하이퍼링크를 활용한 2단계 스팸 메일 필터링 시스템

강신재, 이새봄, 김종완
대구대학교 컴퓨터 · IT공학부

Two-phase Spam-mail Filtering System Applying Hyperlinks

Sin-Jae Kang, Sae-Bom Lee, Jong-Wan Kim
School of Computer and Information Technology, Daegu Univ.

요 약

본 논문은 하이퍼링크를 활용한 2 단계 스팸 메일 필터링에 관한 방법을 제시한다. 일반적으로 스팸 메일의 본문에는 텍스트 문장보다는 그림이 더 많이 포함되어 있기 때문에 단어의 블랙리스트와 같은 전형적인 방법으로 스팸 메일을 구분하기에는 많은 어려움이 따른다. 이러한 문제를 해결하기 위하여 본 논문에서는 스팸 메일에 포함되어 있는 하이퍼링크를 추출하여 해당 웹 페이지를 가져온 후, 이를 확장된 형태의 메일 본문이라 간주하여 텍스트 정보를 추출하였다. 또한 스팸 메일을 구분하기 위한 정보를 두 가지로 구분하여 사용하였는데, 메일 송신자의 정보와 확실한 스팸 키워드 리스트를 확실한 정보군으로 구분하여 먼저 적용하고, 이보다 덜 명확한 정보들은 따로 구분하여 속성벡터를 만들어 SVM 알고리즘을 적용하였다. 실험결과 하이퍼링크를 통하여 웹 페이지를 가져온 방법이 그냥 원본 메일만 사용한 방법보다 F-measure 값이 평균 2.8%의 성능향상을 보였다.

1. 서론

인터넷의 대중화와 적은 비용으로 메시지를 빠르게 전달할 수 있는 편리성 때문에 오늘날 전자우편은 사용자간 의사소통을 하기에 없어서는 안 될 필수적인 통신수단이 되었다. 전자우편은 사용자에게 많은 편리성을 준 반면 매일 많은 양의 스팸 메일을 처리해야 하는 불편함도 주고 있다. 스팸 메일, 즉 원치 않는 상업성 메일의 폐해로는 각 개인의 메일박스가 매일 아침 원치 않는 메일들로 가득 차게 되고, 미성년자에게는 전달되지 않아야 할 부적절한 내용이 전달되며, 또한 네트워크에 부하를 주는 것 등을 생각해 볼 수 있겠다[1]. 대부분의 전자우편 클라이언트 소프트웨어는 송신자 블랙리스트나 키워드 기반의 필터 형태로 스팸 메일을 제거하고 있다. 하지만 이러한 리스트나 필터의 구축은 대부분 수작업으로 이루어지기 때문에 구축비용 및 시간이 많이 필요하며, 또한 실제 상황에서 모든 스팸 메일을 완벽하게 처리할 수는 없다는 문제가 있다. 기본적으로 스팸 메일 필터링 문제는 문서분류의 특별한 한 형태로 볼 수 있기 때문에 여러 다양한 정보검색 기법과 기계학습 알고리즘들이 이 문제의 해결을 위해 사용되어져 왔다[2-10]. Sahami[2]는 나이브 베이지안 분류기(Naive Bayesian classifier)를 스팸 메일 필터링에 사용하였는데, 수작업으로 구축된 구(phrase) 정보와 송신자의 도메인 타입, 제목에서 기호 문자의 비율 등 다양한 비 텍스트(non-textual) 정보를 도메인 속성으로 정의하여 사용하였다. Carreras[3]는

