베이지안을 이용한 인터넷 커뮤니티 상의 유해 메시지 차단 기법

김 범 배[†] · 최 형 기[††]

요 약

스팸의 피해가 이메일 서비스를 넘어 인터넷 전반에 걸쳐 급증하는 현재 인터넷은 익명성을 악용하여 해당 커뮤니티의 공동 관심사와는 무관한 메시지들, 즉 상업적 광고, 상호비방, 종교 홍보 등의 스팸 메시지들을 게재하면서 심각한 사회적 문제를 일으키고 있다. 본고에서는 인터넷 커뮤니티 상의 스팸 메시지를 해결하고자 기존의 스팸 메일 차단에 이용되고 있는 베이지안 접근법을 적용한 인터넷 커뮤니티 상의 스팸 메시지 차단 방법을 소개한다. 나아가 인터넷 커뮤니티 상에서의 스팸 메시지 필터링의 효과를 증대시키기 위한 방편으로 스팸 메시지를 다양한 소분류로 세분화가 가능토록 구성했다. 이는 인터넷 커뮤니티의 다양한 이용자의 요구를 충족시키기 위한 방안이다. 구현된 베이지안 필터링 기법은 현재 운영되고 있는 사이트들을 대상으로 정확도를 측정하였다.

키워드 : 스팸, 학습 기반 필터링 기법, 베이지안 접근법

Spam Message Filtering with Bayesian Approach for Internet Communities

Bumbae Kim[†] · Hyoung-Kee Choi[††]

ABSTRACT

Spam Message has been causing widespread damages on the Internet. One source of the problems is rooted from an anonymously posted message in the bulletin board in Internet communities. This type of the Spam messages tries to advertise products, to harm other's reputation, to deliver religious messages and so on. In this paper we present the Spam message filtering using the Bayesian approach. In order to increase usefulness of the Spam filter in the bulletin board in Internet communities, we made the Spam filter which can divide the Spam message into six categories such as advertisement, pornography, abuse, religion and other. The test conducted against messages posted on the popular web sites.

Key Words : Spam Message, Communitiy Spam Machine-learning Filtering Technique, Bayesian Approach

1. 서 론

인터넷 사용의 급증과 함께, 원격 이용자들 간의 의사소통을 담당하는 서비스들은 인터넷의 필수 요소로 빠르게 자리 잡고 있다. 대부분의 이런 서비스들은 원활한 의사소통 환경의 제공은 물론, 대량의 정보 및 지식의 전파와 공유를 가능케 하는 등, 인터넷의 순기능에 큰 역할을 담당하고 있다. 이런 상황 하에서 의사소통의 저해와, 자원 및 인력의 낭비 그리고, 인터넷 윤리 논란 문제를 야기하는 전자 메시지들은 인터넷에서 반드시 해결해야 할 중대한 인터넷 이슈로 빠르게 대두되고 있다. 스팸이라 불리는 이런 메시지들은 상업적 광고, 청소년 유해물, 비방과 욕설 등을 주제로 수신자의 의사와 상관없이 대량으로 송신되고 있다.

의사소통의 수단이 이메일로 대표되던 인터넷 초기에, 스팸은 이메일 서비스에 국한된 문제였다. 그러나 인터넷 커뮤니티, 인스턴스 메시지 등 의사소통을 위한 다양한 인터넷 서비스가 제공됨에 따라, 스팸은 이런 서비스들을 매개체로 그 양과 피해를 급증시키고 있는 대표적 인터넷 윤리 문제로 발전하였다. 현재 스팸은 송신되는 이메일의 40%이상을 차지하고 있으며, 이로 인해 매해 약 230 억원의 피해가 발생하고 있다[1, 2]. 또한, 스팸은 매해 인스턴스 메시지의 형태 (SPIM)로 30억 개의 스팸 메시지가 송신되고 있으며, 이외의 다양한 형태로 인터넷 이용자에게 송신되는 등, 인터넷 전반에 걸쳐 막대한 피해를 발생시키고 있다[3].

커뮤니티 스팸 (Community Spam)은 인터넷 게시판, 블로그(blog) 등과 같은 인터넷 커뮤니티에 게재되는 스팸 메시지로, 피해가 급격히 증가하고 있는 주목할 필요성이 있는 스팸 중 하나이다. 커뮤니티 스팸은 일반적인 스팸에 비

※ 이 논문은 성균관대학교의 2005학년도 성균학술연구비에 의하여 연구되었음.
† 준 회 원 : 성균관대학교 컴퓨터공학과 석사과정
†† 정 회 원 : 성균관대학교 정보통신공학부 조교수
논문접수 : 2006년 5월 17일, 심사완료 : 2006년 8월 22일

해 적은 비용으로도 불특정 다수의 인터넷 이용자에게 의사소통의 저해, 인터넷 커뮤니티의 신뢰도 저하 등의 문제를 야기하는 경향이 있다. 일반적으로 스팸은, 이메일과 인스턴스 메시지 등의 형태로, 단일 스팸 메시지가 단일 인터넷 이용자에게만 피해를 가하는 형태이다. 이에 반해, 커뮤니티 스팸은 동일한 비용의 단일 스팸 메시지가 해당 인터넷 커뮤니티에 접근하는 다수의 인터넷 이용자와 인터넷 커뮤니티의 관리자에게 피해를 가하고 있다. 이와 함께, 커뮤니티 스팸은 인터넷 커뮤니티의 접근이 용이해지고, 인터넷 이용자의 사회참여와 의견게재가 활발해지는 현 인터넷 세태와 맞물려, 그 양과 피해가 급격히 증가하고 있는 상황이다.

