스팸성 자질과 URL 자질의 공동 학습을 이용한 최대 엔트로피 기반 스팸메일 필터 시스템

공 미 경[†] · 이 경 순[††]

요 약

본 논문에서는 스팸메일에 나타나는 스팸성 자질과 URL 자질의 공동 학습을 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. 메일에 나타난 정상적인 URL과 필터 시스템을 피하기 위해 변형된 비정상적인 URL들이 스팸 메일을 걸러내는데 도움을 줄 수 있기 때문이다. 또한 스팸성 자질과 URL자질을 이용한 공동 학습을 하였다. 공동 학습은 학습 과정에서 두 자질을 독립적으로 이용한 비지도 학습 방법으로 정답을 모르는 문서를 이용할 수 있다는 장점을 갖는다. 실험을 통해 스팸성 자질과 URL을 이용함으로써 스팸 필터 시스템의 성능을 향상시킬 수 있음을 확인하였으며 두 자질 집합을 이용한 공동 학습이 필요한 학습 문서의 수를 감소시키면서, 정확도는 일괄 학습 정확도에 근접한다는 것을 확인하였다.

키워드 : 스팸성 자질, URL 자질, 스팸메일 필터, 최대 엔트로피 모델, 공동 학습

A Spam Filter System Based on Maximum Entropy Model Using Co-training with Spamminess Features and URL Features

Mi-Gyoung Gong[†] · Kyung-Soon Lee[††]

ABSTRACT

This paper presents a spam filter system using co-training with spamminess features and URL features based on the maximum entropy model. Spamminess features are the emphasizing patterns or abnormal patterns in spam messages used by spammers to express their intention and to avoid being filtered by the spam filter system. Since spammers use URLs to give the details and make a change to the URL format not to be filtered by the black list, normal and abnormal URLs can be key features to detect the spam messages. Co-training with spamminess features and URL features uses two different features which are independent each other in training. The filter system can learn information from them independently. Experiment results on TREC spam test collection shows that the proposed approach achieves 9.1% improvement and 6.9% improvement in accuracy compared to the base system and bogo filter system, respectively. The result analysis shows that the proposed spamminess features and URL features are helpful. And an experiment result of the co-training shows that two feature sets are useful since the number of training documents are reduced while the accuracy is closed to the batch learning.

Key Words : Spamminess Feature, URL Feature, Spam Filter, Maximum Entropy Model, Co-Training

1. 서 론

웹 서비스 중 하나인 전자우편은 사용하기 쉬우며 빠르다는 장점을 갖고 있다. 그러나 최근 '스팸메일'의 등장으로 전자우편 이용자들의 불편이 가중되었고 인터넷 서비스 제공업체 서버 관리 능력 또한 저하되어 재정적, 정신적 부담이 늘어나고 있다. 스팸메일이란 '불특정 다수에게 일방적으로 발송되며, 수신자가 원하지 않는 쓸모 없는 정보를 담고 있는 전자 메시지'를 말한다. 이러한 스팸메일을 사전에 차단하고 효율적인 전자우편 사용을 위해 스팸메일 필터 시스템의 필요성이 대두되었다. 스팸메일 필터 시스템은 수신된 이메일을 자동적으로 스팸메일 또는 정상메일로 분류하는 클래스가 두 개인 문서분류 시스템이다[1]. 스팸메일 필터링 관련 연구는 국제적인 정보검색 평가대회인 TREC(Text REtrieval Conference)[2]에서 표준화된 성능평가방법 제공을 목적으로 2005부터 진행하고 있다. TREC2005 참가자들은 베이지안 분류기(Bayesian Classifier), 마코프 랜덤 필드

† 정 회 원 : 전북대학교 컴퓨터공학과 공학석사
†† 정 회 원 : 전북대학교 전자정보공학부/영상정보신기술연구센터 조교수
논문접수 : 2006년 11월 23일, 심사완료 : 2007년 12월 30일

DOI: 10.3745/KIPSTB.2008.15-B.1.61

모델((Markov Random Field Model)과 K-NN 방법(K-nearest neighbor method)을 이용한 스팸 필터 시스템에 관한 연구를 진행하였다.

스팸메일을 살펴보면 두드러진 특징들을 발견할 수 있다. 상품 광고의 메일에서는 가격과 할인율 표현이 많이 나타나며, 스패머들이 필터 시스템을 빠져나가기 위해 인위적으로 변형시킨 단어들, 강조의 목적으로 대문자나 강조부호를 빈번하게 사용하는 것들을 볼 수 있다. 본 논문에서는 학습 문서에 나타난 스팸메일을 관찰하고 자주 나타나는 특징들을 스팸 필터 시스템에 반영하기 위해 스팸성 자질로 정의하고 이들을 문맥 정보로 활용한 최대 엔트로피 모델을 사용하였다. 그리고 정의한 자질들을 이용해서 공동학습을 하였다. 공동학습은 정답이 주어지지 않은(unlabeled) 문서를 학습에 이용함으로써 정답을 아는(label) 문서를 얻기 위한 비용이 발생하지 않는다는 장점을 가지고 있다. 서로 다른 자질 집합(스팸성 자질과 URL 자질)을 이용해서 점진적으로 학습 문서를 추가하는 공동 학습은 두 자질을 독립적으로 학습에 반영할 수 있다.

