한국 퍼지 및 지능시스템 학회 2002, Vol. 12, No. 5, pp. 432-441

자기구성 지도를 이용한 인터넷 FAQ의 자동응답 및 개념적 브라우징

Automatic Response and Conceptual Browsing of Internet FAQs Using Self-Organizing Maps

안준현, 류중원, 조성배

Joon-Hyun Ahn, Jungwon Ryu, Sung-Bae Cho

연세대학교 컴퓨터과학과

Department of Computer Science, Yonsei University

요 약

최근 인터넷상의 정보를 가공하여 사용자에게 효율적으로 제공하는 서비스들이 많아지고 있지만, 컴퓨터에 익숙하지 않은 사용자들은 이러한 서비스를 쉽게 이용하지 못하기 때문에 사용자들을 돕는 시스템이 필요하다. 예를 들어, 웹사이트의 경우 전자우편을 통한 사용자들의 질문에 대해 관리자가 직접 답을 해줘야 하는데, 사용자의 증가로 질의응답 업무의 양이 커지고 있다. 본 논문에서는 이를 해결하기 위하여 사용자의 질의를 자동으로 분류하여 응답하고 사용자가 FAQ를 개념적으로 브라우징할 수 있도록 하는 시스템을 제안한다. 이 시스템은 다양한 크기의 질의 메일을 정형화된 크기로 만들기 위한 키워드 클러스터링 자기구성 지도(SOM)와 이를 실제 해당 답변 클래스로 분류하는 전자 우편 분류 SOM의 이단계 구조로 구성되어 사용자의 질의에 해당하는 답변을 자동으로 전송할 수 있으며, 사용자가 이차원상에 표현된 문서 지도를 이용하여 쉽게 전체 자료의 분포를 파악하여 검색할 수 있다. 실제 한 달간 수집한 2,206개의 한메일넷 질의 데이터에 대한 실험 결과, 95%의 분류율을 보여 그 유용성을 볼 수 있었으며, 단계별 검색이 가능하여 사용자가 효율적으로 검색할 수 있음을 확인할 수 있었다.

ABSTRACT

Though many services offer useful information on internet, computer users are not so familiar with such services that they need an assistant system to use the services easily. In the case of web sites, for example, the operators answer the users' e-mail questions, but the increasing number of users makes it hard to answer the questions efficiently. In this paper, we propose an assistant system which responds to the users' questions automatically and helps them browse the Hanmail Net FAQ (Frequently Asked Question) conceptually. This system uses two-level self-organizing map (SOM): the keyword clustering SOM and document classification SOM. The keyword clustering SOM reduces a variable length question to a normalized vector and the document classification SOM classifies the question into an answer class. Experiments on the 2,206 e-mail question data collected for a month from the Hanmail net show that this system is able to find the correct answers with the recognition rate of 95% and also the browsing based on the map is conceptual and efficient.

Key Words : 문서 분류, Semantic Web, 자기구성 지도, E-mail 자동 응답

1. 서 론

컴퓨터의 보급과 함께 인터넷의 대중화로 다양한 정보가 제공되면서 많은 사람들이 정보통신 기반 서비스를 이용하게 되었다. 하지만 이러한 서비스에 익숙하지 않은 사용자가 자신이 원하는 정보를 찾는 것은 그리 쉬운 일이 아니다. 따라서 ISP (Internet Service Provider)나 PC통신업체 등 정보통신 서비스 업체는 사용자가 접하는 문제들을 해결하기 위해서 전화상담 창구를 운영하고, FAQ나 게시판의 형태로 유형화된 질문에 대한 답을 제공하거나, 전자우편으로 사용자의 질의에 답하기도 한다. 그러나 사용자의 엄청난 증가로 인해 이러한 서비스 제공에 많은 인력과 시간이 필요하게 되었다.

한메일넷의 경우, 2000년 현재 500만명 이상의 사용자가 이용하고 있다. 하루 평균 200통 정도의 사용자 질의를 처리해야 했는데, 이를 실시간으로 자동응답하는 시스템을 구축한다면 사용자에게 만족도 높은 서비스를 제공할 수 있을 뿐만 아니라, 관리자도 중복된 일을 피할 수 있어 업무의 효율을 높일 수 있다. 따라서, 사용자와 관리자의 편의를 위해 질의 자동응답 시스템을 개발할 필요가 있다.

접수일자 : 2001년 10월 29일
완료일자 : 2002년 9월 5일

본 논문에서는 자기구성 지도(Self-Organizing Map)를 이용하여 인터넷 질의의 자동응답과 FAQ를 개념적으로 브라우징할 수 있는 시스템을 제안한다. 이를 통하여 클래스별 데이터(문서)의 개수도 불균등하고 각 데이터의 크기도 다른 전자우편의 자동응답 가능성과 개념적 브라우징 방법의 유용성을 보이고자 한다. 본 논문에서 제안한 시스템에서는 자기구성 지도로 서로 연관된 문서들을 군집화 하여 이로부터 적절한 응답을 보내거나, 사용자 자신이 직접 브라우저 상에 지도 형태로 제시된 검색어들을 단계적으로 선택함으로써 원하는 문서들의 전체적인 분포를 직관적으로 파악하여 검색을 수행할 수 있다. 이때 문서지도의 특정영역에 대한 상세 정보를 계층적으로 디스플레이하여 효율적 검색이 가능하도록 한다.

이 논문의 구성은 다음과 같다. 먼저 2절에서는 인터넷 질의의 특성을 분석하고 자동응답 시스템의 개요를 기술한다. 3절에서는 자기구성 지도를 사용한 질의 분류 방법을 제안하고, 4절에서는 개념적 브라우징에 대해서 설명한다. 5절에서는 실제 질의 데이터에 대한 실험 결과를 보이고, 결론을 맺는다.

2. 인터넷 질의와 자동응답

문서 자동분류는 새로운 문서를 미리 정의된 클래스로 대응시키는 일이다[1]. 일반적으로 문서 자동분류 시스템은 미리 수집한 문서 집합으로부터 이들의 패턴을 학습하는 단계와, 새로운 문서에 대한 분류를 수행하는 단계로 구성된다. TREC, OHSUMED, Reuters와 같은 대규모의 문서 데이터베이스를 사용하여 여러 문서 특징 선택 및 분류 알고리즘에 관련한 연구가 진행되어왔다[2, 3, 4, 6]. Lewis, Yang 등은 통계적 혹은 정보이론 기반의 방법들을 사용하여 문서로부터 최적의 특징을 선택하는 연구를 수행하였고[2, 5], 이를 기반으로 k-Nearest Neighbor[6], 의사결정 트리[4, 7, 8], Least Square Fit[6], 베이지안 방법[4, 8, 6], Support Vector Machine[9, 6], 인공 신경망[10, 6] 등의 다양한 모델들에 대한 문서 분류기로서의 효용성이 연구되어왔다. 위의 연구들은 모든 문서 분류 문제를 해결 할 수 있는 완벽한 모델을 제시하지는 않았지만, 기계학습 방법을 문서 분류 분야에 성공적으로 적용함으로써 실생활에 응용할 수 있는 가능성을 제시하였다.