이런 인터넷 전반에 걸친 스팸 문제의 급증에 비해, 스팸 차단을 위한 연구는 아직까지도 이메일 서비스에 편중되어 있다. 이메일 서비스 상의 스팸 메일 차단 기법들은 높은 정확도와 낮은 오탐지율을 위해 활발한 연구들이 진행되고 있다. 이밖에 인스턴스 메시지 서비스와 같이 최근에 널리 보급된 서비스들에 대한 스팸 차단에 관한 연구는 현재 지속적으로 증가하고는 있으나, 아직까지도 미비한 수준이며, 커뮤니티 스팸의 차단에 관한 연구는 스팸 메일 서비스나 인스턴스 메시지 서비스의 스팸 연구에 비해 전무한 상황이다.

본고에서는 커뮤니티 스팸을 효과적으로 차단하는 베이지안 정리 기반의 필터링 기법을 제안하고자 한다. 베이지안 필터링 기법은 기존의 수집된 스팸 메시지와 정상 메시지들을 학습하고, 학습을 바탕으로 새로 수신될 메시지의 스팸 여부를 판단하는 기법이다. 베이지안 필터링 기법은 이미 스팸 메일의 차단 분야에서 제안되어 이미 널리 이용되고 있는 기법이다[4, 5]. 이는 높은 정확도와 낮은 오류률 그리고 짧은 학습시간 등의 장점을 지닌 기법이다. 이런 베이지안 필터링 기법을 커뮤니티 스팸 차단에 도입함으로써, 급격히 증가하는 커뮤니티 스팸 문제의 해결 가능성을 제시하고자 한다.

또한, 본고에서는 기존의 스팸 메시지와 정상 메시지로의 이중 분류뿐만 아니라, 스팸 메시지를 여러 분야로 분류할 수 있는 다중 분류 기능을 추가로 제공하도록 구성한다. 현재 인터넷 커뮤니티의 관리자들은 등록되는 메시지를 단순히 정상 메시지와 스팸 메시지로 단순 이중 분류하고 있다. 이러한 분류는 다수의 이용자들에게 효과적인 스팸 메시지의 차단을 가능케 하나, 다양한 이용자의 요구를 충족시키지 못하는 문제를 지니고 있다. 예를 들어, 상품 또는 이벤트의 구입을 목적으로 인터넷 커뮤니티에 접근하는 이용자의 경우, 그리고 중고 장터와 같이 상품과 이벤트의 거래를 목적으로 하는 인터넷 커뮤니티의 경우, 상업적 광고를 단순히 스팸 메시지로 분류하는 것은 해당 인터넷 커뮤니티를 접근하는 이용자의 원활한 의사소통을 오히려 방해하는 요소가 되고 있다. 이런 문제를 해결하기 위해, 본고에서는 스팸 메시지의 다중 분류를 통해 인터넷 이용자에게 메시지의 수용 여부를 결정할 수 있도록 구성하고 있다.

일반적으로, 스팸 메일과 인스턴스 메시지 스팸 등은 메시지의 판단 주체와 수용 주체가 동일하다. 판단 주체는 메시지의 스팸 여부를 결정짓는 주체를 말하며, 수용 주체는 판단된 메시지를 실제로 이용하는 주체를 뜻한다. 이런 서비스들에서 수신된 이메일이나 인스턴스 메시지에 대한 스팸 여부는 수신자에 의해 판단되고, 판단된 메시지도 수신자에 의해 이용된다. 이에 반해, 인터넷 커뮤니티는 스팸 메시지의 판단 주체와 수용 주체가 다르다. 인터넷 커뮤니티의 스팸 메시지 판단 주체는 인터넷 커뮤니티의 운영자이나, 메시지의 수용 주체는 인터넷 커뮤니티를 이용하는 다양한 인터넷 이용자들이 된다. 이런 상황 하에서, 운영자에 의한 메시지의 이중 분류는 메시지 수용에 관한 인터넷 이용자들의 다양한 요구를 충족하기 어렵다. 따라서 인터넷 커뮤니티의 관리자는 다중 분류를 통해서 메시지의 판단 정보만을 수용 주체에게 제공하고, 수용 주체인 게시판 이용자들이 스팸 메시지의 판단을 내리도록 구성하는 것이 효과적인 스팸 메시지의 차단 기법이 될 수 있다.

이밖에도 스팸 메시지를 소분류로 세분화하는 것은 인터넷 커뮤니티의 관리자의 추가적인 메시지 분류 과정을 덜어주고, 인터넷 커뮤니티로의 주제나 특성에 구애받지 않고 도입이 가능하다는 장점이 있다. 성인 컨텐츠를 주 대상으로 하는 인터넷 커뮤니티의 경우, 일반적인 스팸 필터의 도입은 해당 인터넷 커뮤니티의 원활한 의사소통을 저해하는 요소가 될 수 있으나 세분화가 가능한 필터는 성인 컨텐츠 이외의 스팸 메시지만을 소분류를 통해 필터링 할 수 있다.

본고에서는 2장에서 기존의 스팸 차단 기법에 관한 연구를 정리하고, 3장에서 베이지안 필터링 기법에 대해서 정리한다. 4장에서는 수집된 스팸 메시지를 그 종류에 따라 소분류로 세분화하고 분석하며, 5장에서는 구현된 필터링 기법의 정확도를 측정하고 비교, 분석한다. 6장에서는 결론 및 향후과제에 대해 언급한다.

