스팸성 자질과 URL 자질을 이용한 최대엔트로피모델 기반 스팸메일 필터 시스템

공미경, 이경순
전북대학교 전자정보공학부
{mggong, selfsolee}@chonbuk.ac.kr

A Spam Filter System based on Maximum Entropy Model Using Spamness Features and URL Features

Mi-Gyoung Gong, Kyung-Soon Lee
Division of Electronics and Information Engineering,
Chonbuk National University

요 약

본 논문에서는 스팸메일에 나타나는 스팸성 자질과 URL 자질을 이용한 최대엔트로피모델 기반 스팸 필터 시스템을 제안한다. 스팸성 자질은 스패머들이 스팸메일에 인위적으로 넣는 강조 패턴이나 필터 시스템을 통과하기 위해 비정상적으로 변형시킨 단어들을 말한다. 스팸성 자질 외에 반복적으로 나타나는 URL과 비정상적인 URL도 자질로 사용하였다. 메일 수신자에게 추가적인 정보 제공을 목적으로 하이퍼링크로 연결시키거나 메일에 직접 타이핑한 URL 중 필터 시스템을 피하기 위해 유효하지 않은 비정상적인 URL들이 스팸 메일을 걸러내는데 도움을 줄 수 있기 때문이다. 또한 스팸성 자질과 URL을 각각 적용한 두 분류기를 통합하였다. 분류기의 통합은 각 분류기에 이용된 자질을 독립적으로 사용할 수 있다는 장점을 가지고 있다. 실험 결과를 통해 스팸성 자질과 URL을 이용함으로써 스팸 필터 시스템의 성능을 향상시킬 수 있음을 확인할 수 있었다.

1. 서 론

전자우편 서비스는 사용하기 쉽고 빠르다는 장점을 갖고 있다. 인터넷 사용이 급속도로 증가하면서 사람들은 이메일 사용에 익숙해졌고, 각 개인이 이메일 주소 하나씩은 갖고 있을 정도로 보편화되었다. 그러나 최근 스팸메일이 증가하면서 이메일 이용자들에게 불편을 가중시키고 있다. 스팸메일은 수신자와 직접적 관련이 없는 사람이 불특정 다수에게 일방적으로 발송하며, 수신자가 원하지 않는 쓸모 없는 정보를 담고 있는 전자 메시지를 말한다. 이러한 스팸메일을 사전에 차단하고 효율적인 전자우편 서비스를 제공하기 위해 스팸 필터 시스템의 필요성이 대두되었다.

스팸 필터 시스템은 클래스가 두 개인 문서분류 시스템으로 볼 수 있다[1]. 스팸 필터 시스템은 수신된 이메일을 자동적으로 스팸메일 또는 정상메일로 분류한다. TREC 2005부터 표준화된 성능평가방법 제공을 목적으로 스팸 필터링 관련 연구를 진행하고 있다.

TREC2005 참가자들은 베이지안 분류기(Bayesian Classifier), 마코프 랜덤 필드 모델((Markov Random Field Model)과 K-NN 방법(K-nearest neighbor method)을 이용한 스팸 필터 시스템을 제안하였다.

불특정 다수에게 대량의 메일을 발송하는 스패머들은 메일에서 강조하고자 하는 단어를 대문자로 표기하거나 느낌표 등의 강조부호를 지나치게 반복해서 사용하는 경향이 있다. 또한 필터 시스템을 빠져나가기 위해 단어를 이루는 문자의 일부를 유사한 기호나 숫자로 변경하기도 한다. 스팸 필터 시스템이 진화함에 따라 스패머들도 끊임없이 필터 시스템을 빠져나갈 수 있는 방법들을 생각해낸다. 스패머들이 사용하는 단어표현이나 변칙들은 스팸메일을 걸러내는데 중요한 자질(스팸성 자질)이 될 수 있으며 이를 필터 시스템에 반영할 필요가 있다.

본 논문에서는 스팸성 자질과 스패머들에 의해 변형된 비정상적인 형식을 포괄하는 URL자질을 각각 이용한 최대 엔트로피 모델 기반 분류기와 이들을 결합한 스팸 필터 시스템을 제안한다. TREC 스팸 필터 테스트 컬렉션을 이용한 실험을 통해 제안 방법의 유효성을 검증하였다[2].

성능 비교를 위해 TREC에서 제공한 공개 시스템 중 가장 좋은 성능을 보인 보고필터[3]를 사용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 스팸 필터링 분야에서 진행되어 온 관련연구들을 중심으로 살펴보고, 3장에서는 본 연구에서 사용한 확률 모델인 최대 엔트로피 모델에 대해 간략히 소개한다. 4장에서는 제안하는 스팸 필터 시스템의 자질함수와 시스템 구조에 대해 기술하고 5장에서는 실험 결과를 토대로 분석하고 결론을 맺는다.