본 논문에서는 스팸성 자질과 스패머들에 의해 변형된 비정상적인 형식을 포함하는 URL 자질의 공동 학습을 이용한 최대 엔트로피 모델 기반 스팸 필터 시스템을 제안한다. TREC 스팸 필터 테스트 컬렉션[2]을 이용한 실험을 통해 그 유효성을 검증하였다. 스팸메일에 나타나는 특징들을 스팸성 자질과 URL 자질 두 개로 분리하고, 각 자질을 이용한 분류기를 생성한 후 두 자질 집합을 이용한 공동 학습을 통해 점진적으로 학습 문서를 추가한다. 제안하는 스팸메일 필터 시스템은 생성된 두 분류기를 결합해서 메일의 클래스를 결정한다.

본 논문의 구성은 다음과 같다. 2장에서는 스팸메일 필터 시스템과 공동 학습에 대한 기존 연구와 최대 엔트로피 모델에 대해 간략하게 살펴본다. 3장에서는 제안하는 스팸메일 필터 시스템의 구조와 정의한 스팸성 자질과 URL 자질에 대해 설명하고 이를 이용한 공동 학습 방법을 설명한다. 4장에서는 실험을 통해 스팸성 자질과 URL 자질의 유용성과 공동 학습 결과를 보인다. 마지막으로 5장에서는 결론과 함께 향후 연구에 대해 서술한다.

2. 관련 연구

필터 시스템에 관한 연구들을 살펴보면 메일이 갖고 있는 특징 또는 정보를 활용해서 필터 시스템을 설계하고 있음을 알 수 있다. 본 논문에서는 스팸메일이 갖고 있는 특징들을 필터 시스템에 적용하기 위해 스팸메일을 관찰하고, 그 결과를 토대로 스팸성 자질과 URL 자질을 정의하였다. 또한 서로 다른 자질 집합을 이용하여 학습 문서를 얻는 비용을 줄일 수 있는 공동 학습에 정의한 두 자질을 적용시켜 보았다. 그 결과 정의한 스팸성 자질과 URL 자질이 공동 학습에서 유효함을 확인하였다. 본 장에서는 증가하는 스팸메일을 차단하기 위한 필터 시스템과 서로 다른 두 자질 집합을 이용한 공동 학습에 관한 연구를 살펴보고자 한다.

2.1 스팸메일 필터 시스템에 관한 연구

스팸메일 필터 시스템은 메일의 정보를 이용해서 스팸메일을 걸러내고 있다. 메일을 분류할 때 본문 내용을 확인하기 전에 블랙리스트/화이트리스트 필터를 사용할 수 있다[3,4]. 메일 헤더의 From 필드를 이용해 블랙리스트(스팸메일의 From 필드)와 화이트리스트(정상메일의 From 필드)를 생성한다. 필터 시스템에 들어온 메일은 보낸 사람이 블랙리스트에 있으면 스팸메일로 화이트리스트에 있으면 정상메일로 분류된다. 스팸 필터링 시스템 연구의 대부분은 베이지안 분류기를 기반으로 하고 있다. 그 밖에 마코프 랜덤 필드 모델(Markov Random Field model)을 이용[5]하거나 K-NN 방법(K-nearest neighbor method)을 이용[6]한 연구가 진행되었다.

가중치가 부여된 베이지안 분류기 [7]는 정보통신부의 개정안을 준수하는 메일을 분류하기 위한 전처리 단계와 사용자의 행동을 학습하여 보다 정확한 분류가 가능하도록 지능형 에이전트가 결합된 형태의 스팸메일 필터 시스템을 제안하였다. SMTP 경로 분석 분류기와 LNB(Less Naïve Bayes)를 결합한 [8]은 시스템에서 이용하는 자질들의 독립을 가정하는 나이브 베이지안 분류기를 확장한 LNB(less Naïve Bayes)와 메일을 발송한 서버 주소를 이용해서 메일을 분류하는 SMTP 경로 분석 분류기의 결합을 제안하였다. 두 독립적인 분류기의 통합은 다양한 자질을 결합함으로써 분류기의 정확도를 향상시킬 수 있다는 것과 하나의 분류기를 사용하는 것보다 시스템의 안정성을 높일 수 있다는 장점이 있다. 문자열(character sequences) 기반 베이지안 분류기 [9]는 전형적인 필터 시스템이 문서에 나타난 단어(BOW: bag-of-word)를 사용하는 것과는 다르게 문자 기반 스팸메일 필터 시스템을 설계하였다. 각 클래스 별로 문자열의 확률을 추정하는 모델을 생성하고 이를 베이지안 분류기처럼 사용하였다. 문서의 클래스는 테스트 문서가 각 클래스에 포함될 조건 확률을 추정해서 확률이 가장 큰 쪽으로 결정된다. 비용 측면에서 필터 시스템의 정확률과 오류율을 계산한 [10]는 필터 시스템에서 손실 비용 따른 가중치를 정확률과 오류율 계산에 반영한다. 정확률에서는 정상메일을 정상메일로 분류한 경우에 가중치를 부여하고 오류율에서는 정상메일을 스팸메일로 분류한 경우에 가중치를 부여한다. 다이그래믹 베이지안 분류기(dbacl: digramic Bayesian classifier)기반 필터 시스템 [11]는 각 클래스에서 최대 엔트로피 원리를 이용한 파라미터 값을 계산한 후 테스트 문서의 클래스를 베이지안 기법을 이용해서 결정한다. TREC에서 성능 비교를 위해 제공한 공개 시스템 보고필터(bogofilter)[12]는 베이지안 스팸 필터로 확률 계산을 할 때 역카이스퀘어함수(inverse chi-square function)를 사용한다. 이 함수는 정상메일일 확률에 민감하며, 그 역함수는 스팸메일일 확률에 민감하다. 보고필터는 두 함수의 결과값을 이용해서 임의의 메일이 스팸일 정도를 0과 1사이의 값으로

나타낸다.