인터넷 질의 자동응답 문제 역시 문서 분류 문제들의 하나로, 흥미로운 연구 과제이다. Ask Jeevs[11] 서비스나 Pragmatech Software 사의 RFP Machine[12] 등이 이러한 서비스를 하고 있는데, Ask Jeevs 서비스는 자연어 형태로 사용자 질의를 받아서 처리한 후, 답변을 발견할 수 있는 사이트를 알려주며, RFP Machine은 데이터베이스에 저장되어 있는 답변과 제안서의 질의를 매칭시켜 자동으로 제안서를 작성해 원하는 형태로 만들어주는 프로그램이다. 그러나 아직까지 일반 고객을 대상으로 한 서비스 회사의 질의 응답 시스템에 응용된 사례는 찾아보기 힘들다. 인터넷 질의 문서를 분류하는 경우, 사용자들은 인터넷을 통하여 정보를 얻기 위해 단순한 몇 개의 키워드를 입력하거나, 홈페이지 관리자에게 메일을 보내기도 한다. 이러한 짧은 길이의 질의를 사용하여 방대한 양의 웹 컨텐트 중에서 사용자가 원하는 것을 찾아내기란 쉬운 일이 아니며 무엇보다 질의 메일의 특성을 잘 파악하는 것이 중요하다.

이를 위하여 본 논문에서는 사용자의 질의 안에 포함된 키워드들의 문맥정보를 사용하였다. 문맥정보로 학습된 자기구성 지도를 사용하여 질의 문서의 특징을 표현하고 이를 해당 답변 클래스로 분류하거나 사용자가 웹브라우저 상에서 답변 클래스를 직접 브라우징 하도록 하였다.

2.1 자기구성 지도

일반적으로 신경망은 입력값과 그에 상응하는 출력값을 가지고 학습을 수행한다. 학습은 각 입력값에 대해 올바른 출력값이 나올 수 있도록 가중치를 변화시키는 것이다. 이러한 학습 방법을 교사 학습이라고 하는데, 인공 신경망(artificial neural network)은 입력 자극에 대한 인간의 두뇌의 신경 세포(neuron)의 자극 전달 과정에 착안하여 고안되었다. 그러나 학습을 위하여 올바른 출력값(desired output)이 미리 주어져야 한다는 점은 생물학적 의미에 맞지 않다. 이 같은 문제에 대하여 튜보 코호넨은 자기구성 지도를 제안하였는데, 그는 자기구성 지도를 사용하여 신경망이 자기 스스로 학습할 수 있음을 비교적 간단하게 제시하였다 [13]. 자기 구성이라는 말은 신경망이 주어진 입력에 대해 올바른 출력값이 제공되지 않고도 학습함을 의미 한다. 뿐만아니라, 노드들이 반응하는 순서나 위치를 통해 데이터의 위상을 보존(topology preserving)하는 특징을 가지고 있다. 이러한 특성 때문에 자기구성 지도는 데이터의 시각화(visualization)나 위상 보존매핑(topology-preserving mapping)이 필요한 산업의 여러 분야에서 이용되고 있다.

자기구성 지도의 구조도는 그림 1과 같고, 다음의 수식을 통해서 경쟁학습(competitive learning)한다. 자기 구성지도는 입력층과 출력층, 두 개의 층으로 구성되어 있다.

$$m_i(t+1) = m_i(t) + \alpha(t) \times n_{ci}(t) \times \{x(t) - m_i(t)\} \quad (1)$$

여기서 $\alpha(t)$는 학습률을 나타내는 함수, $n_{ci}(t)$는 이웃 함수, $m_i(t)$는 노드의 가중치, $x(t)$는 입력 벡터값이다[14]. $n_{ci}(t)$에서 c는 승리자 노드의 인덱스인데, 승리자는 다음의 수식으로 얻을 수 있다[15]. 입력 벡터가 신경망에 들어오면, 입력 벡터와 모든 노드들과의 유클리드 거리(euclidean distance)를 계산한다. 최소 거리를 갖는 노드가 승자로 선택되고 수식에서 m_c로 나타낼 수 있다.

$$\|x - m_c\| = \min_i \|x - m_i\| \quad (2)$$

수식 (1)과 (2)와 같이 승리자 노드를 선택하여 승리자와 그 이웃 노드들의 가중치를 업데이트 한다.

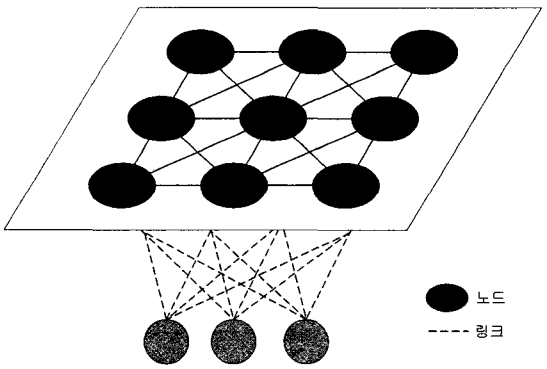

그림 1. 자기구성 지도의 구조도

2.2 인터넷 질의의 특성 분석

일반적으로 인터넷 질의의 형태는 클래스별로 데이터의 수에 편차가 심하고, 각 데이터 내의 키워드 수도 심한 차이를 보인다. 그 예로서 한메일넷 시스템에서 한달간 수집한 질의 형태와 분포를 표 1과 2에 정리하였다. 표 1에서 보면, 전체적으로 클래스별로 데이터의 수가 매우 불균등하며, 각 데이터 내의 키워드수도 2개에서 305개로 차이가 심함을 볼 수 있다. 표2는 질의 빈도에 따른 질의의 분포를 보여준다. 질의의 빈도가 많은 클래스가 빈도수의 절반을 차지함을 볼 수 있다. 여기에서 개별응답 질의는 분류할 필요없이 관리자에게 포워딩해야할 클래스들이다. 그리고 통계적으로 처리하기 힘든 질의란 빈도수가 너무 적은 클래스들을 의미한다.

정리하자면, 인터넷 질의의 특징은 다음과 같다. 먼저, 표 2에서 볼 수 있듯이 빈도수가 높은 특정 클래스에 질의가 편중되는 경향이 있다. 또한, 클래스가 정의되지 않았거나 해당 웹사이트 사용과는 관련없는 다양한 질의가 존재하므로 이들의 패턴 추출에 어려움이 있다. 그리고 일반 사용자들이 작성하므로 통신상의 속어나 약어, 맞춤법에 맞지 않는 표현을 많이 포함하고 있다.

2.3 인터넷 질의 자동응답 시스템

인터넷 질의 자동응답 시스템은 자동분류와 FAQ 브라우징 시스템으로 구성되어 있다. 자동분류 시스템은 비슷한 질문이 들어왔을 때 자동으로 분류하여 답장을 보내고 필요한 경우에만 관리자가 직접 답변을 하도록 한다. FAQ 브라우징 시스템은 이러한 분류구조를 계층적으로 제시하여 질의에 대한 답변을 개념적으로 검색할 수 있도록 한다.

