신경망과 유전자 알고리즘을 이용한 스팸 메일 필터링 기법의 구현과 성능평가

김 범 배[†] · 최 형 기[††]

요 약

스팸 메일의 양의 급증함에 따라, 다양한 스팸 메일 필터링 기법이 제시되고 있다. 이런 필터링 기법 가운데, 학습 기반 필터링 기법은 현재 가장 보편화된 필터링 기법 가운데 하나이다. 본고에서는 신경망과, 유전자알고리즘, 카이제곱통계를 이용한 학습 기반 필터링 기법을 제시한다. 제안된 필터링 기법은 기존 필터링 기법의 문제를 해결하고, 스팸 메일 필터링에 높은 정확도를 제공할 수 있다. 제안된 필터링 기법은 스팸 메일 필터링 정확도와 정상 메일 필터링 정확도에서 각각 95.25%와 95.31%의 높은 정확도를 보인다. 이런 실험 결과는 기존의 규칙 기반 필터링 기법과 베이지안 필터링 기법에 비해 각각 7%, 12% 이상 높은 수치이다.

키워드 : 스팸 메일, 학습 기반 필터링 기법, 신경망, 유전자 알고리즘

Implementation and Experimental Results of Neural Network and Genetic Algorithm based Spam Filtering Technique

Bumbae Kim[†] · Hyoung-Kee Choi[††]

ABSTRACT

As the volume of spam has increased to extreme levels, many anti-spam filtering techniques have been proposed. Among these techniques, the machine-learning filtering technique is one of the most popular filtering techniques. In this paper, we propose a machine-learning spam filtering technique based on the neural network, the genetic algorithm and the χ^2-statistic. This proposed filtering technique is designed to overcome the problems in existing filtering techniques, and to achieve high spam filtering accuracy. It is able to classify spam and legitimate email with 95.25 percent and 95.31 percent accuracy. This accuracy of the spam filtering is 7.75 percent and the 12.44 percent higher than rule-based filtering and the Bayesian filtering technique, respectively.

Key Words : Spam, Machine-learning Filtering Technique, Neural Network, Genetic Algorithm

1. 서 론

인터넷 사용의 급격한 증가와 함께, 원격 이용자 간의 원활한 의사소통을 가능케 하는 이메일 서비스는 인터넷의 필수요소로 자리 잡고 있다. 이와 함께, 상업적 광고, 청소년 유해 자료, 웜 및 바이러스의 전파를 목적으로 하는 이메일도 그 양이 급증하고 있다. 스팸메일이라 불리는 이런 이메일들은 최근 들어 그 양이 급증하고 있으며, 인터넷 전반에 걸친 다양한 피해를 발생시키고 있는 인터넷 이슈로 대두되고 있다. 웹사이트 TopTenReivews에 따르면, 스팸 메일은 2006년 송신된 이메일의 40% 이상을 차지할 것으로 예측되며, 그 양도 지속적으로 증가해 2007년 약 63% 이상을 차지할 것이라 추정하고 있다[1]. 또한 이로 인한 피해도 급증하고 있다.

스팸 메일이 인터넷 전반에 걸쳐 다양한 피해를 발생시킴에 따라, 스팸 메일의 효과적인 차단을 위한 다양한 기법들이 제시되고 있다. 이 가운데서도 현재 Listing, DNS lookup, 그리고 단어 필터링 등의 기법은 널리 이용되고 있는 대표적 차단 기법들이다. Listing 기법은 이미 알려진 스패머의 정보를 등록시킴으로써, 스팸 메일의 차단을 시도하는 기법으로, 수신된 이메일이 등록된 정보와 부합할 경우, 이를 스팸 메일로 간주하는 기법이다. DNS lookup 기법은 송신자의 인증을 통해 스팸 메일을 차단하는 기법이다. 단어 필터링 기법은 수신된 이메일의 단어 또는 문장을 통해 스팸 메일의 여부를 판단하는 기법으로, 규칙 기반 필터링 기법과 학습 기반 필터링 기법으로 세분화된다. 이 가운데

※ 이 논문은 2005년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2005-003-D00373).
† 준 회 원 : 성균관대학교 정보통신공학부 컴퓨터공학과 대학원생
†† 정 회 원 : 성균관대학교 정보통신공학부 컴퓨터공학과 교수
논문접수 : 2006년 2월 8일, 심사완료 : 2006년 3월 10일

서도 학습 기반 필터링 기법은 현재, 안티 스팸 메일 분야에 있어, 가장 널리 사용되고 있는 스팸 메일 차단 기법 가운데 하나이다.

학습 기반 필터링 기법은 기 분류된 이메일 집합으로부터 학습을 시행하고, 이를 통해 수신되는 이메일의 스팸 메일 여부를 판단하는 방법으로, 규칙 기반 필터링 기법이 지닌 문제를 해결 하고, 높은 필터링 정확도를 제공할 수 있다는 장점이 있다. 규칙 기반 필터링 기법은 사전에 정의된 단어, 문장, 이메일 주소를 포함한 이메일을 스팸 메일로 간주하는 기법으로, 구현이 간단하고, 인터넷상의 적용이 용이해 스팸 메일 필터링에 종종 이용되고 있으나, 필터링 정확도가 낮고, 합법적인 이메일을 스팸 메일로 판단할 오류율이 높으며, 스팸 메일을 송신하는 스패머의 간단한 속임수에도 유연히 대처하지 못하는 등의 문제가 있다.

