가중치가 부여된 베이지안 분류자를 이용한 스팸 메일 필터링 시스템

(Spam-Mail Filtering System Using Weighted Bayesian Classifier)

김 현 준 † 정 재 은 †† 조 근 식 †††
(Hyun-Jun Kim) (Jason J. Jung) (Geun-Sik Jo)

요 약 최근 인터넷의 급속한 성장과 더불어 전자메일(E-Mail)은 통신 및 정보, 의사교환의 필수적인 매체로 사용 되어지고 있다. 그러나 편리하고 비용이 들지 않는 장점을 이용해 엄청난 양의 스팸 메일이 메일같이 쏟아져 오고, 그 문제의 심각성에 정보통신부는 '정보통신망 이용촉진 및 정보보호등에 관한 개정안'이라는 새로운 법률까지 만들었다. 본 논문에서는 기존의 문서 분류에 널리 쓰이던 나이브 베이지안 분류자(naive Bayesian classifier)보다 개선된 가중치가 부여된 베이지안 분류자 (weighted Bayesian classifier)와 정보통신부의 개정안을 준수하는 메일을 분류하기 위한 전처리 단계, 그리고 사용자의 행동을 학습하여 보다 정확한 분류를 가능하게 지능형 에이전트(intelligent agent)가 결합된 형태의 스팸 메일 필터링 시스템(spam mail filtering system)을 제안한다. 제안된 시스템에서는 사용자가 직접 규칙을 넣을 필요 없이 학습한 데이타를 가지고 자동적으로 스팸 메일을 분류할 수가 있는데, 특히 이메일의 특징 추출(feature extraction)을 이용하여 상대적으로 스팸/논스팸 판별에 비중이 큰 단어들에 대해 가중치를 부여함으로서 필터링의 성능향상을 도모하였다. 실험에서는 제안된 시스템의 최적의 성능 평가를 위해서 일반 나이브 베이지안 필터링시의 성능과 이메일 헤더정보, 특정 Tag들 그리고 하이퍼링크 부분에 가중치를 준 베이지안 필터링, 마지막으로 4가지를 결합한 상태의 필터링 성능을 각각 비교 분석하였다. 그 결과 제안하는 시스템이 나이브 베이지안 분류자를 이용한 시스템보다 정확도에서는 5.7% 저조한 성능을 보였으나, 재현율에서 33.3%, F-measure에서 31.2% 우수한 성능향상을 보였다.

키워드 : 베이지안 분류자, 메일 필터링, 지능형 에이전트

Abstract An E-mails have regarded as one of the most popular methods for exchanging information because of easy usage and low cost. Meanwhile, exponentially growing unwanted mails in user's mailbox have been raised as main problem. Recognizing this issue, Korean government established a law in order to prevent e-mail abuse. In this paper we suggest hybrid spam mail filtering system using weighted Bayesian classifier which is extended from naive Bayesian classifier by adding the concept of preprocessing and intelligent agents. This system can classify spam mails automatically by using training data without manual definition of message rules. Particularly, we improved filtering efficiency by imposing weight on some character by feature extraction from spam mails. Finally, we show efficiency comparison among four cases - naive Bayesian, weighting on e-mail header, weighting on HTML tags, weighting on hyperlinks and combining all of four cases. As compared with naive Bayesian classifier, the proposed system obtained 5.7% decresed precision, while the recall and F-measure of this system increased by 33.3% and 31.2%, respectively.

Key words : Bayesian Classifier, Mail filtering, Intelligent Agent

† 학생회원 : 인하대학교 컴퓨터정보공학부
dannis@eslab.inha.ac.kr
†† 비 회 원 : 인하대학교 컴퓨터정보공학부
j2jung@eslab.inha.ac.kr
††† 종신회원 : 인하대학교 컴퓨터정보공학부 교수
gsjo@inha.ac.kr
논문접수 : 2003년 3월 29일
심사완료 : 2004년 6월 16일

1. 서 론

인터넷의 급속한 성장과 더불어 전자우편(E-Mail)은 현재 통신 및 정보, 의사 교환의 필수적인 매체로 사용되어지고 있다. 하지만 송, 수신에 있어 편리하고 비용이 들지 않는 장점을 이용해 많은 개인이나 업체들은 자신들의 상업적 광고를 무분별하게 발송하고 있으며, 그 양은 매년 증가하고 있는 추세이다[1]. 이에 따라 메

일 서비스 업체들은 저장장치의 용량부족 등의 문제를 겪고 있고, 일반 사용자들은 쏟아져 들어오는 상업성 광고 및 불법, 음란광고로 인해 자신의 계정부족 및 스팸 메일을 지우는데 시간을 투자하는 불편을 겪고 있다. 연구에 의하면 직장인들은 범람하는 이메일로 인해 하루 평균 30분가량을 소비하고 있었으며, 그중에 약 31%가 스팸 메일 이었다[2].

스팸 메일이 심각한 사회문제로 부각되자 정보통신부는 현재 '정보통신망 이용촉진 및 정보보호 등에 관한 법률 시행령 및 시행규칙 개정안'을 마련해 놓았다[3]. 이 개정안에 의하면 모든 광고메일들은 제목의 처음에 '광고'라는 문구를 넣어야 하며, 끝에는 '@'를 첨가하여, 서비스 제공업체들로 하여금 스팸 메일 필터링을 더욱 효과적으로 할 수 있도록 법제화 하였다. 그러나 이와 같이 단순한 메세지 규칙 기반의 필터링 기법으로는 정보통신부 개정안을 준수하지 않는 스팸 메일을 걸러내는 것이 어려울 뿐 아니라, 논스팸 메일의 경우 본문에 포함된 특정 단어들로 인해 스팸 메일로 처리되는 경우도 발생될 수 있다[4].

