어휘정보와 시소러스에 기반한 스팸메일 필터링†

(Spam-mail Filtering based on Lexical Information and Thesaurus)

강신재*, 김종완*
(Shin-Jae Kang, Jong-Wan Kim)

요 약 본 연구에서는 어휘정보와 개념정보를 기반으로 스팸메일 필터링 시스템을 구축하였다. 스팸메일을 판별할 수 있는 정보를 두 가지로 구분하였는데, 확실한 정보군은 송신자 정보, URL, 그리고 최근 스팸 키워드 리스트이며, 덜 확실한 정보군은 메일 본문에서 추출한 단어목록과 개념코드이다. 먼저 확실한 정보군을 이용하여 스팸메일을 분류하고 그다음 덜 확실한 정보군을 이용하였다. 메일 본문에 포함된 어휘정보와 개념코드는 SVM 기계학습을 한 후 사용된다. 본 연구의 결과, 더 많은 어휘정보를 특징벡터로 사용하였을 때 스팸 정확률이 상승하였으며 추가로 개념코드를 특징벡터에 포함시켰을 때 스팸 재현율이 상승하였다.

핵심주제어 : 정보여과, 기계학습, 형태소분석, 시소러스

Abstract In this paper, we constructed a spam-mail filtering system based on the lexical and conceptual information. There are two kinds of information that can distinguish the spam mail from the legitimate mail. The definite information is the mail sender's information, URL, a certain spam keyword list, and the less definite information is the word lists and concept codes extracted from the mail body. We first classified the spam mail by using the definite information, and then used the less definite information. We used the lexical information and concept codes contained in the email body for SVM learning. According to our results the spam precision was increased if more lexical information was used as features, and the spam recall was increased when the concept codes were included in features as well.

Key Words : Information Filtering, Machine Translation, Morphological Analysis, Thesaurus

1. 서 론

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 스팸메일을 처리해야 하는 불편함도 주고 있다. 스팸메일, 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 스팸메일을 제거하고 있다. 하지만 이러한 리스트나 필터의 구축은 대부분 수작업으로 이루어지기 때문에 구축비용 및 시간이 많이 필요하며, 또한 실제 상황에서 모든 스팸메일을 완벽하게 처리할 수는 없다는 문제가 있다.

기본적으로 스팸메일 필터링 문제는 문서분류의 특별한 한 형태로 볼 수 있다. 여러 다양한

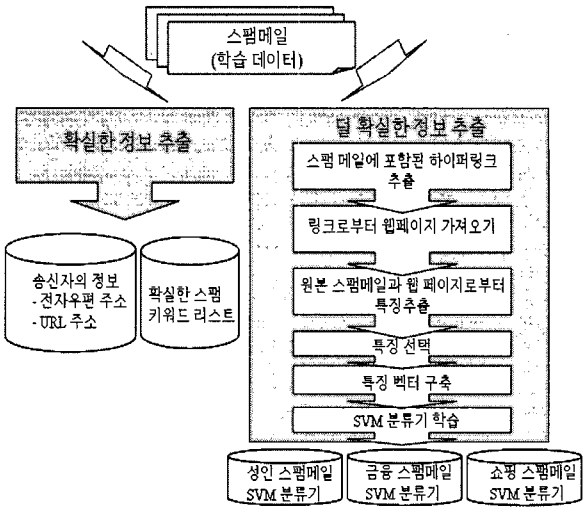
† 이 논문은 2005학년도 대구대학교 학술연구비 지원에 의한 논문임.
* 대구대학교 컴퓨터·IT공학부 교수

