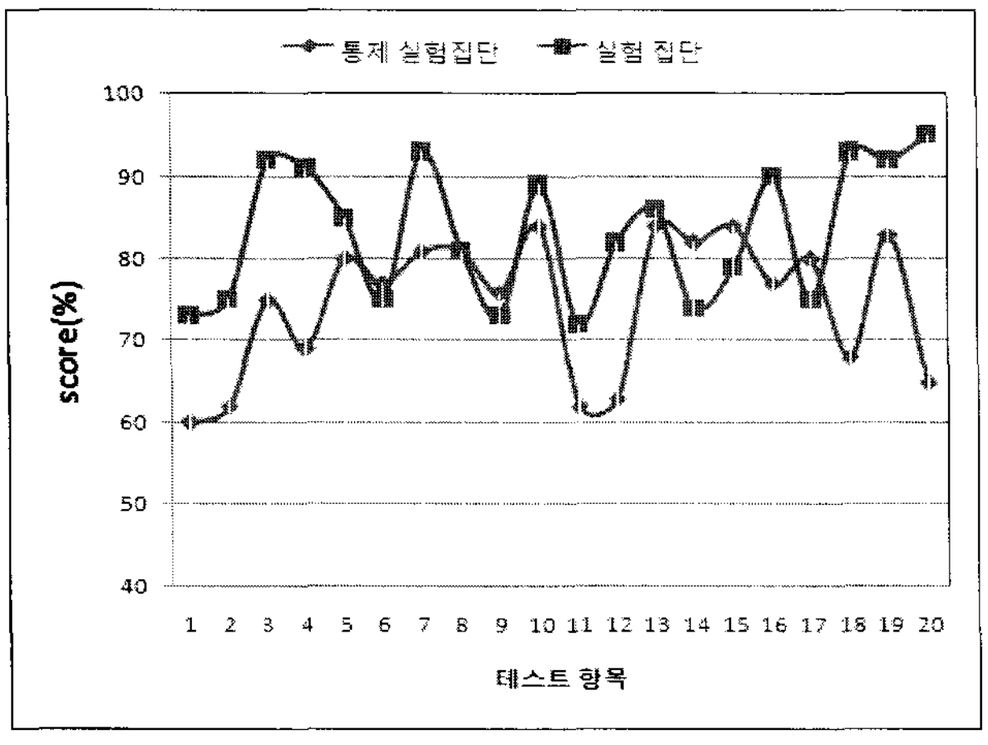


그림 7. 실험집단과 통제 실험집단의 비교
Fig. 7. Comparison of control group and experiment group.

(그림 6)과 (그림 7)의 차트의 결과 텍스트 마이닝을 이용해서 만들어진 Q&A 자료가 학습에 도움이 되었다는 결론을 얻을 수 있다.

V. 결 론

최근에 우리나라는 다양한 교육적인 장점을 갖고 있는 원격 교육 시스템이 많은 확장을 보이고 있다. 오프라인교육의 단점은 시간과 장소적인 한계에 있다. 따라서, 이를 해결할 수 방법이 필요하게 되었다. 이러한 문제점을 해결할 수 있는 방법으로 원격교육의 하나인 이러닝이 등장하게 되었다. 원격 교육은 현장 교육이 아니기 때문에 여러 분야에서 많이 이용하면서 이러닝 분야가 크게 확장되고 있다. 그 결과로서 이러닝 분야의 컨텐츠의 양이 급속히 증가하고 있다. 이 중에서도 텍스트 형태의 교육 자료가 크게 증가하고 있다. 그러나 이들 데이터가 적절히 분류되지 않으면 학습자가 학습에 이용하는데 불편함이 있고 교육적인 효과도 거두기 쉽지 않다. 따라서, 본 논문에서는 이와 같은 데이터가 학습에 도움이 될 수 있도록 분류하여 학습자가 학습하는데 도움이 되도록 하는 것이 목적이다.

이러닝 시스템은 기본적으로 강의 자료를 제공하여 학습에 도움을 주고 있고, 학습 평가문제를 통해서 학습자가 얼마나 이해를 했는가에 대한 측정을 통해서 학습자가 더 높은 수준의 학습을 진행할지 아니면 피드백을 통해서 학습한 내용을 더욱 분명하게 이해하도 록 할지에 대한 결정을 하게 된다.

그러나 때로는 학습자가 강의 자료에는 나오지 않은 질문을 하거나 하여 자신의 학습 욕구를 해소하기를 원할 수 있다. 물론 대부분의 시스템에 이와 같은 간단한 게시판을 제공하고 있지만, 자료가 분류되어 있지 않기 때문에 중복된 질문이 계속 누적될 수 있다. 이것은 매우 낭비일 수 있고, 더욱 심도 있는 질문을 방해할 수 있다. 따라서 본 논문에서는 이를 개선하기 위해서 클러스터링 기술을 게시판에 접목시켜서 다양한 질문을 그룹화 하였다. 그리고 이와 같은 정보를 학습자에게 제공하였고, 그 결과 학습자들은 학습에 있어서 매우 만족스럽게 생각하고 또한, 성적 향상에 도움이 되었음을 결과 데이터를 통해서 보였다. 향후 이 연구는 지능을 이용한 스마트 이러닝에 기초가 될 것이다.

