개인화된 분류를 위한 웹 메일 필터링 에이전트

정 옥 란[†] · 조 동 섭[††]

요 약

인터넷의 발달로 인하여 웹을 통한 문서 송수신이 많아지면서 이메일의 사용자도 기하급수적으로 늘어나고 있다. 또한 일반 사용자나 전자상거래에서 오가는 메일의 양도 갈수록 늘어나고 있다. 편리하다는 점을 이용해서 엄청난 양의 스팸 메일도 매일 같이 쏟아져 나오고 있다. 본 논문에서는 사용자 개인에 맞게 메일을 자동 관리해 주는 즉 개인화된 분류가 가능하고, 또 언제 어디서나 로그인이 가능한 웹 메일 기반인 웹 메일 필터링 에이전트(Web Mail Filtering Agent for Personalized Classification)를 제안한다. 새로운 메일이 오면, 먼저 사용자의 메일 처리과정을 일정 기간 관찰하여 각각 개인에 맞는 룰(Personal rule)을 형성하고, 만들어진 룰을 바탕으로 메시지를 자동 관리 즉 카테고리별 분류·저장 및 개인에게 불필요한 메일이나 스팸 메일을 삭제 해 주는 것이다. 또한 시스템의 정확도를 높이기 위해 동적 임계치를 이용한 베이지안 알고리즘을 적용하였다.

Design and Implementation of Web Mail Filtering Agent for Personalized Classification

Ok-Ran Jeong[†] · Dong-Sub Cho[††]

ABSTRACT

Many more use e-mail purely on a personal basis and the pool of e-mail users is growing daily. Also, the amount of mails, which are transmitted in electronic commerce, is getting more and more. Because of its convenience, a mass of spam mails is flooding everyday. And yet automated techniques for learning to filter e-mail have yet to significantly affect the e-mail market. This paper suggests Web Mail Filtering Agent for Personalized Classification, which automatically manages mails adjusting to the user. It is based on web mail, which can be logged in any time, any place and has no limitation in any system. In case new mails are received, it first makes some personal rules in use of the result of observation ; and based on the personal rules, it automatically classifies the mails into categories according to the contents of mails and saves the classified mails in the relevant folders or deletes the unnecessary mails and spam mails. And, we applied Bayesian Algorithm using Dynamic Threshold for our system's accuracy.

키워드 : 개인화된 분류(Personalized Classification), 웹 메일 필터링 에이전트(Web Mail Filtering Agent), 카테고리별 분류, 동적 임계치 (Dynamic Threshold)

1. 서 론

인터넷 이메일은 사용자들이 가장 많이 애용하는 프로그램이며, 앞으로 오랫동안 이메일의 응용분야는 늘어날 것이다. 이를 위해 인터넷 이메일은 사용자들의 다양한 기능적 요구에 부응해야 하며, 프로그래머와 메일 관리자들은 더 많은 기능을 추가하기 위해 노력해야 한다[1]. 인터넷의 성장으로 대표적인 통신 수단인 이메일은 많은 사람들이 정보를 보내거나 받거나 하는데 이용하고 있다. 이메일은 사용하는데 있어서 비용이 거의 들지 않기 때문에 많은 개인이나 업체들이 자신들의 광고를 위해서 사용하고 있고 이로 인해서 메일 서비스를 운영하는 업체에서는 저장장치 용량의 부족 등의 문제를 겪고 있다. 단지 서버를 운영하는 입장이 아니라 메일을 주고받는 사용자의 입장에서도 쏟아져 들어오는 원하지 않는 메일을 지우는 데도 매일 일정량의 시간을 투자해야 한다[2, 3].

본 논문에서 제안한 웹 메일 필터링 에이전트(Web Mail Filtering Agent)는 이러한 문제점을 해결하기 위해서 개인화된 분류를 가능하게 하였다. 개인화된 분류(Personalized Classification)란 개개인의 메일 내용이나 처리 방법 등을 이용하여 개인적 룰을 만들어 그것을 바탕으로 적합한 카

※ 이 논문은 2003년도 학국학술진흥재단의 지원에 의하여 연구되었음 (KRF-2003-041-D00460).
† 준 회 원 : 이화여자대학교 대학원 컴퓨터학과
†† 종신회원 : 이화여자대학교 컴퓨터학과 교수
논문접수 : 2003년 4월 19일, 심사완료 : 2003년 11월 11일

테고리별 분류 및 저장, 삭제를 자동으로 하는 맞춤 분류를 뜻한다. 그 처리 절차는 새로운 메일이 도착하면, 일정기간 사용자의 메시지 처리과정을 관찰하여 개인적 룰(personal rule)을 형성하고, 만들어진 룰을 바탕으로 메일의 내용에 따라 카테고리별 자동 분류하여 해당 폴더에 저장한다. 사용자에게 불필요한 메일이나 스팸 메일은 삭제한다. 여기에서 학습 및 분류에 이용되는 알고리즘은 베이지안 알고리즘을 적용하였다. 분류에서 가장 핵심이 될 수 있는 정확도를 높이기 위해 기존의 베이지안 알고리즘을 개선하여동적 임계치를 이용한 베이지안 알고리즘을 적용하였다.

2. 일반적인 필터링에 대한 연구 동향

메일 내용을 분류한다 함은 미리 정의되어 있는 여러 범주에 각각의 메일들을 할당하는 것이다. 하지만 메일의 수가 증가할수록 각각의 메일을 효과적으로 검색 및 색인화(indexing)하고, 내용 요약(summarization)과 같은 작업을 수행할 때 많은 어려움을 겪게 된다. 이를 해결하기 위해 각 메일들을 카테고리별로 귀속시키는 작업을 수행하며, 휴리스틱(heuristic)을 이용하는 방법 대신 컴퓨터를 이용하는 자동화된 기계학습 기술이 이용되었다. 대표적인 분류 기법으로는 최근접 이웃분류(nearest neighbor classification), 베이지안 확률분포(Bayesian Probabilistic classification), 결정트리(decision tree), 신경망(neural networks), 결정 규칙(decision rules), 그리고 지지벡터기계(support vector machine)들이 있다[4, 5]. 이러한 분류 알고리즘들은 문서의 특징을 선택하는 여러 방법과 함께, 최근 활발한 연구가 진행중인 문서 분류를 위한 특징 선택 연구에 많이 적용되고 있다. 그러나 지금까지의 연구들은 특징 부분집합(feature subset)을 결정하기 위한 평가, 특징 선택 방법 별 단순 성능 비교, 특징 선택 시 구문적 어구를 선택함으로 인한 효과 등에 초점이 맞추어져 있었다.

