Journal of Korea Multimedia Society Vol. 9, No. 2, February 2006(pp. 255-263)

메일 주소 유효성과 제목-내용 가중치 기법에 의한 스팸 메일 필터링

강 승 식†

요 약

스팸 메일의 특성을 분석해 보면 스팸 메일 발송 프로그램이 메일 헤더에 기록된 주소와 송신자 및 수신자 메일 주소가 일치하지 않는 경우가 빈번하게 발견된다. 또한, 스팸 메일과 정상적인 메일을 비교-분석해 보면 제목만 살펴봐도 스팸 메일인지 여부를 쉽게 판별할 수가 있다. 본 논문에서는 이와 같은 스팸 메일의 특성을 이용하여 스팸 메일 필터링 시스템의 성능을 향상시키는 방안으로 메일 주소 유효성 검사 및 제목과 내용을 구분하여 각각 스팸 확률을 계산하는 기법을 제안하였다. 제안한 방법의 효용성을 검증하기 위하여 단순 베이스 기법에 대해 주소 유효성 검사 및 제목과 내용 등 각 요인의 중요도에 따른 스팸 메일 필터링의 성능 향상 정도를 측정하였다. 그 결과로, 제안한 방법을 적용했을 때 재현율이 11.6%, 정확률은 2.1%의 성능 향상 효과가 있음을 확인하였으며, 스팸 메일 필터링 시스템의 성능 향상에 많은 기여를 하는 것을 알 수 있었다.

Junk-Mail Filtering by Mail Address Validation and Title-Content Weighting

Seung-Shik Kang†

ABSTRACT

It is common that a junk mail has an inconsistency of mail addresses between those of the mail headers and the mail recipients. In addition, users easily know that an email is a junk or legitimate mail only by looking for the title of the email. In this paper, we tried to apply the filtering classifiers of mail address validation check and the combination method of title-content weighting to improve the performance of junk mail filtering system. In order to verify the effectiveness of the proposed method, we performed an experiment by applying them to Naive Bayesian classifier. The experiment includes the unit testing and the combination of the filtering techniques. As a result, we found that our method improved 11.6% of recall and 2.1% of precision that it contributed the enhancement of the junk mail filtering system.

Key words: Junk Mail Filtering(스팸 메일 필터링), Naive Bayes Classifier(단순 베이스 분류자), Mail Address Validation(메일 주소 유효성), Title-Content Weighting(제목-내용 가중치)

1. 서 론

전자 우편은 최소의 비용으로 실시간으로 배달된다는 편의성과 더불어 다수의 수신자에게 동시에 전송하는 특성으로 인해 오프라인 우편을 대체하는 수단으로 사용되고 있으며, 데이터 전송이나 데이터 백업의 목적으로 활용되기도 한다. 그런데 저렴한 비용으로 불특정 다수에게 광고물을 전송하는 마케팅 수

※ 교신저자(Corresponding Author) : 강승식, 주소 : 서울시 성북구 정릉동 861-1(136-702), 전화 : 02)910-4800, FAX : 02)910-4868, E-mail : sskang@kookmin.ac.kr
접수일 : 2005년 12월 7일, 완료일 : 2006년 2월 2일

† 정회원, 국민대학교 컴퓨터공학부 부교수
※ 본 연구는 첨단정보기술연구센터를 통하여 과학기술부/한국과학재단의 지원을 받았음.

단으로 사용되면서 상업적 광고성 메일이 무분별하게 발송되고 있다. 이에 따라 메일 서비스 업체부터 개인 사용자에 이르기까지 많은 피해자가 속출하게 되었고, 그 피해 규모 또한 적지 않다.

스팸 메일이 심각한 사회 문제로 부각이 됨에 따라 정보통신부는 '정보통신망 이용 촉진 및 정보 보호 등에 관한 법률 시행령 및 시행규칙 개정안'을 마련하였다. 따라서 모든 광고 메일들은 제목 처음에 '광고'라는 문구를 넣고 끝에는 '@'를 첨가하도록 하였다. 이에 따라 광고 메일을 차단하는 방안으로 스팸에 주로 출현하는 단어들에 대하여 메시지 규칙 기반의 필터링을 수행하는 방법이 사용된다. 그러나 이 법률을 지키지 않는 광고 메일이 범람하고 있어서 내용 기반 필터링 방법에 의한 스팸 메일 차단 기법이 연구되어 왔다[1,2].

Cohen(1995)은 내용 기반 스팸 메일 필터링 기법 중에서 단순하고 효율적인 기법으로 예제 집합으로부터 if-then 형태의 분류 규칙을 추출하여 규칙베이스로 구축하는 Ripper 방식을 제안하였으며, Shapire(1998)은 학습 문서 벡터를 정규화된 기준 벡터(prototype vector)로 표현하는 Rocchio 방법을 문서 필터링에 적용하였다[3,4]. 그러나 통상적으로 스팸 메일 차단 기법으로 가장 널리 사용되는 방법론은 문서 분류 등에 사용되는 통계적 확률 이론과 최대 엔트로피 모델(Maximum Entropy Model), 지지 벡터 기계(Support Vector Machine) 등 기계학습 이론을 기반으로 하는 기법이다[5-7].