벤치마크용으로 구축된 PU1 전자우편 말뭉치에 대해 세 가지 기계학습 알고리즘을 적용시켜 보았는데, 의사결정 트리(Decision Tree)나 나이브 베이지안보다 부스팅 알고리즘이 보다 나은 성능을 나타냄을 보였다. Yang[4]에서는 텍스트 정보와 송신자 이름, 송신자 소속 등과 같은 메타 데이터를 이용하여 스팸 메일을 구분하고자 하였는데, TFIDF 보다 나이브 베이지안과 SVM(Support Vector Machines)이 훨씬 좋은 결과를 보임을 실험을 통해 입증하였다. 특히 메일의 헤더에서 추출한 속성을 SVM 에 적용하였을 때 가장 좋은 결과를 보였다. 스팸 메일 필터링이나 메일의 자동분류에 관한 최근의 연구들을 대체적으로 살펴보면 TFIDF 나 나이브 베이지안, 의사결정 트리와 같은 기존의 분류 알고리즘보다 Vapnik[11]가 고안한 SVM 이 보다 나은 성능을 보이고 있음을 알 수 있다[4, 5, 6]. 이는 SVM 이 스팸 메일 필터링과 같은 이진 분류 문제(two-class problem)에 적합하기 때문이라고 볼 수 있다.

본 논문에서는 스팸 메일의 특성상 메일의 본문에서 추출할 수 있는 텍스트 정보가 한정되어 있는 문제를 해결하기 위하여, 거의 모든 스팸 메일에 포함되어 있는 하이퍼링크를 활용한다. 메일에서 추출된 하이퍼링크를 따라가서 해당 웹 페이지를 가져오게 되는데, 이것에는 스팸 메일인지 여부를 가릴 수 있는 힌트를 포함하고 있을 확률이 높기 때문에 본 시스템의 성능을 높이는데 큰 도움을 주게 된다. 또한, 본 연구에서는 스팸 메일을 구분하기 위한 정보를 두 가지로 구분하여 사용하였는데, 메일 송신자의 전자우편 주소와 URL 과 같은 정보와 확실한 스팸 키워드 리스트를 *확실한 정보군*(definite information)으로 구분하여 필터링 작성 시 먼저 적용하게 된다. 하지만 송신자의 정보는 위조되거나 누락될 수도 있으며, 모든 스팸 메일을 구분할 수 있는 확실한 정보를 구축하기에는 어려운 문제가 있다. 그래서 텍스트 정보와 같이 이보다 *덜 명확한 정보들*(less definite information)을 따로 구분하여 속성벡터를 만든 후, SVM 알고리즘의 학습을 통하여 필터링에 적용하였다.

2. SVM

Vapnik [11]가 고안한 SVM 은 기본적으로 두 범주를 갖는 객체들을 분류하는 방법이다. 기존 분류기(classifier) 대부분이 경험적 위험(empirical risk)을 최소화하는 개념에 기반한 반면, SVM 은 일반화 에러의 상한(upper bound)을 최소화하는, 구조적 위험 최소화(structural risk minimization)라는 원리에서 동작한다. 다른 학습 알고리즘과는 달리 SVM 에서 사용되는 파라미터의 수는 데이터를 구분하는 마진(margin)에 의존하며, 입력 속성의 수에는 영향을 받지 않는다. 그러므로 오버피팅(over-fitting) 문제를 피하기 위해 속성의 수를 줄이는 과정은 필요치 않다. 이러한 특징은 문서 분류와 같이 고차원의 특성을 가지는 응용분야에서 큰 장점이 될 수 있다 [12].

SVM 은 비선형 패턴 인식 문제, 함수 회귀 문제, HCI(Human-Computer Interaction), 데이타마이닝, Web Mining, 컴퓨터 비전, 인공지능, 의학진단 등의 분야에서 다양하게 활용될 것으로 보이며, 최근 매우 활발하게 연구가 진행되고 있다.

본 연구에서는 Witten[13]이 개발한 WEKA (Waikato Environment for Knowledge Analysis) 패키지에 포함된 SVM 분류기를 이용하여 실험하였다 WEKA 는 실제 응용 프로그램에서 기계학습 알고리즘의 구현을 돕기 위해 만들어진 도구이다.

3. 학습단계

스팸 메일을 효과적으로 가려내기 위하여 본 연구에서는 힌트(속성 또는 특징)를 *확실한 정보*(definite information)와 *덜 확실한 정보*(less definite information)의 두 가지로 구분하였다. 학습단계에서의 전체적인 처리 과정은 그림 1 에 제시되어 있다.

