情報保護學會論文誌
第18卷 第6(A)號, 2008. 12

비정상 문자 조합으로 구성된 스팸 메일의 탐지 방법*

이 호 섭†, 조 재 익, 정 만 현, 문 종 섭‡
고려대학교

An Approach to Detect Spam E-mail with Abnormal Character Composition*

Ho-Sub Lee†, Jae-Ik Cho, Man-Hyun Jung, Jong-Sub Moon‡
Korea University

요 약

인터넷의 활용도가 높아짐에 따라, 스팸메일이 전체 메일에서 차지하는 비중이 점점 커지게 되었다. 전체 인터넷 자원에서 필요에 의해 사용되는 메일의 기능보다, 주로 광고나 악성코드 등의 전파를 위한 목적으로 사용되는 메일의 비중이 점점 커지고 있으며, 이를 방지하기 위한 컴퓨터 및 네트워크, 인적자원의 소모가 매우 심각해지고 있다.

이를 해결하기 위해 스팸 메일 필터링에 대한 연구가 활발히 진행되어 왔으며, 현재는 문맥상의 의미는 없지만 가독상에서 의미를 해석할 수 있는 문장에 대한 연구가 활발히 이루어지고 있다. 이러한 방식의 메일은 기존의 어휘를 분석하거나 문서 분류 기법 등을 이용한 스팸 메일을 필터링 방법을 통해 분류하기 어렵다.

본 연구는 이와 같은 어려움을 해결하기 위해 메일의 제목에 대한 N-GRAM 색인화를 통해 베이지안 및 SVM 을 이용하여 스팸 메일을 필터링 하는 방법을 제안한다.

ABSTRACT

As the use of the internet increases, the distribution of spam mail has also vastly increased. The email's main use was for the exchange of information, however, currently it is being more frequently used for advertisement and malware distribution. This is a serious problem because it consumes a large amount of the limited internet resources. Furthermore, an extensive amount of computer, network and human resources are consumed to prevent it.

As a result much research is being done to prevent and filter spam. Currently, research is being done on readable sentences which do not use proper grammar. This type of spam can not be classified by previous vocabulary analysis or document classification methods.

This paper proposes a method to filter spam by using the subject of the mail and N-GRAM for indexing and Bayesian, SVM algorithms for classification.

Keywords : Spam mail filtering, N-GRAM, Support Vector Machine, Bayesian Decision Theory

접수일: 2008년 9월 1일; 채택일: 2008년 12월 08일
* 이 논문 또는 저서는 2006년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF- 2006-521-D00461).
* 이 연구에 참여한 연구자(의 일부)는 '2단계bk21사업'의 지원비를 받았음.
† 주저자, leehosub@korea.ac.kr
‡ 교신저자, jsmoon@korea.ac.kr

Ⅰ. 서 론

인간 활동이 오프라인에서 온라인으로 이동함에 따라, 사람들에게 편의를 제공하기 위한 많은 서비스가 등

장하고, 발전하였다. 그 중 메일은 오프라인에서의 편지 및 전보, 광고, 음성사서함 등 다양한 역할을 대체하고 있다. 인터넷의 활용도가 높아짐에 따라, 스팸 메일이 전체 메일에서 차지하는 비중이 점점 커지게 되었다.