기계학습 기법 중에서 지지 벡터 기계가 일반적으로 다른 기법들에 비해 더 나은 성능을 보인다고 알려져 있으나, 그렇지 않은 경우도 있다. 그 예로, Drucker(1999)는 에러율면에서 boosting 알고리즘이 더 나은 성능을 보이고 있으며, Brutlag(2000)는 실험 데이터의 특성에 따라 TF-IDF 방법이 더 좋은 성능을 보인다는 결과를 제시하고 있다[8,9]. 문서 분류 방법에서도 지지 벡터 기계의 성능이 우수한 경우가 많으나 데이터 특성에 따라 실험 결과가 달라지기도 한다. 통상적으로 특정 이론에 따라 구현된 시스템들이 서로 다른 결과를 보이는 경우가 많다는 점과 지지 벡터 기계는 정교하게 잘 구현된 엔진만을 사용한다는 특성을 고려할 때 다른 기법에 비해 지지 벡터 기계가 절대적으로 우월한 알고리즘이라고 단정하기는 어렵다.

통계적 확률 기법으로 가장 많이 사용되는 방법론은 단순 베이스 분류자(Naive Bayes Classifier)이다. 단순 베이스 분류자는 이론적 간결성과 연구자들이 쉽게 구현할 수 있다는 장점 때문에 연구자들이 다양한 형태로 변형하여 실험을 하는데 적합한 방법론이다. 따라서 최근린법(k-Nearest Neighbor)과 더불어 가장 문서 분류뿐만 아니라 스팸 메일 필터링 연구에 가장 많이 사용되고 있다[10-13]. 단순 베이스 분류자는 문서 내의 주요어들을 이용하여 분류를 수행한다. 메일의 경우, 메일 헤더에서 얻을 수 있는 정보(제목, 보낸 사람의 이메일 주소, 받는 사람의 이메일 주소 등)와 메일 본문에서 얻을 수 있는 정보를 최대한 활용하여 스팸 메일 분류에 특화된 분류자를 만들기가 용이하다.

본 논문에서는 수신자의 메일 주소 유효성 검사 및 제목과 본문에 대해 각각 스팸 확률을 계산하여 스팸 여부를 판단하는 기법을 제안하였으며, 제안한 방법을 단순 베이스 분류자에 적용하여 각각의 스팸 확률 계산 결과에 대해 가중치를 부여하여 최종적으로 스팸 메일 여부를 결정하는 스팸 메일 필터링 시스템을 구현하여 그 효용성을 평가하였다. 본 논문의 구성은 다음과 같다. 2장에서는 속성 선별 방법과 스팸 메일 필터링 기법들을 소개하고, 3장에서는 주소 유효성 검사 기법과 제목-내용에 의한 스팸 확률 계산 기법을 설명하였다. 4장은 제안한 기법들이 스팸 메일 필터링 성능에 미치는 성능 개선 효과를 실험을 통해 확인하고 그 결과를 기술하였다.

2. 관련 연구

전자 메일 소프트웨어는 특정 용어가 포함된 메일을 미리 정의된 폴더로 이동하는 것과 같은 단순한 메시지 규칙에 기반을 둔 필터링 방법을 지원해 왔다. 이 방법은 스팸 메일을 대표하는 단어들을 수집하여 이를 필터링에 이용하는 방법으로써 단순히 해당 단어가 존재하는지의 여부만을 조사하여 판단하기 때문에 속도가 빠르고 재현율과 정확률에 있어서도 비교적 좋은 성능을 얻을 수 있다. 그러나 스팸 메일의 내용이 빠르게 변화하고 있어서 지속적인 메시지 규칙 갱신이 필요할 뿐만 아니라 특정 단어의 존재 여부만으로 스팸 여부를 판단하기 때문에 정상 메일(legitimate email)이 스팸 메일로 분류되는 경우가 빈번하게 발생한다.

이러한 문제점을 해결하기 위하여 메일의 제목과

내용 등을 분석하여 스팸 메일일 확률을 계산하고 스팸 메일로 판단되는 경우에 특정 폴더로 이동하거나 스팸 메일이라는 표시를 해주는 스팸 메일 필터링 시스템이 도입되고 있다. 스팸 메일 필터링 기법으로는 일반적으로 문서 분류 방법으로 사용되고 있는 통계적 확률 모델을 기반으로 하는 기계학습 기법이 가장 널리 사용되고 있다. 기계학습 기법 중에서 대표적인 방법으로는 단순 베이스 분류자와 지지 벡터 기계, 그리고 최근린법 등이다. 이 기법들을 적용할 때는 분류 기준으로 사용되는 속성 선택(feature selection) 과정이 선행되어야 한다.

2.1 속성 선별 기법

속성 선택 방식은 정보 검색, 문서 분류 분야에서 많은 연구가 이루어져 왔으며 가장 대표적으로 널리 사용되어 온 기법들을 소개한다. 속성을 선별하기 위한 방법으로는 TF-IDF 값에 따라 통계적 빈도에 의한 선택 방법과 상호 정보량(mutual information), 카이제곱 통계량, 정보 이득률(information gain) 등 각 속성들의 상대적인 정보량의 차이에 의한 방법이 있다. 속성 선택과 관련된 시스템의 성능 차이를 비교-분석한 연구 결과들을 살펴보면 실험 집합과 실험 대상이 되는 문제 유형에 따라 결과가 다르게 나타나는데 그 차이가 크지 않아서 속성들을 선별하는 기법이 유효한지 그렇지 않은지를 단언하기가 어렵다. Joachims(1998) 등에 따르면 단순 베이스, 최근린법, Rocchio, C4.5, 지지 벡터 기계 등 다섯 가지 학습 알고리즘을 비교했을 때 방법론에 따라 모든 속성을 사용하면 성능이 더 나은 경우도 있고 그렇지 않은 경우도 있다[14]. 즉, 속성 선별 기법이 시스템의 성능에 미치는 영향은 크지 않으며, 실험 데이터의 특성과 방법론에 따라 다르다는 것을 알 수 있다.