마코프 랜덤 필드 스팸메일 필터 시스템 [5]은 윈도우 사이즈를 5로 하는 OSB(Orthogonal Sparse Bigram) 자질을 사용하였다. 인접한 5개의 단어에서 두 단어씩 묶어 이를 자질로 이용한다. 가중치를 적용한 K-NN 분류기 [6]는 거리에 따른 가중치와 정확도에 따른 가중치를 적용하였다. 가중치가 적용된 유클리디안 거리 함수를 학습 문서와 테스트 문서 사이의 유사도 측정에 사용하였으며, 새로운 문서를 분류할 때 이전 학습 문서의 기여도를 반영해서 정확한 분류에 기여한 학습 문서의 가중치를 높여준다. 텍스트 자질(textual feature)과 비텍스트 자질(non-textual feature)을 이용한 빠른 선형 분류기(Winnow)[13]는 문서 분류에서 자질로 사용되는 일반 단어 외에 메일 본문에 나타난 '$'의 수와 같은 이메일 메시지의 속성을 정의한 비텍스트 자질 총 18개를 사용하였다.

2.2 공동 학습에 관한 연구

문서 분류 시스템은 일정량의 학습 문서를 통해 정보를 학습한다. 학습 방법은 지도 학습(supervised learning)과 비지도 학습(unsupervised learning)으로 나눠진다[14]. 지도 학습은 문서와 문서의 정답(label)을 같이 학습한다. 그러나 사람의 의해 검증된 정답이 주어진(labeled) 학습 문서를 얻는 비용이 발생한다. 반대로 비지도 학습은 학습문서의 정답이 주어지지 않은(unlabeled) 문서를 학습함으로써 정답을 모르는 문서를 그대로 사용할 수 있다는 장점을 갖는다. 공동 학습(co-training)은 비지도 학습에서 자질들을 두 가지 자질 집합으로 분리한 후 서로 다른 두 관점(view)을 이용해 학습 문서를 추가하는 것으로 blum에 의해 제안되었다[14]. 여기서 두 자질 집합을 이용한 분류기는 서로 독립적이며 두 분류기를 통해 서로 다른 정보를 얻을 수 있다. 사용한 자질 집합은 웹 문서 내에 있는 단어와 하이퍼링크로 처리되어 있는 단어이다.

이메일 분류에서 공동 학습을 이용한 [15]는 메일의 제목과 본문으로 이용하였다. 모델은 나이브 베이지안 분류기와 SVM(Support Vector Machine)사용하였으며 SVM이 더 나은 성능을 보였다. 자연어 처리에서 공동 학습을 이용한 [16]은 두 분류기의 결과가 정답과 일치하지 않는 경우 이를 정답으로 교정하여 학습한다. 분류기의 결과가 정답과 일치하지 않을 경우 잘못된 학습으로 성능이 저하될 수 있으므로 이를 학습과정에서 수정하였다.

2.3 최대 엔트로피 모델

최대 엔트로피 모델[17]은 최대 엔트로피 원리를 기반으로 하는 확률모델이다. 최대엔트로피 원리란 알고 있는 정보는 최대한 반영하고 모르고 있는 정보에 대해선 공평한 확률 분포를 형성하는 것으로 이때 엔트로피는 최대가 된다. 예를 들어, 초등학교의 한 반에서 "안경을 낀 여학생의 확률"을 구하고자 한다고 가정해보자. 확률의 합이 1이라는 사실 외에 어떠한 사실도 주어지지 않은 상황에서 구하고자 하는 확률 값은 0, 0.5 등 다양하게 나올 수 있다. 여기서 "안경을 낀 학생이 50%이다."라는 사실을 알게 된다면 이를 고려해서 다시 확률을 구하게 된다. 즉 새롭게 알게 된 사실들은 제약 조건으로 작용하여 이를 만족하면서 최대한 공평한 확률 분포를 구하는 것이 최대 엔트로피 원리이다.