정보검색 기법들이 이 문제를 해결하는데 적합하며 우리가 해결해야 할 문제는 스팸인지 비스팸인지를 분류하는 이진 분류문제이다. 여러 다양한 기계학습 알고리즘들이 전자우편 분류에 사용되어져 왔다[2, 4, 5, 6]. Sahami et al. [2]에서는 나이브 베이지안 분류기(Naive Bayesian Classifier)를 스팸메일 필터링에 사용하였는데, 수작업으로 구축된 구(phrase) 정보와 송신자의 도메인 형식, 제목에서 기호 문자의 비율 등 다양한 비 텍스트(non-textual) 정보를 도메인 속성으로 정의하여 사용하였다. 대부분의 경우, 기존의 분류기보다 성능이 더 우수한 Vapnik[3]의 SVM(Support Vector Machines)을 전자우편 문서 분류에 이용해왔다[4, 5]. Yang et al. [6]은 나이브 베이지안(Naive Bayesian)과 SVM (Support Vector Machines)이 TFIDF보다 훨씬 우수한 성능을 보임을 실험으로 증명하였다. 특히, SVM을 머리말(header)에 적용하였을 때 최고의 성능을 얻었다. 따라서 SVM 분류기가 이진 분류문제에 좀 더 적합함을 결론내릴 수 있다.

본 연구에서는 어휘정보와 시소러스의 개념정보를 기반으로 한 스팸메일 필터링을 위하여 2단계 스팸메일 필터링 시스템을 제안한다. 첫 번째 단계에서는 송신자의 URL, 전자우편 주소, 그리고 스팸 키워드 리스트로 구성된 확실한 정보(definite information)가 적용된다. 두 번째 단계에서는 첫 번째 단계에서 분류되지 않고 남은 메일을 덜 확실한 정보(less definite information)와 전자우편의 머리말과 본문(body)뿐만 아니라 패치(fetch)된 웹 페이지의 내용을 이용하여 분류한다.

2. 학습단계: 특징 선택과 기계학습

특징이나 속성을 추출하여 효율적으로 필터링하기 위하여 두 가지 정보군(확실한 정보군, 덜 확실한 정보군)으로 나누어 처리하였다. 학습단계에서의 전체적인 처리 과정은 <그림 1>에 제시되어 있다.



<그림 1> 스팸메일 필터링을 위한 학습과정

2.1 확실한 정보군

스팸메일 필터링을 위한 확실한 정보군은 송신자의 전자우편 주소, URL 주소와 같은 송신자 정보와 "포르노", "신용대출", "광고" 등과 같은 확실한 스팸 키워드 리스트를 말한다. 만약 새로 도착한 메일의 정보가 송신자의 전자우편 주소나 URL 주소 정보 중 하나와 일치한다면 해당 메일은 스팸메일일 확률이 매우 높기 때문에 다른 처리 과정 없이 바로 스팸메일로 분류되게 된다. 하지만 확실한 스팸 키워드가 제목 부분에 1회 이상 혹은 메일 본문에 3회 이상 나타나는 경우에만 스팸메일로 분류하게 된다. 송신자의 정보는 스팸메일로부터 자동으로 추출하고 확실한 스팸 키워드 리스트는 수작업으로 구축한다.

2.2 덜 확실한 정보군

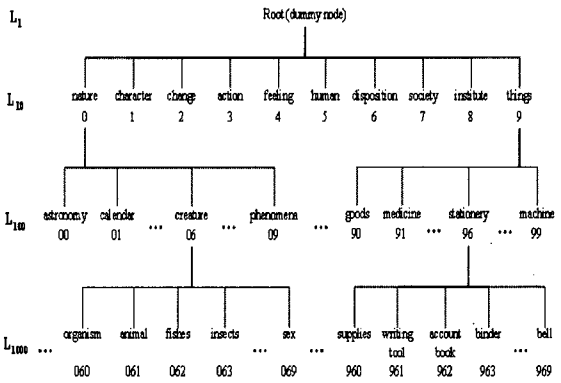
전자우편의 특별한 특징들은 전자우편이 스팸인지 아닌지를 구분하는 힌트가 된다. 예를 들어, 확실한 스팸 키워드가 포함되지 않은 전자우편 내의 단어, 송신자의 도메인 형식(e.g. co, com), 전자우편 수신시각, 또는 제목에 포함된 특수문자의 비율 등이 스팸메일임을 암시한다[2]. 이 단계는 전자우편 문서의 특징들을 어휘정보 혹은 개념코드와 일치하는 부분을 찾아 특징벡터 공간을 만들어 이를 기반으로 SVM