이러한 문제의 해결을 위한 방법의 한가지로 전자문서 분류에 많이 이용되는 나이브 베이지안 분류자(Naive Bayesian Classifier)는 문서 내의 단어들을 대상으로 분류를 하는데, 메일의 경우에는 본문에 포함된 단어들과 HTML Tag들, 그리고 메일을 보낸 사람의 이메일 주소, 제목 등의 헤더정보를 바탕으로 스팸 메일 여부를 확률적으로 판단함으로서 보다 정확하고 환경의 변화에도 적응성이 있는 필터링이 가능해진다[5].

본 논문에서는 스팸 메일에 자주 사용되는 특정 Feature에 가중치를 부여하는 베이지안 분류자(Weighted Bayesian Classifier)를 이용한 스팸 메일 필터링 시스템을 제안하고, 가중치에 따른 최적의 성능을 평가하였다. 이와 더불어 전처리 단계(Pre-Processing)와 지능형 에이전트(Intelligent Agent)와의 결합을 통해 필터링 성능을 향상시켰다.

2. 관련연구

스팸 메일을 걸러내기 위한 방법으로 기존에 연구되어지고 있는 문서분류, 필터링 등의 데이타 마이닝 기법들이 사용되어지고 있다. 현재의 대부분의 메일 시스템들은 메시지 규칙에 기반을 둔 필터링 방법을 주로 적용하고 있으며, 그밖에 많은 확률적인 방법을 적용한 메일 필터링 시스템들이 개발되어지고 있다[24].

2.1 메시지 규칙을 이용한 분류(Rule-based Classification)

메시지 규칙을 이용한 필터링 방법은 스팸의 특징으로 대표될 수 있는 단어들을 찾아내어 스팸 메일의 여부를 판단하는 방법으로서 확률적인 방법들에 비해 단순하며 재현율과 정확도에 있어서도 비교적 좋은 성능을 얻을 수 있다[6]. 반면에 사용자가 직접 메시지 규칙을 입력해야 하며, 스팸 메일의 형태가 변화함에 따라 메시지 규칙도 지속적으로 갱신 시켜야 하는 문제가 발생한다. 뿐만 아니라 참과 거짓 두 가지의 경우만을 결과로 갖기 때문에 보다 정확한 필터링을 하는데 한계가 있다[7,22].

2.2 나이브 베이지안 분류자(Naive Bayesian Classifier)

1763년 Thomas Bayes의 논문에 발표된 베이즈 이론에서 발전된 확률적인 알고리즘으로서[21], 나이브 베이지안 분류자는 2.1절에서 설명한 메시지 규칙의 한계를 극복하기위해 문서분류, 스팸 메일 필터링 시스템에 사용되어지고 있는 대표적인 확률적 방법 중의 하나이다[23]. 기본적으로, 가설 H와 주어진 학습(Training) 데이타 D로부터 사후확률(posterior probability)을 구하고자 할 때 사용하는 베이즈 이론은 식 (1)과 같다.

$$\text{Bayes' theorem} : P(h|D) = \frac{P(D|h)P(h)}{P(D)} \quad (1)$$

$P(H)$는 H의 사전확률(Priori Probability)이고, $P(H|D)$는 D가 주어졌을 때 H의 사후확률이다[8]. 나이브 베이지안 분류자에서 데이타 D는 벡터로서 $\vec{d}=\langle d_1, d_2, \ldots, d_n\rangle$과 같이 표현되며, d_n은 D_n의 속성값을 의미한다. 만일, m개의 클래스 $\langle c_1, c_2, \ldots, c_m\rangle$를 갖는 C가 있다고 가정하고, 임의의 D데이타가 존재할 경우, 분류자는 D에 해당하는 최대의 사후 확률을 갖는 클래스 c_i를 예측하게 된다.

$$P(c_i|D) > P(c_j|D) \text{ 단, } 1 \le j \le m, j \ne i \quad (2)$$

식 (2)를 통해서 최대의 사후확률을 갖는 $P(c_i|D)$를 얻을 수 있으며, 따라서 클래스 c_i는 식 (1)에 의해서 사후확률이 최대인 가설(MAP: Maximum Posterior)이 된다[8].

$$P(c_i|D) = \frac{P(D|c_i)P(c_i)}{P(D)} = P(D|c_i)P(c_i) \quad (3)$$

식 (3)에서 분모항 $P(D)$는 c_i에 대하여 독립적인 상수값을 가지므로 생략될 수 있다. 또한 주어진 데이타 D가 m개의 많은 속성들을 가지고 있는 경우 $P(D|c_i)$의 계산을 위한 비용이 커지는 문제가 발생하는데, 이러한 문제를 해결하기 위해서 나이브 베이지안 분류자에서는 각 속성들이 상호 독립적(Conditionally Independence)이라 가정한다. 즉, 속성들 사이에 서로 영향을 주고받는 관계가 없다고 가정하면 (4)와 같은 식을 얻어낼 수 있다[9].

$$c_{NB} = \underset{c_i \in C}{argmax}\, P(c_i) \prod_{k=1}^{n} P(d_k|c_i) \quad (4)$$

즉, D의 결합 확률은 각 속성들의 확률의 곱으로 계산될 수 있다.

현재까지 *POPFILE*[10] 등의 나이브 베이지안 분류자를 이용한 스팸 메일 필터링 소프트웨어들이 많이 개발되어왔으나[20], 스팸/논스팸 메일에 모두 등장하는 불용어(stopword)로 인한 정확도 저하 등, 필터링 성능에 한계가 존재해 왔다. 따라서 스팸의 특징을 갖는 특정 토큰(Token)들에 대하여 가중치를 부여함으로서 보다 정확한 필터링이 이루어지도록 하는 연구가 진행되어지게 됐다.