학습 기반 필터링 기법에서, 학습은 기 분류된 이메일 집합으로부터 합법적인 이메일과 스팸 메일의 특징을 파악하는 과정으로, 학습 기반 필터링의 정확도를 결정짓는 중요한 과정 가운데 하나이다. 현재 이용되는 대부분의 학습 기반 필터링 기법은 베이지안 필터링 기법을 이용하고 있다[2]. 베이지안 필터링 기법은 학습의 과정이 간단하고, 스팸 메일 판단에서도 높은 정확도를 제공할 수 있다는 장점 때문에, 대부분의 학습 기반 필터링 기법들이 채택하고 있는 기법이다. 그러나 베이지안 필터링 기법은 체계적인 학습 과정을 지니지 못했고, 낮은 효율의 학습 알고리즘을 이용하는 등, 스패머에게 악용될 수 있는 문제들을 내포하고 있다. 이러한 문제는 베이지안 필터링의 정확도를 떨어뜨리는 요인이다. 따라서 본고에서는 베이지안 필터링에서 나타나는 문제들을 해결하고, 보다 높은 필터링 정확도를 지닌 신경망과 유전자 알고리즘을 이용한 학습 기반 필터링 기법을 제안한다. 추가로, 제안된 필터링 기법은 χ^2-통계를 도입하여 유전자 알고리즘의 효율을 높인다. 유전자 알고리즘은 입력 값의 수가 증가함에 따라, 알고리즘의 수행시간이 기하급수적으로 증가하는 문제가 있다.

신경망과 유전자 알고리즘을 이용한 필터링 기법은 Linger라는 이메일 필터링 기법으로 기존에 제시되었다[3]. 그러나 Linger는 스팸 메일의 필터링보다 이메일의 그룹화에 주목적을 두고 있고, 기하급수적으로 증가하는 유전자 알고리즘의 수행시간을 해결하기 위해 도입한 알고리즘의 제시가 불분명하다. Linger는 도입 가능한 알고리즘을 열거하였으나, 실제로 도입한 알고리즘과 그에 대한 성능 평가가 이루어지지 않았다. 또, Linger의 필터링 성능에 대한 고찰도 부족하다. 본 필터링 기법은 Linger의 문제를 해결하기 위해, 스팸 메일 필터링을 목적으로 구현되었으며, χ^2-통계를 도입하여 유전자 알고리즘의 효율을 높이는데 이용한다.

본고에서는, 2장에서 스팸 메일 차단 기법에 대한 관련 연구를 정리하고, 3장에서 학습 기반 필터링 기법의 구성에 대해 설명한다. 4장에서는 제안된 필터링 기법에 이용된 알고리즘에 대해 정리하고, 적용 방법에 대해 설명하며, 5장에서는 제안된 필터링 기법의 정확도를 측정하고, 수행시간 측면과 정확도 측면에서 다른 필터링 기법과의 성능을 비교한다. 6장에서는 결론 및 향후 과제에 대해 언급했다.

2. 관련 연구

스팸 메일을 차단하는 기법들은 다음과 같은 리스트(list)에 의한 차단 기법, DNS(Domain Name System)에 의한 차단 기법, 그리고 워드 필터링 기법들이 이용되고 있다.

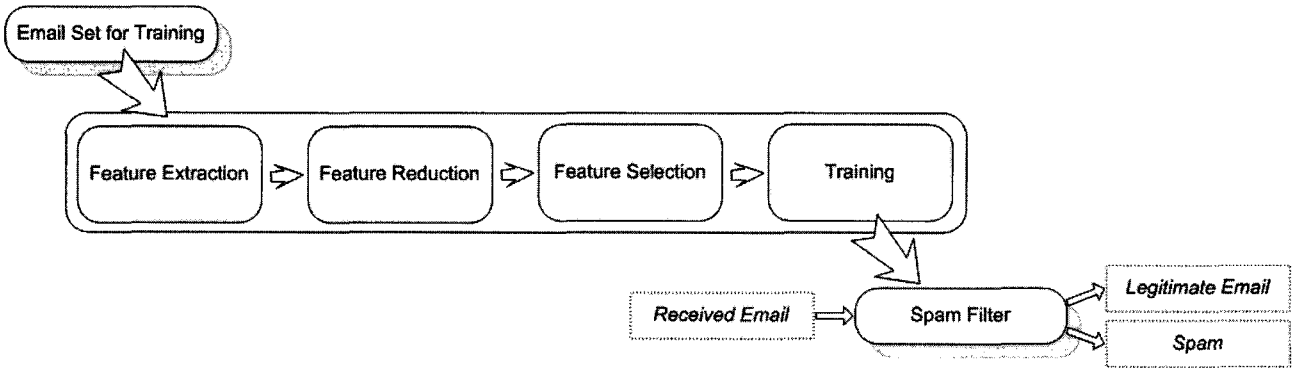
리스트에 의한 차단 기법은 블랙리스트(black list), 화이트리스트(white list), 그레이리스트(grey list) 기법으로 세분화되고, 각각의 리스트들은 이메일의 주소, 이메일 서버의 IP 주소 등을 등록, 관리한다. 블랙리스트 기법은 리스트에 등록된 이메일의 경우, 이메일의 송·수신을 거부하는 기법이며, 화이트리스트 기법은 리스트에 등록된 이메일에 대해서만 송·수신을 허가하는 기법이다. 그레이리스트 기법은 블랙리스트와 화이트리스트의 장점을 혼합한 방식으로 먼저 모든 이메일의 수신을 거부하고, 일정 시간 이내에 재송신되는 이메일에 한해, 수신을 허가하는 기법이다. 그러나 리스트에 의한 차단 기법은 스패머가 자신의 정보를 은닉하기 쉽고, 리스트의 관리가 어렵다는 문제 때문에 스팸 메일 차단에 낮은 정확도와 높은 오류율을 보이는 기법이다.