2. 관련 연구

스팸 필터링 시스템 연구의 대부분은 베이지안 분류기를 기반으로 하고 있다. 그 밖에 마코프 랜덤 필드 모델(Markov Random Field model)을 이용[4]하거나 K-NN 방법(K-nearest neighbor method)을 이용[5]해서 스팸 필터 시스템을 제안한 사례도 있다.

다음은 베이지안 분류기를 사용한 연구들이다. 메일에서 기본적으로 헤더와 본문을 추출한다. [6]에서는 헤더 정보를 사용해서 블랙리스트 필터를 설계하였다. 헤더 정보의 "From:" 필드는 보내는 사람의 이름과 메일 주소를 포함한다. 이를 통해 테스트 메일의 스팸성 여부를 판단한다. 반대로 [7]는 정상메일에서 "From:" 필드를 추출함으로써 화이트리스트를 사용하였다. [8]에서는 메일에 나타나는 표현적 특징이나 메일 본문과 제목에서 느낌표의 발생 빈도, 대문자의 사용 정도를 자질로 포함시켰다. [9]에서는 구두점과 다른 특수기호의 배열이 스팸 필터 시스템에서 유용하게 사용할 수 있는 특징임을 파악하고 문자기반 모델을 제안하였다. [10]은 나이브 베이지안 분류기의 확장된 형태로 볼 수 있는 Less Naïve Bayes(LNB)를 사용한 필터 시스템을 제안하였다. 또한 LNB를 SMTP 경로 분류기와 통합하였다. [11]은 나이브 베이지안 분류기와 키워드 기반의 스팸 필터링 시스템을 비용 측면에서 비교하였다. 보고필터(bogofilter)[3]는 역 카이 제곱 함수를 적용한 베이지안 스팸 필터 시스템이다.

다이그래믹 베이지안 분류기(dbacl;digramic Bayesian classifier) 기반 필터 시스템[12]은 메일의 원문 내용을 MIME 디코딩과 HTML 태그 제거 과정을 거쳐 헤더와 본문을 토대로 자질을 추출한다. 다이그래믹 베이지안 분류기는 자질의 파라미터 값을 최대 엔트로피 원리를 이용해서 계산하고, 각 테스트 문서의 확률은 베이지안 기법을 이용해서 계산한다.

최대 엔트로피 모델은 문맥적 정보를 반영하며 베이지안 모델과 달리 자질들이 독립적이라는 가정을 필요로 하지 않는다. 분류기 경우 문맥적 정보는 정의된 자질과 자질을 포함하는 클래스가 될 수 있다. 자질과 클래스를 함께 고려해서 자질의 파라미터 값을 산출한 후 최종 확률 계산에 이용한다. 문맥적 정보는 자질 함수를 통해 다양하게 반영할 수 있기 때문에 본 논문에서는 최대 엔트로피 원리를 기반으로 하는 최대엔트로피모델(Maxent)[13]에 스팸성 자질과 URL 자질을 적용한 스팸 필터 시스템을 제안한다.

3. 최대 엔트로피 모델(Maximum Entropy Model)

최대 엔트로피 모델[14]은 조건 확률을 이용해서 임의의 문서의 클래스를 추정하는 확률 모델로 최대 엔트로피 원리를 기초로 하고 있다. 최대 엔트로피 원리(Maximum Entropy principle)는 확률 분포를 최대한 균일하게 만드는 것을 의미한다. 주어진 제약 조건을 만족하면서 엔트로피가 최대가 되도록 확률 분포를 형성한다. 즉, 알고 있는 정보는 최대한 반영하면서, 고른 확률 분포를 만드는 것이다.

최대 엔트로피 모델에서는 학습 데이터가 확률 분포의 제약 조건을 결정하는데 이용된다. 즉, 확률 계산에 사용되는 자질들의 값은 근사적으로 학습 데이터에 한정되어 구해진다. 또한 최대 엔트로피 모델은 모델에서 사용되는 자질들을 선택하기 위해 자질 함수를 필요로 한다. 자질 함수에 의해 생성된 각 자질의 파라미터 값은 학습데이터를 이용해서 구해지며, 이 값들은 임의의 메일이 스팸메일이나 정상메일에 속할 확률을 계산하는데 이용된다.

최대 엔트로피 모델에서의 엔트로피 계산과 확률 계산 방법은 아래와 같다.

$$H(p) = -\sum p(c \mid d) \log p(c \mid d) \tag{1}$$

$$p(c \mid d) = \frac{1}{Z(d)} \exp\left(\sum_i \lambda_i f_i(d,c)\right)$$

여기서

$f_i(d,c)$: 자질 함수

λ_i: 자질의 파라미터 값 (2)

$Z(d)$: 정규화

$$Z(d) = \sum_c \exp\left(\sum_i \lambda_i f_i(d,c)\right)$$

문서 분류에서 자질 함수의 일반적 표현은 다음과 같다.

$$f_{w,c'}(d,c) = \begin{cases} 1 & \text{if } w \in d \ \&\ c' = c \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

여기서

w: 단어, d: 문서, c: 클래스

각 자질의 파라미터 값 추정 방법은 1972년 Darroch와 Ratcliff에 의해 고안된 GIS(Generalized Iterative Scaling) 알고리즘을 이용하였다[15].