위와 같이 알고 있는 사실은 최대한 반영하고 모르고 것에 대해 공평한 확률 분포를 형성하는 최대 엔트로피 모델의 장점은 문맥적 정보를 반영하며, 베이지안 모델처럼 자질들의 독립을 가정하지 않는다는 것이다. 자질 함수를 통해 다양한 정보를 결합, 반영할 수 있기 때문에 본 논문에서는 최대 엔트로피 모델을 사용하였다. 학습 문서에 나타난 정보들을 제약조건으로 삼고 이를 바탕으로 공평한 확률 분포를 구성하며 미리 정의한 자질들을 문맥 정보로 활용한다. 최대 엔트로피 모델(Maximum Entropy Model)[17]에서 사용하는 자질의 파라미터 값 계산은 학습 문서를 통해 이루어지며, 확률의 총 합은 1이 되어야 한다는 제약 조건을 가지고 있다. 각 자질의 파라미터 값 추정 방법은 1972년 Darroch와 Ratcliff에 의해 고안된 GIS(Generalized Iterative Scaling) 알고리즘을 이용하였다[18]. 모델에서 사용하는 자질들을 결정하기 위해 자질 함수를 정의한다. 주어진 조건을 만족하면 자질 함수가 1이 되고 그렇지 않으면 0이 된다. 문서 분류에서 자질 함수의 일반적 표현은 아래와 같다. 여기서, w는 단어, d는 문서, c는 클래스를 의미한다. 식1은 단어 w가 문서 d에 나타나고, 클래스가 c로 일치할 때 자질 함수 f가 1이 된다는 것을 의미한다. 즉 이때의 w가 자질로 이용된다.

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

최대 엔트로피 모델에서의 확률 계산 방법은 아래와 같다. 문서d가 클래스 c에 들어갈 확률은 문서에 포함된 각 자질의 파라미터 값을 더해서 정규화 시킨 수치가 된다.

$$p(c\mid d)=\frac{1}{Z(d)}\exp\left(\sum_i \lambda_i f_i(d,c)\right) \quad (2)$$

여기서

$f_i(d,c)$: 자질 함수

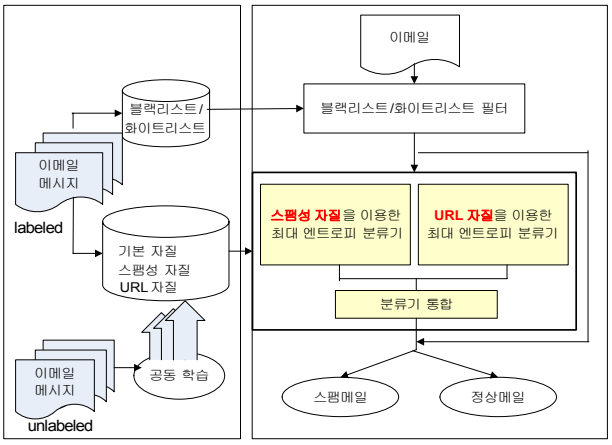
λ_i : 자질의 파라미터 값

$Z(d)$: 정규화

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

3. 스팸성 자질과 URL 자질의 공동 학습을 이용한 필터 시스템

제안하는 스팸성 자질과 URL 자질의 공동 학습을 이용한 최대 엔트로피 기반 필터 시스템은 (그림 1)과 같다. 학



(그림 1) 스팸 필터 시스템 구조

습단계에서는 정답을 아는(labeled) 문서의 보내는 사람을 이용한 블랙리스트와 화이트리스트가 생성되고, 최대 엔트로피 모델을 통해 스팸성 자질, URL 자질, 두 자질의 파라미터 값이 결정된다. 스팸성 자질과 URL 자질을 이용한 공동 학습 과정에서는 정답을 알지 못하는(unlabeled) 문서가 두 분류기에 의해서 점진적으로 학습 문서 집합에 추가된다. 테스트단계에서는 메일의 클래스가 결정된다. 필터 시스템에 들어온 메일은 보내는 사람을 확인하여 동일 주소가 블랙리스트에 있으면 스팸메일로, 화이트리스트에 있으면 정상메일로 분류된다. 블랙리스트/화이트리스트 필터에서 걸러지지 않은 메일들은 스팸성 자질 분류기와 URL 자질 분류기를 각각 통과한다. 각 분류기에서 계산된 스팸메일에 속할 확률 값이 곱해져 메일의 클래스가 결정된다.

3.1 스팸성 자질과 URL 자질 함수 정의

본 논문에서는 문서에 나타난 단어를 기본 자질로 정의하고 스팸에 나타난 특징들을 중심으로 스팸성 자질과 URL 자질을 정의하였다. 정의한 자질 함수들은 다음과 같다.

3.1.1 기본자질

문서분류에서 문서에 나타난 각 단어는 하나의 자질로 간주된다. 본 실험에서 사용한 기본 자질의 자질 함수는 다음과 같다. 여기서 w는 단어, d는 문서, c는 클래스를 의미한다.

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

3.1.2 스팸성 자질

스팸성 자질은 학습 문서에 포함된 스팸메일을 관찰하고 스팸메일에 자주 나타나는 특징들을 스팸성 자질로 정의하였다. 스팸메일에는 정상메일과 달리 대문자나 느낌표와 같은 강조부호가 반복해서 나타난다. 단어를 구성하는 문자를 변경한 무의미한 단어도 나타난다. 이와 같이 스팸메일에 빈번하게 나타나는 특징이나 스패머들이 인위적으로 메일에 삽입하는 다양한 패턴들을 필터 시스템에 반영하였다. 스팸메일의 제목과 본문에 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. <표 1>은 스팸성 자질을 나타내는 조건들을 보여준다. 스팸성 자질 함수는 다음과 같다. 조건 h를 만족하는 경우 자질 함수에 의해 스팸성 자질로 결정된다.

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

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

1) 숫자와 관련된 자질

달러나 퍼센트 기호와 함께 사용된 숫자와 같이 그 단어가 갖는 의미는 동일하나 표현방식이 다른 자질들이 숫자와 관련된 자질에 속한다. 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 1~2번 자질이 여기에 속한다.

