한국어와 영어 스팸메일의 필터링 성능 분석

황운호, 강신재, 김태희, 김희재, 김종완
대구대학교 컴퓨터·IT공학부

Analysis of filtering performance of Korean and English spam-mails

Wun-Ho Hwang, Sin-Jae Kang, Tae-Hee Kim, Hee-Jae Kim, Jong-Wan Kim
School of Computer and Information Technology, Daegu Univ.

요 약

본 연구에서는 한국어와 영어 메일을 대상으로 2단계 스팸 메일 필터링 시스템을 구축하여 성능평가를 수행한다. 2단계 스팸 메일 필터링 시스템은 블랙리스트를 활용하는 1단계와 기계학습을 통한 지능적인 분류를 하는 2단계로 구성된다. 만약 새로 도착한 메일이 블랙리스트의 내용을 포함한다면 이 메일은 스팸 메일로 분류되고 그렇지 않은 메일은 2단계로 넘어가서 스팸 메일 여부를 판단하게 된다. 메일의 본문이 영어로 작성된 영어 스팸 메일을 일반 메일로부터 분류해내기 위해서는 우선 Stemming과 Stopping 기법을 이용하여 본문에서 정형화된 어휘정보들을 추출한다. 추출된 어휘정보들을 대상으로 속성벡터를 구축한 후 SVM 기계 학습을 시켜 SVM 분류기를 생성하여 지능적인 스팸 메일 필터링을 수행한다. 속성벡터를 구축할 때 기준이 되는 자질을 어떻게 선택하느냐에 따라 스팸 메일 필터링 시스템의 성능이 좌우된다. 따라서 SVM 기계 학습을 위한 속성벡터를 구축할 때 기준이 되는 자질을 선택하는 여러 알고리즘들을 적용하여 성능을 비교 분석한다. 그리고 한국어 스팸 메일 필터링 시스템과 비교하여 영어 스팸 메일 필터링 시스템의 전체적인 성능을 비교 분석한다.

1. 서 론

전자우편은 사용자에게 많은 편리성을 준 반면 매일 많은 양의 스팸 메일을 처리해야 하는 불편함도 주고 있다. 스팸 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].

Yang[4]에서는 텍스트 정보와 송신자 이름, 송신자 소속 등과 같은 메타 데이터를 이용하여 스팸 메일을 구분하고자 하였는데, TFIDF보다 나이브 베이지안과 SVM (Support Vector Machines)이 훨씬 좋은 결과를 보임을 실험을 통해 입증하였다. 특히 메일의 헤더에서 추출한 속성을 SVM에 적용하였을 때 가장 좋은 결과를 보였다. 스팸 메일 필터링이나 메일의 자동분류에 관한 최근의 연구들을 대체적으로 살펴보면 TFIDF나 나이브 베이지안, 의사결정 트리와 같은 기존의 분류 알고리즘보다 Vapnik[11]가 고안한 SVM이 보다 나은 성능을 보이고 있음을 알 수 있다[4, 5, 6]. 이는 SVM이 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 확률

이 높기 때문에 본 시스템의 성능을 높이는데 큰 도움을 주게 된다.

추출된 웹 페이지와 메일 본문을 합쳐서 한국어 메일의 경우 형태소 분석과 불용어 제거 과정을 실행하고 영어의 경우 Stemming과 Stopping을 통하여 정제된 단어들을 추출한 후 이를 이용하여 속성벡터를 구성한다. 스팸 메일 필터링을 위한 SVM 알고리즘의 학습에 적용시키기 위한 속성벡터를 어떻게 구성하느냐에 따라 시스템의 성능에 큰 영향을 끼친다. 이 속성벡터를 구성하기 위해서는 우선 기준이 되는 자질을 어떻게 선택하는지가 관건이다. 이를 위해서 본 연구에서는 한국어와 영어를 각각 따로 실험을 진행하며 자질 선정을 위하여 여러 특징 추출 알고리즘을 활용하여 변별력이 높은 순으로 자질을 추출한다. 추출된 자질들을 기준으로 여러 개수로 나누어 실험 한 후 성능을 평가하여 최적의 특징 추출 알고리즘과 자질 개수를 구한다. 이렇게 실험한 결과를 바탕으로 시스템을 구축한다. 그리고 본 시스템은 스팸 메일을 구분하기 위한 정보를 두 가지로 구분하여 사용하였는데, 메일 송신자의 전자우편 주소와 URL과 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 앞에서 추출한 자질을 기준으로 속성벡터를 만든 후, SVM 알고리즘의 학습을 통하여 필터링에 적용한다. 2장에서는 SVM에 대해서 알아보고 3장에서 본 시스템의 전반적인 흐름을 설명한다. 4장에서는 실험을 통하여 한국어 시스템과 영어 시스템의 성능을 비교분석한 후 5장에서 결론을 맺는다.