DNS에 의한 차단 기법은 스패머가 주로 이용하는 정보 은닉을 역이용한 기법으로, 이메일 내의 정보가 변경되었는지를 확인함으로써, 스팸 메일 여부를 판단하는 것이다. DNS에 의한 차단 기법은 최근 대형 IT업체와 ISP(Internet Service Provider)에 의해 발전되고 있는 기법으로, Pobox의 SPF, Microsoft의 SenderID, Yahoo의 DomainKeys, Cisco의 IIM 등의 기법들이 제시되고 있다[4, 5, 6, 7]. 그러나 개발 업체 간의 이해관계로 인하여 발전의 속도가 둔화되고 있으며, 표준이 제시되고 있지 못함으로써 일반적인 솔루션으로 활용되기 어렵다는 문제가 있다.

워드필터링 기법은 이메일 내에 포함되어 있는 단어를 기준으로 스팸 메일 여부를 판단하는 기법으로, 규칙 기반 필터링, 학습 기반 필터링으로 구분될 수 있다. SpamAssassin으로 널리 알려진 규칙 기반 필터링 기법은 사전에 규칙을 설정해두고 이에 부합하는 이메일을 스팸 메일로 간주하는 기법이며, 베이지안 필터링으로 널리 알려진 학습 기반 필터링 기법은 과거의 이메일을 바탕으로 학습을 진행하여, 새로 수신하는 이메일의 스팸 메일 여부를 판단하는 기법이다[8].

3. 학습 기반 필터링 기법의 학습 과정

일반적인 학습 기반 필터링 기법은 (그림 1)의 학습 과정을 지닌다[9]. (그림 1)에 나타난바와 같이, 학습 기반 필터링 기법은 먼저 학습의 대상을 추출하는 특징 추출(feature extraction)과정을 필요로 한다. 추출된 특징은 방대한 양의 데이터를 포함하고 있으므로, (그림 1)에서 보이듯이, 특징



(그림 1) 학습 기반 필터링의 학습 과정

단순화(feature reduction), 특징 선택(feature selection)의 과정을 추가로 거치게 된다. 그 후, 선택된 최적의 특징에 대해서만 학습(training)과정을 진행하며, 학습된 필터는 수신된 이메일에 대해 스팸 메일 여부를 판단하게 된다.

이메일은 이메일 헤더와 메시지 등에 이메일의 특징을 반영하는 정보 -이메일 주소, 제목, 단어, 문장 등 -를 포함하고 있다. 학습 기반 필터링 기법은 이런 정보들을 학습함으로써 스팸 메일의 필터링을 시도하지만, 이런 정보들은 그 양이 매우 방대하기 때문에 학습의 효율을 떨어뜨리는 문제가 있다. 따라서 학습 과정에서는 합법적인 이메일과 스팸 메일의 특징을 잘 반영하는 이메일의 구성 요소를 학습의 대상으로 추출하는 과정이 필요하다. 일반적으로 이메일의 제목, 전달하고자 하는 메시지 등이 학습의 대상으로 추출된다. 추출된 특징들은 바로 학습에 이용될 수 있다. 그러나 이들은 아직까지도 방대한 양의 데이터를 포함하고 있고, 학습에 영향을 미치지 않거나, 학습의 효과를 저해할 수 있는 요소들을 다수 포함하고 있기 때문에, 최적의 학습 효과를 낼 수 없다는 문제가 있다. 이를 위해 특징 단순화, 특징 선택 과정을 통해, 최적의 학습 효과를 낼 수 있는 특징들을 선택해야 한다. 이후, 선택된 특징들을 학습 과정에서 반복적으로 학습한다. 이때, 학습의 효과는 학습 알고리즘의 학습 효율에 따라 결정된다.

제안된 필터링 기법 역시 이와 같은 일반적인 학습 과정을 거친다. 제안된 필터링 기법은 특징 추출에서 이메일의 메시지만을 학습의 대상으로 추출하게 되며, 특징 단순화, 특징 선택 과정에서는 각각 χ^2-통계와 유전자 알고리즘을 도입해, 학습에 이용될 최적의 단어를 선택하게 된다. 학습 과정에서는 신경망을 도입해 선택된 최적의 단어를 학습하게 된다.

4. 이론적 배경 및 구현

본 필터링 기법은 체계적인 학습 과정과 보다 높은 필터링 정확도를 위해 다음과 같은 1) χ^2-통계, 2) 유전자 알고리즘, 3)신경망 등의 알고리즘을 도입한다. 본 필터링이 이용한 알고리즘의 이론적 배경과 실제 구현은 다음과 같다.

4.1 χ^2-통계

χ^2-통계는 범주의 기대 량과 실제 측정되는 통계량과의 상관관계를 바탕으로, 각 범주간의 상관성을 측정하는 기법이다. 예를 들어, χ^2-통계는 나이에 따라 선호하는 약의 종류가 다르다는 가설에 대해서 나이에 따른 약의 선호도를 측정하고, 측정된 통계량과 각각의 범주의 기대 량 간의 비교를 통해, 가설의 가부를 검증한다. 또 나이에 따라 선호도가 높은 약의 종류에 대해서도 검증과정에서 파악할 수 있다.