2.3 가중치가 부여된 베이지안 분류자(Weighted Bayesian Classifier)

나이브 베이지안 분류자에서는 각각의 속성값들 $\langle d, d, \ldots, d_n \rangle$에 대해서 동등한 가중치를 적용하고 있다. 그러나 각각의 속성값들에 대해 사용 빈도를 바탕으로 가중치를 부여할 경우, 단순한 나이브 베이지안에 의한 필터링보다 성능 면에서 향상된 결과를 얻을 수 있다. 예를 들어, 학습한 결과 빈도수가 높은 단어나 메일의 성격을 쉽게 구분할 수 있게 해주는 키워드 등에 대한 가중치를 다른 속성값들 보다 높게 부여함으로서 나이브 베이지안에 의한 필터링 속도 및 정확도를 향상시킬 수 있는 것이다. Gartner의 WBC_{SVM} [11]은 나이브 베이지안 분류자에 가중치를 부여한 메일 필터링 시스템의 한 사례이며, 그 밖에 많은 데이타 마이닝 알고리즘들의 소성값에 가중치가 부여됨으로서 정확도와 재현율이 향상된 결과를 얻을 수 있었다.

$$c_{WB} = \underset{c_i \in C}{\arg max}\, P(c_i) \prod_{k=1}^{n} \left(\frac{d_k w_k + 1}{\sum_{j=1}^{n} d_j w_j + n} \right) P(d_k|c_i) \quad (5)$$

식 (5)는 식 (4)에서 가중치가 부여된 확장 형태이다. 속성값 d_i에 해당하는 가중치 w_i를 부여하고 있으며, 여기에 1을 더해 줌으로서 확률이 0이 되는 것을 예방하였다. 또한 가중치 w_i가 부여됨으로서 속성값 자체의 확률이 커지는 것을 방지하기 위해(Normalization) 속성값 $(d_i w_i + 1)$을 가중치가 부여된 전체 속성값의 합 $(\sum_{j=1}^{n} d_j w_j + n)$으로 나누었다[12,13].

2.2 지능형 에이전트(Intelligent Agent)를 이용한 Incremental Learning

지능형 에이전트(Intelligent Agent)란 그림 1과 같이 환경을 인식하여 문제를 발견하고 해결함으로서 사용자가 원하는 상태(Goal State)에 도달하도록 하는 것이다[14]. 어떤 환경이나 시스템 내에서의 에이전트는 그 내부에서 이루어지는 행동이나 상태의 변화 등을 지속적으로 관찰(Monitoring)하고 학습(Learning)함으로서 보다 지능적이고 사용자의 요구에 만족하는 문제 해결이 가능해지는데, 이를 Incremental Learning Agent라 한다[15].

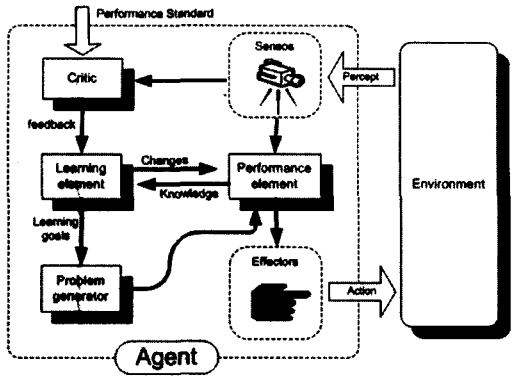


그림 1 Incremental Learning Agent 구조도

Incremental Learning Agent는 일반 소프트웨어 뿐 아니라, 전자상거래(E-Commerce)와 관련된 응용프로그램에 많이 사용되는데, 이는 사용자의 행동을 지속적으로 관찰, 학습하여 최적의 결론을 도출해 내는 업무에 효과적이기 때문이다.

3. 스팸 메일에서의 특징 추출(Feature Extraction)

가중치를 부여한 베이지안 분류자를 이용하여 스팸 메일 필터링을 하기 위해서는 우선 스팸 메일의 특징(Feature)을 추출(Extraction)해 내는 것이 중요하다. 단순한 나이브 베이지안 분류자의 경우 메일에서 공백(Space)으로 구분된 토큰(Token)들을 추출한 뒤, 그 토큰들의 출현 빈도수를 저장하고 이를 바탕으로 필터링이 가능하다. 하지만, 가중치가 부여된 베이지안 분류자에서는 스팸 메일에서의 특징을 추출해 놓아야만 적절한 요소들에 가중치를 부여함으로서 보다 정확한 필터링이 가능해지는 것이다.

3.1 이메일 헤더(E-Mail Header)

이메일을 주고받는 과정에서 메시지를 처리하는 서버는 제일 처음 메시지에서 자신이 작업을 처리하는데 필요한 헤더를 찾는다. 이 메시지 헤더의 정보는 사용자에 의해 직접 입력된 부분도 있으나, 메일/뉴스 프로그램 혹은 서버에 의해 자동으로 기록된 부분도 있다. 따라서 이러한 이메일 헤더의 정보를 이용하여 스팸 메일의 여부를 판단하는데 사용할 수도 있다.

3.2 메세지(Message)