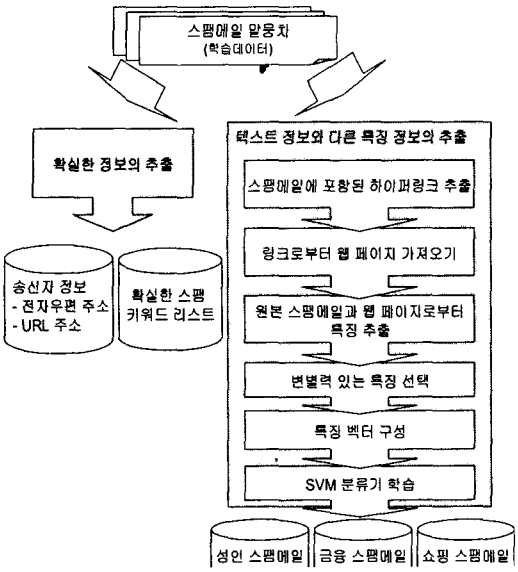

그림 1. 스팸메일 필터링을 위한 학습과정

3.1 확실한 정보

스팸 메일을 걸러 내기 위한 확실한 정보란 송신자의 전자우편 주소, URL 주소와 같은 송신자 정보와 "광고", "포르노", "신용대출", "무보증카드대출" 등과 같은 확실한 스팸 키워드 리스트를 말한다. 만약 새로 도착한 메일의 정보가 송신자의 전자우편 주소나 URL 주소 정보 중 하나와 일치한다면 해당 메일은 스팸 메일일 확률이 매우 높기 때문에 다른 처리 과정 없이 바로 스팸 메일로 분류하게 된다. 또한 메일의 제목에 스팸 키워드 리스트에 등록된 단어나 구가 있는 경우도 스팸 메일로 바로 분류하게 된다. 하지만 메일의 본문에서는 등록된 스팸 키워드가 3 회 이상 나타나는 경우에만 스팸 메일로 분류하게 된다. 확실한 정보는 스팸 메일 말뭉치로부터 수작업으로 추출되었는데, 송신자의 전자우편 주소 1,183 개, URL 주소 488 개, 스팸 키워드 리스트 275 개가 추출되었다. URL 주소는 시간에 따라 변동이 매우 심하기 때문에 주기적으로 유효하지 않은 주소를 확인하여 제거하는 과정이 필요하다.

3.2 덜 확실한 정보

스팸 메일인지 여부를 판단할 수 있는 정보로는 위에서 언급한 확실한 정보 외에도 많이 존재한다. 확실한 스팸 키워드 리스트에는 포함되지 않았지만, 스팸 메일에 종종 나타나는 단어나 구와 같은 텍스트 정보가 그러하다. 이러한 정보들을 추출하여 유사도를 측정하기 위해서 정보검색에서 많이 사용하는 벡터모델을 사용하였다 [14]. 위에서 덜 확실한 정보로 선택된 단어나 구들이 하나의 속성으로 벡터를 구성하게 된다. 보통의 스팸 메일에는 이미지, 그림과 같은 비 텍스트(non-text) 정보만 있는 경우가 허다하다. 이러한 경우에는 텍스트 정보를 추출할 수 없기 때문에 단순한 방법으로는 스팸 메일의 분류가 어렵게 된다. 하지만 대부분의 스팸 메일들은 고객을 유인하여 영업을 하기 위해 웹 사이트의 하이퍼링크가 포함되어 있으므로 이를 이용하면 스팸 메일 분류에 도움을 얻을 수 있다. 확실한 정보에서 사용하였던 URL 의 유무 여부만 가지고 스팸 메일을 분류하기 보다는 해당 웹 페이지를 패치(fetch)한 후 스팸 메일인지 여부를 판단한다면 보다 정확한 판단을 할 수 있을 것이다. 패치한 웹 페이지는 스팸 메일의 일부분으로 간주되어 분석되게 되는데, 텍스트 정보를 추출하기 위해서 형태소 분석을 한 후, 불용어(stopword)를 제거하는 과정을 거쳐서 후보 속성(단어)들을 얻는다. 이렇게 추출된 속성 벡터들을 SVM 분류기에 학습시키게 된다. SVM 의 성능은 입력 속성의 차원에 큰 영향을 받지 않지만 실행시간을 줄이고 시스템의 부하를 줄이기 위해 추출된 후보 속성들 가운데 변별력이 높은 속성들을 수작업으로 선택하였다.