2. SVM (Support Vector Machine)

Vapnik [11]가 고안한 SVM은 기본적으로 두 범주를 갖는 객체들을 분류하는 방법이다. 기존 분류기(classifier) 대부분이 경험적 위험(empirical risk)을 최소화하는 개념에 기반한 반면, SVM은 일반화 에러의 상한(upper bound)을 최소화하는, 구조적 위험 최소화(structural risk minimization)라는 원리에서 동작한다. 다른 학습 알고리즘과는 달리 SVM에서 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 문서분류와 같이 고차원의 특성을 가지는 응용분야에서 큰 장점이 될 수 있다 [12].

SVM은 비선형 패턴 인식 문제, 함수 회귀 문제, HCI(Human-Computer Interaction), 데이타마이닝, Web Mining, 컴퓨터 비전, 인공지능, 의학 진단 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. WEKA는 실제 응용 프로그램에서 기계학습 알고리즘의 구현을 돕기 위해 만들어진 도구이다.

3. 2단계 스팸 메일 필터링

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 처리과정은 그림 1에 제시되어 있다.

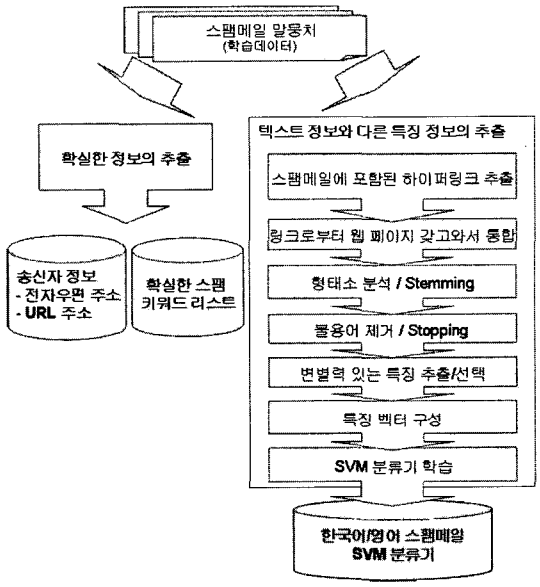

그림 1. 스팸 메일 필터링을 위한 학습과정

1단계 필터링에서는 학습 시킬 스팸메일들에서 송신자 메일주소, 메일 본문에 포함된 URL, 그리

고 제목 및 본문에 포함된 단어들을 스팸키워드로 인식하고 SPAM DB를 구축하였다. 단, 본문에 포함된 단어는 3번 이상의 빈도를 가질 때 포함시켰다. 만약 새로 도착한 메일의 정보가 송신자의 전자우편 주소나 URL 주소 정보, 제목 및 본문에 포함된 단어 중 하나와 일치한다면 해당 메일은 스팸 메일일 확률이 매우 높기 때문에 다른 처리 과정 없이 바로 스팸 메일로 분류하게 된다. 확실한 정보는 스팸 메일 말뭉치로부터 수작업으로 추출되었는데, 자세한 내용은 표 1과 같다.

표 1. 스팸 DB의 데이터 구성

메일 분류 / 데이터	한국어 스팸 메일	영어 스팸 메일
E-mail	1,461	1,259
URL	1,023	1,394
Keyword	603	5,994

URL 주소는 시간에 따라 변동이 매우 심하기 때문에 주기적으로 유효하지 않은 주소를 확인하여 제거하는 과정이 필요하다.

1단계 필터링에서 메일이 확실한 키워드를 가지면 스팸 메일로 분류되게 된다. 하지만 확실한 키워드를 가진다고 판단되지 않은 메일들은 2단계 SVM 분류기 적용단계로 넘어온다. 전체적인 스팸 메일 필터링 과정은 그림 2에 나타나 있다.