4. 스팸 필터 시스템

그림1은 본 논문에서 제안하는 스팸성 자질을 이용한 분류기와 URL 자질을 이용한 분류기를 결합한 필터 시스템의 전체적인 구조를 나타낸다.

학습단계에서는 학습 문서의 보내는 사람 필드를 이용한 블랙리스트와 화이트리스트가 생성되며, 두 분류기(스팸성

자질 분류기와 URL 자질 분류기)의 학습과정에서 스팸성 자질과 URL 자질의 파라미터 값이 결정된다.

테스트단계에서는 필터 시스템에 들어온 메일의 보내는 사람을 확인하여 동일 주소가 블랙리스트에 있으면 스팸메일로, 화이트리스트에 있으면 정상메일로 분류한다. 블랙리스트/화이트리스트 필터에서 걸러지지 않은 메일들은 스팸성 자질 분류기와 URL 자질 분류기를 각각 통과한다. 각 분류기에서 계산된 스팸메일에 속할 확률 값은 분류기 통합과정에서 더해진다. 이 값을 기준으로 메일의 클래스가 결정된다.

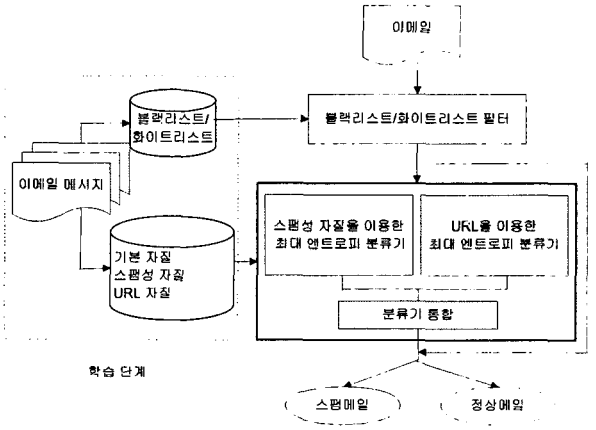

그림 1. 스팸 필터 시스템 구조

4.1 자질함수

1) 기본 자질

문서분류에서 문서에 나타난 각 단어는 단일 자질로 간주된다. 따라서 이메일에 나타나는 단어들은 본 실험에서 기본 자질로 사용하였다.

$$f_{w,c'}(d,c) = \begin{cases} 1 & if\ w \in d\ \&\ c' = c \\ 0 & otherwise \end{cases} \quad (4)$$

여기서

w: 단어, d: 문서, c: 클래스(스팸 또는 정상메일)

2) 스팸성 자질

스패머들이 스팸메일에서 인위적으로 삽입하는 다양한 패턴들을 스팸성 자질로 정의하고 이를 만족하는 자질들을 이메일의 제목과 본문에서 추출하였다. 스팸성 자질들은 학습문서 내 스팸메일 관찰 결과를 토대로 선택하였다. 스팸메일의 제목과 본문에 자주 나타나는 대문자 표현이나 광고성 메일에서 나타나는 가격, 할인율 표현, 의미 없는 특수기호의 사용 등 스팸메일에서 공통으로 묶을 수 있는 패턴들을 찾아 이를 스팸성 자질로 정의한 후 정보이득량을 기준으로 아래 11개 조건에 부합하는 자질을 추출하였다. 표1은 스팸성 자질을 나타내는 조건들을 보여준다. 이들 조건을 만족하는 경우 자질 함수에 의해 스팸성 자질로 결정된다.

$$f_{w,c',h'}(d,c,h) = \begin{cases} 1 & if\ w \in d\ \&\ c' = c\ \&\ h' \in h \\ 0 & otherwise \end{cases} \quad (5)$$

여기서

h: 미리 정의된 조건

스팸성 자질은 사용 목적이나 형식에 따라 크게 숫자와 관련된 자질, 강조 목적으로 사용된 자질, 비정상적인 형식으로 표현된 자질로 나눌 수 있다.

- **숫자와 관련된 자질:**

  달러나 퍼센트 기호와 함께 사용된 숫자와 같이 그 단어가 갖는 의미는 동일하나 숫자 표현방식이 다른 자질들을 말한다. 예를 들면, "$100", "$50" 등이 숫자와 관련된 자질에 속한다. 스패머들은 상품의 판매, 광고, 마케팅과 관련된 이메일에서 잠재적 구매력을 갖는 소비자를 현혹시키기 위해 그들의 상품가격이 싸다는 것을 강조할 때 가격이나 할인율을 표기한다. 그렇기 때문에 "$숫자" 또는 "숫자%"는 광고나 판매 목적의 이메일에서 빈번하게 나타난다. 표 1 의 1~2 번 자질이 여기에 속한다.

