濟州大學校 尖端技術硏究所 論文集, 第14卷, 1號, pp.147-155, 2003. 6
Cheju Nat'l Univ. Res. Inst. Adv. Tech. Jour., Vol.14, No.1, pp.147-155, 2003. 6

역전파 신경망을 이용한 Kornet 사용자 질의 메일 자동 분류에 관한 연구

김 영 민* · 변 영 철** · 이 상 준** · 홍 영 보***

A Study on Automatic Classification of Query Mail using Backpropagation Neural Network

Young-min Kim* · Yung-Cheol Byun** · Sang-Joon Lee** · Young-Bo Hong***

ABSTRACT

For increasing Internet Users, the increasingly asked E-Mail is needed to be classified automatically. This thesis is proposed to automatically classify the asked E-Mail of KORNET queried by users using backpropagation neural network. This experiment is use of the 490 asked E-mail. When selected keyword number is 70, the recognition rate is highest, 73.1%. Even though the Classification rate of the asked E-Mail is less than the iris recognition and fingerprint recognition, it is affirmativefor an asking User to satisfy with a quick reply.

Key Words : backpropagation neural network, query e-mail classification

Ⅰ. 서 론

인터넷 사용 증가에 따른 질의메일 증가에 대해 인터넷 상의 사이트 운영자는 이용자가 질문을 하기 전에 먼저 FAQ나 Q&A를 확인하기를 바라고 있으나, 이용자는 자신이 모르는 부분에 대해서는 사이트 관리자에게 메일을 보냄으로써 답을 손쉽게 얻으려고 한다. 그러나 사이트 운영에 있어서 질의메일 증가는 메일 마스터에게 지나친 업무 부하를 주었고 부수적인 메일 답변을 위해 많은 시간을 투자해야 했다.

이러한 문제로 인하여, 본 연구의 목적은 BP 신경망을 이용하여 Kornet과 관련된 질의메일을 효과적으로 분류하고 그에 따른 업무의 효율성을 향상시킬 수 있는 방안을 제시하는데 있다. 이를 위해, 먼저 한국통신 Kornet에 접수된 기존의 고객 상담 내용을 분석하여 전자메일 내에서 중요하다고 여겨지는 단어를 설문 조사에 근거하여 키워드 사전을 정의한다. 정의한 키워드 사전 지식을 이용하여 미지의 질의 메일 내용 중에서 표준화된 키워드를 추출한다. 생성된 키워드를 특징 벡터로 양자화하여 신경망의 입력으로

* 제주대학교 대학원 컴퓨터공학과 박사과정
Doctor course of Computer Science, Cheju Nat'l Univ.
** 제주대학교 통신컴퓨터공학부, 첨단기술연구소
Faculty of Communication and Computer Eng., Cheju Nat'l Univ., Res. Inst. of Adv. Tech.
*** 한국통신
KT(Korea Telecomm)

사용함으로써 미리 정의한 질의 부류 중 하나로 분류한다. 신경망에 의해 인식된 결과를 바탕으로 적절히 고객에게 답변 메일을 고객에게 전송하고, 실패한 메일은 전문 답변자에게 전송함으로써 효과적인 답변 분류 시스템 체계를 가능하도록 한다.

II. 관련 연구

정보통신기술의 발전은 온라인으로 생성되는 전자 문서의 양을 폭발적으로 증가 시켰으며, 수동으로 문서를 분류하던 종래의 방법 대신에 문서의 자동 분류 기술 개발을 요구하고 있다[1-4]. 문서의 자동 분류란 일반적으로 기계 학습을 이용하여 미리 학습시켜 둔 범주 중 하나로 문서를 분류 처리하는 것을 지칭한다. 이미 분류되어진 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 probability를 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.

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.

이외에도 분류 방법 중 널리 알려진 방법으로 결정 트리 학습법이 있다. 결정 트리는 정보 이론에 기반하여 귀납적 유도 학습방법으로 가장 많이 사용되어지는 것 중 하나로써 1949년 Shannob과 Weaver에 의해 처음으로 소개되었으며, 예는 1986년에 개발된 ID3[15]와 1993년에 C4.5 및 Cubist[16]등이 제안되었다. 잡음이 있는 데이터에 유리하며, 표현을 구별할 수 있는 학습능력이 뛰어난 점이 결정 트리의 특성으로 이 특성을 이용하여 문서의 범주화에 많이 사용되어 진다.