2. 관련 연구

본 장에서는 스팸 차단 기법에 대해 정리한다. 스팸 차단 기법에 관한 연구는 Listing, 송신자 인증, 워드필터링 기법으로 대표된다.

Listing에 의한 스팸 차단 기법은 스패머의 IP 주소, ID, 이메일 주소 등의 정보를 목록화하고 이에 부합되는 정보를 포함한 메시지를 스팸 메시지로 간주하는 기법이다. Listing에 의한 스팸 차단 기법은 높은 차단 효과를 위해 이미 알려진 스패머의 방대한 정보가 필수적이다. 이를 위해 스팸 메일 차단 분야에서는[6-8] 등과 같이 전 세계 스패머의 정보를 공공 또는 사설 기관에서 수집해, 이를 통해 스팸 메일의 차단을 시도하는 Real-time Black list(RBL)이 널리 이용되고 있다. 그러나 이런 listing에 의한 스팸 메시지 차단은 높은 차단 효과를 위해서 지속적으로 정보의 갱신과 관리를 필요로 하고, 메시지의 정보를 은닉하는 스패머의 간단한 트릭에도 유연히 대처하기 어렵다는 단점을 포함하고 있다.

송신자 인증 기법은 스패머들이 메시지내의 정보를 은닉한다는 사실에 기반을 둔 것으로, 메시지니 내의 송신자 정보가 실제 메시지의 송신자와 부합하는 가를 인증하는 기법이다. 이를 위해 스팸 메일 차단 분야에서는 이메일에 포함된 이메일 주소, 메시지를 송신한 IP주소 등을 DNS (Domain Name Service)를 통해 인증하거나, 이메일의 전자서명을 첨부하여 파악하는 등의 기법들이 활용되고 있다. 이런 기법들은 최근 대형 IT업체와 ISP(Internet Service Provider)에 의해 발전되고 있으며, Pobox의 SPF[9], Microsoft의 SenderID [10], Yahoo의 DomainKeys[11], Cisco의 IIM[12]등의 기법들이 제시되고 있는 상황이다. 그러나 커뮤니티 스팸 차단 분야에서는 이메일 서비스와 달리 domain 주소와 같은 공통된 기준이 부재하기 때문에 적용이 어렵다는 단점이 있다.

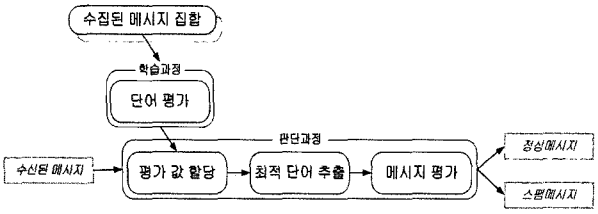
워드 필터링 기법은 listing 기법과 같이 널리 이용되는 스팸 차단 기법으로, 메시지에 포함된 단어 또는 문장을 바탕으로 스팸 메시지를 판단하는 기법이다. 워드 필터링 기법은 규칙 기반 필터링 기법과 학습 기반 필터링 기법으로 구분된다.

SpamAssassin[13] 으로 널리 알려진 규칙 기반 필터링 기법은 사전에 규칙을 설정해두고 이에 부합하는 단어나 문장을 포함한 메시지를 스팸 메시지로 간주하는 기법이다. 이때 규칙이란 스팸 메시지에 자주 등장하는 단어 또는 문장을 뜻한다. 그러나 이 기법은 단어의 조합이나 표현을 변경하는 스패머의 간단한 트릭에도 유연히 대처하기 어렵고, 높은 오탐지율을 지닌다는 단점이 있다. 커뮤니티 스팸 분야에서도 "금지어"라 불리는 규칙을 설정함으로써, 스팸 메시지의 게재를 방지하고 있지만 큰 효과가 없는 상황이다.

학습 기반 필터링 기법은 베이지안 필터링으로 널리 알려진 기법으로, 과거의 메시지를 학습함으로써, 새로 수신되거나 게재되는 메시지의 스팸 메시지 여부를 판단하는 기법이다[4, 5]. 학습 기반 필터링 기법은 학습의 효율에 따라 그 정확도가 좌우된다. 이 때문에 베이지안 필터링 이외에도 보다 높은 정확도와 효율을 지니는 다양한 학습 알고리즘 1) Support Vector Machines (SVMs)[14], 2) Hidden Markov Model (HMM)[15], 3) Neural Network[16] 등이 도입되고 있다. 이런 학습 기반 필터링 기법은 규칙 기반 필터링 기법에 비해 높은 정확도를 지닐 수 있으나 높은 정확도를 얻기 까지 일정 기간의 학습 기간이 필요하다는 단점이 있다.

3. 베이지안 필터링 기법

베이지안 필터링 기법은 학습 기반 필터링 기법의 한 종류로, 가장 널리 알려진 스팸 메일 필터링 기법 가운데 하나이다. 베이지안 필터링 기법은 (그림 1)에서 나타나듯이 일반적인 학습 기반 필터링 기법의 학습과정과 판단과정으로 구성된다.