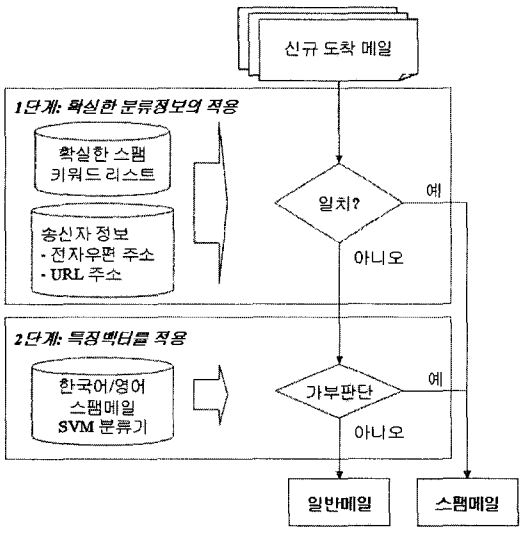

그림 2. 2단계 스팸 메일 필터링 과정

4. 실험

아직까지 한국어를 대상으로 한 전자우편 말뭉치가 공개된 적이 없기 때문에, 본 실험에서는 수작업으로 메일을 수집하여 사용하였으며 영어의 실험 말뭉치는 TREC(Text REtreval Confrence) Spam Track[15]에서 제공하는 spamassasin 말뭉치를 이용하여 실험하였다.

스팸 메일 필터링 시스템의 성능평가를 위해서는 TREC에서 평가지표로 제시한 hm (ham misclassification)과 sm(spam misclassification) 그리고 이 둘을 산술평균한 값을 이용하여 평가하였다. hm은 일반 메일을 스팸 메일로 잘못 분류한 비율이고 sm은 스팸 메일을 일반 메일로 잘못 분류한 비율이다. 시스템이 분류한 스팸 메일의 수와 실제 스팸 메일의 수에 관한 분할표(contingency table)가 표 2와 같이 정의되었을 때, hm과 sm은 각각 다음과 같이 정의할 수 있다.

표 2. 시스템에 의해 분류된 스팸 메일의 수와 실제 스팸 메일의 수에 관한 분할표

실제 / 시스템	ham	spam
ham	a	b
spam	c	d

$$\text{hm\% (ham misclassification)} = \frac{\text{시스템이 잘못 분류한 일반메일의 수}}{\text{실제 일반메일의 총 수}} = \frac{c}{a+c}$$

$$\text{sm\% (spam misclassification)} = \frac{\text{시스템이 잘못 분류한 스팸메일의 수}}{\text{실제 스팸메일의 총 수}} = \frac{b}{b+d}$$

객관적인 성능평가를 위하여 10층 교차 확인법(10-fold cross validation)을 사용하였다. 이는 전체 전자우편 말뭉치를 균등하게 10등분한 다음, 9개는 학습에 사용하고 나머지 한 개는 성능 테스트를 위해 사용하는 방법으로, 각 등분들이 한 번씩 테스트 용도로 사용되도록 10번 반복 실험을 한 후, 그 결과들을 평균 내는 방법이다.

1단계 필터링에서 확실한 스팸 메일로 판단되는 단계를 거친 후 확실한 스팸 메일로 판단되지 않은 메일을 2단계 SVM 분류기에 적용시키게 된다. SVM 학습을 위해서는 어떤 단어나 구의 존재 유무를 가리기 위한 기준이 되는 자질들이 필

요하고 그 자질들에 대해서 각 메일들 내에 포함된 단어나 구의 존재 유무를 나타내는 속성 값이 필요하다. 본 논문에서는 먼저 스팸 메일을 가리는 기준이 되는 자질을 선정하는 여러 알고리즘들을 한국어 스팸 메일과 영어 스팸 메일에 적용하여 최고의 성능을 내는 알고리즘과 자질의 수를 결정한다. 그런 다음 실제 한국어와 영어 스팸 메일 필터링 시스템의 성능을 비교하여 분석해 보도록 한다.

4.1 자질 선정

적용시킬 SVM 분류기를 만드는 과정은 그림 3에서와 같이 먼저 학습시킬 확실한 스팸 메일과 하이퍼링크를 따라가서 추출한 웹 페이지의 형태소를 분석하여 각 메일들에 대한 단어들을 얻은 후 불용어(stopword)를 제거하는 과정을 거쳐서 완벽한 단어들을 얻는다. 이 단어들이 SVM 학습에 사용된 후보 자질들이다. 한국어 메일의 형태소 분석은 포항공과대학교(POSTECH)에서 개발한 KoMA를 이용하였으며 영어 메일의 Stemming은 Porter Stemming Algorithm [16]을 활용하였다. 그런 다음 이 후보자질들을 정보획득량(Information Gain), 카이제곱(Chi Square) 그리고 상호정보(Mutual Information)등과 같은 여러 자질 추출 알고리즘에 적용시켜 변별력 있는 단어들 순으로 내림차순 정렬하여 SVM 학습에 사용할 자질들을 추출하였다. 자질들의 수에 따른 hm과 sm을 비교 실험한 결과는 표 3과 표 4에 나타나 있다.

