Journal of Korea Multimedia Society Vol. 7, No. 3, March 2004(pp. 438-449)

신경망을 이용한 사용자 질의 전자 메일 분류

변영철[†], 홍영보[††]

요 약

인터넷 사용 증가와 함께 질의 메일의 사용이 증가함에 따라 인터넷 사이트 운영자는 이용자가 질문을 하기 전에 먼저 FAQ나 Q&A를 먼저 확인하기를 바라고 있으나 사용자는 간단히 질의 메일을 보냄으로써 답을 손쉽게 얻으려고 한다. 이에 따라 질의 메일 증가는 상담자에게 많은 시간과 비용을 투자하도록 하고 있다. 본 연구는 질의 메일을 자동으로 분류함으로써 담당자가 메일을 효과적으로 처리하도록 하기 위한 방법에 관한 연구이다. 본 연구의 타당성을 검증하기 위하여 현재 한국통신㈜ 코넷에서 받은 질의 메일을 실험 데이터로 사용하였다. 14개의 질의 메일 부류에 대해 210개의 학습 데이터와 280개의 테스트 데이터 등 모두 490개의 데이터를 이용하여 실험을 수행한 결과 신속한 답장을 바라는 사용자의 요구에 부응함을 알 수 있었다.

Classification of Query E-Mail Using Neural Network

Byun Yung-Cheol[†], Hong Young-Bo[††]

ABSTRACT

More and more users are using the query e-mail according to the increment of use of internet. The operator of internet site desires the users to check the FAQ and Q&A contents first before sending the query e-mail to the operator. However the users try to get the solution for a problem easily by simply sending a query e-mail. Therefore the increment of query e-mail is inevitable, and the site operator is suffering from too heavy loads and spending too much time and cost to reply the query e-mail. In this paper, we are proposing an efficient method of classifying the query e-mail of users automatically by using a neural network. To verify the reasonability of our work, the query e-mails of KORNET are used as the test data, which is actually gathered in KT. A total of 210 learning data and 280 test data were used to test the performance of the proposed approach. From the experiments we got the encouraging result from the view point of application in real life. The proposed approach satisfied the request of users who wanted rapid response for their query e-mail.

Key words: E-Mail Classification(전자 메일 분류), Feature Extraction(특징 추출)

1. 서 론

1.1 연구 배경

지식 정보 사회로 급속히 전환되면서 물품 구매 및 각종 티켓 구매 등과 같은 일상적인 일들이 일반 가정에서 인터넷을 통하여 이루어지고 있다. 인터넷을 통한 e-비즈니스가 빠르게 성장하고 있는 것이다. e-비즈니스에 있어서 배너 광고, 온라인 판촉 등과 같은 다양한 온라인 마케팅 커뮤니케이션 수단은 물론 TV, 잡지, 신문광고 등과 같은 기존 오프라인 마케팅 커뮤니케이션 수단, 커뮤니티, 게시판, 메일, 전화 등 다양한 접점에 대한 관리가 필요하다. 특히 고객과의 1 : 1 관계를 유지할 수 있는 게시판, 메일, 전화 등을 이용한 커뮤니케이션은 관심이 많아졌다.

※ 교신저자(Corresponding Author) : 변영철, 주소 : 제주도 제주시 제주대학로 66번지(690-756), 전화 : 064)754-3657, FAX : 064)755-3620, E-mail : ycb@cheju.ac.kr
접수일 : 2003년 8월 1일, 완료일 : 2003년 9월 18일
[†] 제주국립대학교 통신컴퓨터공학부 전임강사
[††] KT 연구원
(E-mail : bluemoon@kt.co.kr)

이처럼 인터넷의 보급으로 통신 매체가 전화 또는 팩스와 같은 아날로그 매체에서 전자 메일을 필두로 한 디지털 환경으로 변하면서 소비자와 기업 간 가상 공간에서 상호 동적인 의사 결정과 양방향 통신이 가능하게 되었다. 고객과의 관계 및 개별화된 커뮤니케이션이 중요한 요소로 자리 매김하였다. 이런 활동 가운데 고객과 1:1 관계를 유지하면서 양방향 커뮤니케이션을 하는 전자 메일의 활용은 매우 중요하다. 커뮤니케이션 수단으로서의 메일은 상품 및 기타 서비스 광고뿐만 아니라 고객의 요구 분석을 위한 메일 서베이, 그리고 사보와 각종 뉴스를 전달하기 위한 수단으로 사용되고 있다.

전자 메일은 지구 반대편에 있는 이용자에게도 자신의 불만 및 요구 사항을 단 몇 분 만에 담당자와 의사소통을 할 수 있도록 해 준다. 따라서 정보에 대한 이용자의 요구가 높아짐에 따라 정확하고 질 높은 정보 제공과 함께 신속한 대응이 어느 때보다도 필수적이 되었고, 이용자의 요청을 어떻게 대응하느냐에 따라 서비스의 평가가 좌우되게 되었다. 이와 함께 서비스의 환경도 인터넷 서비스는 24시간 365일 체제로 운영이 되고 있고, 이용자의 요구는 주로 퇴근 후에 집중적으로 이루어지고 있으며, 이런 요청에 대한 답변이 만족스럽지 못할 경우 부정적인 평가가 급속도로 확산되므로 신속한 답변이 서비스 품질 평가에 있어 중요한 요소가 되었다.

인터넷 사용 증가와 함께 질의 메일의 수가 증가함에 따라 인터넷 사이트 운영자는 이용자가 질문을 하기 전에 먼저 FAQ나 Q&A를 확인하기를 바란다. 하지만 이용자는 자신이 모르는 부분에 대해서는 간단히 사이트 관리자에게 메일을 보냄으로써 답을 손쉽게 얻으려고 한다. 따라서 사이트 운영에 있어서 질의 메일 증가는 메일 상담자에게 지나친 업무 부하를 주었고 자신의 본연의 업무뿐만 아니라 부수적인 메일 답변을 위해 많은 시간을 투자하도록 하고 있다. 이런 업무 부담을 덜어 주기 위해 전담으로 메일 답변자가 생겼고, 전담 메일 답변자는 좀 더 효율적으로 답변을 하기 위해 해당 사이트에 각 담당의 메일 아이디를 게시하여 사이트 이용자가 질문을 할 때 담당 직원 메일로 직접 질문 하도록 유도하였다. 그러나 이용자의 자의적인 판단에 따른 질문으로 질의 메일을 답변하는 담당자는 모든 종류에 대해 답변을 해야 했다. 같은 질문에 대해서도 답변 직원의 성향과 취향에 따라 조금씩 다른 메일답변이 이루어졌으며, 답변자는 일일이 질의 메일 내용을 확인해야 이용자의 질문 내용을 확인할 수 있다.