(그림 1) 베이지안 필터링 기법의 세부 절차

3.1 학습 과정

(그림 1)에서 학습과정에서의 단어 평가는 수집된 메시지 집합(스팸 메시지와 정상 메시지)에 포함된 각각의 단어들이 스팸 메시지와 정상메시지의 특징을 얼마나 잘 반영하는가를 측정하는 과정이다. 이 과정에서 베이지안 필터링 기법은 수집된 메시지 집합의 모든 단어들을 다음과 같이 평가한다. 먼저 수집된 메시지로부터 임의의 단어를 추출한 후, 해당 단어를 포함한 스팸 메시지의 수를 측정한다. 그 후, 측정된 스팸 메시지의 수를 수집된 메시지의 전체 스팸 메시지의 비로 나타내고 이를 단어의 평가 값으로 활용한다. 예를 들어, 'viagra' 란 단어가 수집된 메시지 집합의 스팸 메시지 100개 가운데, 80개의 스팸 메시지에 나타난다고 한다면, 'viagra'란 단어는 0.8 로 평가되는 것이다. 이런 과정은 스팸 메시지에 자주 등장하는 단어를 선별하기 위한 평가 과정으로, 평가된 값은 그림 1의 판단 과정에서 수신된 메시지의 스팸 메시지 여부를 판단할 때 이용되게 된다. 단어의 평가 값이 1에 가까울수록 스팸 메시지의 특성을 잘 반영하는 단어이고, 0 에 가까울수록 스팸 메시지보다는 정상 메시지의 특성을 보다 잘 반영하는 단어이다.

본고에서의 베이지안 필터링 기법은 다중 분류 기능의 제공을 위해 다수의 평가 값을 지니고 있다. 기존의 베이지안 필터링 기법은 스팸 메시지에 대한 평가 값만을 지니지만 본고에서는 상업적 광고, 청소년 유해물 등과 같은 소분류에 대한 평가 값을 단어가 각각 지니게 된다.

3.2 판단 과정

베이지안 필터링 기법의 판단 과정은 (그림 1)에서 나타는 것과 같이 평가 값의 할당, 최적 단어의 추출, 그리고 추출 단어를 통한 메시지의 평가로 구성된다.

평가 값의 할당 과정은 학습 과정에서 평가된 각 단어의 평가 값을 수신된 메시지의 각각의 단어들에게 할당하는 과정이다. 예를 들어, 수신된 메시지에 'viagra' 란 단어가 포함되어 있다면, 이 단어의 할당 값은 학습과정에서의 평가 값 0.8을 할당하는 것이다. 만일 수신된 메시지 가운데, 학습과정에서 학습되지 않아서 할당 값이 없는 단어는 임의의 평가 값 0.4를 할당하게 된다. 0.4 는 [4]에서 반복적 실험을 통해 제안한 값으로, 0.4를 할당 하였을 때 가장 높은 정확도를 지닌 베이지안 필터링 기법이 구성된다고 알려져 있다.

최적 단어의 추출 과정은 수신된 메시지로 부터 메시지의

특성을 가장 잘 반영하는 적정수의 단어를 추출하는 과정이다. 베이지안 필터링 기법은 최적의 단어를 추출하기 위해, 수신된 메시지의 단어 가운데 0.5로부터 가장 멀리 떨어진 할당 값을 지니는 단어들을 순차적으로 추출한다. 이 같은 과정을 통해 단어 추출하는 것은 평가 값 0.5를 기준으로 각각의 단어들을 출현 빈도수가 높은 단어와 낮은 단어로 구분할 수 있기 때문이다. 할당 값은 스팸 메시지가 해당 단어를 포함하고 있을 확률을 나타내기 때문에, 추출된 단어들은 해당 메시지의 특성을 반영할 수 있다. 대체적으로 스팸 메시지는 1에 가까운 평가 값을 지닌 다수의 단어들이 추출될 것이고, 정상 메시지는 0에 가까운 평가 값의 단어들이 주로 추출 될 것이다. 최적 단어 추출 과정에서 추출하는 단어의 수는 반복적인 실험을 통해 최적의 정확도를 지니는 수로 결정하게 된다.

메시지 평가의 과정은 추출된 단어를 통해 수신된 메시지의 스팸 여부를 판단하는 과정이다. 이를 위해, 추출된 단어는 다음의 베이지안 정리 (1) 에 적용되어 수신된 메시지를 평가하게 된다.

$$P(spam|words)=\frac{P(words|spam)P(spam)}{P(words)} \quad (1)$$

식 (1)은 메시지의 스팸 확률을 추론하는 과정으로, 최적 단어 추출 과정에서 추출된 단어들이 메시지에 포함되어 있을 때 해당 메시지가 스팸 메시지 (*spam*) 일 확률 $P(spam|words)$ 을 계산한 것이다. 이때 *words*는 추출된 단어들을 뜻하고, *spam*은 스팸 메시지를 뜻한다. 식(1)의 변수들은 사전에 학습과정에서 간단히 계산될 수 있기 때문에, 이를 통해 메시지의 스팸 확률 P(spam|words)를 도출 할 수 있다.

식 (1)에서의 변수들은 다음과 같이 사전에 계산되었다. $P(words|spam)$는 추출된 단어들을 포함한 스팸 메시지가 나타날 확률을 말하는 것으로, 추출된 단어들이 지니는 평가 값 $P(word|spam)$의 곱으로 계산된다. $P(spam)$는 수집된 메시지 집합 가운데 스팸 메시지가 차지하는 비율로, 학습 집합 가운데서 스팸 메시지의 수를 계산함으로 구하게 된다. 또한, $P(words)$는 전체 메시지 가운데서 추출된 단어들을 모두 포함한 메시지가 나타날 확률로, 이는 식 (2)와 같이 계산한다.

$$P(words)=P(words|spam)\cdot P(spam)+P(words|\sim spam)P(\sim spam) \quad (2)$$

이때, ~spam은 spam을 제외한 나머지 모든 메시지로, 본고에서는 학습 집합의 스팸 메시지 이외에 모든 메시지를 뜻한다. 식 (2)의 모든 변수들은 식(1)의 변수와 같이, 학습 과정에서의 학습 집합을 통해 계산될 수 있다.