제안된 필터링 기법에서 χ^2-통계는 특징 단순화 과정에서 합법적인 이메일과 스팸 메일에 대해 높은 상관도를 지니는 단어를 선택하는데 도입된다. 이때, 높은 상관도를 지니는 단어는 합법적인 이메일과 스팸 메일의 특징을 잘 반영하는 키워드로, 합법적인 이메일과 스팸 메일에 다수 등장하는 단어를 말한다. χ^2-통계는 특징 추출 과정에서 선택된 모든 메시지의 단어에 적용되며, 다음과 같은 식으로 계산된다.

$$X^2(w, spam) = \frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \quad \text{(식 1)}$$

(식 1)은 스팸 메일($spam$)에 대한 임의의 단어(w)의 χ^2-통계 값으로, A는 단어 w을 포함한 스팸 메일의 수를, B는 단어 w을 포함한 합법적인 이메일의 수를, C는 단어 w을 포함하지 않은 스팸 메일의 수를, D는 단어 w을 포함하지 않은 합법적인 이메일의 수를 말한다. N은 합법적인 이메일과 스팸 메일의 총합을 뜻한다.

(식 1)과 마찬가지로, 합법적인 이메일($legitimate\ email$)에 대한 χ^2-통계 값을 계산할 수 있다. 구해진 각각의 χ^2-통계 값은 (식 2)을 통해, 단어 w에 대한 χ^2-통계로 산출된다.

$$X^2 = P(spam)X^2(w, spam) + P(legitimate\ email)X^2(w, legitimate\ email) \quad \text{(식 2)}$$

(식 2)은 각각의 χ^2-통계 값에 스팸 메일과 합법적인 이메일이 나타날 확률을 곱한 값으로, 특징 단순화 과정에서

높은 χ^2-통계 값을 갖는 단어를 순서대로 선택하게 된다. χ^2-통계는 그 과정만으로도 학습을 위한 최적의 특징을 선택할 수도 있으나, 제안된 필터링 기법에서는 χ^2-통계를 특징 선택 과정에서 유전자 알고리즘의 수행시간을 줄이기 위한 선 처리 과정으로만 이용한다.

4.2 유전자 알고리즘

유전자 알고리즘은 자연 생태계에서 진화과정을 모델링 한 알고리즘으로, 현 세대의 우성인자를 다음 세대로 전이, 보다 나은 우성인자의 발생을 목적으로 하는 이론이다.

제안된 필터링 기법에서 유전자 알고리즘은 합법적인 이메일과 스팸 메일의 특징을 가장 잘 반영하는 최적의 단어를 선택하는 특징 선택 과정에 도입된다. 최적의 단어를 선택함에 있어, 특징 선택 과정은 특징 단순화 과정과 마찬가지로, 합법적인 이메일과 스팸 메일에 나타나는 단어들의 빈출 횟수를 그 기준으로 삼는다.

유전자 알고리즘은 입력 값의 수가 늘어날수록 알고리즘의 수행시간이 기하급수적으로 증가하는 문제가 있다. 해당 문제의 해결을 위해 제안된 필터링 기법은 χ^2-통계를 특징 단순화 과정에 도입하여, 유전자 알고리즘이 처리해야 하는 입력 값의 수를 줄이는데 이용한다. 보다 자세한 유전자 알고리즘의 사항은 [10]에 서술되어 있다 .

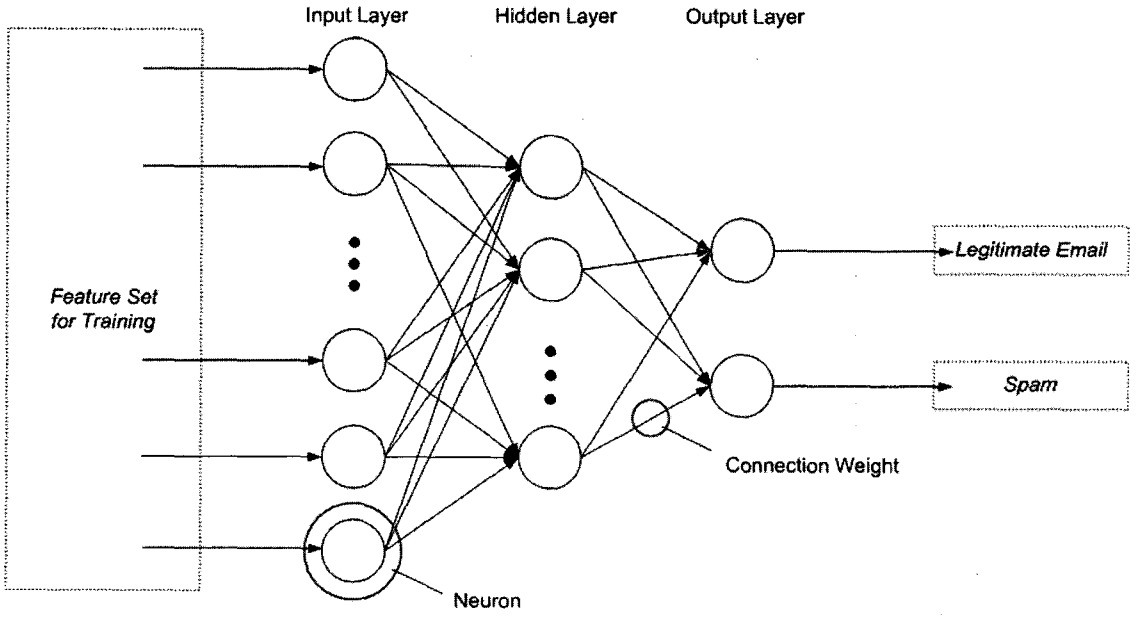
4.3 신경망

학습에 따라 신경세포(neuron)들 사이의 연결 강도(connection weight)가 변화한다는 생물학적인 학습의 원리를 모델링 한 알고리즘으로, 반복적인 학습을 통해 예측과 분류, 군집 분석에 이용되는 알고리즘이다. 신경망에서의 학습이란 신경세포간의 적절한 연결강도를 부여하는 것을 말한다.