참 고 문 헌

[1] Curran, K, A Web-based collaboration teaching environment, IEEE Multimedia, 9(3).2002.

[2] Khalifa, M. & Lam, R.A, Web-based learning : effects on learning process and outcome. IEEE Transactions on Education, 45(4).2002.

[3] Kinshuk & Yang, Khalifa, M., & Lam, R. Web-based asynchronous synchronous environment for online learning. United States Distance Education Association Journal, 17(2), 5-17, 1537-5080.2003.

[3] 최윤정, 박승수, 웹컨텐츠의 분류를 위한 텍스트마이닝 시스템 설계 및 구현, 한국정보과학회 2001년도 봄 학술발표논문집 제28권 제1호(B), 2001. 4.

[4] Wheeler, H. G. WebCT- WebCT clear leader in online learning programs. The Chronicle of Higher Education, 11(Octorber), 34. 2000.

[5] Losiewicz, P.B., Oard, D. W., & Kostoff, R. N., Textual data mining to support science and technology management, Journal of Intelligent Information system, 15(2), 99,-119, 2000.

[6] Yang, Y. An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, 1(1/2), 67-88. 1999.

[7] Aggarwal, C. C., & Yu, P.H., On effective conceptual indexing and similarity search in text data. Proceedings of the 2001 IEEE International Conference on Data Mining (pp.3-10). San Jose., 2001.

[8] Dorre J., P. Gers시], R. Seiffert, "Text Mining : Finding Nuggets in Mountains of Textual Data", In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999.

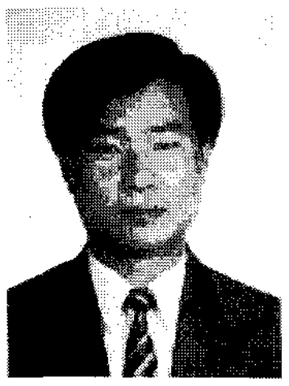
[9] IBM Text Mining, http://www-4.ibm.com/software/data/iminer/fortext/download /whiteweb.html

[10] Salton, G., & McGill, M. J., Introduction to modern information retreival, New York, NY: Mc-Graw Hill, 1983.

[11] 이지행, "FAQ 문서의 자동분류를 위한 다중방법 결합에 관한 연구", 연세대학교 대학원, 2000.

[12] G. Salton, A. Wong, and C. S. Yang (1975), "A Vector Space Model for Automatic Indexing," Communications of the ACM, vol. 18, nr. 11, pages 613 - 620.

저 자 소 개

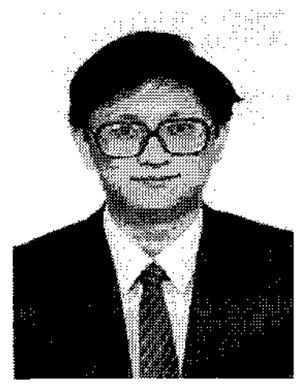


김 천 식(정회원)

1997년 한국외국어대학교 컴퓨터 및 정보통신공학과 (공학석사)
2003년 한국외국어대학교 컴퓨터 및 정보통신공학과 (공학박사)
2000년~2003년 경동대학교 정보통신공학부 교수
2004년~현재 안양대학교 교수
2007년~현재 대한전자공학회 컴퓨터소사이어티 분과위원장
2008년~현재 인터넷 방송통신 TV학회 상임이사
2006년~현재 인터넷 정보학회 학회편집위원
2006년~현재 대한교통학회 정회원
2005년~현재 한국데이터베이스학회 정회원
<주관심분야: 데이터베이스, 데이터마이닝, 유비쿼터스, 텔리매틱스, TPEG, DMB, 홈네트워크, e-Learning>

정 명 희(정회원)

1989년 서울대학교 계산통계학과 졸업.
1991년 U. of Texas, Austin
1997년 U. of Texas, Austin 산업공학과 박사학위
2006년 현재 안양대학교 디지털 미디어공학과 교수.
<주관심분야 : e-learning, 영상, 멀티미디어>



홍 유 식(정회원)

1984년 경희대학교 전자공학과 (학사)
1989년 뉴욕공과대학교 전산학과 (석사)
1997년 경희대학교 전자공학과 (박사)
1985년~1987년 대한항공(N.Y.지점 근무)
1989년~1990년 삼성전자 종합기술원 연구원
1991년~현재 상지대학교 컴퓨터공학부 교수
2000년~현재 한국 퍼지 및 지능시스템학회 이사
2004년~현재 대한 전자 공학회 ITS 분과위원장
2001년~2003년 한국 정보과학회 편집위원
2001년~2003년 한국 컴퓨터 교육산업학회 이사, 편집위원
2004년~현재 건설교통부 ITS 전문심사위원
2004년~현재 원주 시 인공지능신호등 심사위원
2005년~현재 정보처리학회 이사
2005년~현재 인터넷 정보학회 이사
2005년~현재 정보처리학회 강원지부 부회장
2006년~현재 인터넷 방송통신 TV학회 상임이사
<주관심분야: 퍼지 시스템, 전문가시스템, 신경망, 교통제어>