2.1 기존의 웹 문서 분류 시스템

현재까지는 이메일 분류 시스템에 대한 연구는 많이 이루어지진 않았지만 웹 문서 분류 시스템에 대한 연구는 활발하게 진행되어 왔다. 웹 문서를 대상으로 문서 분류 기법을 적용한 대표적인 시스템은 뉴스기사 분류 시스템, 검색엔진 시스템 등이 있다. 다음은 이러한 문서 분류 기법에 적용하며 만들어진 시스템의 사례에 대하여 기술한다.

2.1.1 Personal Webwatcher

카네기 멜론 대학의 Personal Webwatcher는 사용자의 행동을 웹브라우저에서 모니터링하여 사용자에게 편의를 부여하는 시스템이다. 이 시스템은 사용자의 관심도를 학습하는 방법으로 비교사 학습 방식을 이용한다. 이 시스템은 크게 세 부분으로 구성되어 있다. 첫째 관측을 위한 모니터닝 부분, 둘째 모니터닝 결과에 따른 사용자의 프로파일을 만드는 부분, 마지막으로 사용자의 프로파일을 이용하여 사용자에게 관심 웹 문서를 제공하는 부분이다. 먼저 사용자가 검색한 웹페이지를 모니터링하는 부분은 사용자의 컴퓨터 내부에 Proxy를 두어, 사용자가 현재 검색하고 있는 웹 문서의 위치를 관측하고, 사용자가 보고 있는 웹 문서에 대한 행동을 관측하게 된다. 이렇게 모니터링한 결과를 분석하여 사용자의 관심영역을 측정하는 부분이 두 번째 부분인데, 이 부분에서 웹 문서에 대해 분석이 이루어진다. 관심 웹 문서와 웹 문서에 대한 선택된 단어를 이용하여, 사용자의 관심영역으로 웹 문서를 분류하게 된다. 이를 위하여 TFIDF (Term Frequency Inversed Document Frequency), 베이지안 확률(Bayesian Probability)을 이용한 학습이 이용된다[6].

2.1.2 InfoFinder

앤더슨 컨설팅 연구실에서 만든 InfoFinder는 사용자가 검색하는 문서와 사용자의 관심을 관측함으로서 사용자의 프로파일을 만든다. InfoFinder는 관심 문서에 대한 사용자의 직접적인 관심 사항을 입력받아 사용자의 관심을 학습하는 교사 학습 방식을 이용한다. InfoFinder의 교사 학습의 과정은 로터스 노츠를 이용하여 만든 브라우저는 사용자의 관심을 학습하는 에이전트를 포함하고 있어서, 사용자가 온라인 문서를 읽으면 이를 모니터링 한다. 사용자는 관심 문서에 대해 InfoFinder의 메뉴 중 관심 여부를 나타내는 아이콘을 이용하여 사용자의 관심과 문서에 대한 관심분야를 직접 결정한다. 이렇게 관측된 문서를 사용자의 관심 분야 영역별로 저장한다. 에이전트는 사용자의 관심도를 만들기 위해서, 이렇게 형성된 각 카테고리(관심영역)에 대한 키워드를 학습하게 된다. InfoFinder는 각 관심 영역내의 문서들에 대한 중요 키워드를 추출하기 위해서 사용되는 학습법을 ID3을 이용하여 키워드를 학습한다[6].

2.1.3 Webby

Webby는 IBM 도쿄 연구실에서 개발한 웹 에이전트의 일종으로 웹 브라우저를 모니터닝하고 사용자의 기호도를 측정한 후, 결과를 CGI 서비스로 보여주는 웹 에이전트이다. Webby의 기능은 웹 브라우저를 이용해서 필요한 정보를 얻고자 하는 사용자에게 자신이 선호하는 문서에 대한 기호도와 이 문서를 다시 참조하고자 할 때에 편의를 제공한다. Webby의 웹 브라우저 모니터링 기능은 Proxy 서버의 기본적인 기능에 의해서 사용자가 보고자 하는 웹 문서

에 대한 위치를 추출해 낼 수가 있다. 추출된 웹 문서 위치는 간단한 인덱싱 작업을 통해서 사용자 컴퓨터에 저장되고, 이러한 정보는 사용자의 관심을 추출해내는 기초적 자료가 된다. 사용자의 관심을 추출해내는 방법은 단순히 문서에 대한 방문횟수를 이용하는 것이다. 즉, 방문횟수에 대하여 정규화를 이용하여, 현재까지 사용자가 방문한 웹 페이지에 대한 문서의 수치에 대해 백분율로 관심도를 산출한다. Webby는 이러한 사용자 웹 문서에 대한 백분율을 이용하여, 사용자의 관심 분야를 예측하는 시스템이다[6].

2.1.4 NewT

인터넷의 발전으로 수많은 정보가 네트워크로 들어오는 가운데 뉴스분야의 정보는 계속적인 스트림(stream)의 형태로 웹상으로 유입된다. 이러한 뉴스의 스트림 가운데 사용자가 원하는 기사의 선택을 위한 에이전트 시스템으로 NewT가 있다. 4가지 클래스로 필터링 한다. 뉴스기사 문서 분석은 벡터-공간 모델을 사용한 풀-텍스트 분석으로 이루어지며, NewT 에이전트의 특징은 에이전트 협업부분인데, 사용자는 충분히 학습된 에이전트를 복사하여 다른 사용자에게 제공할 수 있도록 유닉스 환경의 C++로 구현한다[7].

2.2 특징 추출(Feature Extraction)

문서 분류 시 기본적으로 미리 잘 정의되어야 할 부분이 특징 추출(Feature Extraction)이다. 문서 전처리 과정을 통하여 학습에 이용될 중요한 속성들을 추출하는 과정에서 신뢰성을 향상시키기 위해서는 해당 문서의 공통적인 특징을 가려내어 이를 기준으로 각 속성마다 가중치를 차별적으로 두어 더욱 정확한 중요 속성을 추출하는 방법이 이용되고 있다. 이러한 속성 추출 방법을 특징 추출(Feature Extraction)이라 한다. 즉, 특징 추출은 학습 데이터들의 특징을 고려하여 구분된 카테고리별로 다시 한번 중요도를 정의하는 특징 추출 가중치 설정 기법이다. 이를 위하여 각 학습 자원들의 특징을 고려하여 구분된 카테고리들을 대상으로 일련의 구별 작업을 두어 이를 기반으로 한 속성 추출 작업을 수행할 필요가 있다. 이러한 가중치 설정 작업은 해당 키워드가 속해있는 카테고리의 정보를 고려하여 이루어지며 이로써 카테고리를 대표하는 키워드에게 더욱 높은 가중치가 설정된다. 이러한 특징추출에 대한 기계 학습 방법은 서로 다른 몇 개의 카테고리가 존재하는 경우, 각각의 카테고리 별 키워드에 가중치를 주는 것이다[8-11].