속성값은 해당 단어나 구의 존재 유무에 따라 이진값을 갖게 된다. 왜냐하면 문서 분류에서 SVM 은 속성의 값을 이진값으로 사용할 때, 최상의 성능을 보임이 증명되었기 때문이다 [5].

본 연구에서 구축한 시스템은 새로 도착한 메일이 스팸 메일인지 아닌지를 최종적으로 가리게 되지만, SVM 분류기를 하나만 두어 이를 구분케 하는 것은 다양한 종류의 스팸 메일의 특성을 잘 활용하지

못하는 결과를 초래하게 된다. 그래서 본 시스템에서는 스팸 메일을 성인, 금융, 쇼핑의 대표적인 세 가지 범주로 구분하였으며, SVM 분류기를 범주별로 각각 구축하여 사용하였다.

4. 2단계 스팸메일 필터링

새로 도착한 메일은 3 장에서 구축된 정보와 SVM 분류기를 통하여 처리하게 된다. 만약 신규 도착 메일이 확실한 정보에 속하는 내용을 어느 정도 가지고 있으면 그 메일은 스팸 메일로 판단하게 된다. 하지만 해당되는 내용이 없다면 SVM 분류기를 적용하는 단계로 넘어가게 된다. 스팸 메일의 범주에 따라 세 가지의 SVM 분류기가 구축되어 있으므로 순차적으로 SVM 분류기를 적용한다. 성인 스팸 메일 분류를 위한 SVM 을 먼저 적용한 결과가 성인 스팸 메일이 아니라고 판단되면, 금융 스팸 메일을 위한 SVM 분류기를 적용하고 여기에서도 아니라고 판단되면 최종적으로 쇼핑 스팸 메일용 SVM 분류기를 적용하게 된다. 이 단계에서도 아니라고 판단되면 해당 메일은 스팸 메일이 아닌 일반 메일이라고 결론지을 수 있다. 전체적인 스팸 메일 필터링 과정은 그림 2 에 나타나 있다.

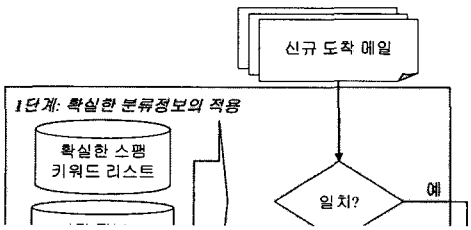

그림 2. 2 단계 스팸메일 필터링 과정

5. 실험

아직까지 한국어를 대상으로 한 전자우편 말뭉치가 공개된 적이 없기 때문에, 본 실험에서는 수작업으로 메일을 수집하여 사용하였다. 학습에 사용된 스팸 메일 말뭉치는 성인 스팸 메일 1,100 개 금융 스팸 메일 1,077 개, 쇼핑 스팸 메일 397 개로 총 2,574 개이다.

스팸 메일 필터링 시스템의 성능평가를 위해서는 정보검색에서 일반적으로 많이 사용하고 있는 재현율(recall)과 정확률(precision), F-measure 를 사용하였다. 시스템이 분류한 스팸 메일의 수와 실제 스팸 메일의 수에 관한 분할표(contingency table)가 표 1 과 같이 정의되었을 때, 재현율, 정확률, F-measure 는 각각 다음과 같이 정의할 수 있다.