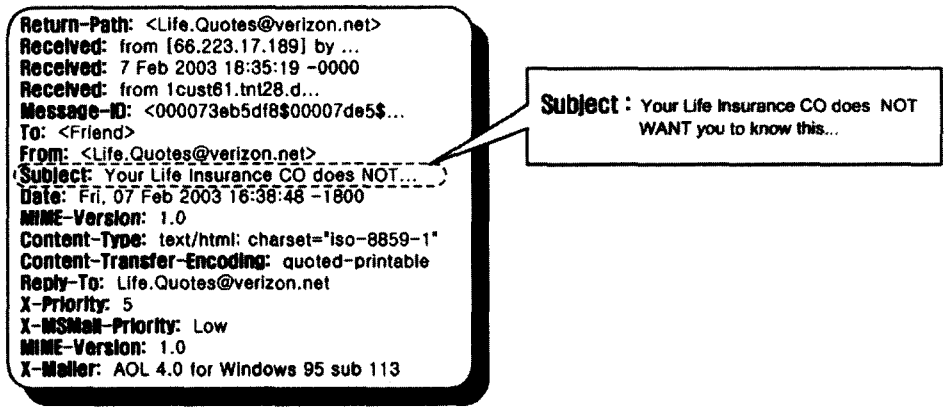


그림 2 이메일 헤더(E-Mail Header) 정보

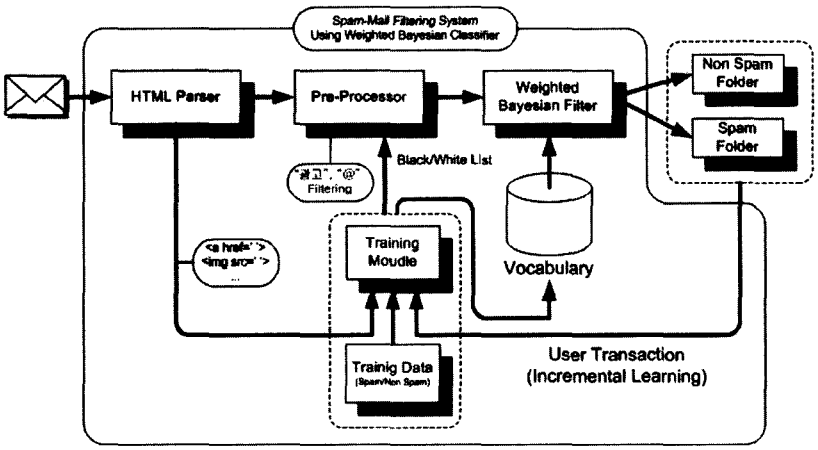


그림 3 시스템 구조도

이메일은 헤더 부분과 메시지 부분으로 이루어져 있다. 메시지 부분은 메일 발송자가 원하는 내용이 포함되는 부분이다. 메시지 부분에는 일반 텍스트와 HTML Tag가 주로 사용되는데, 메시지에 HTML Tag들과 같이 사용된 텍스트들을 관찰함으로서 스팸 메일의 여부를 판단할 수 있다. 즉, 스팸 메일에 자주 사용되는 특정 Tag들(<img src="">, <a href="">)과 같이 사용된 텍스트에 대해 가중치를 부여함으로서 스팸 여부를 보다 정확히 판단할 수 있게 된다[16].

4. 제안하는 스팸 메일 필터링 시스템

스팸 메일 필터링 시스템에 있어 유의해야 할 사항은 잘못된 분류에 대한 비용의 문제이다. 시스템이 간혹 스팸 메일을 읽어야 할 메일, 즉 논스팸 메일로 분류했을 경우에는 사용자가 직접 삭제하는 비용만 들게 된다. 그러나 논스팸 메일을 스팸 메일로 분류한 경우에는 분류된 스팸 메일 중에서 찾아 읽어야 하는 비용이 들고, 스팸 메일을 전부 삭제한 경우에는 메일을 읽을 수 없는 경우도 발생한다. 따라서 잘못된 분류가 존재할 수 있음을 인정할 경우 스팸 메일이 논스팸 메일로 분류될 수는 있되(False-Negative), 그 반대의 경우(False-Positive)는 최대한 발생하지 않도록 시스템을 구성해야 한다[7].

4.1 시스템 구조

제안하는 가중치가 부여된 베이지안 분류자를 이용한 규칙 기반의 스팸 메일 필터링 시스템 구조는 그림 3과 같다.

가중치가 부여된 베이지안 필터링의 수행하는 과정은 학습단계와 분류단계로 구분된다. 우선 학습단계는 학습을 위한 스팸/논스팸 메일을 파싱하여, HTML Tag들과 일반 텍스트들을 단어별로 분류한다. 분류된 단어들은 Training Module에 Training Data로 입력되어 학

```
Training_WBC(S, c_i)                          //S=학습 데이터 셋, c_i={C_spam, C_nonspam}
 Preprocessing(S)                             //학습을 위한 특징단어를 제외한 나머지 제거
 While(S <> EOF)                              //입력값으로 받은 메일의 끝까지 읽어들임
  d_k = parsing(S)                            //메일을 파싱하여 토큰을 d_k에 저장
  If exist_in_vocabulary(c_i, d_k) = true then //d_k가 해당 클래스의 Vocabulary에 존재하면
   freq(c_i, d_k) = freq(c_i, d_k) + 1        //해당 d_k의 빈도수 증가
   total_freq(c_i) = total_freq(c_i) + freq(c_i, d_k) //P(c_i) 계산을 위한 변수
  Else                                        //d_k가 해당 클래스에 존재하지 않을 경우
   vocabulary(c_i) ← d_k                      //해당 클래스의 Vocabulary에 저장
   feature(d_k) ← d_k                         //가중치 계산을 위한 d_k의 속성을 저장
  End If
 Loop
 P(c_i) ← total_freq(c_i) / total_freq(C_spam + C_nonspam)   //P(c_i)와 P(d_k| c_i)를 계산
 P(d_k| c_i) ← freq(C, d_k) / total_freq(c_i)
```

그림 4 가중치가 부여된 베이지안 분류기를 학습하는 모듈

습에 이용된다. 분류단계는 실제 메일을 받게 되면 시스템은 학습된 모듈을 통해 생성된 단어 및 각 단어들의 사용 빈도수 및 가중치가 저장된 Vocabulary를 이용하여 Weighted Bayesian Filtering을 실시하게 되고, 이에 따른 결과로 스팸 메일과 논스팸 메일을 분류하게 된다. 이후에도 사용자에 의해 명확해진 메일에 대하여 시스템은 Training Module을 지속적으로 학습 시키는데, 시스템이 논스팸 이라고 판단한 메일이 결국 사용자에 의해 스팸으로 판단되어 지워질 경우에 시스템은 이를 학습하여 이후에 발생되는 유사한 메일에 대해 보다 지능적이고 사용자의 판단과 유사한 분류를 할 수 있게 되는 것이다.