1.2 연구 목적 및 방법

질의 메일에 대해 답변하는 것은 이용자와의 양방향 채널이고, 이 채널 관리를 어떻게 하느냐에 따라 기업의 이미지에 영향을 미친다. 질의 메일은 특정 기준에 의해서 분류되지 않으며 각종 내용을 내포하고 있어서 질의 내용을 일일이 확인해야 한다. 질의 메일이 분류가 제대로 되지 않을 경우 답변자는 모든 내용에 대한 지식을 필요로 하게 되며, 유사한 질문이 아닌 광범위한 질의 내용을 처리함으로써 업무 집중도가 낮아지고 답변에 많은 비용을 요구하게 된다. 따라서 이를 해결하기 위해서는 질의 메일을 자동으로 분류하여 해당 담당자에게 보내는 효율적인 대안이 필요하다. 즉, 전자 메일을 내용에 따라 적절히 분류하여 해당 담당자에게 전송해주는 일이 매우 중요하게 되었다. 전자 메일을 내용에 따라 분류를 한다면 담당자는 자신의 맡은 영역의 질의 메일만을 답변하므로 전문화로 의해 답변 시간이 줄어들고 비용도 감소시킬 수 있다.

본 연구의 목적은 일반적인 신경망을 이용하여 질의 메일을 효과적으로 분류하고 그에 따른 업무의 효율성을 향상시킬 수 있는 방법을 제시하는데 있다. 메일을 효과적으로 분류하기 위하여 현재 한국통신에서 메일 답변 시스템으로부터 분류를 위한 지식(knowledge)을 추출하여 모형화 하고, 이를 이용하여 신경망의 입력으로 주어지는 특징 벡터를 구성한다. 특징 벡터를 구성하기 위한 measure를 지식에 근거하여 설정하고, 이를 활용하여 양질의 특징 벡터를 구성하고 질의 메일을 분류한다.

구체적으로, 한국통신 코넷에 접수된 기존의 고객 상담 내용을 분석하여 전자 메일 내에서 중요하다고 여겨지는 단어를 상담자를 대상으로 설문을 하여 키워드 사전을 정의한다. 정의한 키워드 사전을 이용하여 미지의 질의 메일 내용 중에서 표준화된 키워드를 추출한다. 생성된 키워드를 특징 벡터로 양자화하여 신경망의 입력으로 사용함으로써 미리 정의한 질의 부류(class) 중 하나로 분류한다. 신경망에 의해 인식된 결과를 바탕으로 적절히 답변 메일을 고객에게 전송하고, 실패한 메일은 전문 답변자에게 전송함으로써 효과적인 답변 분류 시스템 체계가 가능하도록 한다.

1.3 논문 구성

본 연구는 모두 5장으로 구성되어 있다. 다음 장에서는 본 연구와 관련된 연구와 제안하는 방법의 개요 및 연구 범위에 대해 설명한다. 3장에서는 본 연구에서 제안하는 신경망을 이용한 전자 메일 자동 분류 방법에 대해 설명한다. 4장의 실험 결과에서는 실험 방법 및 환경, 그리고 실험 내용에 대해 설명한다. 마지막으로 5장에서는 본 연구에 대한 결론 및 토의에 대해 설명한다.

2. 관련 연구 및 제안하는 방법

2.1 관련 연구

정보 통신 기술의 발전으로 인해 온라인으로 생성되는 전자 문서의 양이 폭발적으로 증가 하게 되었다. 따라서 수동으로 문서를 분류하던 종래의 방법 대신에 문서를 자동으로 분류하여 담당자에게 문서를 배분해 주는 것은 업무 효율성 면에서 매우 바람직한 일이며 이와 관련된 기술 개발이 요구되고 있다[1-4]. 문서의 자동 분류란 일반적으로 기계 학습을 이용하여 미리 학습시켜둔 범주 중 하나로 문서를 인식하는 것을 의미한다. 이미 분류 되어진 문서로부터 각 분류 카테고리에 나타나는 단어들의 출현 빈도에 대한 정보를 추출하여 분류에 이용하는 통계적 분류 방법[5,6,9-12]과 문서가 가지고 있는 뜻을 파악하여 분류에 이용하는 지식 기반 방법[5-8] 등이 있다.

통계적 분류 방법에는 사람에 의해 이미 분류되어 있는 문서들(training set)로부터 각 분류 카테고리에 나타나는 단어들의 출연 빈도에 대한 정보를 추출하고, 분류하고자 하는 문서로부터 주요 단어들을 추출한 후 이를 이용하여 가장 적합한 카테고리를 찾거나 각 카테고리에 대하여 포함 여부를 판단하는 것으로, Bayesian 확률을 이용하여 문서가 각 카테고리에 속할 확률을 계산하는 방법[2,10,12]과, 분류하려는 문서와 각 카테고리에 포함된 문서들 간의 유사도를 계산하는 방법이 제안되었다[3,9,11,13].

지식 기반 방법은 분류 대상 문서의 샘플들을 분석하여 분류 규칙들을 만들고 이러한 규칙을 이용하여 문서 분류를 수행하는 것으로, 문서의 내용에 따라 분류 규칙을 만드는 방법[4,8]과 문서 내용 외의 정보들을 이용하는 방법[1]이 있다. 문서의 내용에 따른 분류 방법으로는 특정 카테고리로의 분류에 결정적인 단서가 되는 핵심 단어들을 추출하고, 이러한 단어들의 출현 여부에 따라 분류를 수행하도록 하는 방법, 그리고 특정 카테고리로 분류되어 있는 문서들에 자주 나타나는 구나 문장 형태를 패턴으로 표현하여 패턴 매칭에 의해 문서를 분류하는 방법, 문서의 내용을 파악하여 문서를 분류하는 방법이 있다. 문서 내용 외의 정보를 이용하는 방법은 문서의 작성 부서와 같은 정보들을 이용하는 규칙을 만들어 문서를 분류하는 방법으로 전문가 시스템의 형태로 구현될 수 있다.

통계적 분류 방법은 단어들의 출현 빈도를 기반으로 각 카테고리로 분류될 확률이나 각 카테고리와의 유사도를 계산하므로 가장 높은 값을 갖는 단일 카테고리로 문서를 분류할 경우, 모든 문서를 분류 할 수 있으나 문서의 내용을 분석하는 것이 아니므로 분류의 정확도에는 한계가 있다[14]. 지식 기반 방법 또한 사람이 분류 대상 문서들에 대해 분석을 수행 한 후 분류 규칙을 만들어 사용하므로 규칙에 따라 분류된 문서의 경우 높은 정확도를 나타내지만 충분한 규칙을 제공하지 못하면 분류되지 못하는 문서들의 비율이 높아질 수 있다. 따라서 이들 방법의 장점과 단점을 해소하기 위해 이들 방법을 병행하는 연구가 시도되었다. 참고문헌 [14]에서는 통계적인 문서 분류 시 문장의 의미를 파악할 수 있는 패턴들을 이용하면서 지식기반 분류 방법을 접목시켜 분류의 정확성을 높이는 방법을 제안하였다. 또한 통계적인 분류와 지식기반 분류를 복합적으로 사용한 분류 실험이 통계적인 방법만을 사용한 경우에 비해 높은 정확도를 나타냄을 보여주었다.