학습 알고리즘을 이용하여 전자우편을 분류하게 된다. 스팸메일의 본문에는 적은 양의 문서 정보가 있으며 최근에는 단지 이미지 데이터만 존재하는 전자우편도 있다. 따라서 본 시스템에서는 전자우편 내에 포함된 하이퍼링크를 따라가서 해당 웹 페이지의 내용을 패치(fetch)하여 그 내용을 스팸메일의 확장된 본문으로 간주한다. 스팸메일 필터링의 힌트가 되는 전자우편내의 단어나 구(phase)의 문자정보를 추출하기 위해서 각각의 전자우편은 심볼 제거, 형태소 분석 그리고 불용어(stop-words) 제거와 같은 전처리 과정을 거치게 된다. 추출된 특징들 중에 변별력이 낮은 특징들은 스팸메일을 필터링하는데 도움이 되지 않기 때문에 특징벡터로부터 제거될 것이다.

최고의 예측을 할 수 있는 특징들의 부분집합을 찾기 위해, 특징집합의 모든 가능한 조합을 평가하여 특징 선택을 하게 된다. 분별력이 높도록 특징을 선택하기 위해서 본 실험에서는 퍼지 추론 방법(Fuzzy Inference Method)을 사용하였다[7].

2.3 가도카와(Kadokawa) 시소러스

가도카와 시소러스[8]는 1,110개의 의미적 카테고리와 4 레벨의 계층 구조를 가진다<그림 2>. L_1, L_{10}, ,L_{100} 레벨에 속해 있는 개념들은 각각 10개의 하위 클래스들로 나뉘며 루트 노드는 더미(Dummy)이다. 명사와 동사는 구분없이 가도카와 시소러스의 계층구조에 공존한다. 동사 카테고리들은 주로 L_{1000} 레벨의 2xx, 3xx, 4xx 코드에 위치한다. 원래 가도카와 시소러스의 계층 구조는 일본어를 대상으로 구성이 되어 있으나 한국어에 대해서도 그대로 활용될 수 있다. 이는 POSTECH(Pohang University of Science and Technology)에서 개발한 COBALT-J/K(Collocation-Based Language Translator from Japanese to Korean) [9]와 COBALT-K/J (Collocation-Based Language Translator from Korean to Japanese)[10]에 활용되어 이미 그 성능을 입증한바 있기 때문이다. 가도카와 시소러스는 COBALT-J/K와 COBALT-K/J 기계번역 시스템에서 어휘의미의 중의성 해소에 유용하다는 것이 증명되었다[11]. 위 시스템에서 사용하고 있는 전자사전의 모든 단어들은 이미 L_{1000} 레벨의 세 자리 숫자 개념코드와 연결되어 있으므로, 본 시스템에서는 전자우편에서 나타나는 단어들의 개념코드를 쉽게 찾을 수 있다.