인공신경망[21, 24]은 인간의 두뇌를 모방하여 두뇌 활동의 매카니즘을 수학적으로 재현한 것으로 학습 경험을 바탕으로 새로운 입력에 대하여 만족 해를 스스로 구할 수 있으며 이는 질의메일 분류에 이용할 수 있다.

Kornet의 질의 메일의 경우 질문 유형이 확실히 구분되고, 메일 답변자가 질의메일을 분석할 때 90% 이상의 메일에 대해 핵심 키워드만을 인지하더라도 쉽게 메일의 유형을 인식할 수 있다는 특성이 있다. 따라서 기존의 범위가 제한되지 않는 일반적인 문서와 같이 문서 유사도를 동적으로 구하기보다는 메일 답변자의 지식에 근거하여 메일을 효율적으로 분석하고자 한다.

본 논문에서는 Fig. 1에서 질의메일 내용에 따른 분류에 기반을 둔 연구 결과를 제시한다. 전자 메일을 자동으로 분류하여 메일 답변자에게 전송하기 위한 시스템 처리 흐름도는 Fig. 2에서 제시하였다.

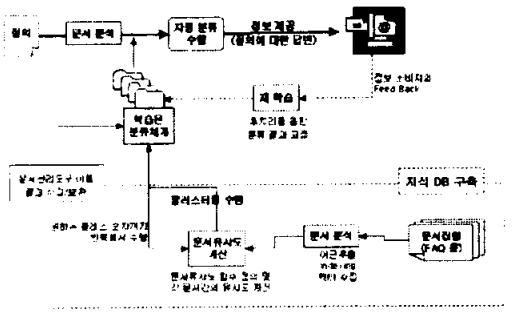

Fig. 1. ERMS process flow diagram

키워드 추출은 질의 메일 중에서 불필요한 용어나 낱말을 삭제하여 가장 중요한 단어나 요약이 가능한 단어로 변형시키는 작업으로 여기에서 생성된 단어는 그 단어 자체로 메일의 내용을 파악할 수 있다. 본 연구에서는 이를 이용하여 특징 벡터를 생성 한 후, 신경망을 이용한 패턴 분류[17-24]를 수행한다.

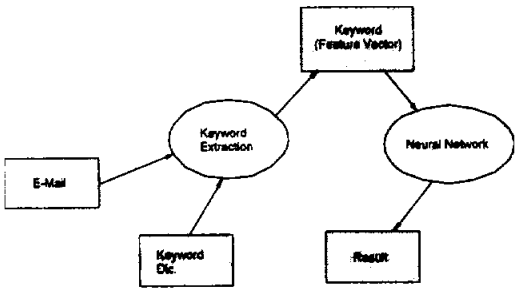

Fig. 2. Flow diagram for mail classification

III. 신경망을 이용한 메일 분류

3.1. 키워드 사전 정의

키워드 사전이란 질의 메일 내용 중에서 반복적으로 나타나거나 중요한 핵심 용어를 정의해 놓은 것으로 키워드 사전을 얼마나 잘 정의 하느냐에 따라서 질의메일을 정확히 분류를 할 수 있다. 문서의 유형이 상대적으로 많고 정형화되어 있지 않은 일반 문서의 경우 문서간 유사도에 의해 자동으로 부류를 나눌 수 있어야 하지만 Kornet 사용자 질의 메일의 경우에는 질문 유형에 따른 메일 유형이 상대적으로 정형화 되어 있었다. 또한 상담원의 설문 결과 키워드가 메일 분류 시 중요한 역할을 수행함을 알 수 있었다. 키워드 사전을 구성하기 앞서 키워드에 의해 분류되어져야 할 부류들이 구성되어야 한다. 우선 부류를 구성하기 위해 다음과 같은 기준에 준하여 설문조사로 분류하였다.

(1) 유사한 내용으로 많이 질의 되는 유형

(2) 메일 내용 중 특정키워드가 최소 10개 이상이 되는 유형

(3) 특정 부류로 지정이 되어야 한다고 생각되는 유형

설문 대상자 30명에게2001년 1월에서 2월 사이의 Kornet '상품 및 개통'과 '일반환경설정' 항목에 대해 위 기준에 따라 세부 항목으로 구분하도록 한 결과 Fig. 3과 같이 분포를 얻을 수 있었다.