이외에도 분류 방법 중 널리 알려진 방법으로 결정 트리 학습법이 있다. 결정 트리는 정보 이론에 기반하여 귀납적 유도 학습 방법으로 가장 많이 사용되어지는 것 중 하나로서 1949년 Shannob과 Weaver에 의해 처음으로 소개되었으며, 예는 1986년에 개발된 ID3[16]와 1993년에 C4.5 및 Cubist[17]등이 제안되었다. 잡음이 있는 데이터에 유리하며, 표현을 구별할 수 있는 학습능력이 뛰어난 점이 결정 트리의 특성으로 이 특성을 이용하여 문서의 범주화에 많이 사용되어 진다. 영국의 Timberlake 사에서 제작한 CART(Classification and Regression Trees) 시스템은 결정 트리 학습 알고리즘을 이용하여 만든 범주 데이터 및 연속 데이터의 분류 시스템이다. 결정 트

리 학습 알고리즘[18]은 트리를 학습에 적용한 것으로 구현이 쉽고, "IF~Then" 형태의 간단한 규칙으로 표현되기 때문에 많이 사용하지만, 깊이나 가지의 수를 제한하기 위해 부가적인 처리가 필요하다는 단점이 있다. 참고문헌[15]에서는 사용자들의 다양한 요구로 인하여 다양한 문서들이 수집되었을 때 이들 문서들 중에서 사용자가 관심 있는 분야의 수만큼 자동적으로 사용자의 프로파일을 생성하고, 이에 따라 문서들을 분류하는 방법을 제시하였다.

위 방법들은 문서 분류 시 문서 중심으로 분류를 수행하였고 모든 문서는 각각이 독립된 분류에 소속된다. 이와는 달리 인공신경망을 이용한 방법[22,25]은 인간의 두뇌를 모방하여 두뇌 활동의 매카니즘을 수학적으로 재현한 것으로 학습 경험을 바탕으로 새로운 입력에 대하여 만족해를 스스로 구할 수 있다. 그럼으로써 문서를 각 분류에 대한 소속의 정도로 표현할 수 있게 하여 여러 주제가 내포된 실생활의 문서를 분류하는데 효율적이다.

앞서 설명한 웹 문서 분류 방법들은 크게 통계적 방법, 구문론적 방법, 신경망을 이용한 방법으로 구분할 수 있으며 각 방법의 특징을 요약 정리하면 표 1과 같다.

2.2 제안하는 방법 개요 및 연구 범위

전자 질의 메일의 경우, 사용자의 질의 습관 및 연령층, 그리고 지식의 정도에 따라 동일한 질의의 경우에도 다양한 형태의 질의 메일이 나타날 수 있다. 따라서 키워드의 빈도수 및 기타 통계 자료에 의한 통계적 분류 방법과 웹 문서로부터 추출한 구문적인 정보의 매칭을 통한 문서 분류 방법으로는 숨겨진 분류 규칙을 찾는데 어려움이 있다. 이러한 단점을 극복하기 위해 본 연구에서는 통계적 방법과 신경망을 이용한 방법으로 질의 메일을 분류한다. 통계적 근거에 의해 키워드 기반의 특징 벡터를 추출하고, 신경망 학습을 통해 다양한 형태의 질의 메일로부터 숨겨진 분류 규칙을 자동으로 찾는다. 그리고 추출한 규칙에 근거하여 효과적으로 질의 메일을 분류한다.

코넷의 질의 메일의 경우 질문 유형이 확실히 구분되고, 메일 답변자가 질의 메일을 분석할 때 90% 이상의 메일에 대해 핵심 키워드만을 인지하더라도 메일의 유형을 분류할 수 있다는 특성이 있다. 따라서 기존의 범위가 제한되지 않는 일반적인 문서와 같이 문서 유사도를 동적으로 구하기보다는 통계적인 방법으로 메일 답변자의 지식을 추출하고 이를 근거로 신경망을 이용하여 질의 메일을 효율적으로 분류한다. 그림 1은 질의 메일을 분류하기 위한 시스템 처리 흐름도이다. 학습 데이터를 위한 질의 메일을 분석하여 지식을 획득한 후 이를 모형으로 이용하여 미지의 질의 메일을 모형으로 등록되어 있는 질의 메일 중 하나로 인식한다.

그림 2는 실제로 미지의 질의 메일로부터 특징을 추출하여 신경망의 입력으로 사용함으로써 질의 메일 내용을 인식하는 과정을 보여준다. 키워드 추출은 질의 메일 중에서 불필요한 용어나 낱말을 삭제하

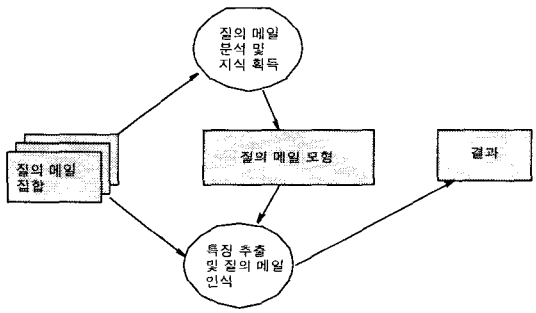


그림 1. 전자 메일 응답 시스템 처리 개요

표 1. 웹 문서 분류 방법 요약

분 류	특 징
통계적 방법	웹 문서로부터 추출한 특징을 N 차원의 특징 벡터로 표현한 후거리 측정이나 혹은 판별(discriminant) 함수를 이용하여 유사도를 측정한다. K-Nearest Neighbor 분류, DTW 정합 알고리즘, K-Means 방법 등이 여기에 속한다.
구문론적 방법	웹 문서로부터 추출한 패턴을 스트링이나 트리, 또는 그래프 형태로 표현하고 구문 분석이나 파싱 절차에 의해 인식이 이루어진다. 유한 오토마타, 결정 트리 분류기, 지식 기반 분류 등이 여기에 속한다.
신경회로망 방법	추출한 특징이 신경망의 입력이 되고 반복되는 학습 과정을 통하여 데이터의 특성을 학습하여 미지의 입력 패턴을 인식하게 한다. 학습 방법으로는 오류 역전파(BP) 알고리즘이 널리 쓰인다.