제안한 질의 자동응답 시스템의 전체적인 구조는 그림 2와 같다. 전자우편으로 수집한 질의 자료를 입력 스트림 추출과정을 통해 전자 우편 분류 SOM의 입력 벡터로 인코딩한다. 인코딩시, SOM을 이용해서 크기가 큰 키워드 벡터들을 일정한 크기의 벡터로 축약한다(3절 참조). 인코딩된 입력 벡터를 이용하여 전자 우편 분류 SOM을 학습시키고, 학습된 지도를 이용하여 질의 메일을 분류하여 적절한 답변 메일과 매칭시켜서 자동응답하거나 FAQ 브라우징에 이용한다.

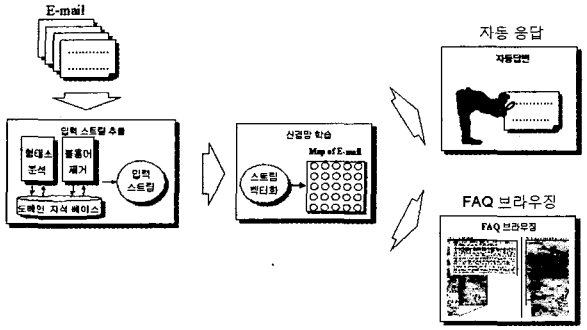

그림 2. 인터넷 사용자 질의 자동응답 시스템의 개요

표 1. 질의 클래스 내용

번호	이 름	질문수	키워드수
0	PCS로메일수신알림	52	27.10
1	POP서버주소	67	10.69
2	POP읽어오기에러	4	26.25

표 1. 질의 클래스 내용(계속)

번호	이 름	질문수	키워드수
3	메일수신알림	20	17.45
4	메일프로그램설정	179	14.89
5	건의사항	15	28.07
6	기타	288	21.00
7	메일수신이안됨	26	13.58
8	비밀번호	67	34.78
9	용량문의	42	22.76
10	접속에러	26	14.88
11	URL에러	4	18.00
12	내용확인불가능	0	·
13	네이버등록	3	6.00
14	메일주소	8	33.38
15	메일친구	2	15.50
16	삼성	42	21.24
17	외국에서사용	1	2.00
18	이용자검색	33	15.64
19	이용자정보요청	21	22.29
20	자동답장에대한항의	4	61.25
21	자동포워딩	5	21.80
22	자신도모르게삭제	2	7.00
23	주소록관리	3	6.67
24	폴더관리	9	13.22
25	홈페이지	14	12.64
26	휴지통비우기시복구	8	23.50
27	DTOP	19	20.00
28	기타	12	16.25
29	스팸	3	305.00
30	인터넷폰	11	20.00
31	카페	33	23.94
32	답장필요없음	37	40.78
33	IE5.0자동저장기능	28	16.46
34	IE버그	18	22.82
35	사용자정보겹침	8	19.63
36	카페접속후메일로그인	1	28.00
37	쿠키설정방법	25	12.00
38	다운로드방법	5	14.20
39	저장방법	2	10.00
40	제공용량	2	12.00
41	메일수신확인여부	4	17.25
42	보낸편지보관함에체크	2	15.50

표 1. 질의 클래스 내용(계속)

번호	이 름	질문수	키워드수
43	복수사용자에게전송방법	1	5.00
44	사용자에게읽지않음으로	0	·
45	외부로의수신여부	4	10.50
46	외부로의수신전송여부	4	13.50
47	외부수신확인여부	3	15.33
48	재전송방법	5	10.80
49	전송제한수	1	24.00
50	전송제한용량	3	11.33
51	전송취소	3	9.00
52	주소확인요망	29	9.76
53	회람	0	·
54	비밀번호확인	321	15.30
55	아이디변경	99	9.91
56	아이디삭제	173	9.88
57	아이디중복	37	22.70
58	아이디확인요청	115	15.39
59	사용자정보수정방법	115	9.77
60	서명수정	8	11.25
61	URL	3	9.33
62	닉네임변경	5	9.60
63	CC, BCC	2	18.00
64	파일덧붙이기	8	31.63
65	Return mail	55	16.07
66	User unknown	6	49.17
67	재질문	49	45.59

표 2.. 빈도수에 따른 질의 분포

클래스 속성	클래스 개수	데이터 개수(비율)
빈도 높은 질의	6	1,002(44.9%)
개별응답 질의	7	585(26.2%)
통계적 처리가 힘든 질의	36	127(5.7%)
기 타	18	492(23.2%)
계	67	2,206(100.0%)

3. 자동응답

본 논문에서 제안한 자동응답 시스템은 키워드 추출 모듈, 키워드 클러스터링 모듈과 질의 분류 모듈으로 구성된다. 키워드 추출 모듈에서는 질의 문서로부터 문서 분류에 의미를 가지는 키워드들을 추출한다. 키워드 클러스터링 모듈은 키워드들의 유사성을 조사하여 유사한 키워드들을 군집화하는 역할을 한다. 즉, 같은 범주에 속하는 키워드들을 하나의 키워드로 그룹화하는 것이다. 마지막으로 전자 우편 분류 모듈은 질의 문서의 키워드 히스토그램을 입력으로 받아 신경망을 학습한다. 시스템의 구성은 그림 3과 같다.

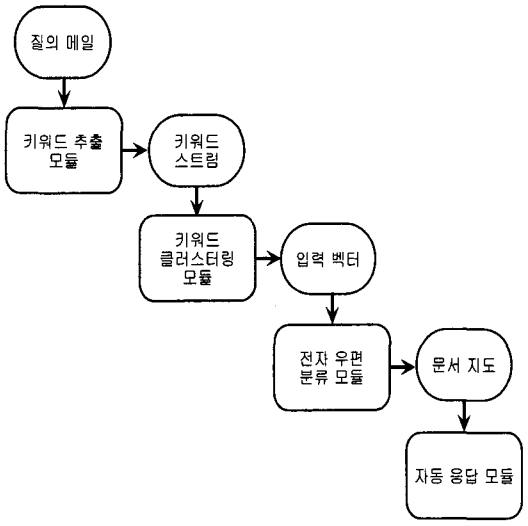

그림 3. 자동응답 시스템의 처리 절차

3.1 키워드 추출

질의 문서는 문자들의 조합으로 이루어져 있기 때문에 이를 신경망에 수정 사용하기 위해서는 문서를 정규화된 수치 벡터 형태로 변환시켜야 한다. 이를 위하여 질의 문서로부터 키워드(단어 및 어구)들을 추출하는 작업이 필요하다. 이 과정을 통해 조사나 어미 등, 문장의 의미에 영향을 미치지 않는 불용어들과 불필요하게 반복되는 키워드들이 제거된다. 그림 4와 5는 키워드의 추출과정과 추출 예를 보여준다.