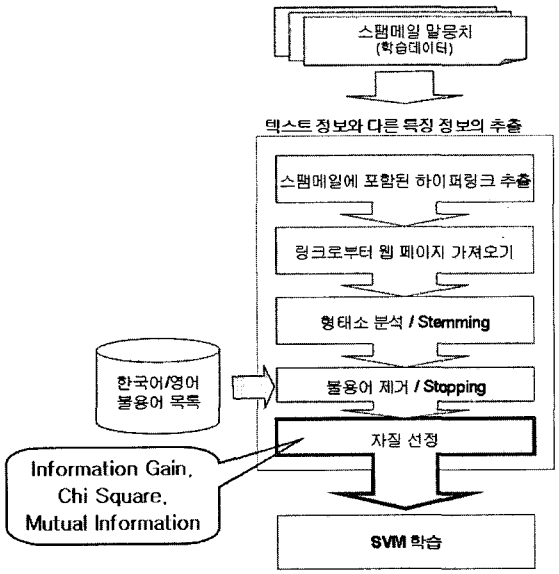

그림 3. SVM 학습을 위한 과정

표 3. 한국어 Feature 수에 따른 hm과 sm

한국어(%)	개수	hm	sm	(hm+sm) /2
정보획득량	200	10.9	1.2	11.5
	400	9.4	1.4	10.2
	600	7.7	1.1	8.2
	800	7.0	0.9	7.4
	1000	6.9	0.9	7.3
	2000	6.4	1.0	6.9
	3000	6.6	0.8	6.9
카이제곱	200	8.8	0.9	9.3
	400	7.4	1.0	7.9
	600	7.0	0.8	7.3
	800	6.4	0.8	6.9
	1000	6.7	0.9	7.2
	2000	6.5	0.9	7.0
	3000	6.2	0.9	6.7
상호정보	200	0.1	20.8	10.5
	400	0.0	20.0	10.0
	600	0.0	20.0	10.0
	800	0.0	19.7	9.8
	1000	0.0	19.5	9.8
	2000	0.1	17.4	8.8
	3000	0.1	17.3	8.7

표 4. 영어 Feature 수에 따른 hm과 sm

영어(%)	개수	hm	sm	(hm+sm) /2
정보획득량	200	2.9	4.9	5.4
	400	2.3	4.4	4.6
	600	2.4	4.1	4.5
	800	2.2	3.5	3.9
	1000	1.9	3.3	3.5
	2000	1.6	3.4	3.3
	3000	1.8	3.0	3.2
카이제곱	200	2.6	4.8	5.0
	400	2.3	4.3	4.4
	600	1.9	3.4	3.6
	800	1.9	3.1	3.4
	1000	2.0	3.7	3.9
	2000	1.6	3.3	3.3
	3000	1.5	2.8	2.9
상호정보	200	0.0	60.7	30.3
	400	0.0	53.6	26.8
	600	0.0	50.6	25.3
	800	0.0	48.8	24.4
	1000	0.0	44.2	22.1
	2000	3.2	24.9	15.6
	3000	3.3	13.9	10.3

실험결과에서 알 수 있듯이 한국어 시스템과 영어 시스템 모두에서 카이제곱으로 3000개의 자질을 선택하였을 때 SVM의 성능이 제일 좋게 나타나는 것을 확인할 수 있다. 이러한 결과가 나타나는 이유는 카이제곱 알고리즘으로 추출된 자질들이 스팸 메일과 일반 메일을 구분할 수 있는

변별력이 높기 때문이다. 따라서 본 시스템은 한국어메일와 영어 메일에서 후보 자질들 중에 가운데 변별력이 높은 자질 3,000개를 카이제곱을 이용해 추출하여 나머지 실험을 진행하였다.

4.2 SVM 학습

실험 데이터는 전체 100%의 메일 데이터 중 90%를 이용하여 학습을 한 후 10%의 메일로 테스트 하는 방식으로 진행한다.