<그림 2> 가도카와 시소러스의 계층구조

2.4 특징벡터 구축

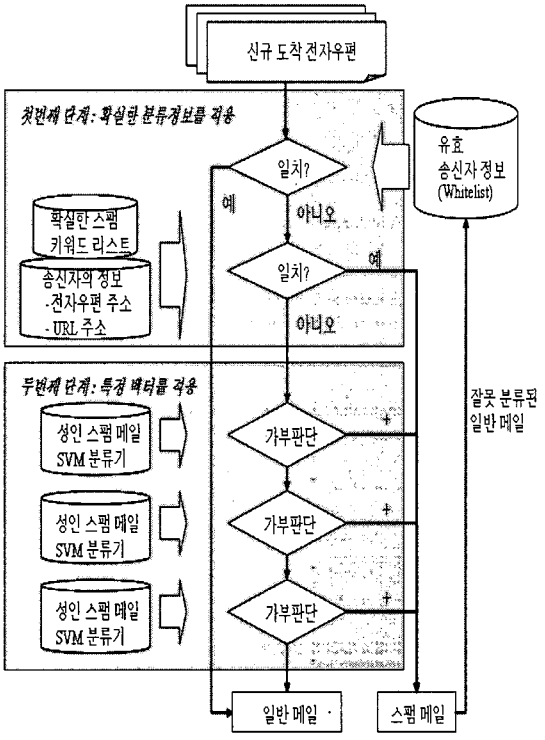
특징벡터는 단어와 개념코드 그리고 송신자의 도메인 타입, 전자우편 수신 시각 등과 같은 특별한 특징들을 이용하여 구성되었다. 즉, 각 메일들은 $\vec{x} = < x_1, x_2, ..., x_n >$과 같이 표현되는데, 각 속성($x_1$, x_2, ..., x_n) 들은 각 특징(X_1, X_2, ..., X_n)들의 속성값이다. 이진값(binary value)을 가지는 특징 벡터를 사용할 때 SVM이 가장 우수한 성능을 보이기 때문에[4], 본 연구에서는 이진 특징 벡터를 사용하였다. 특징벡터에서 단어에 해당하는 부분의 속성값들은 전자우편 본문 내에 그 단어가 나타났는지에 따라 0 또는 1의 값을 가지게 된다. 개념코드의 경우에는 가도카와 시소러스의 L_{1000} 레벨의 1000개의 개념코드로 표현하며 만약 일치하는 개념코드가 있으면 특징 값을 1로 지정하고 그렇지 않을 경우, 0으로 특징 값이 지정된다. 전자우편 수신 시각 특징의 경우, 예를 들어 오후 12시와 오전 5시 사이에 전자우편이 수신되었으면 해당 특징 값이 1, 그렇지 않으면 0으로 설정된다. 이는 스팸메일이 대부분 늦은 밤이나 새벽에 수신되는 특성을 살펴보기 위함이다.

본 시스템에서는 최종적으로 전자우편을 스

팸메일인지 또는 일반메일인지로 이진분류하게 된다. 하지만 스팸메일을 구분하기 위해 하나의 분류기만 만드는 것은 불합리하다고 볼 수 있다. 왜냐하면 스팸메일의 종류에 따라 서로 다른 특징 즉, 단어나 개념을 포함하고 있기 때문이다. 따라서 본 시스템에서는 스팸메일을 크게 성인, 금융, 쇼핑의 세 가지로 구분하고 3개의 SVM 분류기를 각각 구현하였다.

3. 적용단계: 2단계 필터링

수신된 전자우편은 학습 단계에서 구축된 다양한 정보와 SVM 분류기에 의해서 처리된다 <그림. 3>.



<그림 3> 스팸메일 필터링 과정

우리는 이미 스팸메일 필터링을 위한 힌트를 확실한 정보군과 덜 확실한 정보군의 두 가지 정보군으로 나누었다. 수신된 메일이 확실한 정보군의 힌트를 하나 이상 포함할 경우, 이 메일은 스팸메일일 가능성이 매우 높기 때문에 기계학습 알고리즘을 수행할 필요가 없다. 수신된 메일이 확실한 정보군의 힌트를 하나도 갖고 있지 않을 경우, SVM 분류기를 이용하여 검증하게 된다. 부연하면, 만약 메일이 확실한 정보군의 힌트를 하나이상 포함 할 경우 이 메일은 스팸메일로 간주되며, 만약 그렇지 않다면 이 메일은 다음 단계인 SVM 분류기 적용단계로 넘어가게 되는 것이다. 이 단계에서는 성인 SVM 분류기가 먼저 적용되는데, 만약 이 분류기에서 메일이 스팸메일로 분류될 경우 두 번째 필터링 적용단계는 끝이 난다. 그렇지 않을 경우에는 금융 SVM 분류기로 넘어가게 된다. 이 분류기에서 메일이 스팸메일로 분류될 경우 앞의 분류기와 마찬가지로 두 번째 필터링 적용단계는 끝이 난다. 이와 마찬가지로 마지막 쇼핑 SVM 분류기도 순차적으로 적용된다.