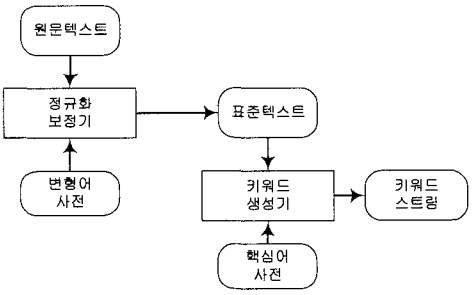

그림 4. 원문 e-mail로부터 키워드를 추출하는 과정

- 입력 메일

017에서 E-mail통보
서비스를 신청했는데
전문이 통보되나요?
아님 통보만 해주나요?
궁금합니다.
메일로 보내주시면
고맙겠습니다.

- 키워드

017 e-mail통보 서비스
신청 전문 통보
통보 해주
메일

그림 5. 키워드 추출의 예

3.2 키워드 클러스터링

그림 5와 같이 키워드 추출 과정이 끝난 후에는 키워드의 집합을 수치화된 벡터로 표현하는 작업이 필요하다. 키워드를 수치화된 벡터로 표현하는 방법에는 벡터 스페이스 모델을 예로 들 수 있으나[16], 질의 메일들은 키워드의 수가 워낙 많기 때문에 계산시 필요로하는 메모리 용량 및 이에 따른 계산량이 매우 크다는 단점이 있다. 따라서 본 논문에서는 SOM을 이용한 인코딩 방법을 사용하여 입력 벡터의 차원을 줄이고자 하였다. 즉, 키워드 클러스터링 자기구성 지도의 크기가 키워드 범주의 개수가 된다. 자기구성 지도의 입력은 각 키워드들에 대한 문맥 정보가 되고 출력은 문맥 정보에 의해 분류된 키워드들이 된다. 이렇게 하면 SOM에 의해서 유사한 키워드들은 지도의 같은 노드나 근접한 위치의 노드에 할당된다[15, 17, 18, 19, 20, 21]. SOM의 학습은 수식 (1)에 의해서 수행된다.

본 논문에서 키워드 x_i의 문맥정보는 다음과 같이 정의하였다[17].

$$X(i) = \begin{bmatrix} E\{x_{i-1}|x_i\} \\ \varepsilon x_i \\ E\{x_{i+1}|x_i\} \end{bmatrix} \quad (3)$$

수식 (3)에서 ε은 크기가 작은 임의의 실수이고, 키워드 x_i에 대하여 생성되는 입력 벡터 $X(i)$는 x_i의 선행자와 후행자의 빈도 평균 및 x_i 자신으로 구성되어있다. 여기서, 선행자의 빈도 평균은 질의 자료의 모든 x_i에 대하여 x_i의 바로 앞에 나오는 키워드들의 벡터 값을 합하여 평균을 구한 값이다. 그리고 후행자는 자료의 모든 x_i의 바로 뒤에 나오는 키워드들의 벡터 값을 합하여 평균을 구한 값이다. 선행자와 후행자는 x_i의 특징을 나타내는 값으로서, 모든 키워드에 대해서 앞과 뒤에 나오는 키워드들을 살펴봄으로써 문맥 정보를 얻을 수 있다[14]. 이 입력벡터를 SOM에 입력하여 문맥 정보에 의해 분류된 키워드의 지도를 얻을 수 있다.

키워드 클러스터링에 대한 예는 다음과 같다. 여기서 문제의 범위를 그림 5로 한정하면, 총 키워드의 개수는 8개가 되고, 8비트로 표현할 수 있다. 8비트는 다음과 같다.

1	2	3	4	5	6	7	8
017	E-mail	통보	서비스	신청	전문	해주	메일

세 번째 키워드 "통보" 에 대해서 선행자 및 후행자 빈도 평균을 구할 경우를 살펴보자. "통보" 키워드는 총 3번 나왔다. 첫 번째 문장에서는 "e-mail" 을 선행자로 가지고, "서비스" 를 후행자로 가진다. 두 번째 문장에서는 "전문" 을 선행자로 가지고 후행자는 가지지 않는다. 그리고 세 번째 문장에서는 "해주" 를 후행자로 가지고 선행자를 가지지 않았다. 세 번의 키워드 빈도에 대한 평균을 구하면, 선행자 빈도 평균의 경우

1	2	3	4	5	6	7	8
0	1/3	0	0	0	1/3	0	0

와 같이 된다. 그리고 후행자 빈도 평균의 경우

1	2	3	4	5	6	7	8
0	0	0	1/3	0	0	1/3	0

와 같이 된다. 그러므로, 키워드 클러스터링 자기구성 지도의 입력 벡터는 다음과 같이 된다.

1~8	1	2	3	4	5	6	7	8	1~8
선행자 빈도평균	0	0	ε	0	0	0	0	0	후행자 빈도평균

3.3 질의 분류

입력 벡터는 키워드 클러스터링 자기구성 지도의 결과를 통해서 생성된다. 일단 키워드 클러스터링 자기구성 지도가 만들어 진 다음에는 자기구성 지도의 각 노드에 어떤 키워드들이 매핑되어 있는지를 알 수 있다. 이 경우, 각 질의 메일의 키워드들이 키워드 클러스터링 자기구성 지도의 몇 번째에 할당되는지를 알 수 있게 되고, 각 키워드들에 대한 히스토그램을 구할 수 있게 된다[14, 22, 23, 24]. 예를 들면, 키워드 클러스터링 자기구성 지도가 3×3이라고 하고, 질의 문서의 키워드들이 각각 (0, 0)에 2개, (1, 2)에 1개가 나타났다고 한다면, 인코딩은

(0, 0)	(0, 1)	(0, 2)	(1, 0)	(1, 1)	(1, 2)	(2, 0)	(2, 1)	(2, 2)
2	0	0	0	0	1	0	0	0

과 같이 된다. 앞에서와 같이 인코딩된 입력 벡터의 차원은 자기구성 지도의 크기와 같다. 즉, $m \times n$의 자기구성 지도에 의해 생성될 수 있는 입력 벡터는 $m \times n$의 1차원 벡터가 되는 것이다. 분류율 향상을 위해서 각 키워드들의 빈도수에 대하여 이들이 얼마나 분류에 중요한지를 샤넌 엔트로피(Shannon's Entropy)[16]를 통해 보완하였다[14, 17, 22]. 샤넌 엔트로피 Ew는 다음과 같이 구해진다.

$$E_w = H_{\max} - H(w)$$
$$H_{\max} = \ln(\text{총 클래스수}) \quad (4)$$
$$H(w) = \sum_i \frac{n_i(w)}{\sum_j n_j(w)} \ln \frac{n_i(w)}{\sum_j n_j(w)}$$

$H(w)$는 키워드 w에 대한 엔트로피 값을 나타내고, $ni(w)$는 클래스 i에서의 w의 빈도수를 나타낸다. 여기에서 $H_{\max} = \ln 67$가 된다.