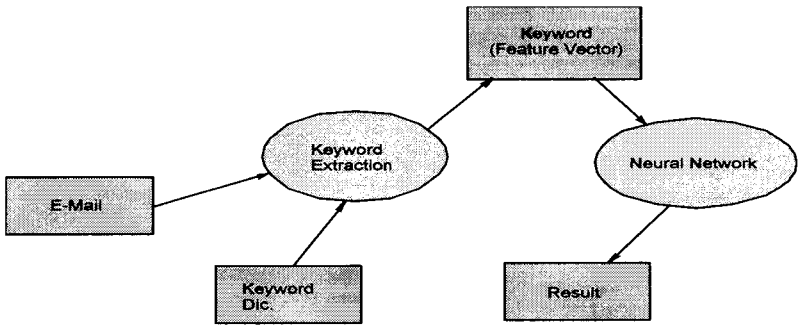


그림 2. 질의 메일 분류 흐름도

여 가장 중요한 단어나 요약이 가능한 단어로 변형시키는 작업이며, 키워드 추출에 의해 신경망의 입력인 특징 벡터가 구성된다. 신경망에 의해 인식된 질의 메일의 경우 자동으로 응답 메일을 전송하거나 응답 메일이 정의되지 않았을 경우에는 메일 담당자에게로 전송된다.

3. 신경망을 이용한 전자 메일 자동 분류

3.1 코넷 사용자 질의 메일 분석 및 실험 방법

현재 한국통신 코넷에서는 사용자로부터 질의 메일을 받을 때 '상품 및 개통관련', '장애 및 고장신고', '고객정보변경', '코넷에 바란다', '일반환경설정', '불량 네티즌 신고센터' 중 하나의 유형을 선택하여 보내도록 하고 있다. 첫 번째 '상품 및 개통관련'은 각종 서비스의 특성 및 이용 시 궁금한 점과 특정 서비스를 개통에 관한 질의를 할 수 있다. 두 번째 '고객정보 변경'은 아이디, 비밀번호 와 주소지 변경 등과 같은 각종 변경에 관한 내용을 질의한다. 세 번째 '일반환경설정'은 접속 프로그램 설정, 네트워크 환경설정, OS별 환경 설정과 메일 설정이 있다. 네 번째 '장애 및 고장신고'는 서비스를 이용하지 못할 경우 그에 따른 궁금한 점을 질의한다. 다섯 번째 '코넷에 바란다'는 코넷 서비스 이용 중에 서비스 개선이나 서비스 이용 시 불편한 사항을 질의할 수 있다. 마지막으로 '불량 네티즌 신고센터'는 코넷에서 제공하는 대화방이나 바둑 또는 게임 서비스 이용 시 악의적인 행위를 하는 네티즌을 신고하는 항목이다.

2002년 3월 한 달 동안 질의 메일을 분석한 결과 그림 3과 같은 분포도를 얻었다. 서비스 행정과 관련된 질의는 '고객정보변경(15%)'과 '상품 및 개통문의(22%)'로 37%를 차지하였고, 코넷 서비스에 대해 불만을 표시하는 메일은 '장애 및 고장신고(16%)'와 '코넷에 바란다(10%)'로 26%를 점유하였다. 이외에 '일반환경설정'이 32%로 가장 많은 점유율을 차지하였고, 점유율이 낮은 '불량 네티즌 신고'가 5%이다.

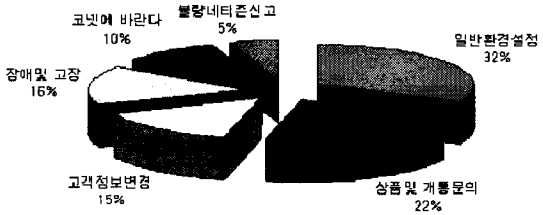


그림 3. 코넷 질의 메일 항목별 분포도

'장애 및 고장신고'와 '코넷에 바란다'는 서비스 이용 중에 발생하는 각종 불만들을 토로하는 질의 메일이 많았고, 각 질의 메일은 주관적이고 자의적인 내용이 주를 이루어 메일 내용에 따라 다른 답변을 해야 하는 경우가 많았다. '장애 및 고장신고'와 '코넷에 바란다'의 비율이 작은 이유는 문제가 생겼을 때 이용자는 담당 부서로 전화를 해서 신속하게 문제 해결을 원하기 때문으로 분석되었다.

본 연구에서는 코넷 질의 메일 내용 중에서 특정 키워드를 이용하여 효과적으로 분류할 수 있는지 확인하기 위하여 질의 메일 답변자의 의견을 조사하였다. 이를 위해 설문을 통하여 질의 메일 내용 중에 특징적인 키워드를 내포하고 있는지, 내포하고 있다면 일반화 시킬 수 있는 항목은 무엇인지를 코넷 메일 답변자 30 명에게 조사하였다. 설문 내용 중 ⓐ '본인은 특정 항목의 질의 메일만을 답변하는가'라는 질문에 응답자중 47%는 특정 항목만 전담해서 하고 나머지는 특정 항목뿐만 아니라 다른 항목까지 같이

처리하는 것으로 나타났다. ⓑ '질의 내용과 질의 항목이 일치 했으면 좋겠는가'라는 질문에 응답자중 93%가 찬성을 나타냈으며 ⓒ '질의 메일이 항목과 연관된 있으면 처리하기 쉬운가' 라는 질문에 응답자 중 87%가 찬성을 나타냈다. ⓓ '답변 시 정형화 된 내용을 활용하는가'라는 질문에 응답자 중 67%가 그라고 답했고 33%가 그저 그렇다라고 답했다. ⓔ '숙달이 되면 메일답변 시간이 줄어드는가'라는 질문에 33%는 찬성했고 50%는 그저 그렇다, 나머지 17%는 그러하지 않다고 하였다. ⓕ '특정 단어로 질의 메일 내용을 파악할 수 있는가'라는 질문에 응답자 중 93%가 찬성을 나타냈다.

조사 내용을 종합해 보면, 특정 항목은 질의 메일이 많아서 여러 명이 답변을 하고 있으며, 질의 메일이 질의 항목과 연관된다면 답변 처리 업무가 쉬워질 것으로 응답하였다. 숙달된 메일 답변자라고 해도 메일 내용에 따라서는 처리 시간이 길어질 수 있다는 것을 보여준다. 거의 모든 응답자들이 답변 시에는 기존의 정형화된 답변을 많이 애용하는 것으로 나타났고, 키워드를 통해서 메일 내용을 파악 가능하고 그 항목은 '상품 및 개통문의'와 '일반환경 설정' 항목이 90% 이상을 차지하였다. 2002년 1/4분기 '상품 및 개통문의'와 '일반환경 설정' 항목 질의 건수는 표 2와 같다.

설문에 의한 데이터 분석을 통해 '상품 및 개통문의'와 '일반 환경 설정'은 키워드 추출을 통한 분류에 적합하고, 또한 두 항목이 54%로 과반수를 넘게 차지하여 본 연구 코넷 사용자 질의 메일 가운데 실험 데이터를 '상품 및 개통문의'와 '일반환경 설정' 항목에 국한해서 처리하고자 한다. 본 실험에 사용된 자료는 2002년 3월 총 4,061개의 질의 메일 중 '상품 및 개통 문의'와 '일반환경 설정' 항목으로 제한하였다. '상품 및 개통 문의'와 '일반 환경 설정' 항목 질의 메일 2,193개 가운데 22.3%인 490개의 메일을 표본으로 하였다.