제안된 필터링 기법에서 신경망은 (그림 2)와 같이 구성된다. (그림 2)에서 나타나듯이, 신경망은 input layer, hidden layer, output layer로 구성되어 있으며, 각 layer는 신경세포(neuron)로 구성되어 있다. 또 각 layer들 간의 신경 세포들은 적절한 연결 강도로 연결되어 있다. 연결 강도란 신경 세포들이 값을 다른 layer로 전달할 때, 부여하는 가중치를 말한다. 제안된 필터링 기법에서 input layer의 신경세포는 특징 선택 과정에서 선택된 최적 단어의 수에 따라 그 수가 결정되며, 각각의 신경세포는 선택된 단어와 연결되어 있다고 가정한다. Hidden layer의 신경세포의 수는 기존의 연구가 제시한 방법에 의해 결정 된다[11]. 해당 연구는 최적의 학습 효과를 낼 수 있는 신경망을 구성하는 방법을 제안한 것으로, input layer와 output layer의 신경세포 수가 결정되었을 때, hidden layer의 신경세포 수를 산출해 낼 수 있다. 제안된 필터링 기법의 output layer는 2개의 신경세포로 구성된다. 이들은 각각 합법적인 이메일과 스팸 메일과 연결되어 있다.

제안된 필터링 기법의 신경망은 각각의 이메일에 대해 다음과 같이 학습한다. 먼저 input layer의 신경세포는 신경세포와 연결된 단어들이 이메일에 포함되어 있는가를 확인한다. 그 후, 각 신경세포는 해당 단어가 이메일 내에 포함되어 있을 경우 1을, 포함되지 않았을 경우 0의 값을 부여받게 되고, 이 값을 hidden layer의 연결 세포들로 임의의 연결강도를 부여해 전달하게 된다. Hidden layer의 신경 세포들로 전달된 입력 값은 최종적으로 output layer의 2개의 신경세포로 동일하게 전달되게 되는데, output layer의 2개 노드 가운데, 더 큰 값의 신경세포가 제안된 필터링 기법의 판단 결과가 되는 것이다.

신경망이 올바른 판단 결과를 갖기 위해서는 먼저 기 분류된 합법적인 이메일과 스팸 메일을 통해, 각 신경세포간



(그림 2) 신경망의 구조

의 연결 강도를 적절히 설정하는 학습 과정이 필수적이다. 학습이 완료되면, 학습된 신경망을 수신되는 이메일의 스팸 메일 여부 판단에 이용하게 된다.

5. 필터링 기법의 정확도 측정 및 성능 평가

본 장에서는 제안된 필터링 기법의 정확도를 측정하고, 필터링 기법의 정확도가 학습에 이용되는 최적의 단어 수가 변화함에 따라 어떻게 변화하는지를 관측하였다. 또, 제안된 필터링 기법의 성능을 수행시간 측면, 정확도 측면에서 심도 있게 평가하였다. 특히, 정확도 측면에서의 성능 평가는 동일한 이메일 데이터 집합에 대한 규칙 기반 필터링 기법, 베이지안 필터링 기법과 비교, 분석하고 평가하였다.

5.1 필터링 기법의 구성

본 필터링 기법의 정확도를 측정하고 성능을 평가하기 위해, 본 필터링 기법은 다음과 같이 구성되었다.

- *특징 추출*: 이메일의 구성 요소 가운데, 메시지만을 학습 대상으로 추출한다.
- *특징 단순화*: 이메일의 메시지 가운데, χ^2-통계를 도입해 2048 개의 단어를 선택한다.
- *특징 선택*: 선택된 2048개의 단어 가운데, 유전자 알고리즘을 도입하여, 학습에 최적화된 100개의 단어를 선택한다. 최적 단어 수의 변화에 따른 필터링 정확도를 알아보기 위해 200, 300개로 최적 단어의 수를 변경, 선택한다.
- *학습*: 신경망은 3장의 나절에서 언급된 기존의 연구 결과에 따라, hidden layer의 신경 세포 개수를 <표 1>과 같이 결정하며, 이메일 집합에 대해, 100회의 반복 학습을 시행한다.

〈표 1〉 신경망의 구성

Input Layer	Hidden Layer	Output Layer
100	6	2
200	3	2
300	3	2

5.2 정확도 측정 및 성능 평가를 위한 이메일 집합의 구성

제안된 필터링 기법의 정확도 측정 및 성능 평가를 위해 본 절에서는 LingSpam의 기 분류된 이메일 집합을 이용 한다[12]. LingSpam은 대량의 합법적인 이메일과 스팸 메일을 수집, 이를 분류해 놓았으며, HTML Tag, 문장 기호, 특수 문자 등, 학습에 큰 영향을 미치지 않는 요소들을 사전에 제거해 놓았기 때문에 전처리 없이 이용이 가능하다는 장점이 있다. 그러나 동일한 메시지를 지닌 다수의 이메일이 중복되어 포함되어 있기 때문에, 정확한 성능 평가를 위해서 중복된 이메일을 제거하는 작업이 추가로 필요하다. 또 합법적인 이메일에 비해 스팸 메일의 양이 매우 적다는 문제가 있다.