- **비정상적으로 변형된 자질:**

  스패머들이 필터 시스템을 빠져나가기 위해 인위적으로 조작 혹은 변형시킨 단어들을 나타낸다. 단어를 이루는 문자들 중 일부를 형태가 비슷한 기호나 숫자로 바꿔 표현하거나 문자 사이에 띄어쓰기를 한 경우가 이에 해당된다. 표1에서 3~5번 자질이 여기에 속한다.

- **강조의 목적으로 사용된 자질:**

  스패머들이 메일에서 강조하고자 하는 부분을 시각적으로 눈에 띄게 만든 자질들이다. 예를 들면, 대문자로 표현된 단어, 느낌표를 포함하는 단어, 특수기호를 여러 번 반복한 단어들이 있을 수 있다. 표1에서 6~11번 자질이 여기에 해당한다.

표1에서 1~10번에 해당하는 자질들은 그 조건을 만족할 때 단일 자질로 표현되며, 11번의 경우는 대문자로 표현된 각각의 단어들이 독립된 자질로 표현된다. 예를 들면 "100%", "50%", "$90" 등은 'NFnumber'라는 단일 자질로 표현되지만, "CLICK", "CASH"는 서로 이질적인 자질로 간주된다.

표 1 스팸성 자질의 조건

조건	예
1 $+ 숫자	$90, $100
2 숫자+ %	100%, 50%
3 'A', 'I', 'a', 'I'를 제외한 단일 문자	Fwd:!
4 문자 사이를 띄어쓰기	M o n e y J u d g e m e n t s
5 변형된 비정상적인 단어	cl!ck, v/agra, 0rder
6 특수 문자의 반복	**********
7 느낌표 반복	Sales!!!!
8 문자+ 기호+ 문자	click....here
9 제목에 대문자가 반 이상	SYSTEMWORKS CLEARANCE SALE
10 특수기호가 제목에 반 이상	Viagra ******20% sales*****
11 대문자로 표현된 단어 (각 대문자가 하나의 자질)	CASH, CLICK

3) URL 자질

URL은 스팸메일의 추가적인 정보를 나타낸다고 볼 수 있다. 스패머들은 스팸메일의 관련된 자세한 사항을 알고 싶을 때 사람들이 그 주소로 연결되는 사이트를 방문할 거라고 예상할 것이다. 따라서 흔히 스팸메일로 분류되는 마케팅, 세일, 성인광고 등과 연관된 URL들이 스팸메일에 반복적으로 나타나고, 이러한 URL들은 스팸 필터에 사용될 수 있는 중요한 자질이 된다.

본 논문에서는 HTML 태그로 링크되어 있는 URL과 메일 본문에 직접 나타난 URL을 추출해서 자질로 사용하였다. 일반적으로 홈페이지가 동일하더라도 하이퍼링크로 연결되면서 추가적인 경로가 뒤에 붙는다. URL전체를 자질로 간주하는 경우 동일 사이트임에도 불구하고 서로 다른 자질로 구분되는 경우가 발생할 수 있기 때문에 이때 URL은 상위레벨까지만 고려한다. 즉, 메일에 나타난 주소 *"http://www.bozomber.com/porno/in=dexhtml"* 를 *"http://www.bozomer.com"* 로 변형해서 자질로 추가한다.

정상적인 형식을 갖는 URL은 그 주소가 각각 서로 다른 사이트를 의미하므로 서로 독립적인 자질로 간주한다.

$$f_{u,c',m'}(d,c,m) = \begin{cases} 1 & if\ u \in d\ \&\ c' = c\ \&\ m' \in m \\ 0 & otherwise \end{cases} \quad (6)$$

여기서

u: URL, m: URL형식(정상 또는 비정상)

스패머들은 URL이 블랙리스트나 화이트리스트로 사용될 경우 필터 시스템에 걸러지는 것을 방지하기 위해 URL형식에 변형을 시도한다. 예를 들면 *"http://ltgjss5p9a@agileconcepts.comr"* 과 같은 비정상적인 형식이 있을 수 있다. 이러한 URL은 웹 상에서 유효하지 않다. 스팸성 자질과 마찬가지로 비정상적인 URL은 인위적으로 변형된 것이기 때문에 필터 시스템에서 중요한 자질이 될 수 있다. 비정상적인 URL에서 일부를 제거하면 본래의 URL을 찾을 수 있는 경우도 있다.