또한, 이렇게 만들어진 히스토그램은 데이터에 민감하다는 단점이 있다. 질의 응답의 경우, 사람마다 질의하는 방식이 틀리기 때문에 유사하지만 편차가 클 수 있는데, 단순히 빈도수로 히스토그램을 정수화시킬 경우 특정 값이 너무 큰 경우가 발생 할 수 있다. 따라서 다음과 같은 가우시안 커널[25]로 2차원 블러링(blurring)을 수행하였다[26].

0.25	0.5	0.25
0.5	1	0.5
0.25	0.5	0.25

따라서, i번째 질의 메일에 대한 입력 벡터 Vi는 다음과 같이 만들어진다.

$$V_i = \sum_w \left(F_w \times \frac{G_i}{E_w} \right) \quad (5)$$

여기서 Gi는 i번째 원소의 커널 값을, Fw는 키워드 w의 빈도수를, Ew는 w에 대한 샤넌 엔트로피 가중치 값을 나타낸다. 엔트로피 값과 블러링에 의한 보완을 통해 만들어진 각 히스토그램은 질의 메일 하나 하나의 고유 입력 벡터가 된다.

입력 벡터를 전자 우편 분류 SOM으로 학습하면, 각 질의 메일은 SOM의 특정 노드에 매핑된다. 이 때, 같은 클래스의 질의들이 같은 노드 혹은 서로 근접한 위치에 있는 노드에 매핑되게 된다. 이렇게 학습된 SOM에 새로이 사용자가 보낸 질의 메일을 인코딩하여 입력하면, 질의 메일은 SOM의 특정 출력 노드에 매핑되게 되고, 이 노드가 가리키는 클래스를 할당받게 된다. 이 클래스에 대한 답변 메일이 최종 응답 메일이 되는 것이다. 학습 데이터에 대해 87.04%였던 인식률이 샤넌 엔트로피 적용한 후 93.47% 까지 크게 향상되었고, 블러링을 사용하였을 때 95.01% 까지 상승하였다.

4. 개념적 브라우징

시스템은 실제 웹사이트 사용자들의 질의 문서들을 자기구성 지도로 인코딩한 질의 문서 지도, 질의 문서 지도의 효율적 검색을 위한 브라우징 인터페이스, 사용자가 찾은 질의 클래스의 답변을 사용자의 웹브라우저에 보내주는 답변 서버로 이루어져 있다. 사용자가 처음 FAQ 검색 페이지에 접속하면 웹브라우저는 브라우징 인터페이스와 질의 문서 지도를 읽어오고, 사용자는 브라우징 인터페이스를 사용해서 자신이 생각하는 질의가 있을 만한 위치를 검색한다. 사용자가 최종적으로 질의 문서를 선택하면, 브라우징 인터페이스는 선택된 질의 클래스를 답변 서버로 보내게 되고, 답변 서버는 해당 질의 클래스에 대한 답변을 사용자의 웹브라우저로 보내준다. 그림 6은 시스템의 구성도를 보여준다.

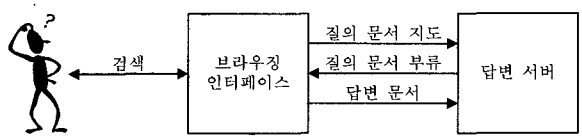

그림 6. FAQ 브라우징 구성도

4.1 질의 문서 지도

자기구성 지도는 입력 데이터를 서로 연관된 것끼리 모아주기 때문에[16], 서로 연관된 질의 문서들끼리 근접한 곳에 모여서 군집을 형성한다. 따라서 사용자는 지도상에 나타난 키워드들의 위치 관계, 군집 등을 고려한 개념적 검색을 할 수 있다.

4.2 브라우징 인터페이스

사용자에게 단순히 키워드만을 보여주는 것이 아니라 각 문서의 특징 키워드들을 자기구성 지도[16]를 통해 인코딩해서 각 키워드 간의 연관성을 표시한 후, 각 키워드와 영역간의 거리 정보를 색깔로 표현해서 각 키워드를 중심으로 한 문서 군집을 제시한다. 사용자는 이 이차원 문서 지도를 통해 키워드 간의 연관도와 문서 군집 정보를 파악해서 개념적 검색을 한다. 또한 문서 지도를 최상위 단계, 중간 단계, 마지막 단계의 3단계 구조로 구축하여 큰 문서 지도를 보다 효율적으로 브라우징 할 수 있도록 하였다.

구현 측면에서 볼 때, 기존 시스템이 대부분 키워드의 링크를 따라가면서 검색을 하도록 하기 때문에 사용자가 링크를 좇아갈 때마다 매번 웹브라우저와 웹서버간의 네트웍 연결 시간을 낭비하게 되고 웹서버 쪽에도 부담을 많이 준다. 하지만 본 시스템에서는 자바 애플릿을 사용해서 시스템 초기화에 많은 시간을 사용하긴 하지만 일단 초기화되고 나면 최종 질의 문서를 선택해서 답변을 받을 때까지 서버와 분리된 상태로 검색을 할 수 있게 되어 여러 번 검색을 할 경우에는 훨씬 빠른 속도로 검색할 수 있다. 또한 검색 과정은 웹서버와의 연결없이 진행되므로 서비스를 제공하는 측면에서 웹서버의 부담이 많이 줄어든다.

브라우징 인터페이스는 사용자가 검색 화면에 접속하면 질의 문서 지도와 답변 서버 정보를 읽어들이고, 질의 문서 지도의 내용을 읽어 문서 지도 화면을 그린다. 사용자는 이 문서 지도 화면을 이용해서 검색을 하면서 자신이 원하는 답변을 찾게 된다. 그림 7은 브라우징 인터페이스의 개요를 보여준다.

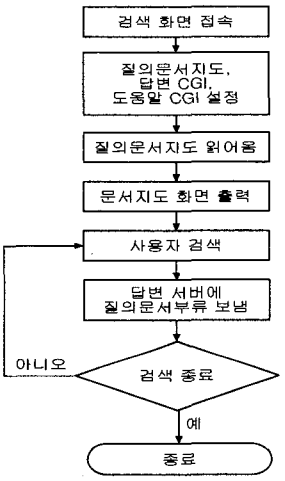

그림 7. 인터페이스 개요

화면구성은 그림 8과 같이 문서 지도 부분과 답변 출력창 부분으로 나뉘어져 있다. 문서 지도 부분은 자기구성 지도를 통해 인코딩된 질의 문서의 지도를 시각화한 것으로 사용자가 검색을 하는 영역이고 답변 출력창은 사용자가 마지막 단계에서 선택한 영역에 대한 질의의 답변을 보여준다.