자질 선정 실험에서 선정된 자질 3000개를 기준으로 한국어 시스템과 영어 시스템에서 각각 실험데이터의 90%를 가지고 SVM 특징벡터를 구성하여 학습을 수행한다. 각 메일의 수는 표 5와 표 6에 나타나 있다.

표 5. 한국어 실험 데이터

	한국어	90%	10%	합계	
SPAM	성인	1092	422	1214	3281
	금융	1355	151	1506	
	쇼핑	504	57	561	
HAM	일반	1563	174	1737	1737
	합계	4514	504	3928	3928

표 6. 영어 실험 데이터

	영어	90%	10%	합계	
SPAM	Spam	450	50	500	1897
	Spam_2	1257	140	1397	
HAM	Esay_ham	2250	250	2500	4150
	Hard_ham	225	25	250	
	Easy_ham2	1260	140	1260	
	합계	5442	605	6047	6047

한국어 메일과 영어 메일 90%로 학습된 한국어 SVM 분류기와 영어 SVM 분류기의 판단을 위한 임계값은 그림 4, 그림 6과 같은 학습결과를 분석하여 결정하였다. 학습결과를 정렬하여 표현하면 그림 5, 그림 7과 같다. 이 분포에서 나타난 값을 기준으로 스팸 메일과 일반 메일을 구분하게 되는 것이며 이 둘을 구분하기 위한 최적의 기준을 선택하여야 한다. 정렬을 하여 표현했을 때 스팸 메일과 일반 메일의 교차점을 확인할 수 있다.

그렇기 때문에 최적의 스팸 분류 기준 값을 찾기 위해 임의의 스팸 분류 기준 값을 정하여 테스트를 시행한다. 한국어와 영어에 대해서 90% 학습데이터의 SVM 반환 값으로 성능을 평가해보면 표 7, 그림 8과 표 8, 그림 9와 같다.

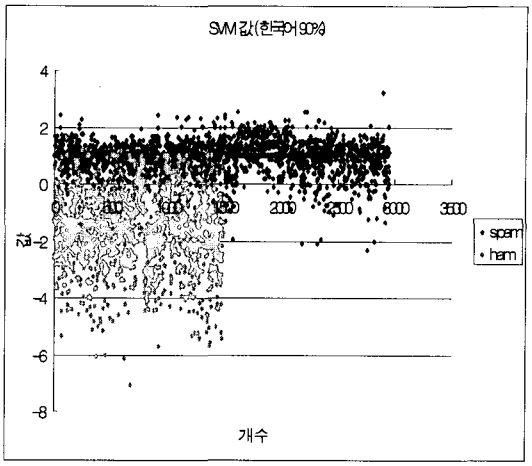

그림 4. 한국어 SVM 반환 값 분포

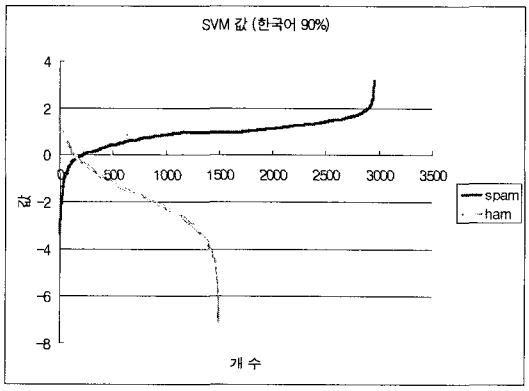

그림 5. 한국어 SVM 반환 값 분포(정렬)