각각의 부류는 개통설치, 요금, 홈페이지, 해지, 정보변경, pc패키지, 이벤트, 로밍, 상품안내, 메일/뉴스, NIC, UMS, 네트워크설정, 접속프로그램 등 상위 14개로 나눌 수 있었다. 기타 부류로는 원클릭CD, 위성인터넷 등이 있었으나 기타 부류를 선택한 사람이 15명을 넘지 않아서 제외했다.

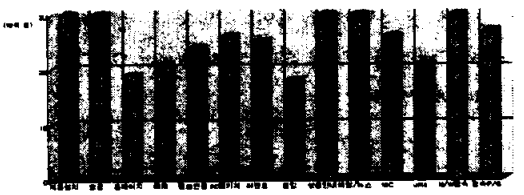

Fig. 3. Mail categories for user query mail researched on the enquetes

14개 부류에 대해 각 부류별 키워드를 선정하기 위하여 2001년 1월에서 2월 사이의 메일 중에서 각 부류에 대해 30개 메일, 전체 30X14개 메일을 30명의 메일 답변자에게 보여주고 키워드를 선별하도록 하였다. Fig. 4는 키워드 선별의 예이다. 이렇게 30명 각자가 선정한 키워드를 조사한 후 빈도수가 상대적으로 많은 키워드를 표시한 결과 Table. 1과 같은 결과를 얻었다.

제 목 : [illegible]

질 문

제가 저번에 [illegible] 이벤트로 응모해서 [illegible] 게임을 여러 번해서 다른 것은 안걸리고 유스호스텔 회원권인가 할인권인가를 2장 당첨되었거든요

근데 무소식이군요.

어떻게 된거죠

Fig. 4. Keyword selection

Table. 1. Keywords for each mail category

1	개통설치	설치,문제,지역,코넷,사용,ADSL,개통,..
2	요금	요금, 고지,월사용,조정,지로,합산,할인,..
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,패스티벌,경품,대잔치,..
8	로밍	해외,로밍,국외,GRIC,출장,..
9	상품	하이텔,b&a,무료,상품권,홈넷,..
10	메일/뉴스	뉴스그룹,메일서버,NNTP,POP,SMTP,..
11	NIC	랜카드,모뎀,외장형,구동,드라이버,..
12	UMS	UMS,FAX,03030,핸드폰,..
13	네트워크설정	TCP /IP,서브넷,IP주소,..
14	접속프로그램	ENTER,WINPOET,NTS,..

여러 부류 중에서 이벤트 부류 내용을 보면 Fig. 5와 같다. 개인별로 보는 관점에 따라 핵심적인 단어 빈도수가 조금씩 다르지만 응답자가 선택한 단어들 가운데 '당첨'을 가장 많이 선택하였고 '참가'는 가장 적게 선택하였다.

응답자 / 키워드	답변자1	답변자2	답변자3	답변자4	답변자5	...	답변자30	합계
당첨	21	23	21	19	24		21	681
패스티벌	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		15	378
특판	14	14	17	14	15		14	543
백만돌파	16	13	21	22	23		19	587
축하	21	19	14	20	20		17	524

Fig. 5. Keyword frequencies for "이벤트" category

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.

3.2. 특징벡터 구성 및 신경망 이용한 분류

신경망 학습을 위하여, 또는 고객의 질의메일을 2장에서 분류한 14 가지의 부류(Class) 중 하나로 분류하려면 키워드를 자동으로 추출한 후 특징 벡터를 구성해야 한다. 이를 위해 먼저 14 가지의 부류 각각에 대하여 빈도수를 기준으로 상위 n개의 키워드를 추출한다. 14 가지의 질의메일 부류(Class) 중 1번째 부류의 키워드 ki에 대한 점수 S^1_{ki}는 식 (1)을 이용하여 계산한다.

$$S^1_{ki} = \frac{f_{ki}}{K} \tag{1}$$

위에서 f_{ki}는 ki 키워드의 빈도수 합계(Fig. 5)를 의미하며, K는 키워드의 빈도수 중 최대 빈도수를 의미한다. 식 (1)의 의미는 특정 부류의 질의 메일에서 자주 나타나는 키워드는 메일 분류 시 양질의 특징을 포함하므로 높은 점수를 부여한다. 식 (1)에 의해 n번째 부류의 i번째 키워드 점수인 S^n_{ki}는 식 (2)을 만족한다.

$$0 < S^n_{ki} \leq 1 \tag{2}$$

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 3>의 경우 한 개의 키워드만을 선택할 경우 '당첨' 키워드의 점수는 681/681로 최대가 되므로 이를 선택한다. 이처럼 14 개의 부류에 각각 대해 S^n_{ki}점수에 근거하여 키워드를 추출한다.