4.2 전처리 과정(Pre-processing)

정보통신부는 '정보통신망 이용촉진 및 정보보호 등에 관한 법률의 시행령 및 시행규칙 개정안 [2]'을 통해 영리목적의 광고성 전자우편(스팸 메일)을 발송할 때에는 제목의 처음에 '(광고)' 또는 '(성인광고)'라는 문구와 함께 제목 끝에는 '@'를 반드시 표시하도록 규정해 놓았다. 전처리 과정에서는 정보통신부 규정을 준수한 스팸 메일을 일차적으로 분류해 주는데, 이 과정은 스팸 여부가 확실한 메일만을 필터링 해주므로 100%의 스팸 정확도(Percision)을 보장하는 과정이며, 베이지안 분류자에 의한 필터링 대상 메일의 수를 줄여줌으로서 전체적인 속도증가의 효과를 얻을 수 있게 된다[17].

4.3 가중치가 부여된 베이지안 분류자를 이용한 필터링(Weighted Bayesian Classifier Filtering)

가중치가 부여된 베이지안 분류자에 의한 학습 및 분류 알고리즘은 그림 4,5와 같다.

그림 4는 기본적으로 Naive Bayesian Classifier와 같으나, 추가적으로 스팸 메일의 분류단계에서 가중치를 부여하기 위해 가중치 테이블에 각 토큰들의 속성을 저장하는 과정을 거친다. 즉, 3절에서 언급한 메일의 특징들(5절의 표 1)을 각 토큰들로부터 추출한 뒤, *feature()* 함수를 통해 테이블 형태로 저장한다. 그 후에 스팸 분류 단계에서는 이 테이블을 이용하여 해당 속성을 갖는 토큰들에게 가중치를 부여하게 된다. 또한, 본 연구에서는 받은 메일이 스팸 메일인지 아닌지에 초점을 두고 있으므로 목적값의 집합인 C는 Cspam과 Cnonspam으로만 이루어진다.

그림 5분에 의해 큰 확률값을 갖는 c_i는 분류 대상 메일 D의 클래스가 된다. 예를 들어, Cspam이 임계값 T보다 더 큰 확률값을 갖게 되면 D는 스팸 메일로 분류 되어지는 것이다. 또한, 하나의 토큰 d_k에 해당하는 속성 $feature(d_k)$의 가중치 w_k는 $feature(d_k)$가 해당 클래스 c_i에서 갖는 빈도수에 의해 정의되어지는데, 이

Classifying_WBC(D)

Preporcessing(S) //학습을 위한 특징단어를 제외한 나머지 제거

While(D <> EOF) //분류 대상 메일 D의 끝까지 읽어들임

d_k = *parsing*(D) //D를 파싱하여 토큰을 d_k에 저장

w_k = *freq*(c_i, *feature*(d_k)) //d_k의 속성이 갖는 빈도수에 의한 가중치 w_k로 부여

Loop

Positions ← *Vocabulary*에 포함된 단어들의 D 안의 위치 c_{WB}를 돌려준다.

$$c_{WB} = \underset{c_i \in C}{argmax}\, P(c_i) \prod_{k \in Positions}^{n} \left(\frac{d_k w_k + 1}{\sum_{j=1}^{n} d_j w_j + n} \right) P(d_k | c_i)$$

그림 5 가중치가 부여된 베이지안 분류기를 이용해 문서를 분류하는 모듈

렇게 정의된 가중치 w_k를 통해 분류자는 스팸 혹은 논스팸 메일의 구분을 보다 빠르고 명확히 할 수 있게 된다[12].

4.4 지능형 에이전트(Intelligent Agent)를 통한 지속적인 사용자 행동 학습

베이지안 분류자를 이용한 필터링은 학습된 모듈을 통한 필터링 후에도 스팸/논스팸의 분류가 정확히 이루어지지 않는 경우가 발행할 수 있다. 이를 보완하기 위해 시스템은 사용자의 행동을 관찰(Monitoring), 학습(Learning)하는 지능형 에이전트를 이용하여 차후에 발생되는 유사한 사건에 대해 보다 정확한 판단을 가능하게 한다. 예를 들어, 그림 6과 같이 에이전트는 시스템

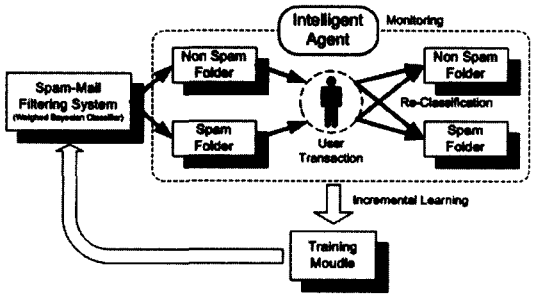


그림 6 지능형 에이전트를 통한 Incremental Learning

이 분류해 놓은 메일에 대한 사용자의 행동을 관찰하는데, 시스템이 논스팸으로 분류해 놓은 메일이 사용자에 의해 스팸으로 판단되어질 경우(사용자가 제목만 보고 지우거나, 내용확인 후 스팸으로 처리한 경우 등), 이러한 메일들은 Training Data로서 Training Module에 의해다시 학습되어지고 이 학습된 모듈을 이용해 차후의 메일에 대한 필터링의 정확성을 높이는 것이다.