"http://www.gardenornaments.adv@hellerwhirligigs.com"
→ "http://www.gardenornaments.com"

본 논문에서는 비정상적인 형태를 갖는 URL 경우 그 주소 자체보다 비정상적인 형태를 갖는다는 것이 더 의미 있으므로 단일 자질로 대체해서 표현하였다.

정상/비정상적인 URL을 다른 자질과 구분하기 위해서 URL만을 자질로 갖는 URL 분류기를 설계하고, 이를 스팸성 자질 분류기와 통합하였다.

4.2 스팸성 자질 분류기와 URL 자질 분류기의 통합

각 이메일의 클래스를 결정하기 위해 필터 시스템은 스팸성 자질 분류기와 URL 자질 분류기에서 계산된 확률을 더한다. 스팸성 자질 분류기는 스팸성 자질을 사용하고, URL 자질 분류기는 URL만 자질로 사용한다.

스팸메일에 속할 확률
= 스팸메일에 속할 확률(스팸성 자질 분류기) (7)
+스팸메일에 속할 확률(URL 자질 분류기)

5. 실험 및 결과

5.1 실험환경

본 논문에서 제안한 방법의 유효성을 검증하기 위해 TREC의 스팸 필터링 테스트 컬렉션[2]을 테스트 문서로 사용하고, SpamAssassin[2]을 학습 문서로 사용하였다. 표2는 학습 문서와 테스트 문서를 구성하고 있는 스팸메일과 정상메일의 개수를 보여준다.

표 2. 학습 문서와 테스트 문서

	테스트 컬렉션 (SpamAssassin and trec05p-1/full)
학습 문서 (스팸메일/정상메일)	6,047 (1,897/4,150)
테스트 문서 (스팸메일/정상메일)	92,189 (39,399/52,790)
총 문서의 개수	98,236

본 실험에서는 TREC 스팸 필터링 트랙과 달리 배치 필터링 방법을 사용하였다. 배치 필터링 방법은 학습 문서와 테스트 문서를 사전에 구분하고 전체 학습 문서에 대해 일률적으로 학습한 후 이를 바탕으로 테스트 문서의 클래스를 결정하는 것이다. TREC 스팸 필터링 트랙에서는 한 번에 하나의 이메일을 필터링 한 후 그 결과를 바로 필터 시스템에

피트백하는 방식으로 실험하였다.

실험 과정은 다음과 같다. 각 이메일의 원문을 MIME 디코딩 한 후 메일에 포함된 HTML 태그를 제거하였다. 스테밍과 불용어 제거 과정은 생략하고 메일 헤더('From:' , 'To:' , 'Subject:')와 본문을 추출한 후 제목과 본문에서 기본 자질과 스팸성 자질, URL 자질을 선택하였다. URL 자질은 학습 문서에 나타난 모든 URL을 사용하였으며 문서 빈도수가 2이상인 스팸성 자질과 기본 자질은 문서 빈도수가 2이상인 것으로 제한하였다.

각 단어의 스팸성 자질 여부는 자질 함수에 의해 결정된다. 메일에 나타난 URL은 스팸성 자질과 마찬가지로 형식이 비정상인지 정상인지 자질 함수를 이용해서 결정한 후 이를 URL 분류기의 자질로 사용하였다.

Jason Baldridge, Tom Morton, Gann Bierner에 의해 설계된 최대 엔트로피 모델 툴킷[13]을 사용하였으며 파라미터 추정 시 사용한 반복 횟수는 학습문서에 대한 정확도가 제일 높게 나온 800번으로 고정하였다.

제안하는 필터 시스템은 기본시스템과 보고필터와 성능 비교를 하였다.

- 보고필터(bogofilter): 베이지언 확률 기반 필터
- 기본시스템: 각 단어를 기본 자질로 이용한 최대 엔트로피 모델 기반 필터
- 제안시스템: 스팸성 자질 분류기와 URL 자질 분류기를 결합한 필터

성능 평가 방법으로는 TREC2005 스팸 트랙[2]에서 사용한 Hm(ham misclassification rate), Sm(spam misclassification rate), Lam(average misclassification rate)과 정확률, 재현률, F_1-측정, 정확도를 사용하였다.

		Gold Standard Judgement	
		ham	spam
Filter Classification	ham	a	b
	spam	c	d

a: ham (correctly classified) [true negative]
b: spam misclassification [false negative]
c: ham misclassification [false positive]
d: spam (correctly classified) [true positive]

그림 2. 성능 평가를 위한 판별법

Hm은 정상메일 오류율로 시스템이 정상메일을 스팸메일로 잘못 분류한 정도를 나타내며, Sm은 스팸메일 오류율로 시스템이 스팸메일을 정상메일로 잘못 분류한 정도를 나타낸다. Lam은 Hm과 Sm의 평균을 나타낸다.

$$Hm(\%) = c / (a+c) \quad (9)$$

$$Sm(\%) = b / (b+d) \quad (10)$$

$$Lam(\%) = \text{logit}^{-1}(\text{logit } Hm\% + \text{logit } Sm\%) / 2 \quad (11)$$

여기서 $\text{logit } x = \log(\text{odds } x)$

$\text{odds } x = x / (100\% - x)$

5.2 실험 결과

표3은 TREC2005 스팸 트랙의 테스트 문서에 대한 비교 실험1의 결과를 보여준다.