2) 비정상적으로 변형된 자질

스패머들이 필터 시스템을 빠져나가기 위해 인위적으로 조작 혹은 변형시킨 단어들이 비정상적으로 변형된 자질에 속한다. 단어를 이루는 문자들 중 일부를 형태가 비슷한 기호나 숫자로 바꿔 표현하거나 문자 사이에 띄어쓰기를 한 경우가 이에 해당된다. <표 1>에서 3~5번 자질이 여기에 속한다.

3) 강조의 목적으로 사용된 자질

스패머들이 메일에서 강조하고자 하는 부분을 시각적으로 눈에 띄게 만든 자질들이 강조의 목적으로 사용된 자질에 속한다. 예를 들면, 대문자로 표현된 단어, 느낌표를 포함하는 단어, 특수 기호를 여러 번 반복한 단어들이 있을 수 있다. <표 1>에서 6~11번 자질이 여기에 해당한다. <표 1>에서 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 다른 자질로 간주된다.

스팸성 자질은 표현 방법이 다양하다. 스패머들은 끊임없이 새로운 변형 방법을 고안하고 이를 스팸메일에 반영할

〈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:! Y
4	문자 사이를 띄어쓰기	Money Judgements
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
10	특수기호가 제목에 반 이상	Viagra ******20% sales*****
11	대문자로 표현된 단어 (각 대문자가 하나의 자질)	CASH, CLICK

수 있다. 본 논문에서 정의한 스팸성 자질은 학습 문서에 나타난 것으로 한정되어 있다. 스패머들이 지능화되어 감에 따라 더 많은 스팸성 자질이 정의될 수 있으며 새로운 스팸성 자질의 지속적인 추가가 필요하다.

3.1.3 URL 자질

메일에 나타난 URL은 메일 내용과 관련 있는 사이트 주소를 나타내므로 스팸메일과 정상메일에 나타난 URL을 메일의 클래스를 구분하는 키워드로 사용할 수 있다. 그러나 두 클래스 사이에 중복되는 URL이 존재하므로 본 논문에서는 키워드 대신 확률 모델을 이용하였다. HTML 태그로 링크되어 있는 URL과 메일 본문에 나타난 URL을 추출해서 자질로 사용하였다. URL전체를 자질로 간주하는 경우 동일 사이트임에도 불구하고 서로 다른 자질로 구분되는 경우가 발생할 수 있기 때문에 상위레벨까지만 고려한다.

"http://www.bozomber.com/porno/in=dexhtml" → *"http://www.bozomer.com"*

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. 변형된 비정상적인 URL은 스팸메일에 포함되어 있을 가능성이 크기 때문에 필터 시스템에서 중요한 자질이 될 수 있다.

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

본 논문에서는 비정상적인 형태를 갖는 URL은 단일 자질로 대체해서 표현하였다. 왜냐하면 그 주소 자체보다 비정상적인 형태를 갖는다는 것이 더 의미 있다고 판단했기 때문이다. 정상적인 형식을 갖는 URL은 그 주소가 각각 서로 다른 사이트를 의미하므로 서로 독립적인 자질로 간주한다. URL 자질 함수는 다음과 같다. 여기서, u는 URL을 m은 URL의 형식 (정상 또는 비정상)을 나타낸다.

초기 집합
L: 레이블(스팸/정상)이 붙은 학습 문서
U: 레이블이 붙지 않은 문서

루프
L에 대하여 **스팸성 자질 분류기**(h1) 생성
L에 대하여 **URL 자질 분류기**(h2) 생성

h1, h2: 신뢰도가 높은 스팸/정상 메일의 레이블 결정(L')

h1과 h2에 의해 레이블이 정해진 문서들을 L에 추가: L ← L+L'
U에서 레이블이 결정된 문서 제거: U ← U - L'

(그림 2) 공동 학습 알고리즘

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

URL 만으로도 분류기를 생성할 수 있다. 본 논문에서는 URL 자질을 다른 자질과 구분하기 위해서 URL만을 자질로 갖는 URL 분류기를 설계하고, 이를 스팸성 자질 분류기와 통합하였다.

3.2 스팸성 자질과 URL 자질을 이용한 공동 학습

본 논문에서는 두 자질(스팸성 자질과 URL 자질)을 이용한 공동 학습을 하였다. 공동 학습 알고리즘은 (그림 2)과 같다.

학습 문서 6,047개에서 75개의 문서를 씨앗(seed) 문서로 랜덤하게 선택하였다. 씨앗 문서에 포함되는 스팸메일과 정상메일의 비율은 학습 문서에서 두 메일의 분포를 고려하여 1:2로 정하였다. 75개의 문서를 제외한 나머지 5,972개는 정답을 알지 못하는 문서이다. 씨앗 문서를 가지고 스팸성 자질을 이용하는 분류기와 URL 자질을 이용하는 분류기를 각각 생성한 후 정답을 알지 못하는 문서에 대해 학습한다. 그 결과 신뢰도가 높은 순서대로 스팸메일 25개, 정상메일 50개를 선택해서 학습 문서로 추가한다. 이때 학습 문서의 클래스는 분류기에 의해 결정된다. 스팸성 자질 분류기와 URL 자질 분류기에서 각각 75개를 선택하며 이 중에서 중복된 문서는 한 번만 추가된다. 문서추가 과정은 50번 반복했다. 씨앗 문서를 랜덤하게 선택한다는 점을 고려하여 공동학습은 10번 수행한 후 그 결과를 평균하였다.