2.2 단순 베이스 분류자

단순 베이스 분류자는 조건부 확률 이론에 따라 입력 문서가 각 범주를 생성할 확률을 계산하여 그 확률이 가장 높은 범주로 입력 문서를 분류하는 방법이다. 예를 들어, m개의 범주로 구성된 범주 집합 C = { c_1, c_2, ..., c_m }에 대해 문서 D가 범주 c_i를 생성할 확률 $P(c_i \mid D)$은 단순 베이스 이론에 따라 식(1)과 같이 계산된다.

$$P(c_i|D) = \frac{P(D|c_i)P(c_i)}{P(D)} \tag{1}$$

이 식에서 $P(D|c_i)$은 범주 c_i가 문서 D를 생성할 확률이다. 그런데 현실적으로 학습 문서 집합으로부터 각 범주들이 문서 D를 생성할 확률값을 직접 구하기는 어렵다. 따라서 문서 D를 해당 문서에 출현한 단어들로 구성된 용어들의 집합 D = { t_1, t_2, ..., t_n }로 표현하고 문서 D의 각 용어들에 대해 단어 독립성 가정(word independence assumption)을 적용하여 $P(D|c_i)$ 를 식(2)와 같이 계산한다.

$$P(D|c_i) = \prod_{k=1}^{n} P(t_k|c_i) \tag{2}$$

단순 베이스 분류자는 식(2)에 의해 각 범주들이 문서 D를 생성할 확률을 계산하여 문서 D를 생성할 확률이 가장 높은 범주 c_{NB}를 구하는 것으로 정의된다. 즉, 이 기법에서는 식(1)과 식(2)를 적용하여 각 범주에 대한 문서 D의 발생 확률을 식(3)과 같이 계산한다.

$$\begin{aligned} c_{NB} &= \arg\max_{c_i \in C} P(c_i \mid D) \\ &= \arg\max_{c_i \in C} \frac{P(D \mid c_i)P(c_i)}{P(D)} \\ &= \arg\max_{c_i \in C} P(c_i)\prod_{k=1}^{n} P(t_k|c_i) \end{aligned} \tag{3}$$

단순 베이스 분류자를 스팸 메일 필터링에 적용할 때 모든 용어에 대해 동일한 가중치를 적용할 경우 불용어를 일반 용어와 동일하게 취급함으로 인해 성능이 저하될 수 있다. 따라서 각 용어마다 그 용어의 중요성을 구분하여 불용어, 범용어, 범주 대표어 등에 따라 차별화된 가중치를 부여하면 단순 베이스 분류자보다 성능이 향상된 결과를 얻을 수 있다. 즉, 분류에 큰 영향을 미치는 용어의 가중치를 다른 용어들의 가중치보다 높게 부여함으로써 단순 베이스 기법의 필터링 성능을 향상시킬 수 있다.

2.3 지지 벡터 기계

지지 벡터 기계는 두 가지 부류의 학습 범주에 대해 각 범주에 속하는 개체들을 문서 벡터로 표현했을 때 두 범주의 구성원들을 가장 잘 구별할 수 있는 범주간

의 경계면을 학습한다. 두 개의 범주 y_i의 값을 각각 -1, +1로 표현할 때 식(4)를 만족하는 벡터 평면 w와 상수값 b를 학습한다. 벡터 평면 w는 두 개 범주의 벡터들을 구별하기 위한 기준이 되며, 이는 각 범주의 벡터들로부터 가장 거리가 먼 경계면을 의미한다.

$$w \cdot x_i - b \geq +1, \text{ if } y_i = +1$$
$$w \cdot x_i - b \leq -1, \text{ if } y_i = -1 \quad (4)$$

입력 문서에 대한 문서 벡터 x를 위 식 $w \cdot x - b$에 의해 계산했을 때 양수이면 +1로 표현된 범주로 분류가 되고, 음수이면 -1로 표현된 범주로 분류한다. 이 방법은 각 범주들의 학습 문서 개수가 다를 경우에 학습 문서의 개수가 많은 쪽으로 편중되지 않는다는 장점이 있다.

3. 주소 유효성 검사와 제목-내용 가중치 기법

스팸 메일을 차단하는데 활용되는 메일의 속성에는 메일 제목과 메일의 내용 등에 나타나는 스팸 메일 관련 특징과 더불어 메일의 형식 등 구조적인 특징이 있다. 예를 들어, 스팸 메일의 제목에는 "$$$ BIG MONEY $$$", " → 드릴세트, 가정집 이거 하나면 끝!!!..."등의 예와 같이 "$$$", "→", "!!!..." 등 강조를 위한 과장된 기호들이 많이 사용된다. 즉, 특히 메일 제목에 텍스트 문자가 아닌 문장 기호들이 나타난다. 스팸 메일의 특징은 스팸 메일 필터링 시스템의 개발과 더불어 변화되기도 한다. 그 예로서 "[貸出] 최 저 금 리 의 은 행 권 << 貸出 >> 이 있습 니 다!!! dkcwnj sk"는 스팸 메일 필터로 차단되는 것을 방지하기 위해 각 문자 사이에 공백을 삽입하였고, 다수에게 다른 제목의 메일을 발성한 것처럼 위장하기 위해 무작위로 생성된 문자열을 제목의 끝부분에 추가한 것이다.