하지만 일반메일이 스팸메일로 분류될 경우 중요한 정보를 잃게 될 수도 있다. 이와 같은 문제를 해결하기 위하여 필터링된 스팸메일의 목록을 사용자에게 주기적으로 알려주고, 만약 사용자가 잘못 분류된 일반메일을 발견하게 되면 해당 메일의 주소를 화이트리스트(white list)에 수작업으로 등록하게 된다. 화이트리스트는 시스템에서 최우선적으로 적용하게 되며 수신된 메일이 이에 일치되는 경우 바로 일반메일로 분류되게 된다.

4. 실 험

실험 평가를 위하여 사용된 전자우편 말뭉치(corpus)는 총 5,018개의 전자우편이며 일반메일은 1,737개, 성인은 1,214개, 금융은 1,506개, 쇼핑은 561개로 총 4개의 카테고리로 나뉜다. 1단계에서 사용되는 확실한 정보군은 전자우편 말뭉치로부터 수작업으로 추출되었다. 송신자 전자우편 주소가 1,461개, URL이 1,023개, 스팸 키워드 리스트 603개가 구축되었다. 2단계 적용에 필요한 특징을 선택하기 위해서 WEKA [12]에서 제공하는 weak.attributeSelection package를 사용하였다. WEKA는 현실 세계의 데이터 집합에 기계학습 기술을 적용하는 목적으로 디자인된 프로그램이다. WEKA는 많은 분류 모델들을 포함하고 있는데, 이번 실험에서 사용된 SVM

분류기 또한 WEKA에서 제공된다. SVM의 학습을 위한 파라미터는 WEKA에 설정되어 있는 기본값을 이용하였다.

전자우편 말뭉치의 필터링 성능을 평가하기 위해서 정보검색 분야에서 많이 이용되는 정확률(Precision), 재현율(Recall)을 사용하였다. 평가척도를 위한 분할표는 <표 1>과 같다.

<표 1> 평가척도를 위한 분할표

시스템 분류 / 실제분류	스팸메일	일반메일
스팸메일	*a*	*c*
일반메일	*b*	*d*

스팸 정확률(spam precision)과 스팸 재현율(spam recall)은 다음과 같이 정의한다.

스팸 정확률:

$$\text{스팸정확률(SP)} = \frac{\text{시스템에의해올바르게분류된스팸메일의수}}{\text{시스템에의해분류된스팸메일의총수}} = \frac{a}{a+b}$$

스팸 재현율:

$$\text{스팸재현율(SR)} = \frac{\text{시스템에의해올바르게분류된스팸메일의수}}{\text{실제스팸메일의총수}} = \frac{a}{a+c}$$

스팸메일 필터링에서 일반메일이 스팸으로 분류되는 것은 심각하고 중요한 문제이다. 반면 스팸메일이 일반메일로 분류되는 에러에 대해서는 위 문제에 비해 그다지 심각하지는 않다고 생각할 수 있다. 따라서 본 실험에서는 정밀도(accuracy)와 오류율(error rate)을 평가 방법으로 추가하여 사용하였다. 정밀도는 전체 수신된 전자우편 중에 분류를 올바르게 예측한 비율을 나타내고, 오류율은 전체 수신된 전자우편 중에 비율을 잘못 예측한 비율을 나타낸다[13]. 정밀도와 오류율은 다음과 같이 정의된다.

$$\text{정밀도(A)} = \frac{\text{올바르게분류한메일의수}}{\text{전체메일의수}} = \frac{a+d}{a+b+c+d}$$

$$\text{오류율(E)} = \frac{\text{잘못분류한메일의수}}{\text{전체메일의수}} = \frac{b+c}{a+b+c+d}$$

바람직한 전자우편 필터링 시스템은 높은 정밀도와 낮은 오류율을 보여주어야 한다. 일반적으로 정밀도를 높이는 것보다 오류율을 낮추는 것이 더 중요하다고 볼 수 있다.