이런 식 (1)을 통해, 본고에서 수신된 메시지는 해당 메시지가 스팸 메시지일 확률을 계산 받게 된다. 본고에서는 스팸 메시지의 다중 분류를 위해 다수의 소분류에 해당하는 스팸 확률을 평가하게 된다. 예를 들어, 하나의 메시지가 상업적 광고, 청소년 유해물 등의 소분류일 확률을 모두 할당 받게 되는 것이다.

하나의 메시지에 각각 계산된 확률을 서로 비교함으로써, 해당 메시지를 소분류로 분류한다. 이때, 가장 높은 확률을 지니는 소분류로 메시지를 분류하나, 이 확률이 임의의 임계값을 넘지 못하는 경우, 해당 메시지는 정상메시지로 분류하게 된다.

4. 메시지의 수집과 세분화

본 장에서는 수집된 스팸 메시지를 소분류로 세분화한다. 인터넷 커뮤니티에서는 스팸 메시지의 정의에 관한 견해차로 인해 인터넷 관리자의 올바른 스팸 메시지 판단에도 불구하고, 이를 수용하는 인터넷 이용자는 이를 그릇된 판단으로 인식하는 문제가 빈번히 발생한다. 따라서 스팸 메시지를 소분류로 세분화함으로써, 인터넷 커뮤니티 상의 다양한 인터넷 이용자들의 요구를 충족시킬 필요가 있다.

4.1 메시지의 수집

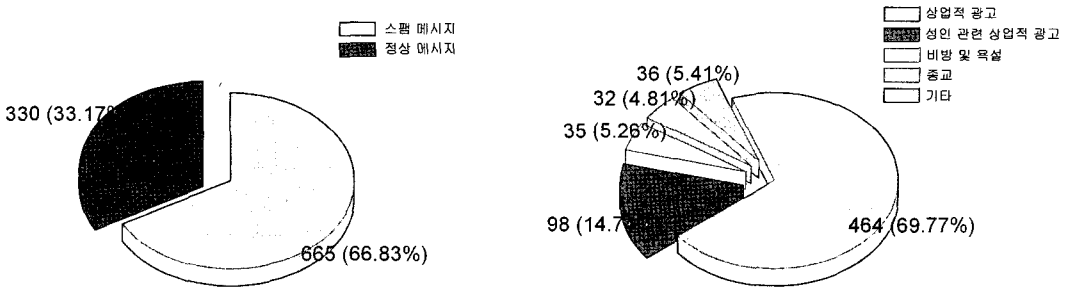
본 장에서는 스팸 메시지의 소분류를 위해 국내 유수의 사이트로부터 상당수의 메시지가 수집되었다. 수집된 메시지는 인터넷 이용자의 의견게재가 활발히 진행 중인 중앙행정기관 및 공공기관과, 국내의 유수 언론 및 미디어 기관의 인터넷 커뮤니티로부터 수집되었다. 메시지 수집에 이용된 웹 사이트의 목록은 <표 1>에 표기하였다.

해당 사이트들은 정부, 공공기관, 대형언론사들에 의해 운영되는 신뢰도가 높은 사이트들로, 인터넷 이용자의 접근이 용이하고 게시물 게재에 별도의 절차가 필요치 않아 다양한 인터넷 이용자들의 의견 게재가 활발한 사이트들이다. 그러나 해당 사이트들은 다수의 관리자들에 의해 관리되고 있기 때문에 실시간으로 게재되는 스팸 메시지에 대해서만 짧은 시간 이내의 수집이 가능하고, 웹 페이지의 다양성 때문에 오프라인 브라우저 등 자동화된 수집 프로그램의 이용이 불가능하다는 단점이 있다.

<표 1>에 나타난 수집된 모든 메시지는 995개로, 스팸 메시지 665개, 정상 메시지 330개로 구성되어 있다. 수집된 메시지는 해당 커뮤니티의 게시판에 등록된 메시지와 그에 대한 덧글로, 등록된 메시지에 대한 이용자 정보, 제목 등의

〈표 1〉 메시지 수집에 이용된 국내 웹 사이트 리스트

수집기간		2005.12.22 ~ 2005.03.06
수집 메시지		총 995 개 메시지 (스팸 메시지 665개, 정상메시지 330개)
수집 웹 사이트	정부 및 공공기관	청와대, 보건복지부, 산업자원부, 문화재청부, 특허청, 국가청렴위원회, 우체국, 서울시청, 노동조합연맹 등 11기관
	언론기관	조선일보, 동아일보, 경향신문, 오마이뉴스, YTN



(그림 2) 수집된 메시지의 구성과 스팸 메시지의 세분화

부가 사항은 모두 제외되었다. 또한, 10개 미만의 단어로 구성된 짧은 메시지도 수집 대상에서 제외되었다. 수집된 메시지는 본고에서 정한 스팸 메시지의 정의에 따라 직접 분류되었다.

4.2 스팸 메시지의 세분화

본 장에서 수집된 메시지는 (그림 2)와 같이 먼저 스팸 메시지와 정상 메시지로 구분되며, 스팸 메시지는 (그림2)의 오른쪽에 나타난 것과 같이 1) 성인 관련 상업적 광고, 2) 상업적 광고, 3) 비방 및 욕설, 4) 종교, 그리고 5) 그 외의 기타 메시지로 세분화된다.