또한 특징을 추출하는 방법은 크게 세 가지로 나누어 생각해 볼 수 있다. 첫째는 분류 방법에 의존하지 않고 특징을 선택하는 방법과 의존하는 방법이고[3, 4, 12], 둘째는 전체의 특징들을 개별적으로 이용하는 방법과 부분집합을 취하여 이용하는 방법, 그리고 셋째는 전체의 특징들을 2차원의 공간상에 위치에 있다고 가정하고 위에서부터 하나씩 특징을 더해가며 최종의 특징 집합을 구하는 방법과 반대로 아래에서부터 전체의 특징 중에서 하나씩 제거해 가면서 적절한 특징만을 취하는 방법 등이 있다.

2.3 베이지안 알고리즘 (Bayesian Algorithm)

메일 필터링을 하는 과정에서 룰을 형성하고, 내용을 분류할 때 학습 알고리즘이 이용된다. 학습 알고리즘으로는 베이지안 알고리즘, k-NN(k-Nearest Neighbor), TFIDF(Term Frequency Inverse Document Frequency)들이 널리 쓰이는데, 본 논문에서는 가장 많이 사용되고 있는 학습 알고리즘인 베이지안 알고리즘(Bayesian Algorithm)을 응용하였다. 이 알고리즘은 모든 문서에서 특정단어의 출현으로 구별되는 이진속성벡터(vector of binary attributes)로 표현된 모델로 문서를 정형화하는데, 모델은 다형성 베르누이 사건 모델(multi-variate Bernoulli event model)을 기초로 하여 각 카테고리의 문서마다 다르게 모델을 만들게 된다. 여기서 이용되는 가설은 문서들의 모든 속성은 주어진 전체 카테고리의 다른 문서의 전후관계에 대해서 독립적이라는 것이다. 기본적인 알고리즘은 다음과 같다. 모델링 작업으로 만들어진 문서의 모델을 사용하여 각각의 카테고리별 문서의 확률 값 중 가장 높은 확률 값을 가진 카테고리에 문서를 분류하게 된다[12].

$$P(d_i \mid C_j) = \prod_{t=1}^{|V|} (B_{it} P(w_t \mid C_j) + (1 - B_{it})(1 - P(w_t \mid c_j)))$$

- B_{IT} : 문서 d_i를 위한 벡터의 값 (d_i = 0, 1)
- $p(d_i|c_j)$: 클래스 c_j에 문서 d_i가 나올 확률
- $P(w_r|c_i)$: 클래스 c_i에 단어 w_t가 나올 확률

(그림 1) 기존의 베이지안 알고리즘

3. 웹 메일 필터링 에이전트(Web Mail Filtering Agent) 인터페이스

3.1 기존의 분류 방법과의 차별화

기존의 분류 방법은 공통적으로 텍스트 위주의 누구에게나 적용되는 분류방법이었다. 물론 웹 상에서 모니터링 하여 사용자 편의를 주는 연구도 이루어졌으나, 사용자가 직접적 관여를 포함하고, 또한 메일이라는 특수성을 살려 메일 사용자에게만 맞는 환경을 구축하는 연구는 아직까지 제안되지 않았다. 개인에 맞는 카테고리 설정을 직접 하는 단계를 만들고, 또한 거기에 따라 개인이 처리하는 과정을

관찰하고, 학습한 후, 그 사람에게만 적합한 개인적 룰(personal rule)을 형성하여, 자동 관리 해 준다는 점을 차별화 하였다. 즉 개인화된 분류(personalized classification)가 가능한 메일 에이전트라는 것이다. 또한 텍스트 분류에 있어서 가장 핵심이 될 수 있는 것은 오분류를 하지 않는 것이다. 여기서 가장 중요시되는 정확도(Precision)의 향상을 위하여 동적 임계치(Dynamic Threshold)를 이용한 베이지안 알고리즘을 이용하였다.

WMFA의 전체적인 인터페이스는 크게 세 가지 모듈로 구성되어 있으며, 각각의 모듈들을 차례로 자세하게 살펴보겠다. 먼저 대략적인 모듈별 역할은 다음과 같다.

① **The Web Mail Interface Module** : 새로운 메일이 도착하면, 먼저 사용자의 메일 처리 과정을 관찰하여 학습한다. 특징 추출 및 규칙(rule) 형성에 도움을 주는 모듈이며, 또한 사용자가 개인에 맞는 카테고리 설정을 할 수 있는 과정이다.

② **The Category Rule Generation Module** : 메일 처리 과정에서 특징을 추출하여 응용된 베이지안 알고리즘을 적용하여 개인에 맞는 룰(rule)을 생성한다.

③ **The Web Mail Classification Module** : 생성된 룰(Rule)을 기반으로 새로운 메시지가 도착하면 카테고리별 분류 및 저장을 한다. 또한 불필요한 메일이나 스팸메일은 자동 삭제한다.

또한 본 논문에서는 기존의 고정된 임계치(threshold)를 동적으로 개선하여 필터링의 정확도를 향상시켰다. 그 알고리즘은 (그림 2)와 같다.