본 실험에서는 보다 객관적인 성능평가를 위하여 10층 교차 확인법(ten-fold cross validation)을 사용하였다. 전자우편 말뭉치는 무작위로 10개의 부분으로 나뉘고, 실험은 10회 반복하여 실시된다. 9개의 부분은 학습을 위해 사용하며, 나머지 한 개의 부분을 평가를 위해 사용하게 되는데, 각 실험마다 서로 다른 부분을 가지고 평가를 한다. 총 10회의 실험이 모두 끝나면 각 결과를 평균내어 최종결과를 구한다.

가도카와 시소러스의 개념코드를 사용하기 전에 올바른 매개변수 설정을 선택하기 위하여 다양한 방법으로 실험을 하였다. 일반적으로, 문장에서 각 단어의 올바른 개념코드를 선택하는 것은 어려운 문제이다. 예를 들어 한국어 "눈"은 사람의 눈, 하늘에서 내리는 눈, 그리고 새싹 등과 같이 다양한 뜻을 갖고 있다. 자연어 처리(Natural Language Processing)에서 이 문제는 단어 의미 중의성 해소(Word Sense Disambiguation)로 불리며, 이 문제는 최근 중요한 이슈가 되고 있다. 단어 의미 중의성 해소 문제를 해결하기 위해서는 의미정보가 추가된 말뭉치, 패턴 집합, 그리고 자연어 처리 프로그램 등과 같은 많은 언어 자원이 필요하다. 따라서 본 실험에서는 전자우편으로부터 개념코드를 추출하기 위한 다음의 방법을 선택하였다. 먼저 2.3절에 언급된 기계번역 사전에 등록되어 있는 모든 표제어들의 가도카와 개념코드를 수집하여 데이터베이스화 한다. 그다음 전자우편에서 나타나는 단어들을 검색하여 개념코드를 수집한다. 이 방법은 단어 의미의 중의성을 해소하지 않고 해당 단어와 관련된 모든 코드를 수집한 것이기 때문에 실제 의미와는 다른 부적절한 코드를 포함하고 있을 수도 있다. 따라서 첫 번째 실험에서는 부적절한 개념코드의 영향에 대한 검사가 이루어졌다. 부적절한 개념코드의 영향을 줄이기 위해 개념코드의 빈도에 따라서 특징의 값을 설정하는 방법을 적용해 보았다. <표 2>는 개념코드의 빈도에 따른 성능을 비교한 것이다. "1번 이상"은 개념이 전자우편 본문에서 1번 이상 나타나면 특징값을 1, 그렇지 않으면 0으로 설정한

다. <표 2>를 보면 "1번 이상"일 때 스팸 정확률과 스팸 재현율이 가장 좋은 성능을 보임을 알 수 있다.

<표 2> 개념코드의 빈도에 따른 실험결과

평가방법 / 빈도	SP	SR	A	E
1번 이상	91.7	77.6	91.2	8.8
2번 이상	87.8	66.2	87.1	12.9
3번 이상	82.9	38.0	79.0	21.0
4번 이상	83.3	19.6	74.6	25.4

<표 3>은 스팸메일 필터링에서 어휘정보와 개념정보의 영향을 보여주고 있다. 만약 같은 특징의 수로 특징벡터를 구축한다면, 어휘정보만을 사용하였을 경우에는 스팸 정확률은 최고 성능을 보이고 어휘정보와 개념정보를 같이 사용하였을 경우에는 비슷한 의미를 갖는 단어들이 같은 개념코드로 분류될 수 있기 때문에 스팸 재현율이 최고 성능을 보인다. 그러므로 특징벡터는 실제 시스템을 사용할 사용자의 개인적인 성향에 따라 구성을 조정하는 것이 바람직하다고 볼 수 있다.