표 2. '상품 및 개통 문의' 와 '일반 환경 설정' 메일 건수(단위: 건)

구분	상품 및 개통문의	일반환경 설정	합계	총 메일 대비 점유율
2002.1	1,057	1,259	2,316	52%
2002.2	1,572	1,354	2,926	51%
2002.3	926	1,267	2,193	54%

3.2 키워드 사전 정의

키워드 사전이란 질의 메일 내용 중에서 반복적으로 나타나거나 질의 메일을 분류하는데 중요한 용어를 정의해 놓은 것으로 키워드 사전을 얼마나 잘 정의 하느냐에 따라서 시스템 성능이 달라질 수 있다. 문서의 유형이 상대적으로 많고 정형화되지 않은 일반 문서의 경우 문서 간 유사도에 의해 자동으로 부류를 나눌 수 있어야 하지만 코넷 사용자 질의 메일의 경우에는 자료 분석 결과 질문 유형에 따른 메일 유형이 상대적으로 정형화되어 있었다. 또한 상담원에 대해 조사한 결과 키워드가 메일 분류 시 중요한 역할을 수행함을 알 수 있었다. 키워드 사전을 구성하기에 앞서 키워드에 의해 분류되어져야 할 부류들이 구성되어야 한다. 2001년 1월에서 2월 사이의 코넷 '상품 및 개통'과 '일반 환경설정' 항목에 대한 질의 메일을 세부 항목으로 구분한 결과, 개통설치, 요금, 홈페이지, 해지, 정보변경, PC 패키지, 이벤트, 로밍, 상품안내, 매일/뉴스, NIC, UMS, 네트워크설정, 접속프로그램 등 14개로 나눌 수 있었다. 기타 부류로는 원클릭CD, 위성인터넷 등이 있었으나 경우의 수가 많지 않아 제외하였다.

14개 부류에 대해 각 부류별 키워드를 선정하기 위하여 2002년 1월에서 2월 사이의 질의 메일 중에서 각 부류에 대해 30개 메일, 전체 30×14개 메일을 30명의 질의 메일 상담자에게 보여주고 키워드를 선별하도록 하였다. 이렇게 30명 각자가 선정한 키워드를 조사한 후 빈도수가 상대적으로 많은 키워드를 표시한 결과 표 3과 같은 결과를 얻었다.

표 3에서 7번 이벤트 부류에 속하는 전체 키워드에 대한 내용을 보면 표 4와 같다. 개인별로 보는 관점에 따라 핵심적인 단어 빈도수가 조금씩 다르지만 응답자가 선택한 키워드들 가운데 '당첨'을 가장 많이 선택하였고 '참가'는 가장 적게 선택하였다. 이벤트 부류의 선택된 단어는 당첨, 페스티벌, 경품, 대잔치, 이벤트, 행사, 무료, 신문광고, 애완견, 참가, 특판, 백만돌파, 축하, 회원 등의 키워드와 이외에도 협찬, 상품권, 온라인 등 다수의 키워드가 있었다. 선정된 키워드를 다시 빈도수가 많은 순으로 정렬하여 빈도수가 높은 순으로 키워드 사전으로 등록하였다.

표 3. 각 질의 메일부류 별 키워드

순서	부류	키 워 드
1	개통설치	설치, 문제, 지역, 코넷, 사용, ADSL, ...
2	요금	요금, 고지, 월사용, 조정, 지로, ...
3	홈페이지	홈페이지, 용량, 저장, 공간, 파일, ...
4	해지	해지, 사용, 방법, 회원, 탈퇴, ...
5	정보변경	변경, 입력, 잘못, 개인정보, ...
6	PC패키지	PC4U, 컴, 구입, 무이자, ...
7	이벤트	당첨, 페스티벌, 경품, 대잔치, ...
8	로밍	해외, 로밍, 국외, GRIC, 출장, ...
9	상품	하이텔, b&a, 무료, 상품권, 홈넷, ...
10	메일/뉴스	뉴스그룹, 메일서버, NNTP, POP, ...
11	NIC	랜카드, 모뎀, 외장형, 구동, 드라이버, ...
12	UMS	UMS, FAX, 03030, 핸드폰, ...
13	네트워크설정	TCP /IP, 서브넷, IP주소, ...
14	접속프로그램	ENTER, WINPOET,NTS, ...

3.3 특징 벡터 구성 및 신경망을 이용한 분류

신경망 학습 및 고객의 질의 메일을 신경망으로 분류하려면 키워드를 자동으로 추출한 후 특징 벡터를 구성해야 한다. 이를 위해 먼저 14가지의 부류 각각에 대하여 빈도수를 기준으로 상위 n개의 키워드를 추출한다. 14 가지의 질의 메일 부류 중 1번째 부류의 키워드 k_i에 대한 점수 s_{ki}^1는 식 3.1을 이용하여 계산한다.

$$s_{ki}^1 = \frac{f_{ki}}{K} \tag{3.1}$$

위에서 f_{ki}는 k_i 키워드의 빈도수 합계(표 3)를 의미하며, K는 키워드의 빈도수 중 최대 빈도수를 의미한다. 식 3.1의 의미는 특정 부류의 질의 메일에서 자주 나타나는 키워드는 메일 분류 시 양질의 특징을 포함하므로 높은 점수를 부여하고자 하는 것이다. 식 3.1에 의해 n번째 부류의 i번째 키워드 점수인 s_{ki}^n는 식 3.2를 만족한다.

$$0 < s_{ki}^n \leq 1 \tag{3.2}$$

특정 부류의 질의에 대해 자주 나타나는 키워드는 질의 메일 분류 시 중요한 역할을 수행하므로 위의 식 3.1의 점수 크기에 근거하여 키워드를 선택한다. 예를 들어, 표 5의 경우 한 개의 키워드만을 선택할 경우 '당첨' 키워드의 점수는 681/681로 최대가 되므로 이를 선택한다. 이처럼 14 개의 부류에 각각 대해 s_{ki}^n 점수에 근거하여 키워드를 추출한다.

표 5는 실제로 앞서 표 4에서 구한 이벤트 부류 키워드에 대해 점수 s_{ki}^7을 구한 결과이다. 실제로 14 개의 모든 부류에 대해 점수 크기를 기준으로 각 부류별 상위 10개의 키워드를 추출한 결과 중복되는 키워드를 한 번만 셀 경우 모두 112개의 키워드를 얻을 수 있었다.