또한, 스팸 메일은 회사나 교육 기관 등 공신력이 있는 기관보다는 무료로 전자우편 서비스를 제공하는 정보 서비스 업체에 등록된 사용자 아이디로부터 발송되는 경우가 많다. 또 다른 스팸 메일의 특징으로 광고 메일이라는 속성 때문에 메일 내용을 이미지로 제작한 경우와 이미지를 포함하여 표 형식으로 치장하는 경우가 있다. 그러나 카드와 연하장, 그리고 요금 청구서 등 스팸이 아닌 정상 메일의 경우에도 표 형식이나 예쁘게 꾸민 메일이 있기 때문에 메일 내용을 예쁘게 치장하였다고 해서 모두 스팸으로 간주할 수는 없다. 따라서 스팸 필터링 성능을 개선하려면 이미지와 표 형식의 메일에 대해서는 그에 적합한 필터링 방식을 적용해야 한다. 본 논문에서는 이처럼 세부적인 속성들과 관련된 필터링 방식을 적용하기에 앞서 보편적으로 적용되는 방식을 우선적으로 고려하였다.

3.1 주소 유효성 검사

스팸 메일의 특성을 찾기 위하여 데이터 집합을 분석하는 중 스팸 메일에 유독 메일 헤더 내에 받는 사람의 주소가 없거나 해당 사용자의 메일 주소가 아닌 경우가 많음을 발견할 수 있다. 이런 경우가 발생하는 원인은 SMTP에게 전달하는 보내는 사람 및 받는 사람 정보와 메일 헤더 간의 관계가 없음으로 인한 것이다. 스팸 메일 발송자들은 이를 이용하여 메일 헤더를 일일이 수정하지 않고도 메일 주소가 다른 여러 사용자에게 같은 메시지를 발송할 수 있다.

스팸 메일 필터링 시스템은 이런 특징을 이용하여 사용자의 메일 주소와 메일 메시지 내의 헤더의 받는 사람 주소를 비교하여 스팸 메일을 효율적으로 필터링한다. 즉, 메일을 파싱하여 헤더를 추출하면 추출된 헤더에 대해 주소 유효성을 검사하고 유효하지 않은 경우 이를 반영하여 필터링함으로써 필터링 성능이 향상된다.

3.2 제목-내용의 중요도 반영 및 임계값 조정

스팸 메일 여부를 판단하기 위해 주소 유효성 검사를 이용하는 방법과 단순 베이스 분류자에 의해 메일 제목과 본문 내용에 대해 각각 스팸 메일 확률을 계산하는 방법을 적용하여 스팸 메일 확률을 계산한다. 그런데 통상적으로 스팸 메일의 제목에 주로 사용되는 '카드 대출' 등 특징적인 용어들이 발견되므로 제목과 본문 내용에 출현하는 용어의 중요도를 차별화함으로써 필터링 성능을 향상시킬 수 있다. 제목과 본문 내용의 상대적인 중요도를 차별화하여 비중을 다르게 반영하기 위하여 식(5)와 같이제목-내용 각각에 대해 계산된 스팸 확률값의 반영 비율을 다르게 적용한다. 제목과 본문의 반영 비율을 α, β라 할 때 $\alpha + \beta = 1$ 이며, 구체적인 반영 비율은 최적화 실험에 의해 결정된다.

$$\alpha\frac{P(c_{sp}|E_T)}{P(c_{sp}|E_T)+P(c_n|E_T)}+\beta\frac{P(c_{sp}|E_B)}{P(c_{sp}|E_B)+P(c_n|E_B)}>\tau-a \quad (5)$$

주소 유효성 검사에 의해 유효하지 않은 주소를 가진 메일은 모두 스팸으로 간주하는 방법도 있으나, 스팸이 아닌 경우도 있으므로 스팸 여부를 판단할 때 임계값을 차등 적용하도록 한다. 정상적인 메일의 임계값 τ라 할 때, 주소 유효성 검사를 통과하지 못한 메일은 스팸 메일일 가능성이 높으므로 임계값을 낮춰서 τ-a 로 적용한다. 상수 a값은 유효 주소인 경우 0이고, 그렇지 않을 때의 임계값은 실험을 통해 최적의 값을 결정한다.

최종적으로, 제목-내용의 중요도를 반영하고 주소 유효성 검사에 따라 임계값을 조절하여 스팸 메일 확률을 계산하는 식은 다음과 같다. 이 식에서 전자우편 E의 제목을 ET, 본문의 내용을 EB 로 분리하여 각각에 대한 스팸 확률을 계산할 때 ET와 EB의 스팸 확률은 스팸 범주에 속할 확률과 정상 메일 범주에 속할 확률을 각각 계산한 후에 스팸으로 분류될 비율로써 스팸 메일인지를 판단한다.