5. 실험 및 결과

본 논문은 표 1과 같이 텍스트만을 기반으로 베이지안 필터링, 메일 헤더에 포함된 Subject의 내용을 기반으로 한 가중치 부여, HTML Tag들을 기반으로 한 가중치 부여, 하이퍼링크를 기반으로 한 가중치 부여, 마지막으로 이 4가지의 경우를 종합하여 실험한 총 5가지의 경우에 대한 필터링 성능을 평가하였다. 가중치 부여는 각각의 4가지 경우에 해당하는 단어들에 대해 일반 텍스트보다 일정크기 이상의 고정된 값을 줌으로서 성능을 실험하였다.

5.1 데이타 집합

본 논문의 실험을 위해서 IIS 5.0, Micorsoft Active Server Page와 MS-SQL Server을 사용해서 구현하였으며, 실험환경은 펜티엄3 1GHz, 256MB RAM의 시스템이었다.

표 1 Weighted Bayesian Classifier 성능 측정을 위한 상황 구분

		내 용
N_T	본문의 텍스트 기반	가중치 부여가 없이 일반적인 베이지안 분류자로 테스트
H_T	HTML Tag 기반	특정 HTML Tag와 함께 쓰인 단어들에 대한 가중치 부여
H_L	Hyper Link 기반	... 와 함께 쓰인 단어들에 대한 가중치 부여
H_S	HTML Subject 기반	E-Mail 헤더의 내용 중 Subject의 내용에 가중치를 부여
ALL	4가지 경우를 종합	N_T, H_T, H_L, H_S의 경우를 종합하여 테스트

트레이닝 및 테스트에 사용된 데이타들은 수집된 실제 영문 메일이며, 데이타들의 구성은 표 2와 같다. 테스트에 쓰인 모든 데이타는 트레이닝 데이타보다 나중에 온 메일이다.

표 2 데이타셋의 구성

Spam	329	148
NonSpam	247	53
합 계	576	201

학습데이타의 정제과정은 불용성 제거(Elimination of Stopwords) 단계와 스테밍(Stemming) 단계로 구성되는데, 불용성 제거 단계에서는 스팸/논스팸 메일 내에서 공백문자와 같이 빈도수가 상당히 높은 단어들을 걸러냄으로서 스팸 분류의 변별력을 높이는 작업이 이루어졌으며, 스테밍 단계에서는 특정 단어에서 파생된 여러 형태의 단어들을 하나의 형태 즉, 스템으로 대치함으로서 질의어와 문헌의 해당 단어 사이에 완전 정합을 저해하는 구문적 변형의 문제를 해결하는 작업이 이루어졌다[18].

5.2 성능 평가 방법

문서분류의 성능을 평가하기 위한 기준은 주로 정확도, 재현율 또는 F-measure 측정식이 사용는데[11], 식 (6)은 F-measure 측정식을 보여주고 있다. 식 (6)에서 P는 정확도(Precision), R은 재현율(Recall)을 의미하며, b는 정확도 P에 대한 재현율 R의 상대적 가중치를 나타내는 수치이다.

$$P(\mathrm{Precision}) = \frac{\text{'스팸' 으로 분류된 실제 '스팸' 메일 수}}{\text{'스팸' 으로 분류된 메일 수}}$$

$$R(\mathit{Recall}) = \frac{\text{'스팸' 으로 분류된 실제 '스팸' 메일 수}}{\text{전체 '스팸' 메일 수}}$$

$$F-\mathit{measure} = \frac{(b^2+1)PR}{b^2P+R} \qquad (6)$$

본 실험에서는 필터링 과정에서 정확도를 재현율보다 높게 해줌으로서 논스팸 메일이 스팸 메일로 잘못 분류되는 False-Positive 문제(식 (7))를 최소화시키기 위해 b를 0.5로 설정하여 분류 결과를 분석하였다.

$$\mathit{False-Positive} = \frac{\text{'스팸' 으로 분류된 '논스팸' 메일 수}}{\text{'스팸' 으로 분류된 메일 수}} \qquad (7)$$

5.3 실험 결과

5.3.1 F-measure 측정을 통한 최적의 임계값 $T_{optimal}$ 산출

각각의 임계값(Threshold) 변화에 따른 스팸 메일의 정확도와 재현율, F-measure 측정식에 의한 스팸 메일 필터링 성능은 그림 7과 같다.

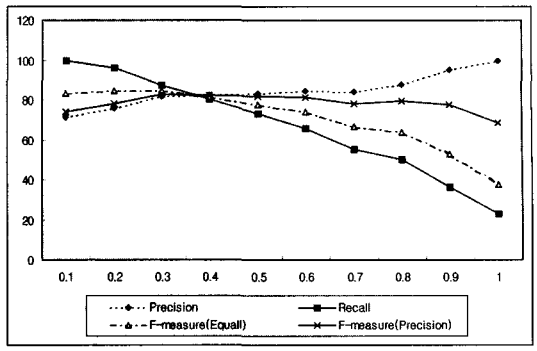


그림 7 임계값의 변화에 따른 '스팸'의 정확도, 재현율, F-measure 측정 결과

그림 7에서 알 수 있듯이 임계값을 0.1에서 1까지 변화시켜가며 성능을 측정한 결과, 스팸 메일의 재현율은 약 67.4%의 차이가 있었으며, 정확도에서는 약 28.3%의 차이를 보였다.

F-measure 측정에서는 b값을 1로 주었을 때 0.2에서 최적의 성능을 보였으나, 스팸 메일 분류의 정확도를 높이기 위해 b값을 0.5로 주었을 때 임계치 0.3에서 최적의 성능을 보였다. 따라서 임계값 $T_{optimal}$은 0.3으로 정하였다.

5.3.2 가중치에 따른 필터링 성능 측정

각 시스템을 구현하여 분류 성능을 비교한 것이 표 3과 그림 8이다.