표 6은 부류의 수가 3이고 점수 크기를 기준으

표 4. 이벤트 부류에 대한 답변자 별 키워드 리스트

구분	답변자1	답변자2	답변자3	답변자4	답변자5	...	답변자30	합계
페스티벌	12	21	14	14	10	...	13	633
경품	15	14	13	13	12	...	21	653
대잔치	22	25	21	15	12	...	24	521
이벤트	12	20	15	14	24	...	26	672
행사	19	21	22	21	12	...	29	629
무료	14	12	22	17	21	...	18	394
신문광고	16	25	13	20	24	...	23	603
애완견	17	17	16	12	22	...	21	594
참가	12	8	11	14	12	...	16	378
특판	14	14	17	14	15	...	14	543
백만돌파	16	13	21	22	23	...	18	587
축하	21	18	14	20	20	...	17	524
회원	23	19	10	24	21	...	17	577

표 5. 이벤트 부류의 키워드별 점수(s_{ki}^{7})

키워드		빈도수 합계	s_{ki}^{7}
k1	당첨	681	1
k2	이벤트	672	0.987
k3	경품	653	0.959
k4	페스티벌	633	0.930
k5	행사	629	0.924
k6	협찬	614	0.902
k7	신문광고	603	0.885
k8	애완견	594	0.872
k9	백만돌파	587	0.862
k10	회원	577	0.847
k11	상품권	569	0.836
k12	특판	543	0.797
k13	축하	524	0.769
k14	대잔치	521	0.765

표 6. s_{ki}^{n} 에 근거하여 추출한 키워드 예

메일 부류	s_{ki}^{n} 크기에 근거한 키워드
C1	당첨, 애완견, 회원, 경품
C2	당첨, 회원, 협찬, 축하
C3	당첨, 애완견, 특판, h

로 상위 4개의 키워드를 추출한 예이다. 이 예의 경우 당첨, 애완견, 회원, 경품, 협찬, 축하, 특판, 대잔치 모두 8개의 키워드가 존재한다. 이 중 당첨은 모든 부류에 나타났으며 애완견은 C1과 C3에 나타났다. 표 6에서 당첨은 비록 각 부류에서 빈도수가 가장 높기는 하지만 모든 부류에 나타나므로 질의 메일을 분류할 수 있는 변별력이 없다. 따라서 변별력이 높은 키워드를 선택함으로써 인식률을 높이고 특징 벡터의 차원을 줄여 신경망에 의한 처리 시간을 줄이기 위하여 식 3.3에 의해 키워드 점수를 계산한다.

이때 f_{Ki}^{A}는 14개 부류에서 특정 키워드(k_i)가 나타난 빈도수를 나타낸다. 예를 들어 당첨의 빈도수 f_{Ki}^{A}는 3이다.

$$S_{Ki}^{A}=\frac{1}{f_{Ki}^{A}} \tag{3.3}$$

위 공식에 의해 S_{Ki}^{A}는 식 3.4를 만족한다.

$$0 < S_{Ki}^{A} \leq 1 \tag{3.4}$$

식 3.4에 의해서 여러 클래스에 공통적으로 나타나는 키워드는 상대적으로 점수가 작게 계산되며, 반대로 오직 한 클래스에만 나타나는 키워드는 가장 큰 점수인 1을 얻을 수 있다. 이제 S_{Ki}^{A} 점수에 근거하여 신경망의 입력으로 주어질 특징 벡터를 구할 수 있다. 위 예에서 당첨, 애완견, 회원, 경품, 협찬, 축하, 특판, h 각각의 점수는 1/3, 1/2, 1/2, 1, 1, 1, 1, 1이다. 이를 근거로 7개의 키워드를 선택할 경우에는 당첨을 제외한 나머지, 즉, 애완견, 회원, 경품, 협찬, 축하, 특판, 대잔치를 특징 벡터 추출을 위한 키워드로 선택할 수 있다. 이처럼 특징 벡터 추출을 위한 키워드가 결정되면 질의 메일에 해당 키워드의 존재 유무에 따라 특징 벡터를 구성할 수 있다. 가령 특정 질의 메일에 키워드 회원, 협찬, 특판, 대잔치가 존재할 경우 신경망의 입력으로 주어지는 특징 벡터는 표 7과 같다.

표 7. 특징 벡터의 예 (특징 벡터 : 0 1 0 1 0 1 1)

키워드	애완견	회원	경품	협찬	축하	특판	대잔치
존재유무	0	1	0	1	0	1	1

위의 경우 0은 질의 메일에 특정 위치의 키워드가 존재하지 않음을 의미하며, 1은 존재함을 의미한다. 본 연구에서는 점수 S_{Ki}^{A}에 근거하여 다양한 차원의 특징 벡터를 구성하여 신경망의 입력으로 사용하였다. 질의 메일에서 추출하는 키워드 수에 따라 신경망 입력 층의 입력 노드 수가 결정되며, 출력 층의 노드의 수는 분류하고자 하는 메일의 부류 수인 14개이다. 은닉 층의 노드 수는 입력 층 노드의 수가 n일 경우 2×n+1개로 하였다.

한편, 입력 노드의 수는 자의적으로 결정하기 보다는 신경망 입력 층의 노드 수에 해당하는 특징 벡터의 차원을 25에서 80까지 다양하게 바꿔가며 실험을 수행하고, 그 결과를 바탕으로 최적의 결과를 얻을 수 있는 벡터 차원을 구하였다.

4. 실험 결과

4.1 실험 환경 및 실험 데이터

본 연구에서는 제안하는 메일 분류 방법을 검증하

기 위하여 표 8과 같이 2002년3월 한국통신 코넷에 접수된 질의 메일을 이용하여 특징 추출 및 분류 실험을 수행하였다.

2002년 3월 메일 표본 중에서 490개 메일을 다시 2개 부분으로 나누어 먼저 15×14개의 메일은 학습 데이터(training set)로 사용하였고 나머지 20×14개의 메일은 학습된 신경망을 통하여 결과를 얻기 위한 실험 데이터(test set)로 사용하였다. PENTIUM III 650에서 C++를 이용하여 키워드 추출 및 특징 벡터 구성 알고리즘을 구현하였으며, 추출된 특징 벡터는 MATLAB을 이용하여 학습 및 분류 실험을 수행하였다.

키워드가 제대로 추출되고 이에 따른 특징 벡터가 효과적으로 구성되는지는 질의 메일 인식률을 이용하여 평가할 수 있다. 따라서 본 연구에서는 앞서 점수 계산 방법에 의해 추출하는 키워드 수를 변경하면서 분류율의 변화를 살펴보았다. 그리고 질의 메일 인식률 및 처리 시간 관점에서 효율적인 키워드 및 특징 벡터 차원을 평가하였다. 즉, 가급적 짧은 시간 내에 메일 인식률을 높일 수 있는 특징 벡터를 구성하여 실험하였다.

표 8. 실험에 사용된 코넷 관련 질의 메일

범주	질의부류	학습 데이터 수	테스트 데이터 수
상품 및 개통	개통설치	15	20
	요금	15	20
	홈페이지관련	15	20
	해지	15	20
	정보변경	15	20
	PC패키지	15	20
	이벤트	15	20
	로밍	15	20
	상품 안내	15	20
일반 환경 설정	메일/뉴스	15	20
	랜카드	15	20
	UMS	15	20
	네트워크설정	15	20
	접속프로그램	15	20

4.2 신경망 학습 및 질의 메일 분류

서식 분류를 수행하기에 앞서 먼저 표 8의 학습 데이터를 이용하여 신경망 학습을 수행하였다. 그림 4는 학습율과 오류 목표 값이 각각 0.05, $1*e^{-5}$이고 최대 600 epochs 만큼 학습을 수행할 경우 오류가 특정 값으로 수렴됨으로써 학습이 수행되는 과정 및 결과를 보여준다.