3.3 스팸 메일 필터링 시스템

스팸 메일 필터링 시스템의 구조는 그림 1과 같이 학습 모듈과 분류 모듈로 구성된다. 학습의 전제 조건으로, 학습 메일 문서들은 사용자에 의해 스팸 메일과 정상 메일 집합으로 분류되어 있어야 한다. 학습 모듈은 스팸-정상 메일을 분석하고 색인어 추출기를 이용하여 제목과 본문에 대해 각각 주요어(keyword)를 추출한다. 주요어 추출에 색인어 추출기를 이용한 것은 의미가 없는 주요어를 제거하고 어근 추출(stemming)을 하기 위한 것이다. 이 때, HTML 태그가 있는 경우 주요어 추출이 용이하지 못하므로 추출 이전에 HTML 태그를 제거한다. 추출된 주요어들은 {주요어, 스팸에 출현한 빈도수, 정상 메일에 출현한 빈도수} 쌍으로 저장 및 갱신되어 차후 단순 베이스 분류자의 데이터로 사용된다.

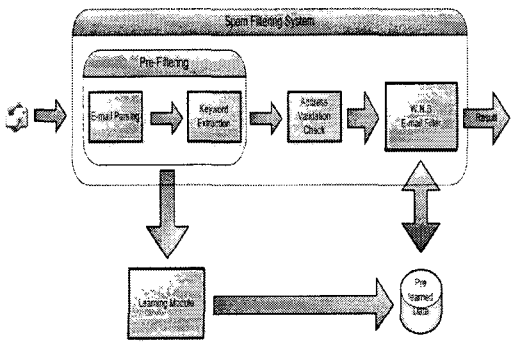

그림 1. 스팸 메일 필터링 시스템 구조도

분류 모듈의 입력으로 주어진 메일은 학습 모듈에서 학습 메일을 분석한 것과 동일한 방법으로 메일의 구조를 분석한다. 메일의 구조 분석 결과로부터 메일 헤더에서 주소 유효성 검사를 수행하여 메일 주소의 유효성을 검사한다. 입력 메일로부터 문서 벡터를 구성하기 위해 HTML 태그를 제거하고 제목과 내용에 대해 각각 주요어들을 추출한다. 입력 메일에서 추출된 주요어들에 대해 통계적인 확률 기법을 기반으로 단순 베이스 분류자를 적용하여 스팸 확률을 구하여 필터링을 수행한다.

본 논문에서는 정보 검색과 문서 분류에서 사용되는 TF-IDF를 각 단어의 가중치로 사용한 단순 베이스 분류자를 사용하여 필터링을 수행한다. TF-IDF는 공통으로 포함된 단어에 대해서 그 단어들이 전체 문서에서 문헌 빈도가 낮을 경우 특징적인 주요어라고 추측하여 단순 출현 빈도의 단점을 보완하여 상대적으로 높은 값으로 전환하기 위한 방법이다. 문서 D_i에서 단어 t_k가 나온 횟수를 tf_{ik}, 문서 전체의 수를 N, t_k를 포함하는 문서의 수를 n_k라고 할 때, TF-IDF를 이용한 t_k의 가중치 w_{ik}는 식(6)과 같이 계산된다.

$$w_{ik}=tf_{ik}*\log(N/n_k) \quad (6)$$

4. 실험 및 성능 평가

4.1 데이터 집합 및 평가 방법

스팸 필터링 성능의 향상 방안으로 제시한 방법들의 효용성에 대해 실험을 수행하였으며, 각 요인들을 복합적으로 조합한 시스템의 성능을 측정하였다. 학습 및 실험에 사용된 메일들은 회사내 사용자들이 수신한 전자 메일로부터 수집하였다. 실험에 사용된 학습 데이터와 실험 데이터의 크기는 표 1과 같이 총 5,155개이고 정상 메일의 개수가 2,319개, 스팸 메일의 개수는 2,736개이다.

표 1. 데이터 집합의 구성

	학습 문서	실험 문서
스팸 메일	1,915	821
정상 메일	1,623	696
합계	3,538	1,517

스팸 필터링 성능을 평가할 때 사용되는 평가 방식은 10분 교차 평가 방식(10-fold cross validation)이 많이 사용된다. 그 이유는 실험 데이터 집합의 크기가 충분히 크지 않기 때문이다. Tretyakov(2004)의 경우에 메일 개수가 총 1,099개이고 스팸 메일의 개수가 481개인 PU1 코퍼스를 사용하였다. Androutsooulos (2000)의 실험에 사용된 메일의 개수는 2,893개인 Ling-spam 코퍼스를 사용하고 있다. 이 실험 데이터는 스팸 메일의 개수가 481개로 전체 메일의 약 16.6%이다. 본 논문에서 수집한 메일의 개수는 5,155개로 실험 집합의 크기가 크기 때문에 2분 교차 평가 방식을 사용하였다.

스팸 메일 필터링의 성능은 통상적으로 스팸 메일 필터링의 성능 평가에 사용되는 아래 평가 기준들을 사용하여 필터링 성능을 측정하였다. 아래 성능 평가 수식에서 $N_{S\to L}$은 스팸 메일을 정상 메일로 판단한 오류(false-negative)의 개수이고, $N_{L\to S}$은 정상 메일을 스팸 메일로 판단한 오류(false-positive)의 개수이다.