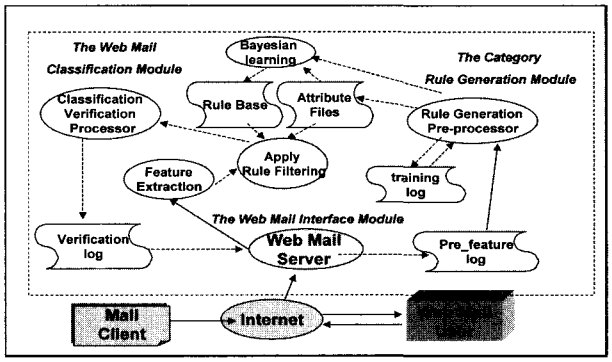
$$
\begin{aligned}
&Category\quad Set\ C = \{c_0, c_1, c_2, c_3, \ldots, c_k\}\ ,\ C_0 = unknown\quad category\\
&Document\quad Set\ D = \{d_1, d_2, d_3, \ldots, d_i\}\\
&\Re(d_i) = \{p(d_i \mid c_1), p(d_i \mid c_2), p(d_i \mid c_3), \ldots, P(d_i \mid c_k)\}\\
&P_{max}(d_i) = \max(\ p(d_i \mid C_t))\qquad ,\ t = 1, \ldots, k\\
&C_{bew}(d_i) = \begin{cases} \{c_i \mid P(d_i \mid c_i)) = P_{max}(d_i),\ if\ P_{max}(d_i) \geq T \\ \qquad\qquad\qquad\qquad whereT\quad = 1 - \dfrac{P_{max}(d_i)}{\sum_{i=1}^{k} P(d_i \mid C_i)} \\ c_0\quad ,\qquad otherwise \end{cases}
\end{aligned}
$$

(그림 2) 동적 임계치를 이용한 베이지안 알고리즘

기존의 베이지안 알고리즘은 임계치 값인 T값을 고정을 하여, 적용하게 되어있었으나, 본 연구에서는 동적 임계치를 이용하여 데이터들의 값에 따라 임계치를 정할 수 있는 방법을 이용하였다. 이는 본 연구의 실험 결과에서 좀 더 나은 정확도를 보여주었다.

3.2 WMFA의 전체적인 구조

본 시스템의 목적은 사용자가 메일을 처리하는데 도움을 주는 에이전트 개발에 있다고 할 수 있으며, 또한 각 사용자의 메일 관리가 편리하도록 인터페이스 환경을 제공하는데 있다고 볼 수 있다. 전체적인 시스템의 구조는 (그림 3)과 같다.



(그림 3) WMFA 시스템의 모듈별 구조도

이 시스템의 가장 큰 특징은 에이전트의 모듈별 구성이며, 그 모듈들은 공유된 파일들에 의해서 서로 전달된다. 모듈 간 가장 중요한 공통 요소는 특징 추출(Feature Extraction) 이라 할 수 있으며, 이는 룰 형성과 카테고리별 분류 시 중요한 역할을 하게 된다.

3.3 제안한 분류 방법에 의한 모듈별 설계

3.3.1 The Web Mail Interface Module

이 모듈은 사용자의 관찰과 교섭을 통해서 이루어진다. 메일 메시지에 대한 사용자의 행동은 나중에 룰 형성에 이용되기 위해 기록된다. 이러한 과정 때문에 일정 기간 학습이 필요하다. 에이전트는 사용자와 메일 툴 사이에 투명하게 놓여있다고 생각할 수 있으며, 먼저 사용자의 행동을 캡쳐링(capturing)해서 인식한 다음, 새로운 메시지가 도착했을 때 카테고리별 분류를 해서 가장 적정한 폴더에 저장시키고, 불필요한 메일들은 삭제하게 된다. 메일 메시지를 파일링(filing)하는 동안 에이전트는 나중에 메시지로부터 특징 추출을 위해 메일 툴에서 메일 메시지를 요구한다. 이 메시지와 사용자의 명령은 *Pre_feature log file* 에 저장된다. 일단 관찰이 이루어지면, 명령이 실행된다. 이것은 에이전트가 메시지 도착함과 동시에 명령을 자동으로 수행하고, 나중에 결과를 사용자가 보게 되는 것이다. 여기에서 가장 중요하게 고려할 점은 에이전트가 빈약한 트레이닝 데이터로 인한 잘못 분류가 이루어 졌을 때 사용자가 룰 형성 시 교섭하기가 쉽게 만들어져야 한다는 것이다. 즉 User Feedback 부분도 고려해야 한다는 것이다. 이를 위해서 사용자가 직접룰 형성에 교섭할 수 있도록 고려하였다. 관찰을 통

해서 얻게 되는 특징 추출은 Rule Generation Module과 Classification Module에서 사용하게 된다. 메일 메시지는 메시지 헤더 부분과 메시지 바디부분으로 크게 두 부분으로 나눌 수 있다. 본 연구에서는 일단 두 부분으로 나눠서 단어기준으로 분석이 이루어진다. 그 단어들은 형성된 룰에 의해서 소트(sort) 된다.

3.3.2 The Category Rule Generation Module

메일 메시지에 대한 사용자의 행동은 메일 인터페이스 모듈에서 기록된다. 그 기록은 룰을 형성하는데 있어 주어진 시간동안 트레이닝(training) 자료로 사용되며 두 가지 로그 파일-*Pre_feature log file, training log file*로 저장된다. 여기에 사용되는 룰을 형성하는 기계 학습 알고리즘은 동적 임계치를 이용한 베이지안 알고리즘을 적용하였다. 먼저 파일을 스팸여부를 판단하는 함수 CheckRule()에 적용한 후, 스팸이 아닐 경우 MergeRule()과 bayesfilt()함수에 적용하여 카테고리별 폴더에 분류하여 저장한다. 또한 이 모듈에서는 attribute files를 생성한다. 이 파일은 메시지를 분류하기 위해 필터링(filtering)하는데 사용된다[13, 14].

3.3.3 The Web Mail Classification Module

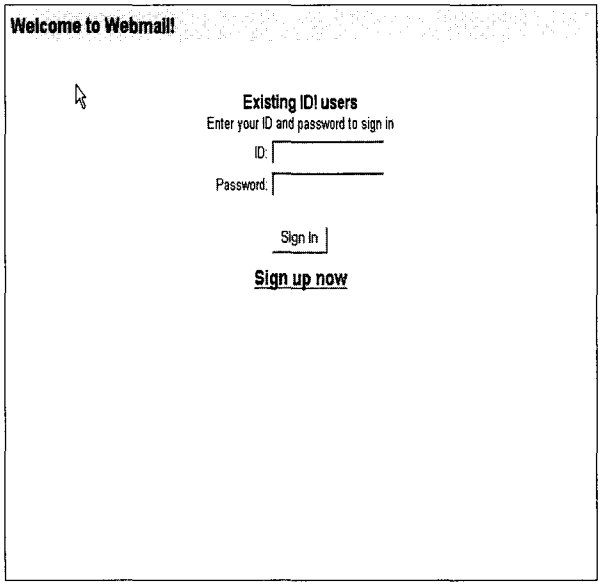
이 모듈은 새로운 메일이 도착하면 룰 베이스(Rule Base)에 적용하여 테스트하고, 결과를 분석한다. 주어진 메일에 적용되는 룰에 의해 실제 분류된 결과는 사용자를 대신해서 카테고리별 자동 분류 한다. 도착한 메시지는 룰에 적용돼서 분류됨과 동시에 다음에 도착할 메시지들을 위한 룰 형성을 위해 사용된다. 분류된 메일은 카테고리별 설정된 폴더에 저장된다[15, 16].