수집된 스팸 메시지는 나타난 빈도수에 따라 상위의 소분류로 세분화 되었다. 본장에서 수집된 스팸 메시지는 상업적 광고에 69.77%, 성인 관련 상업적 광고 14.74%, 비방 및 욕설 5.26%, 종교 관련 스팸 메시지 4.81%, 그리고 어떤 소분류에도 포함되지 않는 기타 스팸 메시지 5.41%로 구성되어 있다.

제시한 스팸 메시지의 소분류는 인터넷 커뮤니티의 특성에 따라 변화 할 수 있다. 본 장에서는 자주 등장하는 주제에 따라 스팸 메시지를 소분류로 분류하였다.

5. 필터링 기법의 구현 및 정확도 측정

본 장에서는 스팸 메시지를 세분화하는 베이지안 필터링 기법을 구현하고, 구현된 베이지안 필터링 기법의 정확도를 측정하고, 다른 필터링 기법과 비교, 분석한다.

5.1 베이지안 필터링 기법의 구현

본 장의 베이지안 필터링 기법은 CPAN[18]에서 제공하는 모듈을 이용하여 구현되었다. 해당 모듈은 베이지안 필터링 기법을 위해 추상적으로 구현된 모듈로써, 학습 과정에서 계산된 각각의 단어 평가 값을 데이터베이스화 하여 보관하고, 메시지가 수신되었을 시 데이터베이스의 정보를 바탕으로 스팸 메시지 여부를 판단한다.

구현된 베이지안 필터링 기법은 3장에서 설명한 바와 같이, 다양한 소분류로 세분화하기 위해 다수의 단어 평가 값을 데이터베이스화하도록 구현하였다. 구현된 베이지안 필터링 기법은 상업적 광고, 성인 관련 상업적 광고, 비방 및 욕설, 종교, 그 외 등의 소분류 항목에 대한 단어의 평가 값을 계산하고 보관한다. 예를 들어, 'viagra' 란 단어는 상업적 광고에 대한 평가 값 0.7, 성인 관련 상업적 광고에 대한 평가 값 0.8, 비방 및 욕설에 대한 평가 값 0.3, 종교에 관한 평가 값 0.1, 그 외의 스팸 메시지에 대한 평가 값 0.1로 평가되고 데이터베이스화되어 보관되는 것이다.

판단 과정에서는 수신된 메시지를 각각의 소분류 베이지안 필터에 반복 적용함으로써, 스팸 메시지 여부를 판단하게 된다. 예를 들어, 인터넷 커뮤니티에 임의의 게시물이 게재되면, 해당 메시지는 먼저 상업적 광고에 속하는지 여부를 판단 받게 되고, 그 후 성인 관련 상업적 광고, 비방 및 욕설, 종교, 기타 스팸 메시지에 포함되는 내용인가에 대해 차례로 판단 받게 된다. 이 가운데 하나의 범주라도 해당 메시지가 포함된다면, 메시지는 스팸 메시지로 간주되고, 포함되는 소분류의 스팸 메시지로 표기되는 것이다.

5.2 전처리 과정

본 절에서는 보다 정확한 베이지안 필터링 기법의 정확도 측정을 위해 1) 메시지 분류, 2) 메시지 내의 조사 제거 등의 작업을 진행한다.

필터링 기법의 정확도 측정을 위해서는 구현된 필터를 학습시키기 위한 학습 집합과 학습의 결과와 정확도를 측정하기 위한 실험 집합이 필요하다. 학습 집합이란 3장의 가절에서 언급된 일련의 과정에 이용되는 메시지 집합이며, 실험 집합은 3장의 나절에서 언급된 일련의 과정에 이용되는 메시지 집합이다. 실험 집합은 실제 인터넷 커뮤니티에 등록되는 다양한 이용자의 메시지를 대신한다. 이를 위해 본 절에서는 4장에서 수집된 995개의 메시지 가운데 임의의 20%를 학습의 결과와 정확도를 판단하기 위한 메시지 집합으로 활용하고, 나머지 80% 의 메시지를 필터의 학습을 위한 메시지 집합으로 이용한다. <표 2>는 4장에서 수집된 메시지를 학습 집합과 실험 집합으로 구분 한 것이다.

추가로 본 절에서 이용된 메시지들은 필터링 기법의 정확도를 하락시킬 수 있는 단어를 사전에 제거하였다. 제거된 단어는 대명사, 한글자로 구성된 단어, 그리고 조사 등으로 스팸 메시지의 분류에 영향을 끼치지 않는 쓰임이 모호한 단어들이다.

〈표 2〉 정확도 측정을 위한 학습 집합과 실험 집합의 구성

	Spam Messages	Acceptable Messages
학습집합 (Training Set)	548	256
실험집합(Test Set)	117	74

5.3필터링 기법의 정확도 측정

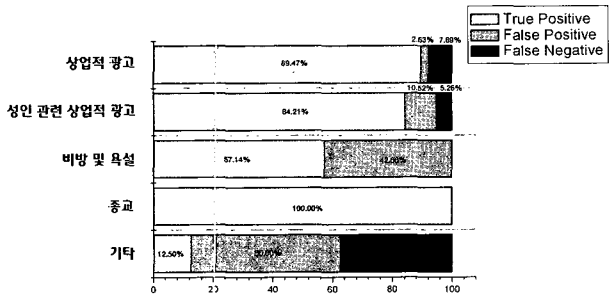
본 절에서는 베이지안 필터링 기법을 이용한 커뮤니티 스팸 메시지의 분류 정확도를 측정한다. 정확도 측정은 정상 메시지와 스팸 메시지의 단순 분류와 스팸 메시지의 세분화로 나누어 측정한다. <표 3>은 앞 절의 실험 집합을 스팸 메시지와 합법적인 메시지로 단순 분류 했을 때의 정확도를 나타내며, (그림 3)는 스팸 메시지를 소분류로 세분화 했을 때의 정확도를 나타낸다.