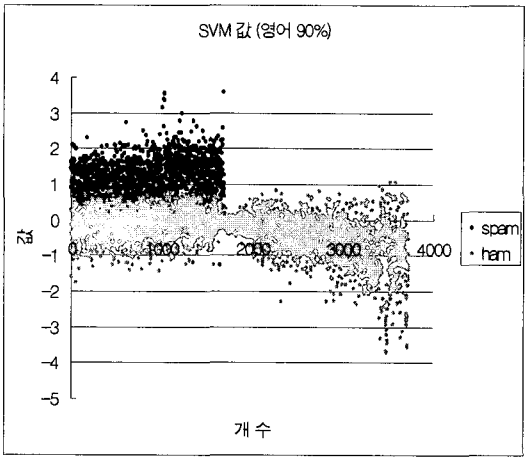

그림 6. 영어 SVM 반환 값 분포

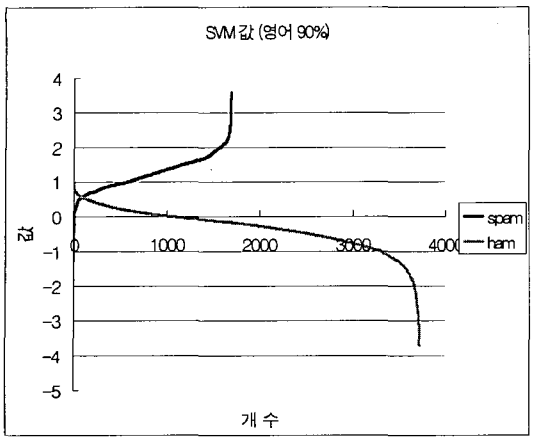

그림 7. 영어 SVM 반환 값 분포(정렬)

표 7. 한국어 SVM 스팸 분류 기준 값 평가

기준	정확률	에러율	hm	sm	(hm+sm)/2
-0.40	92.6	7.4	15.3	3.3	9.29
-0.35	92.6	7.4	14.4	3.8	9.08
-0.30	92.9	7.1	13.0	4.0	8.48
-0.25	92.9	7.1	12.2	4.3	8.28
-0.20	93.0	7.0	11.3	4.8	8.04
-0.15	93.1	6.9	10.5	5.0	7.77
-0.10	92.6	7.4	10.0	6.0	8.02
-0.05	92.5	7.5	9.6	6.4	7.98
0.00	92.3	7.7	8.6	7.2	7.93
0.05	92.2	7.8	8.1	7.7	7.89
0.10	91.7	8.3	7.6	8.6	8.13

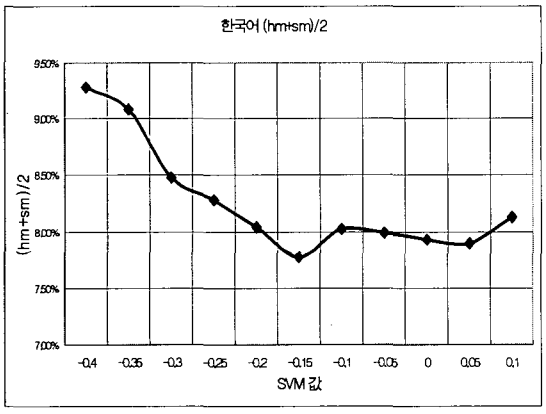

그림 8. 한국어 SVM 스팸 분류 기준 값에 따른 결과

실험 결과를 보면 한국어 시스템의 경우 -0.15를 기준으로 분류하였을 때 최적의 성능을 보였으며 영어 시스템의 경우 0.55를 기준으로 분류하였을 때 최적의 성능을 보임을 알았다.

이를 바탕으로 스팸 메일과 일반 메일을 분류하는 스팸 분류 기준 값을 지정한 후 1단계만 실험, 2단계만 실험 그리고 1단계와 2단계를 모두 구축한 후 실험하여 성능을 평가하였다. 성능 평가 결과는 표 9, 표 10과 같다.

표 8. 영어 SVM 스팸 분류 기준 값 평가

기준	정확률	에러율	hm	sm	(hm+sm)/2
0.30	92.5	7.5	10.2	1.6	5.97
0.35	93.6	6.4	8.4	1.9	5.17
0.40	94.8	5.2	6.6	2.3	4.42
0.45	95.7	4.3	5.1	2.6	3.85
0.50	96.2	3.8	3.8	3.7	3.75
0.55	96.7	3.3	2.6	4.6	3.63
0.60	96.8	3.2	1.7	6.4	4.07
0.65	96.8	3.2	1.0	8.0	4.49
0.70	96.5	3.5	0.6	9.6	5.13
0.75	95.0	5.0	0.4	14.9	7.65
0.80	94.3	5.7	0.3	17.4	8.86