4. 시스템 구현 및 평가 분석

본 시스템은 언제 어디서나 로그인이 가능하고 시스템에 제한이 없으며, 또한 별도의 메일 클라이언트 프로그램이 필요 없는 장점을 가지고 있는 웹 메일을 기반으로 하였다. 구현 환경으로는 Windows 2000 professional, 데이터 베이스 컨트롤을 위해 MS SQL 2000, 룰 형성 및 알고리즘 실행을 위해서 MS Visual C++ 6.0과 ASP, ASP 콤포넌트를 이용하였다.

4.1 웹 기반 메일 시스템

웹 기반 메일 시스템은 계정과 서비스를 제공하는 서버와 사용자와의 인터페이스 역할을 한다. 브라우저를 통해서 메일 서비스를 통해서 전자우편 서비스를 제공하는 웹사이트를 접속한 사용자는 ID와 암호를 가지고 자신의 정보를 관리하게 된다. 또한 사용자는 서버에 자신의 전자우편 주소를 가지게 되며 이를 이용하여 다른 사용자와 전자우편을 주고받을 수 있다. 현재 많은 회사에서 일반 사용자들에게 무료로 서비스하고 있으며, 동작원리는 다음과 같다. 사용자는 일반 웹 브라우저를 통해 서버를 접속한 후 보내고자 할 내용을 입력하고 POST 방식으로 서버에 전송하면 서버는 전송 받은 메시지를 CGI 프로그램에서 파싱한다. 파싱된 메시지는 연속적인 8-bit의 흐름으로 이루어져 있지만, 대개의 전자우편 시스템은 ASCII 문자만을 인식하므로 3개의 8-bit를 4개의 ASCII 문자로 변화시키는 Base64 인코딩 작업을 거친 후, sendmail 프로그램을 구동해서 목적지 서버로 전송하게 된다. 목적지 서버에 도착한 메시지는 서버에 저장되어 있다가 수신자가 웹 브라우저를 통해 서버에 접속해서 요청을 하게 되면 반대과정을 거쳐 수신자의 브라우저에 보이게 된다[19, 20]. 이러한 기능을 할 수 있는 웹 메일 서버의 실제 구현된 로그인 화면은 그림 4와 같으며, 영문 인터페이스로 작성하였다.

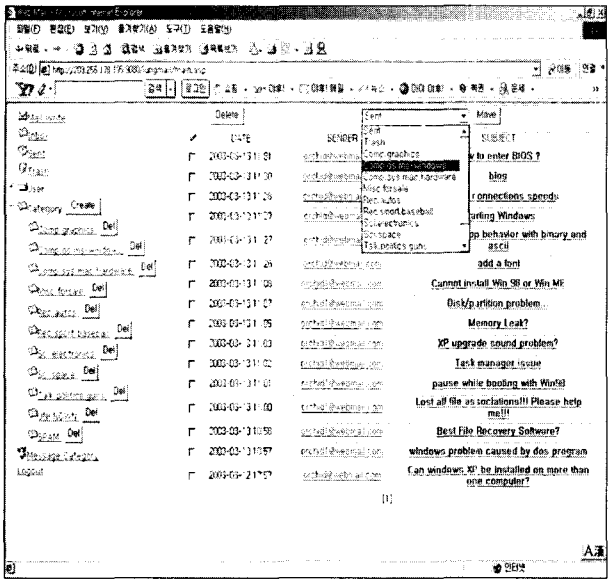


(그림 4) 구현된 웹 메일 시스템의 로그인 화면

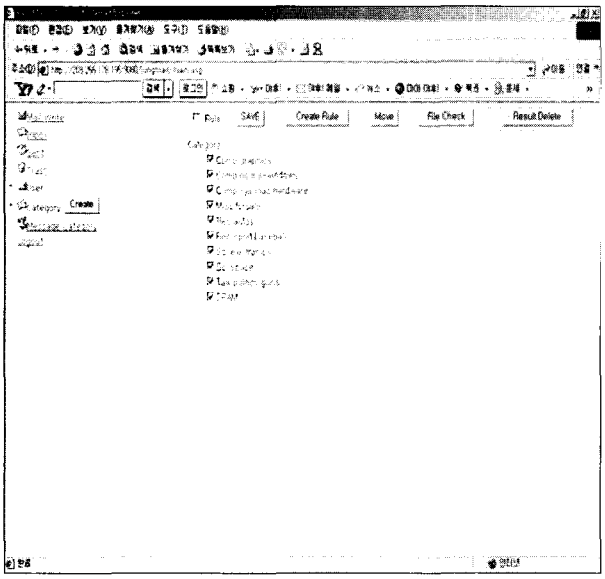
4.2 사용자 인터페이스

다음 (그림 5)는 사용자의 실제 구현된 WMFA의 사용자 인터페이스 화면을 보여준다.

사용자 인터페이스는 사용자 관찰과정에 이용되며, 실제 카테고리 생성 및 저장을 할 수 있다. 본인이 자주 쓰는 카테고리를 만들 수 있으며, 학습하는 과정을 거친 후 관련 메일을 카테고리별 분류를 하게 되는 것이다. 또한 진행되는 과정에서 필요한 카테고리 생성 및 더 이상 필요 없는 카테고리를 삭제하는 기능이 있어, 사용자와 에이전트 사이에 교섭이 가능하다고 볼 수 있다.