4.3 레이블링

레이블링은 각 영역을 가장 잘 대표하는 하나의 키워드를 선택하는 과정이다. 이를 위하여 각 영역에 매핑된 질의 문서들에 나타나는 키워드들의 빈도수를 기반으로 하여 해당 영역을 레이블링할 키워드를 결정하였다[15]. 그러나, 어떤 키워드는 단 한번 나타나더라도 한 클래스의 특징을 나타낼 수 있고, 어떤 키워드는 빈도수는 높지만 그 클래스의 특징을 잘 나타내지 못하는 경우가 있다. 그래서 빈도수와 함께 다음과 같이 키워드 w의 중요도 값 Iw를 계산해서 각 영역의 레이블 값을 계산하였다.

$$I_w = F \times \frac{DocFin}{DocFout} \quad (6)$$

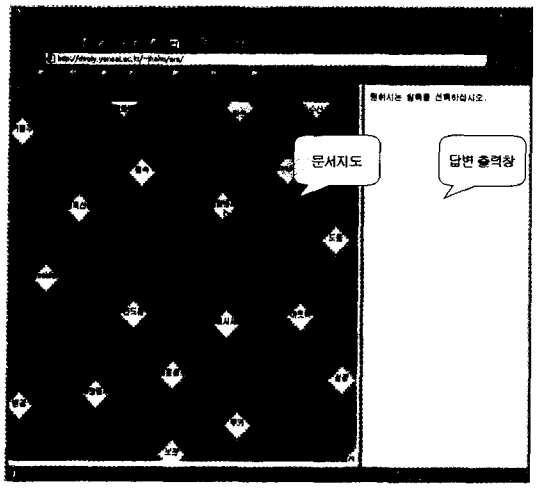

그림 8. 브라우징 인터페이스

여기서, F는 노드 내의 키워드 w의 총 출현 빈도수, $DocFin$은 노드 내에서 키워드 w를 가진 문서들의 수, $DocFout$은 노드 밖에서 키워드 w를 가지는 문서들의 수를 나타낸다. 이 중요도 값은 각 단계별로 군집 영역의 크기를 정해서 각 영역 내에서 계산을 하게 되고 각 영역에서 가장 큰 중요도값을 가진 키워드가 화면에 보여지게 된다 (그림 8 참고). 단계별 군집 영역의 크기는 최상위 단계 30×30, 중간 단계 15×15, 마지막 단계에서는 1×1의 크기로 정하였다.

노드의 색깔은 키워드와 해당 영역의 거리 정보를 나타내는데, 밝은 영역일수록 키워드와 연관성이 높은 문서들을 나타내고 어두운 영역일수록 키워드와 연관성이 낮은 문서들임을 나타낸다. 또한 지도상의 검은 영역은 각 군집간의 경계를 나타낸다.

4.4 답변 서버

답변 서버는 사용자가 검색한 영역의 질의 문서 클래스의 답변을 보여준다. 사용자가 문서 지도의 마지막 단계에서 선택한 질의 문서 클래스의 정보를 받으면 사용자가 보낸 데이터 중에서 질의 문서 클래스를 파싱하고 이 클래스에 대한 답변이 존재하는지를 검사한다. 만약 답변이 존재하지 않으면 관리자에게 직접 문의하라는 도움말을 보여주고 그렇지 않으면 해당 답변을 보여준다. 이 CGI는 Perl[27]을 이용해서 작성하였는데, Perl은 문자열 처리가 쉽고 CGI를 작성하는데도 용이하다.

그림 9는 답변 서버의 개요를 보여준다.

5.1 자동응답

시스템 구축을 위한 데이터로는 국내 최대 규모의 포탈 사이트 중의 하나인 한메일넷의 질의 메일을 사용하였다. 실제 약 한달간 한메일넷을 통해 2,232개의 질의 문서를 수집하였다. 하지만 전체 데이터 중에서 분류가 필요없는 2개 클래스, 26개의 데이터를 제외한 67개 클래스, 2,206개의 질의 메일만 실험의 대상으로 삼았다. 키워드 클러스터링 SOM의 크기는 10×10으로 하였으며, 키워드 동의어의 개수와 전자 우편 분류 SOM의 입력 벡터의 차원을 100으로 하였다. 먼저 2,206개의 질의 메일 전체를 학습 데이터로 사용하여, 전자 우편 분류 SOM의 크기를 최적화 하였는데, 그 결과는 표 3과 같다.

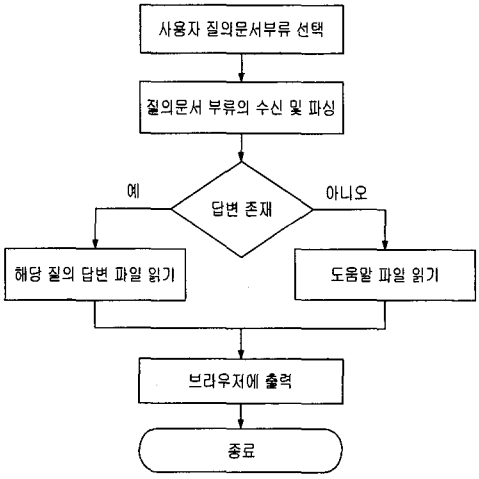

그림 9 답변 서버 개요

5. 실험 결과

표 3에서 150×150의 SOM을 사용한 경우가 가장 높은 인식률을 보였다. 이 경우의 결과를 클래스 별로 분석하면 표 4와 같다. 질의의 빈도수가 많은 클래스의 인식률과 기타의 인식률이 낮음을 알 수 있다. 이는 빈도수가 많은 특정 4개의 클래스가 낮은 인식률을 보였기 때문인데 구체적으로, 아이디 변경, 아이디 삭제, 아이디 중복, 아이디 확인요청 클래스가 여기에 해당된다. 이 클래스들은 내용이 서로 유사하므로 키워드로 특징을 추출해내는 것이 어려워 분류가 잘 되지 않았던 것으로 생각된다. 그리고 기타에는 18개의 클래스들이 있는데, 이들도 역시 나머지 49개 클래스와 연관된 매우 유사한 내용을 가지고 있었다.

표 3. 학습 데이터에 대한 인식률

지도 크기	인식률	
100×100	1,553/2,206	70.40%
120×120	2,014/2,206	93.74%
150×150	2,098/2,206	95.01%
160×160	2,075/2,206	94.06%

표 4. 클래스별 인식률

종 류	해당 클래스 수	인식률	
빈도가 높은 질의	6	954/1,002	94.0%
개별 응답 질의	7	561/585	95.9%
통계적으로 처리하기 힘든 질의	36	124/127	97.6%
기 타	18	459/492	93.3%
합 계	67	2,098/2,206	95.01%

전자 우편 분류 SOM의 크기를 150×150으로 하여 학습 데이터와 테스트 데이터로 나누어서 두 번째 실험을 하였다. 학습 데이터는 1,545개로 하였고 테스트 데이터는 661개로

하였다. 실험은 자기구성 지도의 양자화 오류(Quantization Error)의 임계치 변화에 따라 진행되었고, 기각률은 양자화 오류값에 의해 결정된다. 즉, 그림 10과 표5와 같이 임계치 값의 한계를 0.5, 0.4, 0.3으로 낮춰가며 실험하였는데, 각각 307, 397, 482개의 데이터를 기각하였다. 기각되지 않은 데이터들로부터 인식률을 계산하였다. 임계치가 낮을수록 기각률은 높아지지만 기각되지 않은 데이터에 대해서는 높은 인식률을 보여준다.