<표 3> 특징 구성에 따른 실험결과

평가방법 / 특징벡터구성	SP	SR	A	E
어휘정보 2000개	88.7	62.9	95.0	5.0
어휘정보 3000개	88.8	66.1	95.3	4.7
어휘정보 2000개 + 개념코드 1000개	82.8	70.6	95.1	4.9

총 5,018개의 전자우편 중 4,514개 (90%)의 전자우편은 SVM 분류기 학습을 위하여 사용됐고 나머지 504개 (10%)의 전자우편은 본 스팸메일 시스템의 성능을 테스트하기 위해서 사용됐다. 테스트용 전자우편은 학습단계에서 구성된 정보와 분류기를 사용하여 스팸인지 아닌지를 결정하는데 사용된다.

<표 4>를 통하여 본 논문에서 제안한 2단계 필터링 방법이 각 단계를 따로 적용하는 방법보다 더 효과적이라는 것을 알 수 있다.

<표 4> 제안된 시스템 성능(개념코드 포함시) (%)

평가방법 / 적용단계	SP	SR	A	E
첫 번째 단계만	100	83.0	88.9	11.1
두 번째 단계만	86.0	37.3	55.0	45.0
첫 번째 + 두 번째 단계	96.0	94.5	93.8	6.2

첫 번째 단계에서는 높은 정확률을 가지고 메일을 처리하지만 저장된 정보와 일치하지 않는 경우 판단을 할 수 없는 것처럼 적용률이 낮은 특징을 가지고 있으며, 두 번째 단계에서는 기계학습 방법에 의해 모든 메일을 판단할 수 있는 즉, 적용률은 높지만 정확률은 첫 번째 단계보다 떨어지는 특징이 있다. 그러므로 이 두 가지 방법을 접목한 2단계 방법이 전반적으로 좋은 성능을 보여주게 된다. 2단계 방법은 첫 번째 단계 혹은 두 번째 단계만 수행하는 것보다 스팸 재현율의 성능이 각각 13.9% 와 153.4% 개선되었다. 첫 번째 단계와 두 번째 단계의 평균을 내보면 제안한 2단계 방법의 스팸 재현율이 57.1% 상승함을 알 수 있다. 여기서 개념정보의 활용이 스팸 재현율을 상승시키는데 중요한 역할을 한다는 사실을 알 수 있다.

5. 결 론

본 연구에서는 어휘정보와 개념정보를 기반으로 한 스팸메일 필터링 시스템을 구축하고 어휘정보와 개념정보의 시스템 성능에의 영향을 분석하였다. <표 3>을 보면 더 많은 어휘정보를 특징으로 사용하였을 때 스팸 정확률이 상승하고, 개념정보를 특징 벡터에 포함시켰을 때 스팸 재현율이 상승한다는 사실을 알 수 있다.

또한, 어휘정보와 개념정보 그리고 하이퍼링크에 기반한 스팸메일 필터링을 위하여 2단계 방법을 제안하였다. 최근 스팸메일의 본문은 적은 양의 텍스트 정보를 갖고 있기 때문에 정상메일로부터 스팸메일을 구별해 내기 위한 힌트가 부족하다. 이 문제를 해결하기 위해서 메일 본문에 포함된 하이퍼링크를 활용하였고 패치

(fetch)된 웹 페이지와 원래의 메일 본문에서 모든 가능한 힌트를 추출하였다. 이렇게 추출된 힌트들은 SVM 분류기를 구축하는데 사용된다. 힌트는 확실한 정보군과 불확실한 정보군의 두 가지 종류로 나뉘게 되는데, 확실한 정보군을 먼저 적용하고, 그다음 덜 확실한 정보군을 적용하였다.