그림 5는 600 epochs까지의 학습을 수행한 후 표 7의 280개 테스트 데이터에 대해 분류 실험을 수행한 결과이다. 실험 결과 키워드의 수를 증가할수록 대체적으로 인식률은 증가하였다. 하지만 키워드의 수가 65개와 80개일 경우에는 이전의 인식률에 비해 인식률이 상대적으로 다소 감소하였는데, 이는 몇 가지 메일 부류에 공통적으로 존재하는 키워드도 선택되어 오히려 혼동을 초래하였기 때문이었다.

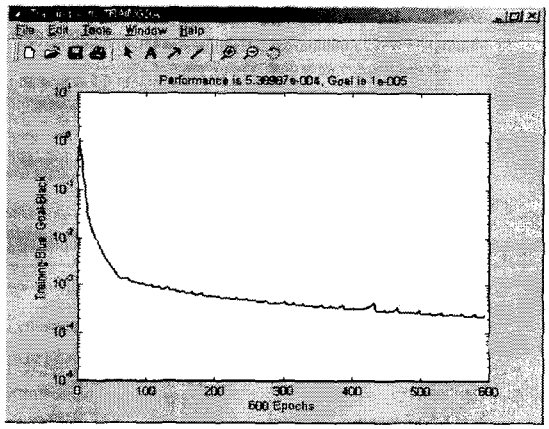


그림 4. 역전파 학습에 따른 오류 값의 변화

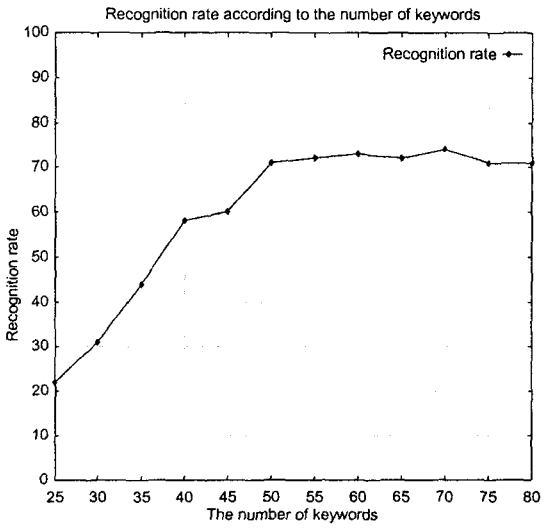


그림 5. 키워드 수에 따른 질의 메일 인식률

오인식의 가장 큰 원인은 키워드가 추출되지 않음으로 인하여 잘못 구성된 특징 벡터에 기인하였다. 키워드가 추출되지 않은 이유는 키워드가 등록되지 않은 이유가 오인식의 67%로 가장 많았고, 비록 키워드가 등록되어 있지만 키워드 변형에 기인한 오인식은 28%였고, 기타 신경망 결정 경계의 중복에 의한 오인식이 5%였다.

실험 결과 키워드가 70개일 경우 인식률은 73.1%였는데, 이는 비록 '상품 및 개통문의'와 '일반 환경설정'에 한정된 질의 메일을 분류한 것이지만 응용면에서 긍정적인 결과를 얻을 수 있었다. 왜냐하면, 전자 질의 메일의 분류는 지문 인식 혹은 홍채인식과 같이 높은 신뢰도를 필요로 하기 보다는 신뢰도가 다소 떨어지더라도 신속한 답장을 바라는 사용자의 요구에 부응하였기 때문이다. 실험 결과 얻은 인식률을 바탕으로 질의 메일 분류의 만족도를 조사하기 위하여 데이터(test set) 280 개의 메일을 분류한 후 각 유형의 질의를 답변하는 답변자를 통해 알아보았다. 즉 개통설치, 요금, 홈페이지, 해지, 정보변경, pc 패키지, 이벤트, 로밍, 상품안내, 매일/뉴스, NIC, UMS, 네트워크설정, 접속프로그램 등에 대해 모두 30명의 답변자에게 본 실험 결과를 업무에 적용할 경우의 만족도를 알아보았다. 본 실험 결과를 바탕으로 메일 답변자 30 명에게 업무의 만족도를 조사하였다. 본 연구에서 제안한 메일 자동 분류 방법 및 실험 결과가 업무에 도움을 주는 정도에 따라 '매우만족', '만족', '보통', '불필요', 그리고 '전혀필요 없음' 등으로 답변하였다. 분류된 후 메일 답변에 대해 긍정적인 반응을 나타낸 사람이 30명중 27로 90%를 나타내고 부정적인 반응을 보인 사람이 3명에 불과하였다.

선택한 키워드 수가 70 개일 경우 최고 73.1%라는 인식률에 대해 설문 응답자 중 90%가 질의 메일 업무처리에 있어서 긍정적인 반응을 보였다. 또한 빠른 답변을 원하는 사용자 입장에서 볼 때 업무의 효율성 증대라는 관점에서 긍정적인 결과로 판단된다. 이 결과를 바탕으로 정인식된 70%는 부류별 지정 답변자가 담당을 하고, 오인식된 30%에 대해서는 모든 부류에 대해서 알고 있는 답변자가 전담하는 시스템으로 바꿀 수 있어 인원 배치를 효율적으로 할 수 있다.

한편, 이와 같은 실험 결과는 Chakrabarti[26] 등이 시도한 특정 단어를 이용하여 Fisher 분포 테이블에서 문서의 범주를 분류하는 방법이 보인 평균 인식률 64.4% 보다 약 9% 정도 인식률이 높음을 확인하였다. 다만 Chakrabarti 등의 방법은 질의 메일과 같은 비교적 정형화된 데이터 보다는 웹 문서처럼 일반적인 데이터를 분류할 목적으로 개발 되었다는 차이점이 있다. 비록 동일한 데이터는 아니지만 본 연구가 인식률이 상대적으로 높은 이유는 문서 인식에 필요한 특징을 추출하는데 있어서 사용자의 지식을 이용함으로써 보다 양질의 특징을 추출할 수 있다는 데 기인한다.

5. 결 론

본 논문에서는 한국통신 코넷 관련 질의 메일 중 '상품 및 개통문의'와 '일반 환경 설정'에 국한하여 메일을 분류하는 방법을 제시하였다. 질의 메일 분류는 다른 일반 문서의 분류와는 달리 질문 유형이 어느 정도 정형화되어 있으므로 메일을 분류할 수 있는 키워드를 효과적으로 찾을 수만 있다면 질의 메일 분류 또한 효율적으로 수행할 수 있다.