표 5. 클래스별 인식률

임계치	기각률		인식률	
0.5	307/661	46.4%	207/354	58.5%
0.4	397/661	60.1%	169/264	64.0%
0.3	482/661	72.9%	148/179	82.7%

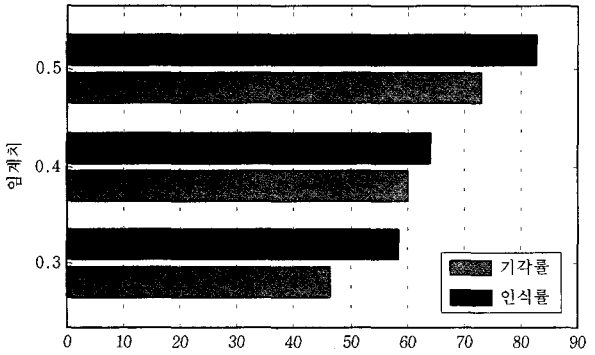

그림 10. 테스트 데이터의 인식률 [%]

5.2 개념적 브라우징

그림 11~13은 "비밀번호 변경" 방법을 찾는 브라우징 인터페이스의 단계별 검색 화면을 보여준다. 그림 11은 문서 지도의 최상위 단계로 150×150 크기의 지도에 전체 질의 문서 데이터 중에서 가장 큰 클래스를 이루는 키워드들과 각 키워드를 중심으로 한 군집을 색으로 표현하였다. 이 단계에서는 전체 질의 문서의 분포를 파악하고 각 키워드와 군집 정보를 보고 개념적으로 자신이 찾고자하는 정보가 있을 것 같은 영역을 선택하게 한다.

실험에서는 "비밀번호" 키워드가 표시된 영역을 선택하였다. 그러면 시스템은 문서 지도 중간 단계로 넘어가면서 그림 12와 같은 화면을 보여준다. 문서 지도 중간 단계는 최상위 단계에서 선택한 "비밀번호" 영역을 중심으로 50×50 크기의 지도를 보여준다. 최상위 단계보다 확대된 문서 지도를 통해 상세한 정보를 보여주게 된다. 여기서 "변경" 영역을 선택하면 그림 13과 같이 문서 지도 마지막 단계로 넘어간다. 마지막 단계에서는 이전 단계에서 선택한 "변경" 영역을 중심으로 10×10 크기의 문서 지도를 보여주고, 질의 문서가 존재하는 모든 영역의 대표 키워드를 보여준다. 사용자가 이 키워드를 선택하면 관련 답변이 출력된다. 실험에서는 "비밀번호" 영역을 선택해서 답변창에 답변을 얻었다. 객관적인 평가를 위해서는 좀더 체계적인 분석이 이루어져야 하나, 몇 가지 예에 대한 사용 결과를 통해 사용자가 문서 지도를 계층적으로 검색해서 개념적 검색이 가능함을 알 수 있었다.

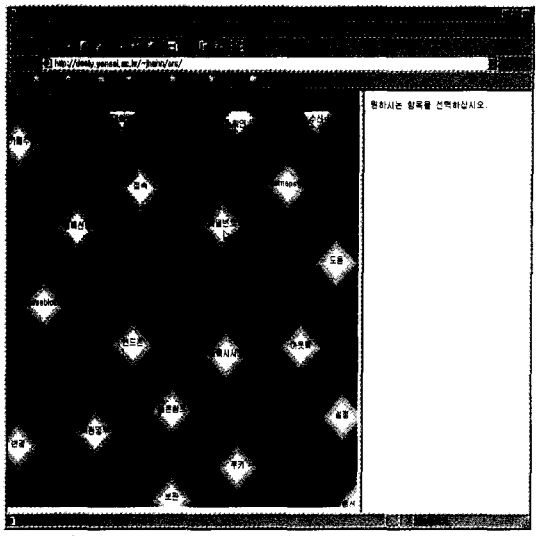

그림 11. 문서 지도 최상위 단계

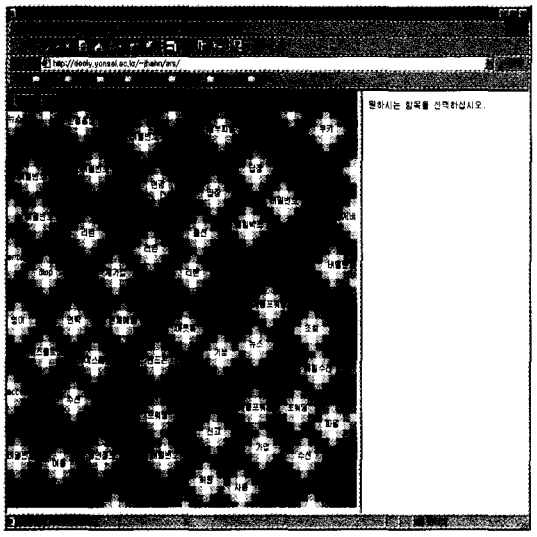

그림 12. 문서 지도 중간 단계

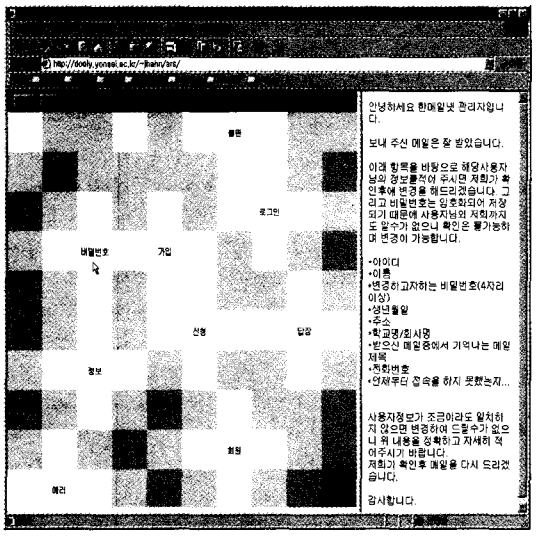

그림 13. 문서 지도 마지막 단계

6. 결 론

본 논문에서는 인터넷 질의 메일의 자동응답과 브라우징을 가능하게 하는 시스템을 제안하였다. 자동응답 시스템의 경우 한메일넷 데이터를 사용한 실험 결과, 학습 데이터에 대해서는 높은 분류율을 보이는 반면, 테스트 데이터에 대한 분류율은 받아들일 만 하지만 기각률이 너무 크다. 본 논문에서는 자동응답 시스템에 대한 가능성을 제시하였지만, 향후 연구에서는 대분류를 이용한 다단계 분류, 다중 답안 마련 등의 성능 향상을 위한 방법을 연구할 것이다.