스팸성 자질 분류기와 URL 자질 분류기는 서로 독립적이므로 통합 시스템의 확률은 두 분류기의 확률 곱으로 계산한다[14]. 식 6은 확률 계산 방법을 나타낸다. p(spam), $p_{spamminess}$(spam), p_{url}(spam)은 각각 통합 시스템, 스팸성 자질 분류기, URL 자질 분류기에서 스팸메일에 속할 확률이다. 스팸성 자질 분류기는 기본 자질과 스팸성 자질을 사용하고, URL 자질 분류기는 URL 자질만 사용한다.

$$p(spam) = p_{spam\,\min ess}(spam) \times p_{url}(spam) \quad (6)$$

4. 실험

본 논문에서 제안한 방법의 유효성을 검증하기 위해 TREC의 스팸 필터링 트랙(spam filtering track)에서 제공하는 테스트 컬렉션[2]과 SpamAssassin[2]을 이용하여 실험하였다.

4.1 실험 환경

학습 문서와 테스트 문서를 구성하고 있는 메일의 개수는 <표 2>와 같다. 본 실험에서는 TREC 스팸 필터링 트랙과 달리 배치 필터링 방법을 사용하였다. 배치 필터링 방법은 학습 문서와 테스트 문서를 사전에 구분하고 전체 학습 문서에 대해 일괄적으로 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 HTML 태그를 제거하였다. 스테밍과 불용어 제거 과정은 생략하고 메일 헤더('From:', 'Subject:')와 본문을 추출한 후 제목과 본문에서 기본 자질과 스팸성 자질, URL 자질을 추출하였다. 각 단어의 스팸성 자질 여부와 메일에 나타난 URL 형식(비정상 또는 정상)은 자질 함수에 의해 결정된다. 최대 엔트로피 모델 툴킷[17]을 사용하였으며 파라미터 추정 시 사용한 반복 횟수는 학습문서에 대한 정확도가 제일 높게 나온 800번으로 고정하였다. 성능 평가 방법으로는 TREC2005 스팸 트랙[2]에서 사용한 Hm, Sm, Lam과 정확률, 재현률, F_1-측정, 정확도를 사용하였다.

Hm (ham misclassification rate)은 정상메일 오류율로 시스템이 정상메일을 스팸메일로 잘못 분류한 경우를 나타내며, Sm (spam misclassification rate)은 스팸메일 오류율로 시스템이 스팸메일을 정상메일로 잘못 분류한 경우를 나타낸다. Lam (average misclassification rate)은 Hm과 Sm의 평균을 나타낸다. 필터 시스템이 사용하는 임계값에 따라 Hm과 Sm은 달라진다. 임계값을 높게 설정하면 Hm은 낮아지고 Sm은 높아진다. 반대로 임계값을 낮게 설정하면 Sm은 낮아지고 Hm은 높아진다.

〈표 2〉 학습 문서와 테스트 문서의 개수

	스팸메일 개수	정상메일 개수	전체 개수
학습 문서	1,897	4,150	6,047
테스트 문서	52,790	39,399	92,189
총 문서	54,687	43,549	98,236

사람 \ 시스템	정상메일	스팸메일
정상메일	a	c
스팸메일	b	d

a: 정상메일을 정확하게 분류한 것
b: 스팸메일을 정상메일로 잘못 분류한 것(Sm)
c: 정상메일을 스팸메일로 잘못 분류한 것(Hm)
d: 스팸메일을 정확하게 분류한 것

(그림 3) 성능 평가를 위한 판별법

$$Hm(\%) = c / (a+c) \quad (7)$$

$$Sm(\%) = b / (b+d) \quad (8)$$

$$Lam(\%) = \text{logit}^{-1}(\text{logit } Hm\% + \text{logit } Sm\%) / 2 \quad (9)$$

여기서, logit x = log(odds x), odds x = x / (100% − x)

4.2 실험 결과

스팸성 자질과 URL 자질의 배치(batch) 학습 결과와 공동 학습 결과는 다음과 같다.

4.2.1 스팸성 자질과 URL 자질을 이용한 실험

제안하는 필터 시스템 성능은 기본시스템, 보고필터와 비교하였다. 결과는 <표 3>과 같다.

- 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

기본시스템에서는 41,689개, 스팸성 자질 분류기에서는 35,410개의 자질이 사용되었다. 단일 자질로 대체되는 자질들이 있기 때문에 전체 자질의 수는 감소한다. URL 자질 분류기에서 사용된 자질은 총 5,265개이다. 제안시스템을 정확도 측면에서 기본시스템, 보고필터와 비교했을 때 각각 9.1%와 6.9%의 성능 향상을 보였다.