Keyword		Total Frequency	s_{ki}^{7}
k1	당첨	681	1
k2	이벤트	672	0.987
k3	경품	663	0.959
k4	제스티벌	653	0.930
k5	행사	629	0.924
k6	응모	614	0.902
k7	신문광고	603	0.885
k8	해당권	594	0.872
k9	백만불짜	587	0.862
k10	회원	577	0.847
k11	상품권	569	0.836
k12	특판	543	0.797
k13	마이	524	0.769
k14	대전지	521	0.765

Fig. 6. Keyword points for "이벤트" category(S^{7}_{ki})

Fig. 6은 실제로 앞서 Fig. 5에서 구한 이벤트 부류 키워드에 대해 점수 S^{7}_{ki}을 구한 결과이다. 실제로 14 개의 모든 부류에 대해 점수 크기를 기준으로 각 부류별 상위 10 개의 키워드를 추출한 결과 중복되는 키워드를 한 번만 셀 경우 모두 112 개의 키워드를 얻을 수 있었다.

Fig. 7은 부류(Class)의 수가 3이고 점수 크기를 기준으로 상위 4 개의 키워드를 추출한 예이다.

Mail Category	Keywords based on S^{n}_{ki}
C1	a, b, c, d
C2	a, c, e, f
C3	a, b, g, h

Fig. 7. Extracted keywords using S^{n}_{ki}

Fig. 7의 경우 a, b, c, d, e, f, g, h 모두 8 개의 키워드가 존재한다. 이 중 a는 모든 부류에 나타났으며 b는 C1과 C3에 나타났다. 이 때 Fig. 7에서 a는 비록 각 부류에서 빈도수가 가장 높기는 하지만 모든 부류에 나타나므로 질의 메일을 분류할 수 있는 변별력이 없다. 따라서 변별력이 높은 키워드를 선택함으로써 인식률을 높이고 키워드 수를 줄여 신경망에 의한 처리 시간을 줄이기 위하여 식 (3)에 의해 키워드 점수를 계산한다. f^{A}_{ki}는 14개 부류에서 특정 키워드(ki)가 나타난 빈도수를 나타낸다. 예를 들어 Fig. 7에서 a의 빈도수(f^{A}_{k1})는 3이다.

$$S^{A}_{ki} = \frac{1}{f^{A}_{ki}} \tag{3}$$

위 공식에 의해 S^{A}_{ki}는 식 (4)을 만족한다.

$$0 < S^{A}_{ki} \leq 1 \tag{4}$$

식 3.3에 의해서 여러 클래스에 공통적으로 나타나는 키워드는 상대적으로 점수가 작게 계산되며, 반대로 오직 한 클래스에만 나타나는 경우에는 가장 큰 점수인 1을 얻을 수 있다. S^{A}_{ki}점수에 근거하여 신경망의 입력으로 주어질 특징 벡터를 구할 수 있다. 가령 Fig. 7의 a, b, c, d, e, f, g, h 각각의 점수는 1/3, 1/2, 1/2, 1, 1, 1, 1, 1이다. 이를 근거로 7개의 키워드를 선택할 경우에는 a를 제외한 나머지, 즉, b, c, d, e, f, g, h를 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 질의 메일 F에 키워드 c, e, g, h이 존재할 경우 신경망의 입력으로 주어지는 특징 벡터는 Fig. 8과 같다.

Keyword	b	c	d	e	f	g	h
count	0	1	0	1	0	1	1

Feature Vector : "0 1 0 1 0 1 1"

Fig. 8. Sample feature vector string

위의 경우 0은 질의 메일에 특정 위치의 키워드가 존재하지 않음을 의미하며, 1은 존재함을 의미한다.