기 분류된 이메일 집합을 <표 2>와 같이 구분된다. 학습 집합(Training Set) 은 필터링 기법의 절차에 따라 신경망의 학습에 이용되는 이메일 집합이며, 테스트 집합(Test Set)은 학습이 완료된 필터링 기법의 정확도와 성능을 비교 측정하기 위해, 이용되는 이메일 집합을 말한다. 각 집합은 합법적인 이메일과 스팸 메일이 유사한 비율로 구성되어 있다.

〈표 2〉 기 분류된 이메일 집합의 구성

	Legitimate Email	Spam
Training Set	901	188
Test Set	569	128

5.3 필터링 기법의 정확도 측정

제안된 필터링 기법의 정확도는 <표 3>과 같다. <표 3>에서 스팸 메일 정확도(Spam Accuracy)는 실제 스팸 메일을 스팸 메일로 판단할 확률을 말하며, 합법적인 이메일 정확도(Legitimate Email Accuracy)는 실제 합법적인 이메일을 합법적인 이메일로 판단 할 확률을 말한다.

<표 3>에서 볼 수 있듯이, 제안된 필터링 기법은 신경망이 학습에 이용할 수 있는 단어의 수가 증가할수록 필터링의 정확도가 증가한다. 학습 단어를 100개 사용했을 때, 스팸 메일 필터링의 정확도는 78.43%이나, 학습 단어를 300개로 늘렸을 경우, 제안된 필터링 기법은 스팸 메일 필터링의 정확도가 95.25%로 증가한다. 그러나 제안된 필터링 기법은 학습에 이용되는 단어의 수를 늘릴수록 제안된 필터링 기법의 정확도는 높아지나, 지속적으로 정확도가 증가하지는 않는다. 본 절의 측정에서는 제안된 필터링 기법의 정확도가 학습에 이용되는 단어의 수가 300개 까지는 높아졌으나, 그 이후부터는 기존의 정확도에 비해 크게 변화하지 않았다. 오히려 학습할 단어의 수가 많아짐으로써, 신경망의 학습 효율이 떨어지는 문제가 발생했다. 이는 제안된 필터링 기법은 LingSpam의 이메일 집합에 대해 300개의 최적 단어가 제공되면 최적의 학습을 시행할 수 있음을 뜻한다. 그 이상의 단어들은 합법적인 이메일과 스팸 메일의 특징을 기존의 단어들에 비해 잘 반영하지 못하기 때문에, 학습에 큰 영향을 미치지 못한다.

〈표 3〉 신경 세포에 따른 필터링 기법의 정확도 변화

Input Neurons	Legitimate Email Accuracy	Spam Accuracy
100	98.43	78.43
200	95.31	93.14
300	95.31	95.25

<표 3>에서, 제안된 필터링 기법의 정확도는 학습에 참여하는 단어의 수가 100개일 때 합법적인 이메일의 필터링 정확도가 98.43%로, 200개, 300개일 때 보다 3.12%정도 높은 것으로 나타난다. 그러나 이 현상은 제안된 필터링 기법이 테스트 집합의 대다수 이메일을 합법적인 이메일로 잘못 판단하기 때문에, 합법적인 이메일의 필터링 정확도가 98.43%로 높게 나타나는 것이며, 스팸 메일 필터링의 정확도가 78.43%로 낮게 나타나는 것이다. 스팸 메일과 합법적인 이메일을 잘못 판단할 확률은 100% 로부터 정확도의 차를 각각 계산 할 수 있으며, <표 3>에서 단어의 수가 300개 일 때, 각각 4.75% 그리고 4.69%이다.

5.4 필터링 기법의 성능 평가

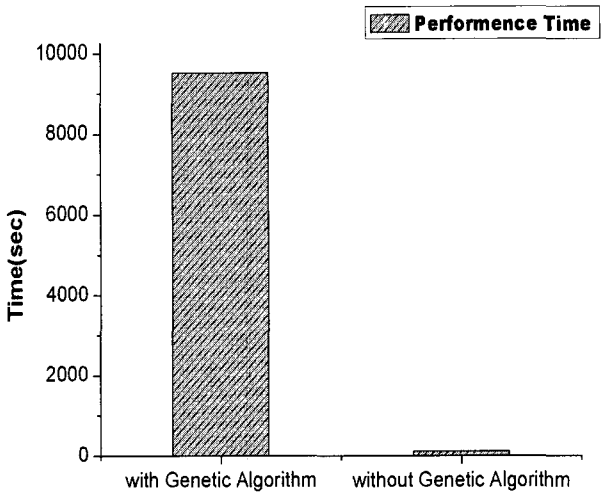
제안된 필터링 기법의 성능은 다음과 같은 1) 수행 시간, 2) 정확도 측면에서 평가한다. 수행 시간 측면에서의 성능 평가는 동일한 이메일 집합을 학습하는데 필요한 수행 시간을 유전자 알고리즘의 유무에 따라 평가한 것이며, 정확도 측면에서의 성능 평가는 필터링의 정확도를 규칙 기반 필터링 기법, 베이지안 필터링 기법과 비교, 평가 한 것이다.