실험한 결과 제안시스템의 재현률은 감소하고 정확률은 증가하였다. F_1-측정을 통해 전체적으로 성능이 향상됨을 확인할 수 있다. 정상메일로 분류되는 오류율(Hm)은 증가하고, 스팸메일로 분류되는 오류율(Sm)은 감소하면서 전체적인 오류율은 감소하였다. URL 분류기 성능은 학습 문서에 나타난 URL 자질을 포함하는 테스트 문서에 대해서만 계산된 결과이다. Spamassasin을 학습 문서로 사용했을 때 URL을 포함하는 테스트 문서는 5,101개이며 URL 분류기의 정확도는 88.47%을 보였다.

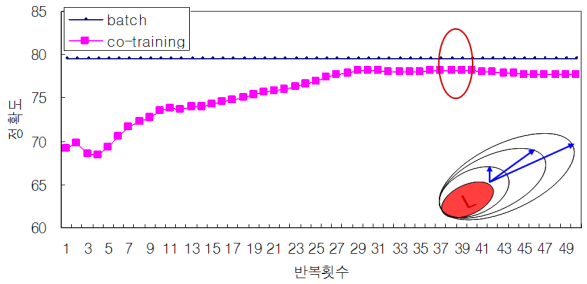
〈표 3〉 실험 결과

평가방법 (%)	보고필터	기본 시스템	스팸성자질 분류기	URL 분류기	제안시스템
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

4.2.2 스팸성 자질과 URL 자질의 공동 학습 실험

공동 학습은 사람에 의해 검증된 정답(label)을 모르는 학습 문서를 이용할 수 있다는 장점이 있다. 공동 학습은 스팸성 자질과 URL 자질을 이용해서 정답을 모르는 학습 문

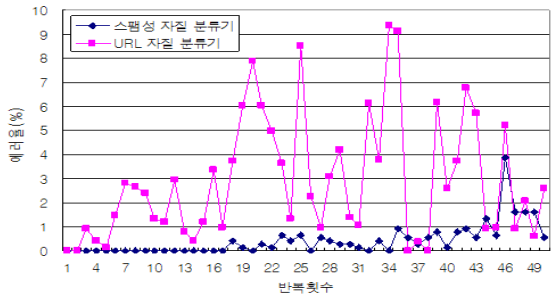
서를 점진적으로 추가한다. 공동 학습 결과 정확도는 일괄 학습의 98.2%에 도달하였으며 학습 문서의 수는 6,047에서 3,714개(labeled: 75개, unlabeled: 3,639개)로 줄일 수 있었다. 공동 학습 결과 그래프는 (그림 4)와 같다. batch는 일괄학습 결과를 나타내며 co-training은 공동 학습 결과를 나타낸다. 반복횟수가 증가함에 따라 정확도가 향상되며, 37번 반복했을 때 스팸성 자질 분류기와 통합 분류기에서 가장 높은 정확도를 보였다. <표 4>는 기본시스템, 일괄학습, 공동 학습 결과를 비교한 것이다. 공동 학습 과정에서 추가되는 문서의 클래스는 스팸성 자질 분류기와 URL 자질 분류기에 의해 결정된다. 문서의 클래스가 잘못 할당된 경우 시스템은 오히려 현저하게 저하될 수 있다. (그림 5)은 공동 학습 과정에서 스팸성 자질 분류기와 URL 자질 분류기에서 선택되는 문서에 대한 오류율 변화를 나타낸다. 두 분류기에서의 에러는 모두 10%이내였으며 스팸성 자질 분류기보다 URL 자질 분류기의 오류율이 높았다. 이는 추가되는 문서의 수를 75개로 고정함으로써 학습 과정에서 신뢰도가 높지 않은 문서가 추가되었을 가능성이 있을 것으로 본다.



(그림 4) 반복횟수에 따른 스팸성 자질과 URL 자질의 공동 학습 결과

〈표 4〉 일괄 학습과 공동 학습 결과 비교

평가방법(%)	기본 시스템	제안 시스템	
		일괄 학습	공동 학습
재현률	84.64	78.63	79.53
정확률	63.87	74.89	72.34
F-측정	0.728	0.767	0.758
Hm	15.36	21.37	20.47
Sm	35.73	19.68	22.70
Lam	24.11	20.51	21.56
정확도	72.97	79.60	78.24



(그림 5) 공동 학습에서 각 분류기의 오류율

4.3 결과 분석

실험을 통해 스팸성 자질과 URL 자질을 사용한 필터 시스템의 성능이 향상된 것을 확인하였다. 두 자질의 유효성을 확인하기 위해 시스템에 의해 할당된 문서의 클래스 변화를 살펴보고자 한다.

4.3.1 스팸성 자질의 유효성

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 있다.

4.3.2 URL 자질의 유효성

URL 자질을 적용한 결과 270개의 이메일 클래스가 변경되었으며, 이들 중 238개가 정확하게 분류되었다. 클래스가 변경된 문서들에 대해서 정확도는 약 88.15%를 보였다. 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.

메일에 나타난 URL 자질만을 이용한 실험(URL 자질 분류기)은 약 88.47%의 정확도를 보였다. 이를 통해 스팸메일에 포함되어 있는 URL이 스팸메일을 분류하는데 의미 있는 자질로 사용될 수 있음을 확인할 수 있다.

4.3.3 오류율 (Hm 과 Sm)

블랙리스트/화이트리스트 필터를 통해 사전에 걸러진 메일은 총 129개(스팸메일 105개, 정상메일 24)이며 오류율은 0%였다.