(그림 5) WMFA의 사용자 인터페이스

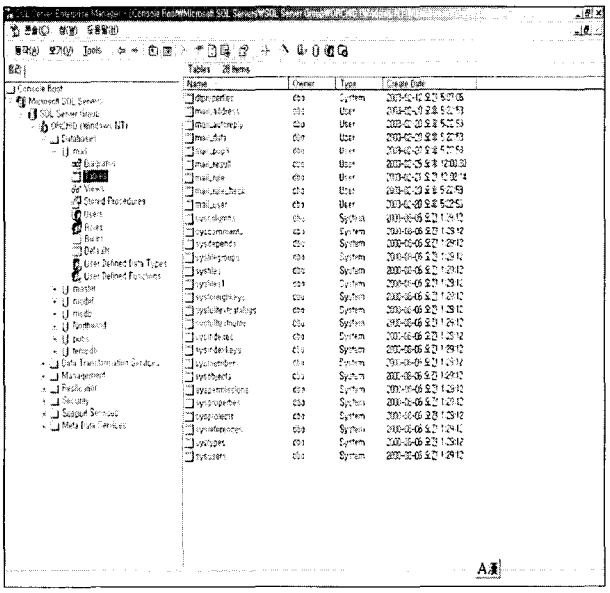


(그림 6) The Web Mail Classification Module

(그림 6)은 본 시스템에서 실제 카테고리별 분류를 하는 기능이 있는 Message Category 모듈을 보여준다. 사용자가 사용하는 메일 쓰기·읽기·보내기·받기·휴지통 등의 기능은 일반 메일 시스템과 별 차이가 없으나, 웹 메일 개인화를 위한 본 메일 시스템만이 가지고 있는 기능들은 아래와 같다.

- Rule Check Box : 체크박스의 기능은 로그인 시 자동으로 Inbox 안에 있는 메일이 분류될 수 있도록 하는 기능을 하며, 체크 시 자동분류가 된다. 체크가 안 되어있을 때는 일반 웹 메일과 동일하게 사용되며, 또한 학습하는 과정이나, 정확률을 체크 할 때 이용된다고 볼 수 있다.
- Save Button : 이 버튼은 첫 번째 Rule 체크박스나 하단의 메일 폴더 트리에서 체크 등 수정작업을 한 후 현재 체크된 상태를 저장할 때 사용한다. 즉 설정된 카테고리 중에 메일 분류시 실제 이용하고자 하는 카테고리만 체크하고 이 버튼을 누르면 그 상황이 효력을 발휘하게 된다.
- Create Rule Button : 이 버튼은 모든 폴더에 있는 룰을 초기화(룰 데이터 삭제)하고, 그 후 하단의 메일 폴더 트리에서 체크를 한 폴더 안에 있는 메일의 내 용을 기준으로 새로운 룰을 생성한다.
- Move Button : Rule 체크 박스는 Inbox의 내용을 로그인 시 자동으로 분류하는 반면, Move Button은 누른 시점에서 Inbox 안의 메일을 분류하게 된다.
- File Check Button : 본 시스템에서 일정기간동안 형성된 룰에 대한 정확도를 체크할 때 이용할 수 있다. 즉 시뮬레이션 할 경우 이용할 목적으로 만들었으 며, 실제 일반 사용자에게는 신뢰성을 줄 수 있다.
- Result Delete Button : File Check 사용 후 다시 초기화 할 때 사용된다.
- Category Folder Tree : 폴더 트리는 생성된 카테고리별 폴더들을 트리 형태로 보여주며, 폴더의 체크 박스를 체크 시 실제 메일이 Inbox에서 이동 시 옮겨지는 폴더의 룰에 옮겨지는 메일의 룰 데이터가 추가된다. 체크를 하지 않을 시는 메일만 옮겨지고 룰의 추가는 일어나지 않는다. Create Rule 버튼을 사용할 때도 적용이 된다.

4.3 DB의 구현



(그림 7) DB 테이블 구성

본 논문에서 제안하는 WMFA(Web Mail Filtering Agent) 시스템의 DB는 MS의 SQL 2000 Enterprise로 구현하였다. 사용자 정보 검색 및 분석이 용이하도록 DB 테이블의 구성은 다음과 같다.

4.3 실험 및 결과 분석

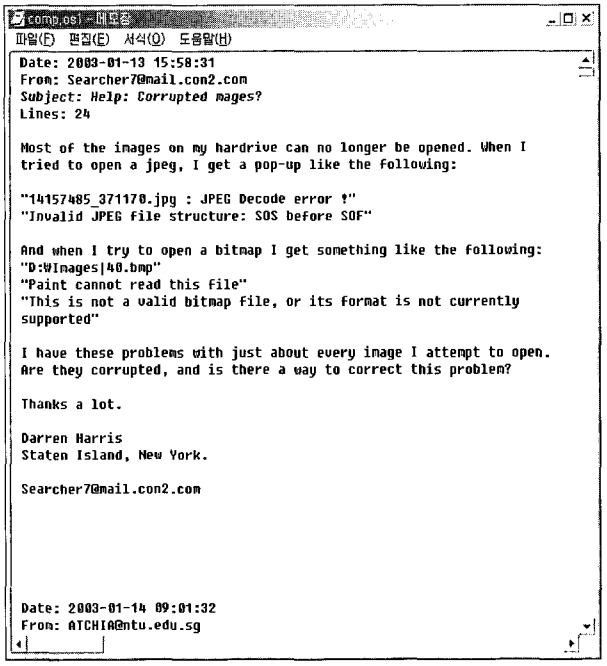
본 연구의 성능 평가를 위해 시뮬레이션을 하였다. 먼저 카테고리를 정하고, 각각 카테고리별 학습을 진행시킨 후 카테고리별 자료를 수집하여 정확률을 계산하였다. 텍스트 분류에 있어서 가장 핵심이 될 수 있는 것은 오분류를 하지 않는 것이다. 그러므로 분류에 있어서 가장 성능 평가 기준이 되는 것은 정확률이 될 것이다. 이 실험을 위하여 10가지 카테고리를 미리 설정하고, 룰을 위한 샘플 데이터와 성능 평가를 위한 데이터를 수집하여 실행하였다. 설정된 카테고리는 다음과 같다.