<표 3>에서 스팸 메시지 정확도 (Spam message Accuracy)는 실제 스팸 메시지를 베이지안 필터링 기법이 스팸 메시지로 판단할 확률을 나타내며, 정상 메시지 정확도 (Acceptable message Accuracy)도 실제 정상 메시지를 정상 메시지로 판단할 확률을 나타낸다.

<표 3>에서, 베이지안 필터링 기법은 <표 2>의 실험 집합에 대해 스팸 메시지 정확도가 91.45%, 정상 메시지 정확도가 86.48%로 나타난다. 이는 이전의 실험[17] 에 비해서도 적은 양의 학습 집합으로도 높은 정확도를 지니는 것이다. 이전의 실험에서는 약 600여개의 스팸 메시지가 학습에 이용됐을 때, 베이지안 필터가 높은 정확도를 지닌다고 알려져 있다. 이는 수집된 메시지가 유사한 성향의 언론기관, 공공기관 등 인터넷 커뮤니티에서 수집되었기에 메시지들의 주제 범주, 성향, 단어 등의 차이가 크지 않기 때문이다. 그러나 1장에서 언급한 것과 같이, 이런 단순 분류는 다양한 이용자의 요구 사항을 충족하기 어렵기 때문에 실제 인터넷 커뮤니티에 활용하기 어려운 단점이 있다. 인터넷 커뮤니티

〈표 3〉 Bayesian Filtering 기법의 정확도

Spam Message Accuracy	Acceptable Message Accuracy
91.45%	86.48%



(그림 3) 스팸 메시지의 소분류 정확도

의 관리자 관점에서는 높은 정확도로 메시지를 분류하나 다양한 이용자의 관점에서는 스팸 메시지 분류에 큰 효과가 나타나지 않는다.

(그림 3)은 스팸 메시지를 세분화하는 각각의 베이지안 필터의 정확도를 측정한 것이다. (그림 3)의 True Positive는 스팸 메시지가 스팸 메시지로 판단되었고, 세분화된 소분류로도 정확히 판단되었을 경우를 나타낸 것이며, False Positive는 스팸 메시지로 판단은 되었으나 잘못된 소분류로 판단된 경우를 나타낸다. False Negative는 스팸 메시지임에도 불구하고, 어느 소분류에도 포함되지 못해 정상 메시지로 오판단 된 메시지를 나타낸다.

(그림 3)에서, 구현된 베이지안 필터링은 스팸 메시지를 추가로 소분류로 세분화하였다. 소분류의 세분화는 <표 3>과 달리 다양한 이용자의 요구사항에 부응할 수 있는 여지를 제공한다. 각각의 소분류는 (그림 3)의 true positive 만을 고려했을 때, 상업적 광고 메시지가 89.47%, 성인 관련 상업적 광고가 84.21% 등으로 높은 정확도로 분류가 가능하다. 그러나 본 필터링 기법의 소분류는 비방 및 욕설, 기타 등이 각각 57.14%, 12.50%로 낮은 정확도를 지닌다.

본 장에서 (그림 3)의 정확도가 소분류에 따라 큰 차이가 나타나는 것은 학습 집합의 1) 메시지의 양, 2) 주제 범주의 차이 등을 원인으로 분석된다.

구현된 베이지안 필터링 기법은 스팸 메시지와 합법적인 메시지의 단순 분류를 위해 546개의 스팸 메시지와 256개의 합법적인 메시지 등 충분한 양의 메시지에 의해 학습되었다. 이에 반해 (그림 3)의 각각의 소분류는 이에 미치지 못하는 양으로만 학습되었다. 4장에서 구분한 소분류 가운데 상업적 광고가 388개의 스팸 메시지, 256개의 합법적인 메시지로 가장 많은 메시지를 학습에 이용되었고, 기타를 위해서는 오직 36개의 스팸 메시지와 256 개의 합법적인 메시지가 이용되었다. 이로 인해 구현된 각각의 베이지안 필터들은 충분히 학습되지 못했고, <표 3>에 비해 낮은 결과를 나타낸다.

그러나 본 실험치와 달리 베이지안 필터링 기법을 도입하는 인터넷 커뮤니티에서는 해당 문제가 큰 쟁점이 되지 않는다. 본 실험에서는 메시지 수집에 어려움이 있었으나 실시간으로 인터넷 커뮤니티를 관리하는 관리자의 입장에서는 다양한 종류의 스팸 메시지를 짧은 시간동안 대량으로 수집 가능하기 때문에, 본 실험치와 같이 적은 양의 스팸 메시지에 의한 낮은 정확도는 문제가 되지 않을 것으로 추정된다.