표 3. 비교 실험1

평가방법 (%)	보고필터	기본 시스템	스팸성자질 분류기	URL 분류기	통합시스템
재현률	86.70	84.64	78.52	96.10	78.63
정확률	65.14	63.87	74.58	76.30	74.89
F-측정	0.744	0.728	0.765	0.851	0.767
Hm	13.30	15.36	21.48	3.90	21.37
Sm	34.63	35.73	19.97	13.53	19.68
Lam	22.18	24.11	20.72	7.38	20.51
정확도	74.49	72.97	79.38	88.47	79.60
변화률	-	-	+8.78		+9.09

기본 시스템에서 사용된 자질은 총 41,689개, 스팸성 자질 분류기 실험에서는 총 35,410개의 자질이 사용되었다. 단일 자질로 대체되는 자질들이 있기 때문에 전체 자질의 수는 감소한다. URL 자질 분류기에 사용된 자질은 총 5,265개이다.

통합시스템을 기본 단어들만 고려한 기본시스템과 베이지안 확률 기반 공개 시스템인 보고필터와 정확도 측면에서 비교했을 때 각각 9.1%와 6.9%의 성능 향상을 보였다. 재현률은 감소하고 정확률은 증가하면서, F_1-측정을 통해 전체적으로 성능이 향상됨을 확인할 수 있다. 오류율을 기준으로 살펴보면 정상메일 오류율은 증가하고, 스팸메일 오류율은 감소하면서 전체적인 오류율은 감소하였다.

URL 분류기 성능은 학습 문서에 나타난 URL 자질을 포함하는 테스트 문서에 대해서만 계산된 결과이다. 따라서 Spamassasin을 학습 문서로 사용했을 때 테스트 문서는 5,101개이다.

스팸메일로 구분하는 기준은 모든 메일 이용자들에게 동일한 기준으로 적용되지 않을 수 있다. 이는 메일 이용자들의 개인적인 성향들이 반영되기 때문이다. 예를 들면 동일한 광고 메일 수신자들 중에는 스팸메일이 아닌 정상메일로 분류하는 경우가 있을 수 있다. 그 메일의 유용성은 사람에 따라 다르게 평가될 수 있기 때문이다. 따라서 스팸 필터 시스템은 개인화된 시스템으로 구축될 필요성이 있다. 이러한 관점에서 TREC 테스트 문서가 일관성 있는 메일집합이라

고 보고 임의로 10,000개의 문서를 뽑아 학습 문서로 사용하고 나머지 82,189를 테스트 문서로 사용한 실험(비교 실험2)을 하였다. 그 결과는 표4와 같다.

표 4. 비교 실험2

평가방법 (%)	보고필터	기본 시스템	스팸성자질 분류기	URL 분류기	통합시스템
재현률	96.46	89.08	88.18	93.28	88.84
정확률	91.29	90.28	96.27	98.18	96.37
F-측정	0.938	0.897	0.921	0.957	0.925
Hm	3.54	10.92	11.82	6.72	11.16
Sm	7.64	7.96	2.83	0.69	2.78
Lam	5.22	9.33	5.88	2.19	5.65
정확도	94.22	90.70	93.09	97.58	93.42
변화율	-	-	+2.64		+3.00

기본 시스템에서는 87,884개, 스팸성 자질 분류기 실험에서는 72,394개, URL 자질 분류기에서는 3,567개의 자질이 사용되었다.

통합시스템은 기본시스템과 비교했을 때 정확도 측면에서 각각 3%의 성능 향상을 보였으나 보고필터와 비교했을 때는 성능이 약 0.85% 감소하였다. 기본시스템과 비교해 보면 비교 실험1과 마찬가지로 재현률은 감소했으나 정확률은 증가하면서 전체 성능은 향상되었고 오류율 측면에서는 정상메일 오류율은 증가했으나 스팸메일 오류율은 감소하면서 전체적인 오류율은 감소하였다.

URL 분류기에서 사용한 학습문서는 TREC2005 테스트 데이터의 앞에서부터 순서대로 추출한 10,000개 메일이며, 테스트 문서는 7,474개이다.