- comp.graphics : 이미지, 컬러, 사진, 컴퓨터 그래픽스 관련
- comp.os.ms-windows : Ms-Windows, 운영체제(OS) 관련
- comp.sys.mac.hardware : 기계, 컴퓨터 하드웨어
- misc.forsale : 광고, 쇼핑 관련
- rec.auto : 자동차, 탈것 관련 모든 분야
- rec.spot.baseball : 스포츠 관련 모든 분야
- sci.electronics : 전자제품
- sci.space : 천문학, 우주, 환경, 자연 현상
- talk.politics.guns : 정치, 사회 관련
- spam : 스팸 메일

```
Date : 날짜 (yyyy-mm-dd hh : mm : ss)
From : 보낸사람 이메일주소 (test@webmail.com)
Subject : 메일제목 (test subject)
Lines : 메일 내용의 총라인수 (20)
빈라인
실제 메일 내용 시작
.
.
.
실제 메일 내용 끝
빈라인
빈라인
빈라인
빈라인
```

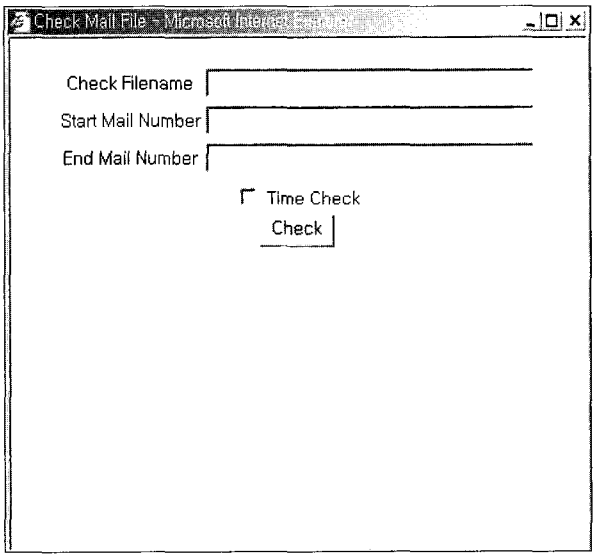
(그림 8) 테스트를 위한 데이터 포멧

본 실험은 시스템의 기능 중 Filecheck 기능을 이용하여 성능 분석을 위하여 정확률을 체크하였다. 실제 본 실험을 위한 메일 데이터 텍스트 포맷은 다음과 같으며, 많은 양의 데이터를 실험해야 하기 때문에 하나의 데이터 포맷으로 만들어, 샘플 데이터 메일을 파일로 작성한 후 테스트하였다.



(그림 9) 실제 테스트용 메일 데이터

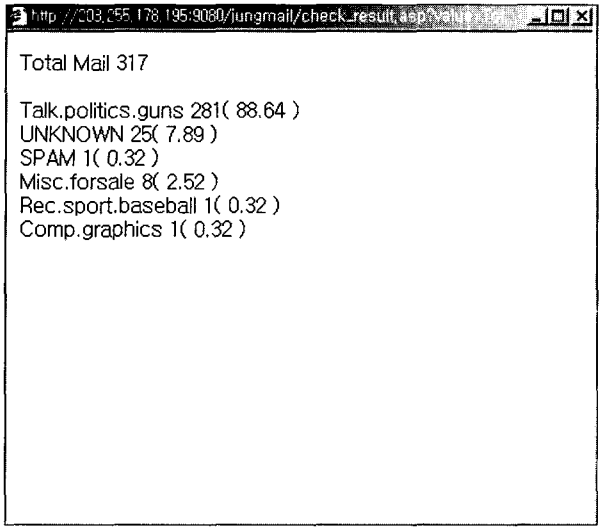
(그림 10)에서는 Filecheck 기능을 이용하여 데이터를 파일로 만들어 분류 시간과 분류 현황을 알아보게 하고, 결과가 해당 카테고리에 적합하게 분류되었는지를 알기 위해 실험하였으며, 결과는 (그림 11)과 같이 보여준다.



(그림 10) Filecheck 기능을 이용한 실험

(그림 11)에서는 Talk.politics.guns라는 카테고리를 테스트 중 한 결과화면을 보여주는 것이며, 해당 카테고리가

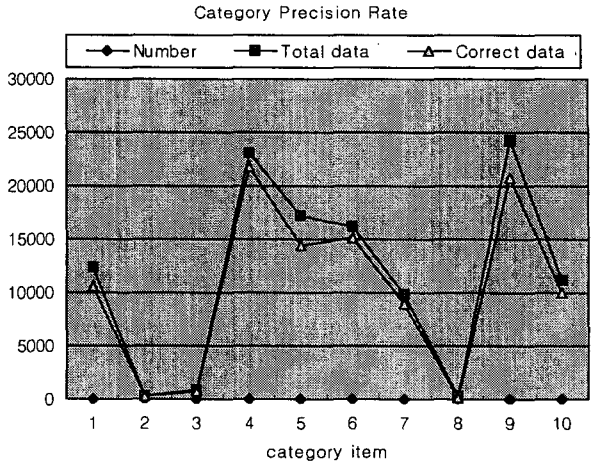
88.64%를 보여주고 있다. 이런 식으로 각 카테고리를 테스트했을 때 최종적인 실험 결과는 각 <표 1>과 도식화한 (그림 12)와 같다.



(그림 11) 실험 결과 화면

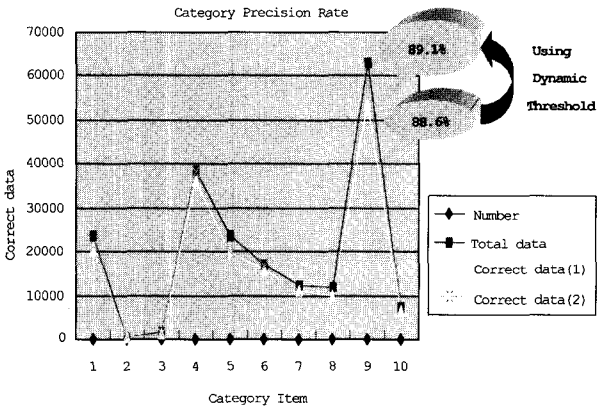
<표 1> 카테고리별 테스트 정확률

Number	Category Item	total data	correct data	Precision(%)
1	comp.graphics	12310	10587	86
2	comp.os.ms-windows	321	295	92
3	comp.sys.mac.hardwore	840	748	89
4	misc.forsale	23100	21714	94
5	rec.autos	17180	14431	84
6	res.sport.baseball	16210	15075	93
7	sci.electronics	9820	8936	91
8	sci.space	240	199	83
9	talk.politics.gus	24310	20664	85
10	spam	11210	9977	89
	Average			88.6



(그림 12) 카테고리별 정확률 분포

평균 정확률은 88.6%로 측정되었으며, 좀 더 많은 학습 데이터 및 학습 기간을 갖게 될수록 정확도는 더 높아질 것이다. 또한 본 연구에서 제안한 동적 임계치를 이용한 베이지안 알고리즘을 적용한 후 동일한 데이터에 대한 실험 결과는 <표 2>와 도식화한 (그림 13)과 같다.