나이브 베이지안 분류자만을 사용한 시스템(N_T)보다 제안하는 시스템이 대체로 같거나 높은 성능을 보였다. 스팸의 경우, 제안하는 시스템이 정확도 면에서

표 3 각 시스템의 성능 비교

스팸 정확도(P)	100%	88.89%	84.62%	100%	94.44%
스팸 재현율(R)	52.38%	76.19%	52.38%	61.9%	85.71%
스팸 F-measure	55.51%	81.60%	59.17%	65.20%	86.74%

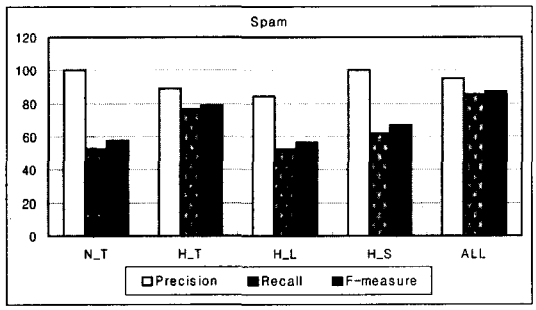


그림 8 스팸 메일에 대한 필터링 성능의 각 가중치별 측정 결과

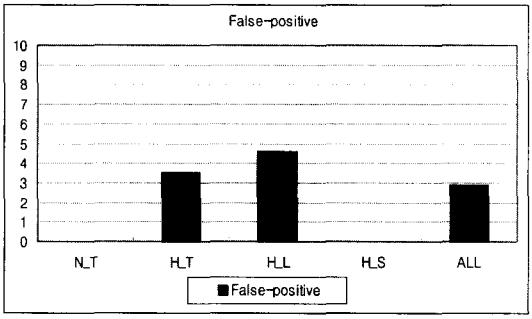


그림 9 각 가중치별 False-Positive 측정 결과

N_T 보다 5.7% 저조한 결과를 보였으나, 재현율에서 33.3%, F-measure 측정 결과에서 31.2%의 우수한 결과를 보였다.

정확도에서 N_T 보다 저조한 결과가 나온 원인으로는 첫째, 학습데이타의 부족, 둘째, 스팸 메일의 Feature를 갖는 일부 논스팸 메일이 스팸으로 잘못 분류되어진 것을 들 수 있다.

그림 9는 논스팸이 스팸으로 잘못 분류되어지는 False-Positive에 대한 측정 결과이다. 총 5가지의 가중치에서 H_T, H_L, ALL의 경우 False-Positive 문제가 발생하는데, 이는 메일서비스 업체에서 제공하는 E- Card, 음악메일 등에서 스팸 메일의 Feature로 규정된 HTML Tag, Hyper Link 등의 등장이 원인이었다. 또한 가중치를 부여하지 않은 시스템(N_T)에서 False_positive 문제는 평균 약 3.9% 적었으나, 반면에 가중치를 부여한 시스템보다 재현율이 평균 20% 저조한 성능을 보였다.

5.3.3 Incremental Learning에 의한 학습 및 필터링

스팸/논스팸 구분이 잘못 되어진 메일에 대한 에이전트의 재학습 후 필터링 성능 및 F-measure 측정값의 변화는 각각 그림 10, 그림 11과 같다.

그림 10에서 알 수 있듯이, Incremental Learning 후의 정확도는 거의 변함이 없었으나, 재현율은 4.76% 향

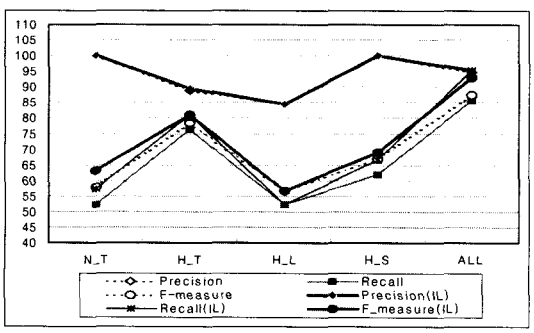


그림 10 스팸 메일에 대한 Incremental Learning 후의 필터링 결과

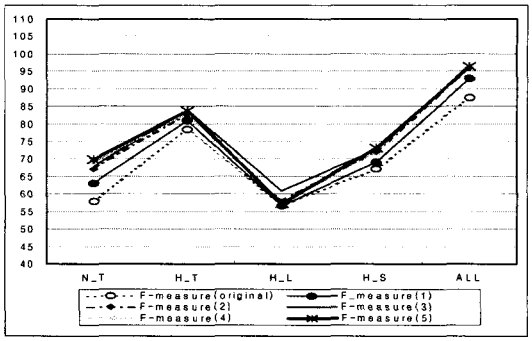


그림 11 5회의 Incremental Learning 후의 F-measure에 의한 필터링 성능 증가 그래프

상된 결과를 얻을 수 있었으며, 일정기간동안 Incremental Learning을 통한 다섯 차례 스팸 메일 필터링 결과는 그림 11과 같이 가중치를 부여한 다섯 경우에 대하여 평균 약 7.3%의 성능 향상 결과를 얻을 수 있었다. 특히 N_T의 경우 10.4%로 가장 큰 성능 향상을 보였으며, H_L의 경우 약 4.3%로 가장 작은 성능 향상을 보였다.