본 연구에서는 점수 S^{A}_{ki}에 근거하여 다양한 차원의 특징 벡터를 구성하여 신경망의 입력으로 사용

하였다. Fig. 9에 기본적인 신경망 구조를 제시하였다. 질의 메일에서 추출하는 키워드의 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+1개가 된다.

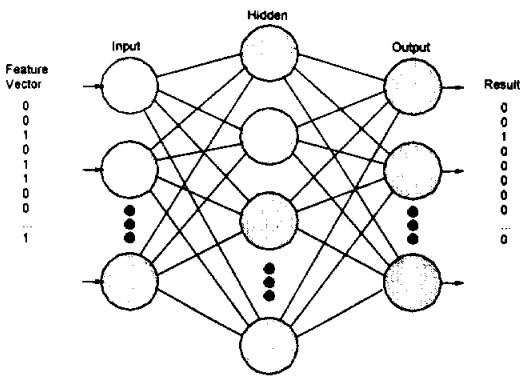

Fig. 9. Neural network structure for mail classification

본 연구에서는 구체적으로 신경망 입력 층의 노드 수에 해당하는 특징 벡터의 차원을 25에서 80까지 다양하게 바꿔서 실험을 수행하였다. 실험 결과에 대해서는 다음 장에서 설명한다.

IV. 실험 결과

본 연구에서는 제안하는 메일 분류 방법을 검증하기 위하여 Fig. 10과 같이 한국통신 Kornet에 접수된 질의 메일을 이용하여 특징 추출 및 분류 실험을 수행하였다.

메일 표본 중에서 490매일을 다시 2개 부분으로 나누어 먼저 15X14개의 메일은 학습 데이터(Training Set)로 사용을 하였고 나머지 20X14개의 메일은 학습된 신경망을 통하여 결과를 얻기 위한 실험 데이터(test set)으로 사용하였다. PENTIUM III 650에서 C++를 이용하여 키워드 추출 및 특징 벡터 구성 알고리즘을 구현하였으며, 추출된 특징 벡터는 MATLAB을 이용하여 학습 및 분류 실험에 이용하였다.

Topics	Query Category	Learning Data Number	Test Data Number
상품 및 서비스	계좌업무	15	20
	요금	15	20
	홈페이지업무	15	20
	해지	15	20
	정보망업	15	20
	PC메카지	15	20
	이벤트	15	20
	요청	15	20
	상품 안내	15	20
일반 광범 설정	메일/뉴스	15	20
	펜카드	15	20
	UMS	15	20
	네트워크업무	15	20
	접속프로그램	15	20
Total		210	280

Fig. 10. Query mail of Kornet for experiment

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

서식 분류를 수행하기에 앞서 먼저 Fig. 10의 학습 데이터를 이용하여 신경망 학습을 수행하였다. Fig. 11은 학습율과 오류 목표값이 각각 0.05, $1* e^{-5}$이고 최대 600 epochs 만큼 학습을 수행할 경우 오류가 특정 값으로 수렴됨으로써 학습이 수행되는 과정 및 결과를 보여준다.

Fig. 12는 600 epochs 까지의 학습을 수행한 후 Fig. 10의 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 키워드로 인하여 혼동을 초래하였기 때문이었다.

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 키

워드가 등록되어 있지만 키워드 변형에 기인한 오인식은 28%였고, 기타 결정 경계의 중복에 의한 오인식이 5%였다.