<표 1> 시스템에 의해 분류된 스팸 메일의 수와 실제 스팸 메일의 수에 관한 분할표

	시스템이 분류한 스팸 메일	시스템이 분류한 일반 메일
실제 스팸메일	a	c
실제 일반 메일	b	d

$$\text{재현율(R)} = \frac{\text{시스템에 의해 맞게 분류된 스팸메일의 수}}{\text{실제 스팸메일의 총 수}} = \frac{a}{a+c}$$

$$\text{정확률(P)} = \frac{\text{시스템에 의해 맞게 분류된 스팸메일의 수}}{\text{시스템에 의해 분류된 스팸메일의 총 수}} = \frac{a}{a+b}$$

$$\text{F-measure} = \frac{(\beta^2+1)*\text{정확률}*\text{재현율}}{\beta^2*\text{정확률}+\text{재현율}} = \frac{2PR}{P+R}$$

F-measure 에서 β는 정확률에 대한 재현율의 가중치를 의미하는데, 본 실험에서는 동일한 가중치를 부여하기 위해 β에 1 을 사용하였다.

객관적인 성능평가를 위하여 10 층 교차 확인법(10-fold cross validation)을 사용하였다. 이는 전체 전자우편 말뭉치를 균등하게 10 등분한 다음, 9 개는 학습에 사용하고 나머지 한 개는 성능 테스트를 위해 사용하는 방법으로, 각 등분들이 한 번씩 테스트 용도로 사용되도록 10 번 반복 실험을 한 후, 그 결과들을 평균 내는 방법이다. 선택된 속성을 가지고 스팸 메일 필터링 시스템 2 단계에서 실험한 결과가 아래의 표 2에 제시되어 있다.

<표 2> 스팸 메일 필터링 실험결과(%)

패치 여부	SVM 분류기	속성수	R	P	F
원본 메일만 분석한 경우	성인	513	96.9	39.8	56.4
	금융	468	99.6	46.1	63.1
	쇼핑	286	91.9	45.8	61.2
원본 메일 + 패치한 웹페이지	성인	513	94.3	47.0	62.7
	금융	468	97.2	48.8	65.0
	쇼핑	286	91.5	46.3	61.5

실험결과를 보면 원본 메일만 분석한 방법보다 원본 메일과 웹 페이지를 패치하여 모두 분석한 방법이 재현율이 훨씬 좋아지는 것을 볼 수 있으며, 반대로 정확률은 다소 떨어진 것을 볼 수 있다. 이는 하이퍼링크를 통해 수집된 정보들이 100% 확실한 정보는 아니기 때문에 정확률 측면에서는 다소 부정적인 영향을 주지만, 보다 많은 스팸 메일을 구분해 낼 수 있다는 측면에서는 긍정적인 영향을 준 것으로 분석해 볼 수 있다.

하이퍼링크를 활용한 시스템의 전체적인 성능은 사용하지 않은 시스템보다 F-measure 값이 평균 2.8% 정도 향상됨을 확인할 수 있다.

실험에서 1 단계의 결과가 제시되지 못한 이유는 스팸 메일이 아닌 일반 메일 말뭉치를 구하기가 어려웠기 때문이다. 본 실험에서 사용한 스팸 메일 말뭉치로는 1 단계에서의 성능이 100%이다. 왜냐하면 1 단계의 정보는 모든 학습말뭉치를 대상으로 하여 추출한 정보이기 때문이다. 1 단계의 객관적인 성능평가를 위해서는 상당량의 일반 메일이 필요하다. 하지만 일반 메일은 개인의 사생활이 노출되는 문제가 있기 때문에 쉽사리 구하기가 어렵다. 이의 해결을 위해선 한국어를 대상으로 PU123A 말뭉치[15]와 같이 단어들을 숫자로 바꾼 형태로 메일들을 암호화하여 전자우편 말뭉치를 구축하는 방법이 필요할 것으로 보인다.

6. 결론

스팸 메일 필터링은 각 개인의 전자우편 관리에 대한 부담을 덜어주고, 미성년자에게는 유해한 메일을 차단해 주는 등 일상생활에 직접적인 영향을 줄 수 있는 매우 중요한 연구이다. 본 연구에서는 하이퍼링크를 활용한 2 단계 스팸 메일 필터링에 관한 방법을 제시하였다.