- 에러율(error rate)

$$E = \frac{N_{S\to L} + N_{L\to S}}{N}$$

- 정상메일 fallout

$$F_L = \frac{N_{L\to S}}{N_L}$$

- 스팸 메일 fallout

$$F_S = \frac{N_{S\to L}}{N_S}$$

- 스팸 정확률과 스팸 재현율

$$P = \frac{\text{스팸으로 분류한 실제 스팸 수}}{\text{스팸으로 분류된 메일 수}}$$

$$R = \frac{\text{스팸으로 분류된 실제 스팸 수}}{\text{전체 스팸 메일 수}}$$

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

이 식에서 알 수 있듯이, 정확률에는 정상 메일이 스팸 메일로 분류되는 경우가 포함되어 있고, 재현율에는 스팸 메일이 정상 메일로 분류되는 경우가 포함되어 있다. 스팸 메일이 정상 메일로 분류되는 경우에는 사용자가 필터링해야 하는 불편이 따른다. 그러나 정상 메일이 스팸 메일로 분류되는 경우에는 중요한 메일이 유실되는 최악의 경우가 발생할 수 있으므로 필터링은 전자보다 후자를 더 중시하는 방향으로 이루어져야 한다.

4.2 임계값 및 요인별 가중치 실험

실험 및 성능 평가에 앞서 성능에 영향을 미치는 여러 요인들의 유효성 평가를 위한 모듈별 실험을 위해 각 요인들을 결합하여 구성한 실험 모델들은 표 2와 같다.

4.2.1 F-값 측정에 의한 임계값 계산

실험 전체에서 사용할 임계값은 기본 단순 베이스 기법 F_N의 F-값 측정을 통해 구하였다. 이 실험에 사용된 데이터는 학습 데이터를 사용하였는데 각각의 임계값 변화에 따른 스팸 메일 필터링 성능은 그림 2와 같다.

그림 2에서 임계값을 0.1에서 1까지 변화시키면서 F-값을 측정한 결과, b값을 1로 주었을 때는 임계값이 0.4일 때 94.3%가 최대이고, b값을 0.5로 주었을 때에는 임계값이 0.5일 때 92.6%의 성능을 보이고 있다. 본 논문에서는 b값을 0.5로 취하였으므로 임계값을 0.5로 정하였다.

표 2. 필터링 성능 실험 모델

모 델	설 명
F_N	기본적인 단순 베이스 분류자
F_A	주소 유효성 검사 사용 (단순 베이스 분류자 사용하지 않음)
F_NA	주소 유효성 검사 + 단순 베이스 분류자
F_T	단순 베이스 분류자 + TF-IDF 가중치
F_AT	주소 유효성 검사 + 단순 베이스 분류자 + TF-IDF 가중치
F_ATR	단순 베이스 분류자 + TF-IDF 가중치 + 제목-내용-주소 검사 비중 조절

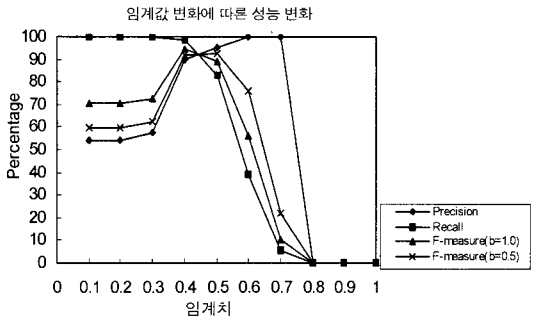

그림 2. 임계값 변화에 따른 필터링 성능

4.2.2 주소 검사에 따른 필터링 성능 측정

이 절에서는 우선 주소 유효성 검사 기법만 적용한 필터링 성능을 측정하여 단순 베이스 분류자를 사용할 때 주소 유효성 검사가 어느 정도의 성능 향상을 보이는지를 실험하였다. 주소 검사의 효용성만을 평가하기 위해 단순 베이스 분류자는 TF-IDF 가중치를 사용하지 않았으며 측정 결과는 표 3과 같다.

주소 유효성 검사를 통해 주소가 잘못된 경우에 무조건 스팸으로 필터링해 보았다. 실험 결과를 살펴보면, 정확률이 98.8%로 향상되었고 재현율은 77.0%였다. 이를 통해, 1.3%의 정확률 하락으로 약 77%의 스팸 메일을 단순 베이스 분류자를 사용하지 않고 빠르고 확실하게 필터링할 수 있음을 알 수 있었다. 또한, 주소 유효성 검사를 혼합하여 사용할 경우 세 가지 중 가장 높은 성능을 보임으로써 주소 검사의 사용이 필터링 성능 향상에 기여함을 확인하였다.

4.2.3 TF-IDF 가중치 적용에 따른 필터링 성능 측정

단순 베이스 분류자에 TF-IDF를 적용하는 경우와 적용하지 않는 경우의 필터링 성능을 측정해 보았다. 그 효용성을 평가하기 위해 주소 유효성 검사는 하지 않았으며 실험 결과는 표 4와 같다.

표 3. 주소 유효성 검사에 따른 필터링

모델	정상→스팸	스팸→정상	F_L	F_S	스팸 정확률	스팸 재현율	F-measure (b=0.5)
F_N	34	135	4.9%	16.4%	95.3%	83.6%	92.7%
F_A	8	189	1.2%	23.0%	98.8%	77.0%	93.5%
F_NA	42	78	6.0%	9.5%	94.7%	90.5%	93.8%

표 4. 가중치 적용에 따른 필터링 성능

모델	정상→스팸	스팸→정상	F_L	F_S	스팸 정확률	스팸 재현율	F-measure (b=0.5)
F_N	34	135	4.9%	16.4%	95.3%	83.6%	92.7%
F_T	21	110	3.0%	13.4%	97.1%	86.6%	94.8%

단순 베이스 분류자에 TF-IDF 가중치를 적용할 경우 적용하지 않은 경우보다 정확률과 재현율 모두 향상되었다. 또한, F-값에서 2.6% 높은 성능을 보임으로써 TF-IDF 가중치 적용이 필터링 성능 향상에 도움이 됨을 확인하였다.