객관적인 입장에서 질문 유형(부류)을 결정하고 합리적인 방법으로 키워드를 선택하기 위하여 한국통신 코넷에서 수개월 동안 사용자의 질의 메일을 답변해오고 있는 메일 답변자들을 대상으로 통계적인 방법으로 키워드를 분석하였으며, 동일한 클래스 내에서는 빈번하게 나타나는 키워드를 선택함으로써 안정적인 특징을 추출하려 하였고, 서로 다른 클래스 내에서는 상대적으로 중복되지 않는 키워드 위주로, 즉 각 클래스마다 상대적으로 유일하게 나타나는 키워드를 위주로 특징을 추출하여 신경망 입력으로 사용하였다.

14개의 질의 메일 부류에 대해 210개의 학습 데이터와 280개의 테스트 데이터 등 모두 490개의 데이터를 이용하여 실험을 수행한 결과 제안한 방법에 의해 선택한 키워드의 수가 70개일 경우 73.1%의 인식률을 얻을 수 있었다. 이는 응용 측면에서 고무적인 결과로 여겨진다. 왜냐하면, 전자 문의 메일의 분류는 지문 인식 혹은 홍채 인식과 같이 높은 신뢰도를 필요로 하기 보다는 신뢰도가 다소 떨어지더라도 신속한 답장을 바라는 사용자의 요구에 부응하기 때문이다. 또한 답변자 90%가 질의 내용에 따라 분류되어 담당자에게 보내어 지면 업무 처리가 수월해 지는 것으로 응답했다.

비록 본 연구에서는 가급적 객관적인 데이터에 근거하여 키워드를 추출하고자 하였지만 시스템이 적

응적으로 키워드를 결정할 수 있는 능력을 가지고 있지 않다. 따라서 질의 메일 간 유사도 계산 및 적응적 키워드 결정 방법에 관한 연구가 필요하다. 또한 앞서 설명한 방법에 의해 키워드를 선택한 후 유전자 알고리즘 방법에 의한 최적의 키워드 선택 방법에 관한 연구가 필요하다.

참 고 문 헌

[1] M. Blosseville. G. Hebrail, M. Monteil, and N. Penot., "Automatic Document Classification : Natural Language Processing, Statistical Analysis, and Expert System Techniques used together," SIGIR'92, pp.185-192, 1992.

[2] N. Fuhr, "Models for Retrieval with Probabilistic Indexing," Information Processing and Management, Vol. 25, No 1, pp.207-218, 1989.

[3] D. Harman, "Ranking Algorithms," in Information Retrieval : Data Structures and Algorithm, Prentice Hall, 1992.

[4] P. Hayes and S. Weinstein, "CONSTRUE/TIS: A System for Content Based Indexing of a Database of News Stories," Second Annual Conference on Innovative Applications of Artificial Intelligence, pp.193-204, 1990.

[5] P. Hayes, P. Anderson, I. Nirenburg, and L. Schmardt. "TCS: A Shell for Context-based Text Categorization," Proceedings of the 6th IEEE Conference on Artificial Intelligence Applications, Santa Monica, March, pp.254-261, 1990.

[6] J. Hobbs, D. Appelt, M. Tyron, J. Bear and D. Israel, "FASTUS : System Summary," Proc. of Fourth Message Understanding Conference, pp.169-174, 1992.

[7] R. Hoch., "Using IR Techniques for Text Classification in Document Analysis," SIGIR'94, pp.163-172, 1994.

[8] P. Jacobs., "Using Statistical Methods to Improve Knowledge Based News Categorization," IEEE Expert, April, pp.154-163, 1993.

[9] L. Larkey, W. Croft, "Combining Classifiers in Text Categorization," SIGIR'96, pp.189-197, 1996.

[10] D. Lewis, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task," SIGIR'92, pp.257-263, 1992.

[11] B. Masand, "Classifying News Stories Using Memory Based Reasoning," SIGIR'92, pp.192-204, 1992.

[12] M. Maron, "Automatic Indexing : An Experimental Inquiry," Journal of the ACM, pp.267-274, 1961.

[13] G. Salton. "Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer," Addison Wesley, 1989.

[14] 한정기, 박민규, 조광제, 김준태, "구문패턴과 키워드 집합을 이용한 통계적 자동 문서분류의 성능 향상," 한국정보처리 학회 논문지, 제7권 제4호, pp.1151-1158, 2000.

[15] 신진섭, 이창훈, "단어의 연관성을 이용한 문서의 분류," 한국정보처리학회 논문지 제6권 제9호, pp.2422-2429, 1999.

[16] W. Cohen., "Learning Trees and Rules with Set-Valued Features," AAAI-96, pp.199-215, 1996.

[17] T.Mitchell, "Machine Learning," McGraw-Hill, 1997.

[18] S. Weiss, C. Kulikowski, "Computer Systems That Learn : Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems," Morgan Kaufmann, 1991.

[19] C. M. Bishop, "Neural Networks for Pattern Recognition," Clarendon Press, OXFORD, 1995.

[20] M. T. Hagan, H. B. Demuth, M. H. Beale, "Neural Network Design," PSW Publishing Company, 1996.

[21] J. T. Tou, R. C. Gonzalez, "Pattern Recognition Principles," Addison-Wesley Publishing Company, 1981.

[22] J. A. Freeman, D. M. Skapura, "Neural Net-

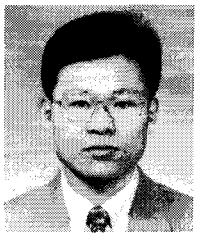
works, Algorithms, Applications, and programming Techniques," Addison-Wesley Publishing Company, 1992.

[23] R. O. Duda, P. E. Hart, D. G. Stork, "Pattern Classification 2nd Edtion," Wiley-Interscience, 2001.

[24] M. Nadler, E. P. Smith, "Pattern Recognition Engineering", Wiley-Interscience, 1993.

[25] L. Fausett, "Fundamentals of Neural Networks, Architectures, Algorithms, and Application", Prentice Hall, 1994.

[26] S. Chakrabarti, B. Dom, R. Agrawal, P. Raghavan, "Scalable Feature Selection, Classification and Signature Generation for Organizing Large Text Databases into Hierarchical Topic Taxonomies," Proc. VLDB'98, pp.163-178, 1998.



변 영 철

1993년 제주대학교 정보공학과 졸업(학사)
1995년 연세대학교 대학원 컴퓨터과학과 졸업(석사)
2001년 연세대학교 대학원 컴퓨터공학과 졸업(박사)
2001년 9월~2002년 11월 한국전자통신연구원 선임연구원
2002년 12월~현재 제주국립대학교 통신컴퓨터공학부 전임강사
관심분야 : 문서인식, 홍채인식, 신경회로망, 시맨틱 웹, 모바일응용서버 등

홍 영 보

1996년 제주대학교 정보공학과 졸업(학사)
2001년 서강대학교 정보통신대학원 정보통신과 졸업(석사)
2001년 8월~현재 KT 연구원
관심분야 : MMS, 신경회로망, 모바일응용서비스 등