5.3 결과 분석

1) 스팸성 자질의 유효성

비교 실험1에서 기본시스템에 스팸성 자질을 적용한 결과 11,835개의 이메일의 클래스가 바뀌었으며, 그것들 중 클래스가 정확하게 할당된 이메일은 8,889개였다. 스팸성 자질을 적용한 후 클래스가 변경된 문서에 대해서 약 75.1%의 정확도를 보였다. 이를 통해 스팸성 자질이 스팸 필터 시스템에서 유용한 자질로 사용되었음을 확인할 수 있었다.

표 5는 스팸성 자질의 조건을 보여주는 표1의 5번 자질의 예를 보여주고 있다.

표 5. 변형된 비정상적인 스팸성 자질

변형 패턴	예
'o' → '0'	0ffer, st0ck, l0w, p0pular
'a' → '@'	re@d, ple@se, @dvisor
'i' → '1'	v1agra, PR1CE, med1cat10n
'i' → '!'	pr!ce, !ncred!bLy, P!I|S
'l' → '/'	Do//ars
'l' → '|'	disc|@imer:, P|ILS
'w' → '₩'	₩/iaagra

비교 실험2에서 표5에 나타난 비정상적인 스팸성 자질을 추가한 후 통합 시스템의 정확도가 92.60에서 93.09로 약 0.53% 향상 되었다. 이를 통해 비정상적인 스팸성 자질이 성능 향상에 도움을 준다는 것을 확인할 수 있다.

2) URL 자질의 유효성

비교 실험1에서 URL 자질을 적용한 결과 270개의 이메일의 클래스가 변경되었으며, 이들 중 238개가 정확하게 분류되었다. 클래스가 변경된 문서들에 대해서 정확도는 약 88.15%를 보였다. 비교 실험2에서는 378개의 클래스가 변경된 문서들 중 331개의 문서가 정확하게 분류되면서 약 87.57%의 정확도를 보였다.

비교 실험2에서 정확하게 분류된 문서들을 통해 URL 자질이 필터 시스템 성능에 영향을 준다는 것을 알 수 있다. 예를 들면 학습 문서의 20개 스팸메일에서만 나타난 URL *"http://www.longlife1004.com"* 을 포함한 메일은 스팸성 자질만 고려할 경우 정상메일로 할당되었으나 URL 자질 분류기와 통합되면서 스팸메일로 클래스가 변경되었다. 반대로 학습 문서의 10개 정상메일에서만 나타난 URL *"http://phonecard.yahoo.com"* 을 포함한 메일은 스팸메일에서 정상메일로 클래스가 변경되면서 정확한 분류가 이루어지고 있었다. 이러한 URL의 영향으로 스팸성 자질 분류기와 URL 자질 분류기를 통합한 전체 필터 시스템 성능은 향상되었다.

또한 메일에 나타난 URL 자질만을 이용한 URL 자질 분류기 실험은 비교 실험1과 비교 실험2에서 각각 약 88.47%, 약 97.58%의 정확도를 보였다. 이를 통해 스팸메일에 포함되어 있는 URL이 스팸메일을 분류하는데 의미 있는 자질로 사용될 수 있음을 확인할 수 있다.

3) 오류율 (HM과 SM)

정확도 측면에서 보면 필터 시스템의 성능은 향상되었으나 HM은 높아지고 SM은 낮아지는 결과를 보였다. 스팸메일 필터 시스템의 오류율은 HM과 SM으로 나누어 생각할 수 있는데 HM이 SM보다 손실 비용이 더 크다. 스팸메일을 정상메일로 잘 못 분류할 경우 이는 사람에 의해 걸러질 수 있지만 정상메일이 스팸메일로 분류될 경우 수신자에 중요한 메일이 필터 시스템에 의해 자동적으로 삭제가 될 수 있기 때문이다. 본 실험에서는 스팸성 자질에 초점을 맞추고 있다. 시스템이 학습하는 정보가 스팸메일에 중심으로 이루어지기 때문에 상대적으로 정상메일을 분류하는데 필요한 정보가 부족해지는 것으로 생각된다. 스팸메일을 잘 걸러내

기 위해 스팸메일이 갖는 특징을 반영하는 것도 중요하지만 비용 측면을 고려한다면 정상메일의 오류를 줄이는 방법을 반영할 필요가 있다.

또한 필터 시스템이 스팸메일을 분류할 때 사용하는 임계값도 HM과 SM에 영향을 준다. 임계값을 높게 설정하면 HM은 낮아지고 SM은 높아진다. 반대로 임계값을 낮게 설정하면 SM은 낮아지고 HM이 높아진다. 본 실험에서는 정확도가 비교적 높게 나오는 0.5를 임계값으로 설정하였다. 스팸메일에 속하는 확률이 0.5를 초과하면 그 메일을 스팸메일로 분류한다. 이는 테스트 메일이 스팸메일과 정상메일에 속할 확률이 각각 0.5로 동일할 경우 정상메일로 분류함으로써 중요한 메일 손실을 줄일 수 있다고 판단했기 때문이다.