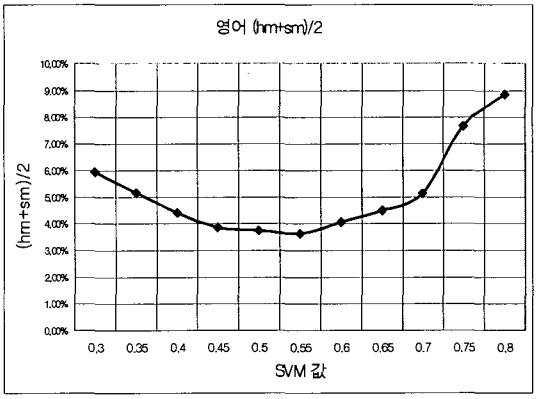

그림 9. 영어 SVM 스팸 분류 기준 값에 따른 결과

표 9. 한국어 SVM 스팸 분류 기준 값 평가

단계	hm	sm	(hm+sm)/2
1단계	0.0	24.8	12.42
2단계	21.4	4.5	12.97
1+2단계	21.4	2.1	11.30

표 10. 영어 SVM 스팸 분류 기준 값 평가

단계	hm	sm	(hm+sm)/2
1단계	0.0	71.6	35.79
2단계	4.3	8.9	6.64
1+2단계	4.3	4.7	4.54

한국어와 영어 메일의 실험 데이터 개수가 다르기 때문에 절대평가를 하기는 어렵다. 한국어 시스템의 경우 sm의 비율이 영어 시스템의 비율보다 낮은 이유는 한국어 시스템에서 실험을 할

때 스팸 메일의 비율이 총 데이터의 83.5%를 차지하기 때문에 일반 메일보다 학습 레코드의 개수가 많아서 학습이 더 잘 되어 성능이 우수하게 나온 것이다. 이와 마찬가지로 영어 시스템의 경우 hm의 비율이 한국어 시스템의 비율보다 낮은 이유는 일반 메일의 비율이 총 데이터의 68.6%를 차지하기 때문이다. 또한 각 메일의 특성이 다름도 하나의 이유가 된다. 학습데이터의 양을 더 많이 늘린다면 스팸 메일을 분류하는 정확률이 더 올라갈 것으로 기대된다.

본 연구의 목적은 한국어 시스템과 영어 시스템의 우열을 가림이 아니라 각 시스템의 성능을 객관적으로 분석하는 것이다.

5. 결론

스팸 메일은 국적을 불문하고 전 세계적인 이슈가 되며 해결해야할 문제로 인식되고 있다. 따라서 한국어 메일과 영어 메일을 대상으로 SVM 기계학습 기법을 이용하여 효율적인 스팸 메일 필터링 시스템을 구성하여 필터링 성능을 살펴보았다.

본 연구에서는 SVM 기계학습에 필요한 속성벡터를 구성할 때의 기준이 되는 자질을 선정하는 여러 알고리즘과 자질의 개수를 다르게 하여 한국어 시스템과 영어 시스템에서 성능을 비교하였으며 카이제곱 알고리즘으로 3000개의 자질을 추출하여 이용하였을 때 한국어 시스템과 영어 시스템 모두에서 최고의 성능을 보임을 확인하였다.

따라서 우리는 카이제곱 알고리즘으로 3000개의 자질을 추출하여 한국어 시스템과 영어 시스템 각각에 SVM 분류기를 만들어 적용시켰다.

본 시스템은 2단계 스팸 메일 필터링 방법을 사용하였으며 2단계의 경우 최적의 성능을 낼 수 있도록 학습 시 이용되었던 메일을 이용하여 SVM 반환 값의 분포를 확인하여 스팸 분류 기준 값을 정하였다. 최종적으로 1단계 필터링, 2단계 필터링 그리고 1+2단계 필터링으로 성능을 분석하였다.

실험 결과 1단계만 적용시킬 때 혹은 2단계만 적용시켰을 때 보다 두 단계를 모두 적용시켰을 때 최고의 성능을 보임을 확인하였고 한국어 시스템과 영어 시스템 모두에서 스팸 필터링 성능이 우수함을 확인하였다. 실제 사용될 시스템에서는 hm과 sm의 가중치를 달리하여 스팸 분류 기준 값을 정할 필요가 있을 것으로 생각된다.

향후 연구에는 기계학습 기법과 더불어 시멘틱 웹 기술의 하나인 온톨로지(ontology)와 추론(reasoning)기법을 도입하면 좀 더 효율적인 필터링이 가능할 것으로 생각된다. 온톨로지를 구성한 후 추론을 수행하면 사실을 바탕으로 더 많은 정보를 컴퓨터가 이해하고 판단할 수 있기 때문이다.

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://plg.uwaterloo.ca/~gvcormac/spam/

[16] http://www.tartarus.org/~martin/PorterStemmer/