브라우징 시스템의 경우, 기존의 단순한 키워드 나열을 통한 검색이 아니라 사용자가 문서의 전체적인 분포를 파악하고 능동적으로 개념적 검색을 할 수 있는 계층적 브라우징 인터페이스를 설계하고 구현해 보았다. 또한 기존 시스템과 달리 자바 애플릿을 사용해 인터페이스를 구현함으로써 시스템 초기 구동시간은 다소 소요되나, 검색 시에는 서버와의 접속 없이 검색을 할 수 있게 되어서 여러 번 검색을 할 경우에 좋은 성능을 보일 수 있을 것이다. 또한 FAQ와 같은 서비스에 익숙하지 않은 사용자나 검색 데이터의 양이 방대하지 않은 경우에 좋은 성능을 기대할 수 있다. 개념적 브라우징의 효용성을 객관적으로 보이기 위해서는 기존 FAQ 검색 방식과의 구체적인 비교 실험이 필요하다. 검색 과정 중의 클릭 횟수나 검색 시간 등의 객관적인 지표를 조사하거나, 만족도나 검색의 용이함 같은 주관적인 지표를 Sheffe의 쌍비교법을 통해서 비교해 볼 수 있지만, 현재 연구 중에 있으며, 향후 연구에서 두 방식의 객관적인 비교 부분을 실험을 통해 보완할 것이다.

참 고 문 헌

[1] 정영미, *정보검색론*, 구미무역출판부, 1993.

[2] D. D. Lewis, "Feature selection and feature extraction for text categorization," *Proc. of Speech and Natural Language Workshop*, Morgan Kaufmann, San Francisco, pp. 212-217, 1992.

[3] 홍진혁, 류중원, 조성배, "실세계 FAQ 메일 자동분류를 위한 문서 특징추출 방법의 성능 비교," *한국정보과학회 춘계 학술대회 발표 논문집(B)*, vol. 28, no. 1, pp. 271-273, 2001.

[4] D. D. Lewis and M. Ringuette, "A comparison of two learning algorithms for text categorization," *Proc. of ACM Intl. Conf. Research and Development in Information Retrieval*, pp. 281-282, 1999.

[5] Y. Yang, "A comparative study on feature selection in text categorization," *Proc. of Intl. Conf. Machine Learning*, Morgan Kaufmann, San Francisco, 1997.

[6] Y. Yang and X. Liu, "A re-examination of text categorization methods," *Proc. ACM Intl. Conf. Research and Development in Information Retrieval*, pp. 42-49, 1999.

[7] C. Apte, F. Damerau and S. Weiss, "Automated learning of decision rules for text categorization," *ACM Trans. Information Systems*, vol. 12, no. 3, pp. 233-251, July 1994.

[8] I. Moulinier and J. G. Ganascia, "Applying an existing machine learning algorithm to text categorization," S. Wermter, E. Riloff and G. Scheler (Eds.), *Connectionist, Statistics and Symbolic Approaches to Learning for Natural Language Processing*, Springer Verlag, Published in the "Lecture Notes for Computer Science" series, no. 1040, pp. 343-354, 1996.

[9] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *Technical Report, Univ. of Dortmund, Dept. of Informatics, Dortmund*, Germany, 1997.

[10] M. E. Ruiz and P. Srinivasan, "Hierarchical neural networks for text categorization," *Proc. of ACM Intl. Conf. Research and Development in Information Retrieval*, pp. 81-93, 1994.

[11] http://www.ask.com

[12] http://www.pragmatech.com

[13] T. Kohonen, "The self-organizing map," *Procee dings of the IEEE*, vol. 78, no. 9, pp. 1464-1480, Sep 1990.

[14] S. Kaski, T. Honkela, Krista Lagus and T. Kohonen, "Creating an order in digital libraries with self-organizing maps," *World Congress on Neural Networks*, pp. 814-817, 1996.

[15] K. Lagus and S. Kaski, "Keyword selection method for characterizing text document maps," *Proc. of Intl. Conf. on Artificial Neural Networks*, vol. 1, pp. 371-376, IEE, London 1999.

[16] T. Kohonen, "Self-organization of very large document collections: State of the art," L. Niklasson, M. Boden and T. Ziemke (Eds.), pp. 343-354, *Proc. of Intl. Conf. on Artificial Neural Networks*, vol. 1, pp. 65-74. Springer, London, 1998.

[17] G. Salton, Automatic *Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, 1988.

[18] J. C. Scholtes, "Kohonen feature maps in full-text database: A case study of the 1987 Pravda," *Proc. of Informatiewetenschap*, pp. 203-220, STINFON, Nijmegen, Netherlands, 1991.

[19] J. C. Scholtes, "Unsupervised learning and the information retrieval problem," *Proc. of Intl. Joint Conf. on Neural Networks*, pp. 95-100, IEEE Service Center, Piscataway, NJ, 1991.

[20] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biol. Cyb.*, vol. 43, pp. 59-69, 1982.

[21] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin Heidelberg, 1995.

[22] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen, "Exploration of full-text databases with self-organizing maps," *Proc. of Intl. Conf. on Neural Networks*, vol. 1, pp. 56-61, IEEE Service Center, Piscataway, NJ, 1996.

[23] K. Lagus, T. Honkela, S. Kaski, and T. Kohonen,

"WEBSOM for textual data mining," *Artificial Intelligence Review*, vol. 13, pp. 345-364, 1999.
[24] S. S. Kaski, T. Honkela, K. Lagus, and T. Kohonen, "WEBSOM: Self-organizing maps of document collections," *Neurocomputing*, vol. 21, pp. 101-117, 1998.
[25] H. Ritter and T. Kohonen, "Self-organizing semantic maps," *Biol. Cyb.*, vol. 61, pp. 241-254, 1989.
[26] E. Gose, R. Johnsonbaugh and S. Jost, *Pattern Recognition and Image Analysis*, Prentice Hall PTR, 1996.
[27] L. Wall, T. Christiansen, R. L. Schwartz, and S. Potter, *Programming Perl*, 2nd Edition, 1996.

저 자 소 개

안준현(Joon-Hyun Ahn)
1998년 2월 연세대학교 컴퓨터과학과 졸업 (학사)
2001년 2월 연세대학교 컴퓨터과학과 석사 과정 졸업
2001년 3월 ㈜ 매직하우스테크놀러지

관심분야 : 신경망, 종분화, 패턴인식, 진화 알고리즘

E-mail : jhahn@candy.yonsei.ac.kr

류중원(Jung-Won Ryu)
2002년 2월 연세대학교 컴퓨터과학과 석사 과정 졸업

관심분야 : 패턴인식, 성별인식, 문서분류, 바이오 인포매틱스

E-mail : rjungwon@candy.yonsei.ac.kr

조성배(Sung-Bae Cho)
1998년 : 연세대학교 전산학과 (학사)
1990년 : 한국과학기술원 컴퓨터과학 (석사)
1993년 : 한국과학기술원 컴퓨터과학 (박사)
1993년~1995년 : 일본 ATR 인간정보통신 연구소 객원연구원
1995년~1998년 : 연세대학교 컴퓨터과학과 조교수
1998년 : 호주 University of New South Wales 초청연구원
1999~현재 : 연세대학교 컴퓨터과학과 부교수

E-mail : sbcho@cs.yonsei.ac.kr