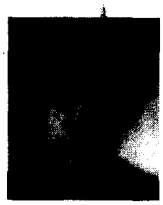
6. 결 론

본 논문에서는 스팸 메일 필터링을 위해서 나이브 베이지안 분류자의 개선된 형태인 가중치가 부여된 베이지안 분류자를 사용한 시스템을 제안하고 구현하였다. 제안한 시스템의 성능을 나이브 베이지안 분류자만 사용한 시스템, 메일 헤더의 Subject, HTML Tag, Hyper Link, 그리고 앞의 모든 것을 통합한 시스템, 총 5가지 순으로 비교 평가하였을 때 전반적으로 스팸 재현율과 정확도 등, 수신된 메일들 중에서 스팸 메일을 분류해 내는데 있어 단순한 나이브 베이지안 분류자에 의한 필터링 시스템보다 우수한 성능을 보였으며, 특히 ALL의 경우가 가장 높은 성능을 보였다. 또한 Incremental Learning을 이용한 지속적인 학습을 통해 스팸메일에 대한 필터링 성능은 시간이 경과에 따라 점차적으로 향상됨을 알 수 있었다. 향후 과제로는 수신된 메일들을 실시간으로 학습하여 최적의 가중치를 시스템 스스로 찾아낼 수 있는 방법을 연구하거나, False- positive 문제를 해결하는 방안을 연구하여 필터링 성능을 극대화시키는 방법, 또한 Support Vector Machine[19] 등의 다양한 이론들을 접목한 시스템을 통해 메일 필터링에 적합한 최적의 시스템을 제시할 수 있을 것이다.

참 고 문 헌

[1] 한국전산원, "국가정보화백서(National Informatization White Paper)", pp. 23, 2002.

[2] Internet E-mail Corporate Usage Report, www.securitymanagenment.com/library/worldtalk0200.html
[3] 정보통신부, 정보통신망 이용촉진 및 정보보호 등에 관한 법률 시행령 제11조 (영리목적의 광고성 전자우편의 명시방법), 2002.
[4] Ricardo, B.-Y. and Berthier, R.-N., Modern Information Retrieval, pp.27, Addison-Wesley, 1999.
[5] Provost, J.. "Naive-Bayes vs. Rule-Learning in Classification of Email," Technical report, Dept. of Computer Sciences at the U. of Texas ay Austin, 1999.
[6] Diao, Y., Lu, H. and Wu, D., "A Comparative Study of Classification Based Personal E-mail Filtering," Proc. of PAKDD-00, 4th Pacific-Asia Conference on Knowledeg Discovery and Data Mining, 2000.
[7] Cohen, W.W., "Learning Rules that Classify E-Mail," Proc. of the AAAI Spring Symposium on Machine Learning in Information Access, 1996.
[8] Sahami, M., Dumais, S., Heckerman, D. and Horvitz, E., "A Bayesian Approach to Filtering Junk E-Mail. In Learning for Text Categorization," Proc. of the AAAI Workshop, pp.55-62, Madison Wisconsin. AAAI Technical Report WS-98-05, 1998.
[9] Androutsopoulos, I., Koutsias, J., Chandrinos, K. V. and Spyropoulos, C. D., "An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-Mail Messages," Proc of the 23rd Annual International ACM SIGIR Conference on Reach and Development in Infromation Retrieval, 2000.
[10] Rev. T. B., "An essay toward solving a problem in the doctrine of chances," Philosopical Transactions of London, vol. 53, pp.370-418, 1763.
[11] Androutsopoulos, I., Koutsias, J., Chandrinos, K. V., Paliouras, G. and Spyropoulos, C. D., "An Evaluation of Naive Bayesian Anti-Spam Filtering," Proc of the 11th European Conference on Machine Learning, pp.9-17, 2000.
[12] Mitchell, T. M., Machine Learning, Chapter 6: Bayesian Learning, McGraw-Hill, 1997.
[13] Han, J., "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2001.
[14] http://popfile.sourceforge.net/
[15] Thomas, G. and Peter, A. F., "Weighted Bayesian Classification based on Support Vector Machine," Proc. of the 18th International Conference on Machine Learning, pp.207-209, 2001.
[16] 고수정, 이정현, "Apriori 알고리즘에 의한 연관단어 지식 베이스에 의한 가중치가 부여된 베이지안 자동 문서 분류", 멀티미디어학회 논문지 제4권 제2호, 2001.
[17] Ferreira, J. T. A. S., Denison, D. G. T., Hand, D. J., "Weighted Naive Bayes modelling for data mining," Technical report, Dept. of Mathematics at Imperial College. 2001.
[18] Russell, S. I. and Norving, P., Arificial Intelligence - A Modern Approach, Prentice Hall, pp.525-529, 1995.
[19] Denzinger, J. and Ennis, S., "Being the new guy in an experienced team - enhancing training on the job," Proc. of the 1st International Joint Conference on Autonomous Agents and Multiagent Systems, Pt3, 2002
[20] Graham, P., "Better Bayesian Filtering," Article of Spam Conference, 2003.
[21] 조한철, 조근식, "나이브 베이지안 분류자와 메세지 규칙을 이용한 스팸메일 필터링 시스템", 한국정보과학회, 제29회 춘계학술대회, 2002.
[22] Fox, C., "Lexical analysis and stop lists. In Information Retrieval: Data Structures and Algorithms," Prentice-Hall, 1992.
[23] Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," European Conference on Machine Learning, 1998.
[24] http://email.about.com/cs/bayesianspamsw/



김 현 준

2000년~2003년 인하대학교 전자계산학과 학사 졸업. 2003년~현재 인하대학교 대학원 석사과정. 관심분야는 인공지능, 데이타 마이닝, 협력적 필터링, 기계학습

정 재 은

1995년~1999년 인하대학교 기계공학과/전자계산학과 졸업(복수전공). 1999년~2002년 인하대학교 전자계산학과 석사 졸업. 2002년~현재 인하대학교 대학원 박사과정. 관심분야는 Intelligent Agent, CSP, Semantic Data Mining, Intelligent data analysis, Semi-supervised Learning

조 근 식

1978년~1982년 인하대학교 전자계산학과 졸업. 1983년~1985년 Queens College/CUNY 전자계산학 석사. 1985년~1991년 City University of New York 전자계산학 박사. 1992년~현재 인하대학교 컴퓨터정보공학과 교수. 관심분야는 지능형 소프트웨어 에이전트, 전자상거래, 전문가 시스템, 지식기반 스케줄링, Contraint Logic Programming 언어