본 논문에서 제안하는 필터 시스템의 성능은 정확도 측면에서는 향상되었다. 그러나 정상메일을 스팸메일로 분류하는 오류율(Hm)은 높아지고, 스팸메일을 정상메일로 분류하는 오류율(Sm)은 낮아지는 결과를 보였다. 스팸메일 필터 시스템의 오류율은 Hm과 Sm으로 나누어지는데 Hm이 Sm보다 손실 비용이 더 크다. 스팸메일을 정상메일로 잘 못 분류할 경우 사람에 의해 2차적으로 걸러질 수 있지만 정상메일이 스팸메일로 분류될 경우 수신자에게 중요한 메일이 필터 시스템에 의해 자동적으로 삭제가 될 수 있기 때문이다. 본 실험에서는 스팸성 자질에 초점을 맞추고 있다. 시스템이 스팸메일을 중심으로 학습하기 때문에 상대적으로 정상메일을 분류하기 위한 정보가 부족해지는 것으로 생각된다. 비용 측면을 고려하면서 스팸메일을 잘 걸러내기 위해서는 정상메일을 스팸메일로 분류하는 오류를 줄일 필요가 있다.

5. 결론 및 향후 연구

본 논문에서는 스팸성 자질과 URL 자질의 공동 학습을 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 결과에서 제안 방법의 유효성을 확인할 수 있었다. 스팸성 자질과 URL 자질을 이용한 통합 시스템이 기본시스템에 비해 약 9.1%의 성능 향상을 보였으며, 보고필터와 비교하여 약 6.9%의 성능 향상을 보였다. 또한 두 자질 집합을 가지고 공동 학습을 함으로써 서로 다른 정보 즉, 스팸성 자질과 URL 자질을 공동 학습에 이용할 수 있음을 보였다.

본 논문에서는 각 분류기를 통과한 메일이 스팸메일에 속할 확률을 곱하여 분류기 통합을 시도하였다. 두 분류기를 좀 더 효율적으로 통합한다면 필터 시스템의 성능 향상에 도움을 줄 것으로 기대된다. 손실 비용 측면에서는 정상메일의 오류율을 낮출 수 있는 방법도 고려해야 한다. 또한 본 논문에서 제시한 스팸성 자질을 보완하고 일반화 시킬 수 있는 방법들이 필요하다. 스패머들은 필터 시스템을 통과하기 위해 끊임없는 변화를 시도할 수 있으며 이는 스팸성 자질이 무한대로 만들어질 수 있음을 뜻한다. 이러한 스팸성 자질을 자동으로 추출할 수 있는 방법에 관한 연구도 필요하다.

참 고 문 헌

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

[2] Cormack, B., Lynam, T. “TREC2005 Spam Track Overview”, Text REtrieval Conference, 2005.

[3] 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”, Text REtrieval Conference, 2005.

[4] Keselj, V., Milios, E., Tuttle, A., Wang, S., Zhang, R. “DalTREC 2005 Spam Track: Spam Filtering Using N-gram-based Techniques”, Text REtrieval Conference, 2005.

[5] Assis, F., Yerazunis, W., Siefkes, C., Chhabra, S. “CRM114 versus Mr. X: CRM114 Notes for the TREC 2005 Spam Track”, Text REtrieval Conference, 2005.

[6] Cao, W., An, A., Huang, X. “York University at TREC 2005: SPAM Track”, Text REtrieval Conference, 2005.

[7] 김현준, 정재은, 조근식 “가중치가 부여된 베이지안 분류자를 이용한 스팸 메일 필터링 시스템”, 정보과학회논문지, 제 31권, 제8호, pp 1092~1100, 2004.

[8] Segal, R. “IBM SpamGuru on the TREC 2005 Spam Track”, Text REtrieval Conference, 2005.

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

[10] Ion Androutsopoulos et al, “An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages”, International ACM SIGIR conference on Research and development in information retrieval, pp. 160-167, 2000.

[11] Breyer, L. A. “DBACL at the TREC 2005”, Text REtrieval Conference, TREC 2005.

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

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

[14] Blum, A. and Mitchell, T. M. “Combining labeled and unlabeled data with co-training”, Annual Conference on Computational Learning Theory, pp. 92－100, 1998.

[15] Kiritchenko, S. and Matwin, S. “Email classification with co-training”, Conference of the Centre for Advanced Studies on Collaborative Research, page 8, Toronto, Ontario, Canada, 2001.

[16] Pierce, D. and Cardie, C. “Limitations of Co-Training for natural language learning from large datasets”, Conference on Empirical Methods in NLP, pp. 1-9, 2001.

[17] 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)

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

공 미 경

e-mail : mggong@chonbuk.ac.kr

2005년 전북대학교 컴퓨터공학과(학사)

2007년 전북대학교 컴퓨터공학과 (공학석사)

관심분야 : 정보검색, 정보 마이닝

이 경 순

e-mail : selfsolee@chonbuk.ac.kr

1994년 계명대학교 컴퓨터공학과(학사)

1997년 한국과학기술원 전자전산학 (공학석사)

2001년 한국과학기술원 전자전산학 (공학박사)

2001년~2003년 일본 국립정보학연구소 (National Institute of Informatics) 연구원

2004년~현재 전북대학교 전자정보공학부/영상정보신기술연구센터 조교수

관심분야 : 정보검색, 정보 마이닝, 자연언어처리