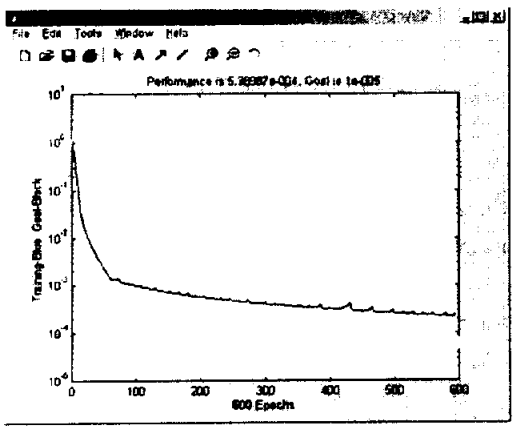

Fig. 11. Error rate according to times of BP learning

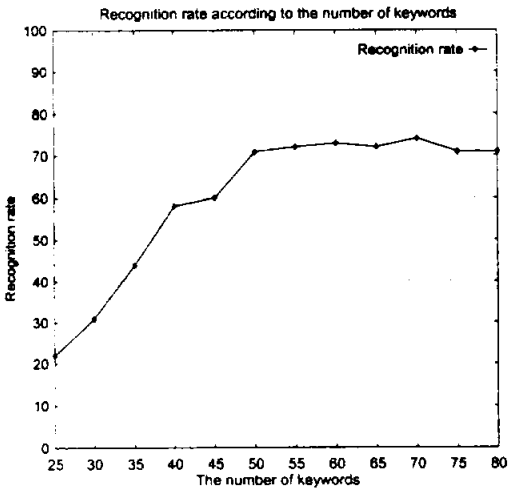

Fig. 12. Mail recognition rate according to keyword number

실험 결과 키워드가 70개일 경우 인식률은 73.1%였는데, 이는 비록 '상품 및 개통문의'와 '일반 환경설정'에 한정된 전자메일을 분류한 것이지만 응용 면에서 긍정적인 결과를 얻을 수 있었다. 왜냐하면, 전자 질의메일의 분류는 지문 인식 혹은 홍채 인식과 같이 높은 신뢰도를 필요로 하기 보다는 신뢰도가 다소 떨어지더라도 신속한 답장을 바라는 사용자의 요구에 부응하기 때문이다

V. 결 론

본 논문에서는 한국통신 Kornet 관련 질의 메일 중 '상품 및 개통문의'과 '일반 환경 설정'에 국한하여 메일을 분류하는 방법을 제시하였다. 질의메일 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 Kornet에서 수 개월 동안 사용자의 질의 메일을 답변해오고 있는 메일 답변자들을 대상으로 설문 조사 방법을 통해 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.

14 개의 질의 메일 부류에 대해 210 개의 학습 데이터와 280 개의 테스트 데이터 등 모두 490 개의 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.

본 연구의 오인식의 원인은 다음과 같다. 우선 하나의 메일에 여러 가지 서로 상이한 내용을 질문할 경우, 낯선 용어, 혹은 예상치 못한 키워드를 이용하여 질문할 경우 오인식 가능성이 높아질 수 있다. 또

한 최근의 잘못된 어법과 단어의 오용도 오인식의 원인이 된다.

비록 본 연구에서는 가급적 객관적인 데이터에 근거하여 키워드를 추출하고자 하였지만 시스템이 적응적으로 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.

참고문헌

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

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

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

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

5) P. Hayes, P. Anderson, I. Nirenburg, and L. Schmardt. 1990, TCS: A Shellfor Context-based Text Categorization, Proceedings of the 6thIEEE Conference on Artificial Intelligence Applications, Santa Monica, March, pp.254-261

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

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

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

9) L. Larkey., W. Croft, 1996, Combining Classifiers in Text Categoriza tion, SIGIR'96, pp.189-197

10) D. Lewis, 1992, 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

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

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

14) 한정기, 박민규, 조광제, 김준태, 2000, 구문패턴과 키워드 집합을 이용한 통계적 자동 문서 분류의 성능 향상, 한국정보처리 학회 논문지, 제7권 제4호, pp.1151-1158

15) W. Cohen., 1996, Learning Trees and Rules with Set-Valued Features, AAAI-96, pp.199-215

16) T.Mitchell., 1997, Machine Learning, McGraw-Hill

17) C. M. Bishop, 1995, Neural Networks for Pattern Recognition, Clarendon Press, OXFORD

18) 김대수, 1993, 신경망 이론과 응용 II, 하이테크정보

19) M. T. Hagan, H. B. Demuth, M. H. Beale, 1996, Neural Network Design, PSW Publishing Company

20) J. T. Tou, R. C. Gonzalez, 1981, Pattern Recognition Principles, Addison-Wesley Publishing Company

21) J. A. Freeman, D. M. Skapura, 1992, Neural Networks, Algorithms, Applications, and programming Techniques, Addison-Wesley Publi-

shing Company
22) R. O. Duda, P. E. Hart, D. G. Stork, 2001, Pattern Classification 2nd Edtion, Wiley- Interscience
23) M. Nadler, E. P. Smith, 1993, Pattern Recognition Engineering, Wiley-Interscience
24) L. Fausett, 1994, Fundamentals of Neural Networks, Architectures, Algorithms, and Application, Prentice Hall