(그림 3)에서 종교 관련 스팸 메시지에 대한 베이지안 필터링 기법의 정확도는 여타의 결과와 다른 결과를 나타낸다. 종교 관련 스팸 메시지의 필터링 정확도는 다른 스팸 메시지들에 비해 적은 양의 스팸 메시지로 학습 되었음에도 불구하고, 100%의 높은 정확도를 나타낸다. 해당 실험치는 유사한 양의 스팸 메시지로 학습된 기타와 차이가 크다. 이는 각각의 소분류가 다루는 메시지의 주제가 상이하기 때문이다. 일반적으로 종교와 관련된 주제는, 기독교, 불교 등과 같이 매우 제한적이다. 이 때문에 메시지에 종종 등장하는

단어 역시도 다른 소분류에 비해 극히 제한적이며, 소수의 단어가 종교 관련 스팸 메시지의 특성을 잘 반영하고 있다. 따라서 구현된 베이지안 필터링 기법도 비록 적은 양으로만 학습 되었지만, 해당 소분류에 대해 높은 정확도를 지니는 것이다. 이에 반해, 비방 및 욕설 그리고 기타의 스팸 메시지들은 이벤트, 재정, 서비스 등은 물론 정치, 경제 등 사회 전반에 걸친 다양한 주제들을 다루고 있기 때문에 적은 양의 메시지로는 해당 메시지의 특성을 충분히 파악하기 어렵다. 본 실험에서도 기타의 스팸 메시지들은 각 소분류에 포함되지 않지만 스팸 메시지로 분류된 다양한 형태의 스팸 메시지가 분류되어 있기 때문에, 이들의 특성을 충분히 학습하지 못해 낮은 정확도가 나타나는 것이다.

이런 점으로 미루어 볼 때, 본 실험에서 구현된 베이지안 필터링 기법은 비록 모든 소분류에 대해 매우 높은 정확도를 지니지는 못했지만, 기존의 단순 분류의 문제를 해결하고 이용자의 다양한 요구를 반영하는 스팸 메시지 필터링을 가능케 한다. 또한, (그림 3)의 종교 관련 스팸 메시지의 소분류 정확도와 같이, 특징이 분명한 소분류로 스팸 메시지를 세분화함으로써 적은 양의 스팸 메시지로도 높은 정확도를 지닐 수 있음을 추정 할 수 있다.

6. 결론 및 향후 과제

스팸으로 인한 인터넷 피해가 날로 증가하고 있는 현재, 커뮤니티 스팸 에 대한 차단에 관한 연구는 인터넷의 필수 연구 분야로 자리 잡고 있다. 커뮤니티 스팸 차단에 최적화된 차단 기법이 없는 현 상황에서, 본고에서는 스팸 이메일의 차단에서 널리 적용되는 기법을 활용한 효과적인 커뮤니티 스팸 필터링 기법을 제시하였다. 이를 위해 본고에서는 베이지안 필터링 기법을 도입하였으며, 커뮤니티 스팸 차단에서 나타나는 이용자들의 다양한 요구 사항을 만족시키기 위해, 커뮤니티 스팸의 소분류 기능을 추가로 구현하였다.

본고에서 구현한 베이지안 필터링 기법은 스팸 메시지의 단순 분류에 있어 적은 메시지들의 학습만으로도 91.45%란 높은 정확도를 지닌다. 더구나 스팸 메시지의 소분류에 있어서도 상업적 광고에 대해 89% 그리고 성인 관련 상업적 광고에 대해 84% 등으로 높은 정확도를 지닌다. 그러나 구현된 베이지안 필터링 기법은 다양한 주제를 다루는, 비방 및 욕설, 기타 등, 소분류에 대해서는 57%, 12%로 낮은 정확도를 지닌다. 이는 본 실험과 달리 대량의 스팸 메시지를 학습 집합으로 이용하고 보다 세분화된 소분류로 스팸 메시지를 분류하는 실제 인터넷 커뮤니티에서는 해결 가능하다.

향후 연구계획으로는 대량의 메시지를 수집함으로써, 베이지안 필터링 기법이 지닐 수 있는 최대 정확도에 대해 정확도의 공정성을 지닌 추가 연구를 수행할 예정이며, 실제 인터넷 커뮤니티에 이용 가능한 자동화된 필터링 어플리케이션의 구현을 목표로 하고 있다. 또한 보다 적은 양의 학습 집합으로도 높은 정확도를 지니는 대체 알고리즘의 연구도 동시에 수행 할 예정이다.

참 고 자 료

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

[2] Paulson, L.D, "Spam hits instant messaging," IEEE Computer, IEEE Computer Society, Volume 37, Issue 4, Aprl 2004 pp. 18

[3] The Radicati Group Inc., "Email Sent and Received Growth Statistic, 2003-2005", Jul. 2003.

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

[5] Graham Paul, "Better Bayesian Filtering," available at http://paulgraham.com/better.html, Jan. 2003

[6] Trend Micro Inc., "Nominations", available at http://www.mail-abuse.com/nominats.html

[7] SpamCop, "SpamCop Blocking List," available at http://www.spamcop.net/bl.shtml

[8] Spamhaus, "The Spamhaus Block List," available at http://www.spamhaus.org/sbl/index.lasso

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

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

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

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

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

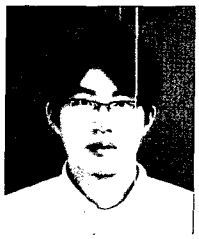
[14] Thornsten Joachims, "Text categorization with support vector machines: learning with many relevant features," Proc. European Conference on Machine Learning, Springer-Verlag, pp.137-142, 1998.

[15] Hongrak Lee and Andrew Y. Ng, "Spam Deobfuscation using a Hidden Markov Model," Second Conference on Email and Anti-Spam (CEAS2005), 2005, available at http://ww w.ceas.cc/papers-2005/166.pdf.

[16] Ian Stuart, Sung-Hyuk Cha, Charles C. Tappert, "A Neural Network Classifier for Junk E-Mail," Proc. Document Analysis System VI, 6th International Workshop, Springer-Verlag, pp.442-450, 2004.

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

[18] Roger Burton, "Mail::SpamTest::Bayesian," available at http://search.cpan.org/~firedrake/Mail-SpamTest-Bayesian-0.02/Bayesian.pm



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