(그림 3)은 유전자 알고리즘의 유무에 따른 필터링 기법의 수행 시간을 평가한 것이다. (그림 3)의 유전자 알고리즘의 존재(with genetic algorithm)는 특징 선택 과정에서 유전자 알고리즘을 도입한 경우의 수행시간을 나타내며, 유전자 알고리즘 부재(without genetic algorithm)는 유전자 알고리즘을 이용한 특징 추출 과정을 거치지 않고, 특징 단순화 과정에서 도입된 χ^2-통계만을 이용해 특징을 선택한 경우의 수행 시간을 나타낸다. (그림 3)에서 볼 수 있듯이, χ^2-통계를 이용해 특징을 선택한 경우에는 평균적으로 104.64초를 필요로 하나, 유전자 알고리즘을 도입한 경우에는 평균적으로 9501.853초를 필요로 한다. χ^2-통계를 이용하여 특징을 선택하는 경우, 표 4와 같이, 최적 단어의 수가 300개일 때 합법적인 이메일과 스팸 메일의 정확도는 각각 88.28%, 93.32%로, 유전자 알고리즘을 이용한 경우에 비해 7.08%, 1.93% 떨어지나 수행 시간 측면에서는 91배의 효과를 나타낸다.

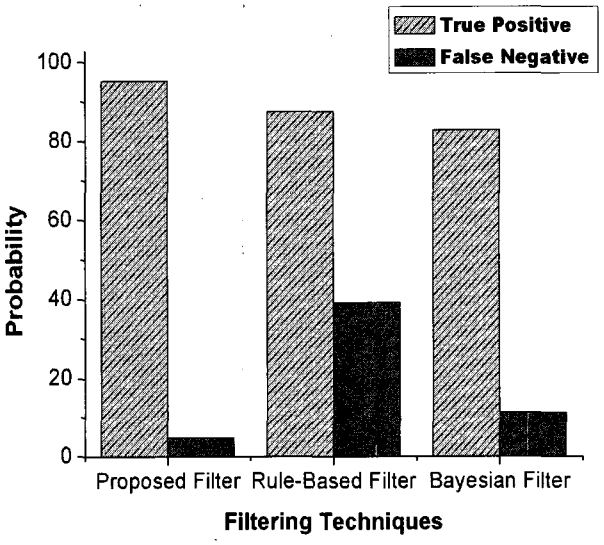
(그림 4)는 제안된 필터링 기법의 성능을 정확도 측면에서 비교, 분석한 결과이다. (그림 4)의 탐지율(true positive)은 실제 스팸 메일을 스팸 메일로 판단할 확률을 나타내며, 오탐지율(false negative)은 실제 합법적인 이메일을 스팸 메일로 오판단할 확률을 나타낸다. 스팸 메일 필터링 분야에서는 일반적으로 스팸 메일을 오탐지 할 확률(false positive)보다 합법적인 이메일을 오탐지 할 확률(false negative)에 더 많은 관심을 두고 있다. 따라서 본고에서도 이에 대한 실험치 만을 언급하고 이를 주로 살펴본다.

(그림 4)에서 볼 수 있듯이, 제안된 필터링 기법의 탐지율은 다른 필터링 기법에 비해 탁월하다. 제안된 필터링 기법은 95.25%의 정확도를 나타내는 반면, 규칙 기반 필터링 기법은 87.5%의 정확도를 보이며, 베이지안 필터링 기법은 약 82.81% 밖에 미치지 못한다. 또한, 스팸 메일 필터링 기법에서 가장 문제가 되는 오탐지율에 대해서도 제안된 필터링 기법은 낮은 수치를 보인다. 제안된 필터링 기법은 오탐지율이 약 5%가 미치지 않는 반면, 베이지안 필터링 기법은 11.25%, 규칙 기반 필터링 기법은 무려 39.20%에 육박한다.

일반적으로 규칙 기반 필터링 기법은 사전에 정의 해 놓은 규칙에 따라 그 정확도가 크게 좌우되므로, 해당 이메일 데이터 집합에 맞는 규칙을 적용할 경우, 그 정확도가 크게



(그림 3) 필터링 기법의 수행시간 성능평가



(그림 4) 필터링 기법의 정확도 성능평가

<표 4> χ^2-통계를 이용한 필터링 기법의 정확도 변화

Input Neurons	Legitimate Email Accuracy	Spam Accuracy
100	92.96	82.95
200	90.62	91.56
300	88.28	93.32

증가 할 수 있으며, 때에 따라서는 매우 낮게 나타날 수 있다. 본 절에서 LingSpam 이메일 집합에 대해서 측정한 (그림 4)의 값 역시도 규칙 기반 필터링 기법의 규칙이 본 절에서 이용한 이메일 데이터 집합과 잘 부합하지 않은 것을 원인으로 들 수가 있다.

그러나 베이지안 필터링 기법의 결과는 일반적인 결과와 다르다. 일반적으로 베이지안 필터링 기법은 약 98% 이상의 스팸 메일 필터링 정확도를 지닌다고 알려져 있다[13]. 그러나 (그림 4)의 결과는 이와 같은 정확도를 만족 시키지 못한다. 이는 베이지안 필터링 기법이 아직까지 충분한 학습이 진행되지 않았음을 뜻한다. Sam Holden의 실험에서는 높은 필터링 정확도를 위해 약 600개의 스팸 메일을 베이지안 필터링 기법의 학습에 이용했고, 높은 정확도를 얻을 수 있었다[14]. 그러나 본 절에서 이용한 이메일 집합은 그 양이 매우 제한적이기 때문에, 약 188개의 스팸 메일을 학습에 이용했고, 이를 측정했다.