4.2.4 제목-내용-주소 검사 비중에 따른 필터링 성능

제목-내용-주소 검사의 비중은 제목이 스팸일 확률에 부여되는 비중, 내용이 스팸일 확률에 부여되는 비중, 주소 체크를 통과하지 못하였을 경우에 부여되는 비중으로 총 3 가지이다. 실험은 앞에서 살펴본 필터링 성능에 영향을 미치는 요인들을 모두 사용하였다. 성능 측정을 단순화하기 위해, 단순 베이스 분류자와 관련이 있는 제목과 내용의 비중 최적값을 먼저 구한 후에 최적값을 적용하여 주소 검사 비중의 최적값을 구하였다.

먼저 제목-내용 비중의 최적값을 찾는 실험을 진행하였으며 두 값이 서로 상대적인 값을 갖도록 하기 위해 제목 비중과 내용 비중의 합이 1이 되도록 하였다. 제목 비중을 0.05 단위로 증가시키면서 필터링 성능의 변화를 관찰하였다. 그림 3은 제목과 내용 부분의 비중 변화에 따른 필터링 성능 변화를 그래프로 표현한 것이다.

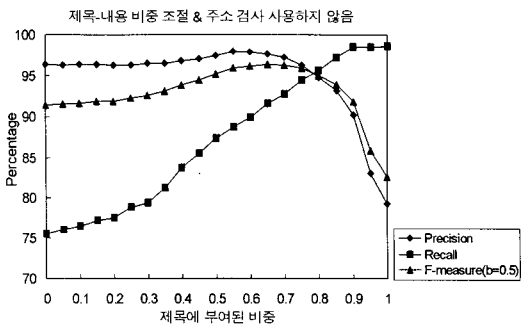

그림 3. 제목-내용 비중에 따른 필터링 성능 변화

실험 결과에 의하면, 정확도가 약간의 상승세를 보이다가 제목에 0.6 이상의 비중을 부여하면서 다시 하락세를 보였으며 재현율은 지속적인 상승을 보였다. 이 중 제목에 0.65의 비중을 부여하였을 때 F-값이 최고치인 96.39%의 결과를 보여 최적의 제목 비중을 0.65, 내용은 0.35로 정하였다. 다음 단계로, 앞에서 최적의 제목과 내용의 비중을 이용하여 주소 검사 비중의 최적값을 구하기 위한 실험을 진행하였다. 주소 검사 비중은 0에서 0.1 사이에서 0.01 단위로 변화를 주었으며 그 결과는 그림 4와 같다.

주소 검사에 대한 비중을 증가시킴에 따라 정확도는 조금씩 하락하고 재현율은 상승하였다. 그러나 0.09부터 정확도가 크게 떨어졌고, 재현율은 0.06부터 적은 폭으로 상승하는 결과를 보여 전체적인 F-값은 최고치 이후로 계속 떨어지는 것을 확인할 수 있었다. 이 실험 결과로부터 0.1 이상의 비중은 의미가 없음을 알 수 있었다. 주소 검사 비중을 0.08로 주었을 때 F-값이 최고치인 97.2%의 성능을 보여 최적의 주소 검사 비중을 0.08로 정하였다. 이는 본 논문에서 제안하고 구현한 스팸 메일 필터링 시스템의 최적의 성능이기도 하다.

4.3 스팸 필터링 실험 및 성능 평가

주소 유효성 검사 및 제목-내용 가중치 차별화 등 필터링 요인들을 모두 사용한 시스템의 필터링 성능을 측정하였다. 측정 결과를 비교할 수 있도록 각각에 대한 필터링 성능을 결과에 포함하였는데 그 결과는 표 5 및 그림 5와 같다.

실험 결과에 의하면, 모든 요인들을 사용했을 때 정확률은 F_A와 F_T보다 약간 낮은 성능을 보이지만 재현율이 91.6%로 가장 높은 성능을 보임으로써

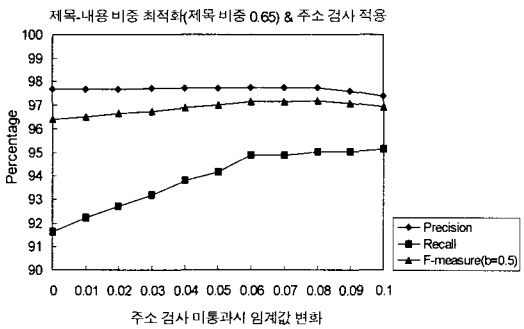

그림 4. 주소 비중에 따른 필터링 성능 변화

표 5. 스팸 메일 필터링 성능 비교

모델	정상→스팸	스팸→정상	F_L	F_S	스팸 정확률	스팸 재현율	F-measure (b=0.5)
F_N	34	135	4.9%	16.4%	95.3%	83.6%	92.7%
F_A	8	189	1.2%	23.0%	98.8%	77.0%	93.5%
F_NA	42	78	6.0%	9.5%	94.7%	90.5%	93.8%
F_T	21	110	3.0%	13.4%	97.1%	86.6%	94.8%
F_AT	29	69	4.2%	8.4%	96.3%	91.6%	95.3%
F_ATR	18	41	2.6%	5.0%	97.7%	95.0%	97.2%