6. 결 론

본 논문에서는 스팸성 자질과 URL 자질을 이용한 최대 엔트로피 모델 기반 스팸 필터 시스템을 제안하였다. 스패머들은 필터 시스템을 빠져나가기 위해 단어를 다양한 방식으로 변형시킨다. 이를 필터 시스템에 반영하기 위해 스팸메일에 나타나는 강조된 단어나 비정상적인 형식을 갖는 단어를 스팸성 자질로 정의하였다. 스패머들이 변형시키는 내용 중에는 URL도 포함된다. 이를 고려해서 정상적인 형식을 갖는 URL외에 비정상적인 형식의 URL을 자질로 사용하였다. TREC 스팸 필터 테스트 컬렉션을 이용한 실험 결과 제안 방법의 유효성을 확인할 수 있었다. 스팸성 자질과 URL 자질을 이용한 분류기를 통합한 결과 기본시스템과 비교하여 약 9.1%의 성능 향상을 보였으며, 보고필터와 비교하여 약 6.9%의 성능 향상을 보였다. 스팸성 자질 분류기와 URL 자질 분류기를 결합한 결과 필터 시스템의 성능은 88.2%의 정확도를 나타내며 향상되었다. URL 자질 분류기 실험을 통해 메일에 포함된 URL이 스팸 필터 시스템에서 의미 있는 자질이 됨을 확인할 수 있었다.

본 논문에서 사용한 분류기 통합 방법은 스팸메일에 속할 확률을 단순히 더하였다. 두 분류기를 좀 더 효율적으로 통합한다면 필터 시스템의 성능 향상에 도움을 줄 것으로 기대된다. 손실 비용 측면에서 스팸메일을 잘 분류하는 것도 중요하지만 정상메일을 스팸메일로 분류하는 경우가 없어야 하므로 정상메일의 오류율을 낮출 수 있는 방법도 고려해야 한다. 또한 본 논문에서 제시한 스팸성 자질을 보완하고 일반화 시킬 수 있는 방법들이 필요하다.스패머들이 필터 시스템을 통과하기 위해 끊임없는 변화를 시도함으로써 스팸성 자질은 무한대로 변형될 수 있기 때문이다.

참고 문헌

[1] Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. A Bayesian Approach to Filtering Junk E-mail. In Proc. AAAI-98 Workshop on Learning for Text Categorization (1998.)

[2] Cormack, B., Lynam, T. TREC2005 Spam Track Overview. Proc. of Text REtrieval Conference (TREC 2005)

[3] Robinson, G. A Statistical Approach to the Spam Problem. Linux Journal, vol. 107 (2003) http://bogofilter.sourceforge.net/

[4] Assis, F., Yerazunis, W., Siefkes, C., Chhabra, S. CRM114 versus Mr. X: CRM114 Notes for the TREC 2005 Spam Track. Proc. of Text REtrieval Conference (TREC 2005)

[5] Cao, W., An, A., Huang, X. York University at TREC 2005: SPAM Track Proc. of Text REtrieval Conference (TREC 2005)

[6] Yang, K., Yu, N., George, N., Loehrlen, A., McCaulay, D., Zhang, H., Akram, S., Mei, J., Record, I. WIDIT in TREC 2005 HARD, Robust, and SPAM Tracks. Proc. of Text REtrieval Conference (TREC 2005)

[7] Keselj, V., Milios, E., Tuttle, A., Wang, S., Zhang, R. DalTREC 2005 Spam Track: Spam Filtering Using N-gram-based Techniques. Proc. of Text REtrieval Conference (TREC 2005)

[8] Wang, S., Wang, B., Lang, H., Cheng, X. CAS-ICT at TREC 2005 SPAM Track: Using Non-Textual Information to Improve Spam Filtering Performance. Proc. of Text REtrieval Conference (TREC 2005)

[9] Bratko, A., Filipic, B. Spam Filtering Using Character-Level Markov Models: Experiments for the TREC 2005 Spam Track. Proc. of Text REtrieval Conference (TREC 2005)

[10] Segal, R. IBM SpamGuru on the TREC 2005 Spam Track. Proc. of Text REtrieval Conference (TREC 2005)

[11] Androutsopoulos, I., et al. An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages, In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (2000)

[12] Breyer, L. A. DBACL at the TREC 2005. Proc. of Text REtrieval Conference (TREC 2005)

[13] Ratnaparkhi, A. Maximum Entropy Models for Natural Language Ambiguity Resolution. Ph.D. Dissertation. University of Pennsylvania (1998) http://maxent.sourceforge.net/(http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html)

[14] Berger, A., Della Pietra, V., Della Pietra, S. A maximum entropy approach to natural language processing. Computational Linguistics (1996)

[15] Darroch, J.N. and Ratcliff, D. Generalized iterative scaling for log-linear models. The Annals of Mathematical Statistics (1972)