이러한 점에서 미루어 볼 때, 제안된 필터링 기법의 성능은 정확도 측면에서 매우 탁월하다. 제안된 필터링 기법은 규칙 기반 필터링 기법, 베이지안 필터링 기법에 비해서도 높은 필터링 정확도를 지니며, 베이지안 필터링에 비해 적은 양의 스팸 메일로도 높은 정확도의 필터링을 가능토록 할 수 있다.

6. 결론 및 향후 과제

스팸 메일로 인한 인터넷 피해가 날로 증가하고 있는 현재, 학습 기반 필터링 기법은 필수적인 스팸 메일 차단 기법으로 자리 잡고 있다. 본고에서는 베이지안 필터링 기법이 가진 문제점을 극복하고, 보다 높은 정확도의 필터링을 제공 할 수 있는 신경망과 유전자 알고리즘을 이용하는 필터링 기법을 제안하고 구현하였다. 또한 수행 시간 측면과 정확도 측면에서의 제안된 필터링 기법의 성능을 평가하였다. 제안된 필터링 기법은 수행 시간에 약 9000 초가 필요한 문제는 있으나, 스팸 메일 필터링에 있어 95.25%의 정확도를 지니며, 규칙 기반 필터링 기법과 베이지안 필터링 기법에 비해 필터링 기법에 비해 각각 7.75%, 12.44% 높은 정확도를 지니는 것으로 나타났다.

향후 연구계획으로는 유전자 알고리즘으로 인한 본 필터링 기법의 수행시간 문제를 해결할 수 있는 대체 알고리즘에 관한 연구와, 신경망 보다 효율이 높은 학습 알고리즘과 그의 적용에 관해 연구를 수행 할 예정이다. 추가로, 영문, 국문 이메일에 나타나는 스패머의 트릭을 연구하고, 이에 대한 효과적인 필터링 기법에 관한 연구도 병행할 예정이다. 대량의 이메일의 수집을 통해 보다 신뢰도 높은 실험치의 도출을 목적으로 한다. 이러한 학습 기반 필터링 기법은 현재, 범람하는 스팸 메일의 양의 효과적으로 줄이고, 나아가 인터넷 전반을 비롯해 휴대폰 상에서도 발생하는 스팸의 효과적인 필터링을 가능할 것으로 예측된다.

참 고 문 헌

[1] TopTenReviews, "Spam Statistics 2006", available at http://spam-filter-review.toptenreviews.com/spam-statistics.html

[2] Graham Paul, "A Plan For Spam", available at http://www.paulgraham.com/spam.html, 2002.

[3] James Clark, Irena Koprinska and Josiah Poon, "E-mail classification : A hybrid approach combining genetic algorithms with neural networks"

[4] Pobox, SPF, "How it works", available at http://spf.pobox.com/howworks.html

[5] Microsoft SenderID, "Sender ID Framework Overview", available at http://www.microsoft.com/mscorp/safety/technologies/senderid/overview.mspx

[6] Yahoo! DomainKeys, "DomainKeys : Proving and Protecting Email Sender Identity", available at http://antispam.yahoo.com/domainkey

[7] Jim Fenton, "Identified Internet Mail", Cisco System, 2004 available at https://antiphishing.kavi.com/events/Conference_Notes/Jim_Fenton_on_Cisco_Internet_Identified_Mail.pd

[8] SpamAssassin, "The Apache SpamAssassin Project", available at http://spamassassin.apache.org/

[9] William S. Yerazunis, Shalendra Chhabra, Christian Siefkes, Fidelis Assis and Dimitrios Gunopulos, "A Unified Model Of Spam Filtration", 2005 MIT Spam Conference, Jan., 2005.

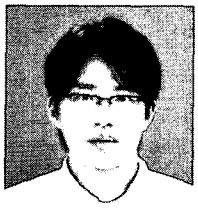
[10] Darrell Whitley, "A Genetic Algorithm Tutorial", Statistic and Computing, Vol.4, 1994, pp.65~85.

[11] T.A. Andrea and Hooshmand Kalayeh, "Application of Neural Networks in Quantitative Structure-Activity Relationships of Dihydrofolate Reductase Inhitors", J. Med. Chem. 34, pp.2824~2836, 1991.

[12] Internet Contents Filtering Group, "Ling-Spam", available at http://www.iit.demokritos.gr/skel/i-config

[13] William S. Yerazunis, "The Spam-Filtering Accuracy Plateau at 99.9% Accuracy and How to Get Past It", 2004 MIT Spam Conference, Jan., 2004.

[14] Sam Holden, "Spam Filters", available at http://freshmeat.net/articles/view/964, Aug., 2003.



김 범 배

e-mail : panic01@ece.skku.ac.kr

2005년 성균관대학교 정보통신공학부(공학사)

2005년~현재 성균관대학교 컴퓨터공학과 석사과정

관심분야 : 스팸 메일, 트래픽 측정 및 특징 분석, 인터넷 보안 등

최 형 기

e-mail : hkchoi@ece.skku.ac.kr

1992년 성균관대학교 전자공학과(공학사)

1996년 Polytechnique University 전기전자(공학석사)

2001년 Georgia Institute of Technology 전기전자(공학박사)

2001년~2004년 미국 Lancope. Inc. 연구원

2004년~2006년 성균관대학교 정보통신공학부 전임강사

2006년~현재 성균관대학교 정보통신공학부 조교수

관심분야 : 인터넷 보안, 모바일 커뮤니케이션 등