(그림 13) 동적 임계치를 적용 후 카테고리별 정확률 분포

<표 2> 동적 임계치를 적용 후 카테고리별 테스트 정확률

Number	Category Item	Total data	Correct date(1)	Precision1(%)	Correct date(2)	Precision2(%)
1	Comp.graphics	23579	20278	86	21457	91
2	comp.os.ms-windows	573	527	92	470	82
3	comp.sys.mac.hardwore	1578	1404	89	1215	77
4	Misc.forsale	38434	26128	94	38050	99
5	rec.autos	23712	19918	84	20155	85
6	res.sport.baseball	17124	15925	93	16610	97
7	sci.electronics	12354	11242	91	10624	86
8	sci.space	11694	9706	83	9940	85
9	talkpolitics. gus	62915	53478	85	59140	94
10	Spam	7260	6461	89	6897	95
	total	199223	176512		184558	
	Average			88.6		89.1

동적 임계치를 이용한 결과 89.1%의 정확률을 보여 주었으며, 기존의 알고리즘을 사용하였을 때 보다 0.5%의 향상을 보였다.

5. 결 론

본 연구에서는 이메일 사용자의 개인화된 분류가 가능한 웹 메일 필터링 에이전트를 설계 및 구현하였다. 현재 이메일을 통해 많은 양의 정보들이 오가고 있고, 사용자들은 또한 이 중에서 본인에게 맞는 맞춤 이메일 인터페이스를 요구하게 될 것이다. 특히, 스팸 메일의 양이 갈수록 늘어나므로 이에 대한 방법들이 쏟아져 나오고 있다. 하지만 그 차단방법은 스팸 메일의 제목들에서 공통적 요소를 필터링하는 방법들이다. 본 연구에서는 스팸도 개인이 판단하는 학습 과정을 거쳐 본인에게 스팸 일 경우 추후부터 차단되는 방법이라고 할 수 있다.

또한 분류에 있어서 가장 핵심이 될 수 있는 정확률을 높이기 위해 기존의 베이지안 알고리즘에 동적 임계치(Dynamic Threshold)를 응용하였다. 사용자들이 많은 양의 메일을 관리할 때 WMFA(Web Mail Filtering Agent)는 매우 유용하게 활용될 수 있을 것이다. 향후 연구 방향으로는 카테고리를 사용자가 직접 설정하는 방법으로 되어 있는데 이 방법과 자동 카테고리 설정을 에이전트가 할 수 있는 방법으로 확장시켜 나갈 것이다.

참 고 문 헌

[1] David Wood. 최규혁역, "Internet e-mail programming," 한빛미디어, 2000.

[2] Dunja Mladenic, Marko Grobelnik, "Feature selection for classification based on text hierarchy," *Proc. of the workshop on Learning for Text and the Web*, Pittsburgh, USA, 1998.

[3] George H. John, Ron Kohavi, Karl Rfleger, "Irrelevant Features and the Subset Selection Problem," *Proc. of ICML 94*, Morgan Kaufmann Publishers, San Francisco., CA, pp. 121-129, 1994.

[4] Ian H. witten and Eibe Frank, Data Mining, Morgan Kauf mann Publishers, Inc., 2000.

[5] Yiming Yang, Jan O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," *Proc. of ICML97*, pp.412-420, 1997.

[6] 백혜정, 박영택, 윤석환, "사용자 관심도를 이용한 웹 에이전트", 정보처리학회지, 1999, *http://sslab1.chosun.ac.kr/~chaehwan/study/agent/makeagent_favoriate.htm.*

[7] Jeffrey M. Bradshaw, "Software agent," AAAI Press/ The MIT Press, pp.151-161.

[8] 이상섭, 오재준, 박영택, "웹 에이전트 핵심 기술", http://member:tripod.lycos.co.kr/ironjohn/agent/agent.html.

[9] Andrew D. May, "Automatic Classification of E-mail Message by Message Type," Journal of the American Society for Information Science, 48(1), pp.32-39, 1997.

[10] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, "Modern Information Retrieval," Addison-wesley, 1999.

[11] Willian W.Cohen, "Learning Rules that Classify E-Mail," AAAI Spring symposium on Machine Learnning in Information Access, pp.18-25, 1996.

[12] McCallum, A. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," In AAAI-98 Workshop on Learning for Text Categorization, 1998, http://www.cs.cmu.edu/~mccallum.

[13] P. Maes, "Agent that Reduce Work and Information Overload," Communications of the ACM, Vol.37, No.7, pp.30-40, 1994.

[14] P. Resnic, N. Iacocou, M. Sushak, P. Bergstrom and J. Riedl, "groupLens : An Open Architecture for Collaborative Filtering of Netnews," Proceedings of the American Association of Artificial Intelligence, pp.439-446, 1999.

[15] D. Golfberg, D. Nichols, B. M. Oki and D. Terry, "Using Collaborative Filtering to Weaves an Information TAPESTRY," Communications of the ACM, Vol.35, No.12, pp. 61-70, 1992.

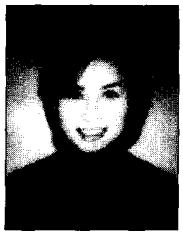
[16] B. Mirkin, "Mathematical Classification and Clustering," Kluwer Academic Publisher, p.428, 1996.

[17] K. Alsabi, S. Ranka and V. Singh, "An Efficient K-Means Clustering Algorithm," IPPS/SPDP Workshop on High Performance Data Mining, Orlando, 1998.

[18] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data : an Introduction to Cluster Analysis," Wiley Series in Probability and Mathematical Statistics, p.342, 1990.

[19] S. Sol and G. Berznieks, "CGI/PERL : Web Scripts," M&T Books, 1997.

[20] W. Stallings, "Network and Internetwork Security : Principles and Practices," Prentice Hall, 1995.



정 옥 란

e-mail : orchung@ewha.ac.kr

1993년 전북대학교 전자계산학과(이학사)

1998년 전북대학교 대학원 정보과학과 (이학석사)

1999년~현재 이화여자대학교 과학기술 대학원 컴퓨터학과 박사과정

관심분야 : 웹 마이닝, 텍스트 마이닝, 지식 공학

조 동 섭

e-mail : dscho@ewha.ac.kr

1979년 서울대학교 전기공학과(공학사)

1981년 서울대학교 대학원 전기공학과 (공학석사)

1986년 서울대학교 대학원 컴퓨터공학과 (공학박사)

1996년 University of California, Irvine Visiting Scholar

1985년~현재 이화여자대학교 컴퓨터학과 교수

관심분야 : 컴퓨터구조 및 인터넷 공학, 컴퓨터비젼, 컴퓨터그래픽스, 가상교육