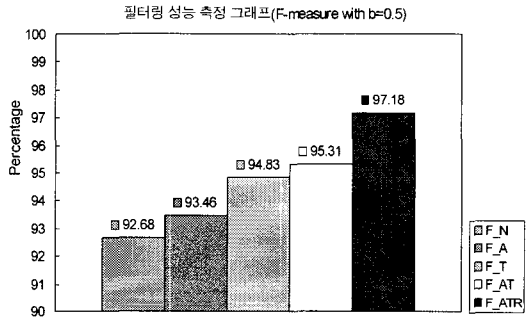

그림 5. 시스템의 필터링 성능 측정

F-값에서 95.3%로 가장 높은 성능을 보였다. 이는 단순 베이스 기법을 사용한 경우에 비해 재현율이 11.6% 향상되고, 정확률은 2.1% 향상되는 효과가 있었다. 또한, 일반적으로 많이 쓰이고 있는 F_T방법과 비교했을 때 재현율이 8.4%, 정확률은 0.6% 향상된 것이다.

5. 결 론

본 논문에서는 주소 유효성을 검사하는 방법과 메일의 제목-내용의 중요도에 따른 비중을 반영하여 스팸 필터링 시스템의 성능을 향상시키는 방안을 제안하였다. 그 성능 향상 정도를 분석하기 위해 단순 베이스 분류 방법에 제안한 방법론을 적용하였고 성능 평가를 실시하였다. 그 결과로, 주소 유효성 검사 및 제목-내용의 중요도를 반영하는 방법이 스팸 메일 필터링 시스템의 성능 향상에 기여함을 밝혔으며, 각 요소들을 조합한 방법이 단순 베이스 분류 기법에 매우 효율적으로 적용될 수 있음을 확인하였다. 본

논문의 연구 결과는 단순 베이스 분류자뿐만 아니라 다른 방법론에도 적용될 것으로 기대되며, 향후 연구로는 메일의 특성을 분석하여 필터링 성능 향상에 기여하는 새로운 요소들을 발견하고자 한다.

참 고 문 헌

[1] 김현준, 정재은, 조근식, "가중치가 부여된 베이지안 분류자를 이용한 스팸 메일 필터링 시스템," 한국정보과학회 논문지:소프트웨어 및 응용, 제31권 8호, pp. 1092-1100, 2004.

[2] 박정선, 김창민, 김용기, "퍼지 관계곱을 이용한 내용 기반 정크 메일 분류 모델," 정보과학회논문지: 소프트웨어 및 응용, 제29권 10호, pp. 726-735, 2002.

[3] W. Cohen, "Fast Effective Rule Induction," *Proceedings of 12th International Conference on Machine Learning*, pp. 115-123, 1995.

[4] R. Schapire, Y. Singer, and A. Singal, "Boosting and Rocchio Applied to Text Filtering," *Proceedings of 21th Annual International Conference on Information Retrieval*, SIGIR, 1998.

[5] K. Tretyakov, "Machine Learning Techniques in Spam Filtering," *Data Mining Problem-oriented Seminar, MTAT. 03. 177*, pp. 60-79, 2004.

[6] L. Zhang, J. Zhu, and T. Yao, "An Evaluation of Statistical Spam Filtering Techniques," *ACM Transactions on Asian Language Information Processing (TALIP)*, Vol. 3, No. 4, pp. 243-269, 2004.

[7] L. Zhang and T. Yao, "Filtering Junk Mail with A Maximum Entropy Model," *Proceeding of 20th International Conference on Computer Processing of Oriental Languages (ICCPOL03)*, pp. 446-453, 2003.

[8] H. Drucker, D. Wu, and V. N. Vapnik, "Support Vector Machines for Spam Categorization," *IEEE Transactions on Neural Networks, Vol. 20, No. 5*, pp. 1048-1054, 1999.

[9] C. Brutlag and J. Meek, "Challenges of the Email Domain for Text Classification," *Proceedings of the 17th Internatinal Conference on Machine Learning*, 2000.

[10] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C. D. Spyropoulos, and P. Stamatopoulos, "Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach," *Proc. 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2000)*, pp. 1-13, 2000.

[11] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian Approach to Filtering Junk E-Mail," *Proceedings of the AAAI Workshop*, pp. 55-62, 1998.

[12] M. Salib, "MeatSlicer: Spam Classification with Naive Bayes and Smart Heuristics," *Proceedings of the Spam Conference*, MA, Jan., 2003.

[13] K. Schneider, "A Comparison of Event Models for Naive Bayes Anti-Spam E-Mail Filtering," *Proceedings 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, pp. 307-314, 2003.

[14] T. Joachims, "Text Categorization with Support Vector Machine: Learning with Many Relevant Features," *Proceedings of 10th European Conference on Machine Learning*, pp. 137-142, 1998.

강 승 식

1986년 서울대학교 컴퓨터공학과 (학사)
1988년 서울대학교 컴퓨터공학과 (석사)
1993년 서울대학교 컴퓨터공학과 (박사)
1994년~2001년 한성대학교 정보전산학부 부교수
2001년~현재 국민대학교 컴퓨터학부 부교수
관심분야 : 한국어 정보처리, 정보검색, 텍스트마이닝 등