제안된 2단계 방법은 첫 번째 단계 혹은 두 번째 단계를 단독으로 사용하는 것보다 스팸 재현율이 57.1% 개선되는 결과를 얻었다. 즉 스팸메일 필터링에서 기계학습 알고리즘, 블랙리스트 또는 키워드기반 필터만 사용하는 방법보다 더 효과적이라는 것을 알 수 있다.

본 연구는 스팸메일로 인한 미성년자의 성인물 접근을 예방할 수 있고, 사용자가 전자우편의 스팸메일 여부를 확인하는 시간을 절약할 수 있다는 점에서 매우 중요하다. 향후 연구로는 전자우편내의 이미지에서 더 많은 특징들을 찾아 내는 연구와 시맨틱 웹(semantic web) 기법을 적용하기 위한 스팸 도메인 온톨로지를 구축하고 활용하는 방법을 연구하고자 한다.

참 고 문 헌

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

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

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

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

[5] Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features,"ECML, Claire Nedellec and Celine Rouveirol (ed.) (1998)

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

[7] Kim, J. W., Kim, H. J., Kang, S. J., and Kim, B. M., "Determination of Usenet News Groups by Fuzzy Inference and Kohonen Network," Lecture Notes in Artificial Intelligence, Vol.3157, Springer-Verlag (2004) 654-663

[8] Ohno, S. and Hamanishi, M., New Synonyms Dictionary, Kadokawa Shoten, Tokyo, (1981) (Written in Japanese)

[9] Park, C. J., Lee, J. H., Lee, G. B., and Kakechi, K., "Collocation-Based Transfer Method in Japanese-Korean Machine Translation," Transaction of Information Processing Society of Japan, Vol.38, No.4, (1997) 707-718 (Written in Japanese)

[10] Moon, K. H. and Lee, J. H., "Representation and Recognition Method for Multi-Word Translation Units in Korean-to-Japanese MT System,"In the 18th International Conference on Computational Linguistics (COLING 2000), Germany, (2000) 544-550

[11] Li, H. F., Heo, N. W., Moon, K. H., Lee, J. H., and Lee, G. B., "Lexical Transfer Ambiguity Resolution Using Automatically-Extracted Concept Co-occurrence Information," International Journal of Computer Processing of Oriental Languages, World Scientific Pub., Vol.13, No.1, (2000) 53-68

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

[13] Resnick, P. J., Hansen, D. L., and Richardson, C. R., "Calculating Error Rates for Filtering Software," Communications of ACM, Vol.47, No.9 (2004) 67-71

강 신 재 (Shin-Jae Kang)

종신회원
- 1995년 2월 : 경북대학교 컴퓨터공학과 (공학사)
- 1997년 2월 : 포항공과대학교 (POSTECH) 컴퓨터공학과 (공학석사)
- 2002년 2월 : 포항공과대학교(POSTECH) 컴퓨터공학과 (공학박사)
- 1997년 1월 ~ 1998년 2월 : SK Telecom 정보기술연구원 주임연구원
- 2002년 3월 ~ 현재 : 대구대학교 컴퓨터·IT 공학부 조교수
- 관심분야 : 시맨틱웹, 온톨로지, 추론엔진, 정보검색, 자연어처리

김 종 완 (Jong-Wan Kim)

종신회원
- 1987년 2월 : 서울대학교 컴퓨터공학과 (공학사)
- 1989년 2월 : 서울대학교 컴퓨터공학과 (공학석사)
- 1994년 2월 : 서울대학교 컴퓨터공학과 (공학박사)
- 1995년 3월 ~ 현재 : 대구대학교 컴퓨터·IT 공학부 부교수
- 1999년 1월 ~ 2000년 2월 : 미국 U of Massachusetts 방문교수
- 2006년 1월 ~ 현재 : 미국 U of Oregon 방문교수
- 관심분야 : 인공지능, 지능형 에이전트, 퍼지 시스템, 정보검색, 온톨로지