스팸 메일을 구분하기 위한 정보들을 두 가지로 구분하여 단계별로 적용하였는데, 몇 가지 특징만으로도 확실하게 스팸 메일을 구분할 수 있는 정보들을 1 단계에서 먼저 적용하였다. 그 외의 정보들은 속성 벡터의 형태로 추출되어 2 단계에서 SVM 분류기를 통해 적용되었다. 이는 스팸 메일의 판단을 위해 모든 메일의 특징 정보를 SVM 분류기에 적용시키는 일괄적인 방법에 비해 시스템의 부하 및 실행시간을 줄이는 효과를 가져올 수 있다.

또한, 대부분의 스팸 메일은 텍스트 정보보다는 그림 등 이미지 정보가 많이 포함되어 있는 특징이 있기 때문에 단순히 단어의 유무를 통한 필터링 방법을 적용하기에는 어려운 문제가 있다. 이의 해결을 위해 거의 모든 스팸 메일에 포함되어 있는 하이퍼링크를 활용하였다. 메일에서 추출된 하이퍼링크를 따라가서 해당 웹 페이지를 가져온 후 속성의 추출에 사용하게 된다. 웹 페이지에는 스팸 메일인지 여부를 가릴 수 있는 속성들이 보다 많이 포함되어 있기 때문이다.

본 연구에서 구현한 스팸 메일 필터링 시스템의 2 단계는 수작업 없이 자동적으로 정보가 수정될 수 있기 때문에, 갈수록 지능적인 방법으로 보내어지는

스팸 메일에 효과적으로 대처하기 위한 방안이 될 수 있겠다.

향후에는 사용자의 피드백을 기억하고 이를 학습단계에서 재이용하여 1 단계 정보도 자동적으로 수정하는 방법과, 각 개인별로 차별화된 필터링 기능을 제공해 주는 방법에 대해 연구할 계획이다.

참 고 문 헌

[1] L. F. Cranor and B. A. LaMacchia, "Spam!," Communications of ACM, Vol.41, No.8, pp.74-83, 1998.

[2] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk e-mail," In AAAI-98 Workshop on Learning for Text Categorization, pp.55-62, 1998.

[3] X. Carreras and L. Marquez, "Boosting trees for anti-spam email filtering,", In Proceedings of ANLP-01, 4th International Conference on Recent Advances in Natural Language Processing, Tzigov Chark, BG, 2001.

[4] J. Yang, V. Chalasani, and S. Park, "Intelligent email categorization based on textual information and metadata," IEICE Transactions on Information and System, Vol.E86-D, No.7, pp.1280-1288, 2003.

[5] H. Drucker, D. Wu, and V. Vapnik, "Support Vector Machines for Spam Categorization," IEEE Trans. on Neural Networks, 10(5), pp.1048-1054, 1999.

[6] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," European Conference on Machine Learning (ECML), Claire Ndellec and Cline Rouveirol (ed.), 1998.

[7] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras, and D. Spyropoulos, "An evaluation of naive Bayesian anti-spam filtering," In Proc. of the workshop on Machine Learning in theNew Information Age. 11th European Conference on Machine Learning. pp.9-17 2000.

[8] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, and D. Spyropoulos, "An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages,"In Proc. of the 23 rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 160-167, Athens, Greece, 2000.

[9] C. Apte, F. Damerau, and S. M. Weiss, "Text Mining with Decision Trees and Decision Rules," in Conference on Automated Learning and Discovery, Carnegie-Mellon University, June 1998

[10] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," AAAI-98 Workshop on Learning for Text Categorization, 1998.

[11] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995

[12] O. de Vel, "Mining E-mail Authorship," Proc Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD'2000), Boston, 2000.

[13] I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and Techniques with java implementations, Morgan Kaufmann, 2000

[14] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," Communication of the ACM, 18(11), pp.613-620, 1975.

[15] http://www.iit.demokritos.gr/skel/i-config/