현재 스팸 메일 필터링 연구에서 문제가 되는 점은, 문맥상의 의미가 없으면서 사용자가 메일을 읽을 때, 즉, 가독상에서 의미를 해석할 수 있는 파괴된 문장에 대한 연구가 이루어지고 있다[1-6]. 예를 들어, “하이” 라는 말은 “ㅎr ol”, “ㅎㅏㅇl” 라고 표현하거나, “100%” 를 “ㅣㅇㅇ%” 로 표현하는 등, 하나의 단어 및 문장에 대해 생성할 수 있는 방법은 무한히 많다. Andrej Bratko에 의하면 “viagra”라는 단어는 “V-I-A-G-R-A”, “V|i|a|g|r|a” 등을 포함해서 약 1.3×10^{21} 가지로 표현될 수 있다고 한다[1]. 이것은 일반적인 단어에 욕설과 은어, 타국어, 특수문자, 기호, 채팅 용어 등과 결합되었을 때, 무한대에 가까운 수의 변종 단어가 존재할 수 있다는 것을 의미하며, 기존 단어 사전 기반의 스팸 메일 필터링 시스템의 효율 및 분류율에 대해 문제를 야기할 수 있는 요소로 작용한다. 따라서 기존의 어휘 분석 방법이나 문서 분류 기법 등을 이용한 스팸 메일을 필터링 방법은 위와 같은 가독상의 해석 문제를 해결하기 어렵고, 파괴된 문장을 복원하기 위해 추가적인 단어 인식 및 매칭 과정이 필요하다[3-6]. 본 논문에서는 이러한 추가적인 과정 없이 메일의 제목에 대한 N-GRAM 색인화와 받는 사람 ID 리스트에 대한 Levenshtein distance[7]를 사용하여 ID 간의 유사도를 사용하는 것을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 스팸메일 필터링을 위해 참조한 관련 연구에 대하여 설명하고, 3장에서 스팸메일 필터링을 위해 제안하는 방법에 대해 설명한다. 4장에서는 제안하는 방법을 이용하여 도출된 실험 데이터에 베이지안 및 Support Vector Machine (SVM) 등의 분류 알고리즘[8-12]을 적용하여 스팸 필터링 결과를 도출한다. 마지막으로 5장에서 본 연구에 대한 결론 및 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

2.1 기존 연구

2.1.1. 베이지안 결정 이론을 이용한 스팸 메일 필터링 연구

김현준, 정재은, 조근식이 제안한 스팸 메일 필터링 방법의 경우 정보통신망 이용 촉진 및 정보보호 등에 관한 법률의 시행령 및 시행규칙 개정안을 기반으로 메일 제목에 ‘(광고)’ 또는 ‘(성인광고)’라는 문구와 함께, 제목의 끝에 ‘@’가 포함된 광고성 전자 우편을 미리 필터링 하고, 이외의 필터링 되지 않은 메일에 대해 연구를 진행하였다[13]. 메일 필터링에 사용한 방법은 가중치를 이용한 베이지안 분류이며, 메일을 파싱하여 토큰의 특징을 저장하고, 해당 클래스에서 특징이 갖가지는 빈도수에 의해 가중치가 정의된다. 이 방법은 단순한 베이지안 분류기에 의한 필터링 보다 우수한 성능을 보였으며, 분류기에 의한 필터링 후에 지능형 에이전트 시스템을 통해 사용자의 행동을 관찰하여, 최종 메일 분류에 따른 분류 모델의 재학습 방법을 시행하였다. 이 방법을 이용하여 시간이 경과함에 따라 메일의 학습 데이터가 많아질수록 점점 필터링 성능이 향상되었다.

강승식이 제안한 방법은 메일 주소의 유효성검사와 정보 검색 및 문서 분류에서 주로 사용되는 tf*idf를 각 단어의 가중치로 사용한 나이브 베이지안 분류자에 의해 메일 제목과 본문 내용에 대해 각각 스팸 메일 확률을 계산하는 방법을 적용하였다. 이 때, 메일 제목과 내용에 사용되는 동일한 단어에 대한 중요도를 차별화함으로써 필터링을 시행하였다. 이 방법 역시 단순 나이브 베이지안을 적용한 메일 분류에 비해 높은 성능을 보였다[14].

두 방법 모두 메일의 제목 및 내용을 파싱 및 분석하여 스팸메일을 분류하기 위한 특징을 생성하였지만, 비정상 문자 조합으로 이루어진 문자열에 대한 검사나 문자열 복원 및 인식 등의 과정은 존재하지 않는다.

2.1.2 SVM분류기를 이용한 스팸메일 필터링 연구

스팸 메일 필터링을 위해 SVM 분류기를 사용한 연구 사례로는 서정우의 연구가 있다[15]. N-GRAM을 적용하여 생성된 색인어와 메일 제목에서 사용되는 단어의 빈도수를 조사하고, 그 중 빈도수가 높은 단어 m개를 추출하여 생성한 단어 사전을 검색하여 데이터 그룹을 생성하고, SVM을 통해 스팸 메일을 분류하였다. 이 연구에서 실험을 위해 도출한 데이터 셋의 특징은, 단어사전과 N-GRAM 색인어를 검색하여, 단어사전과의 일치 여부를 표현하는 2진 결과 값을 갖는다. 앞서 언급한 언어 파괴에 관련된 문제는 다루지 않고 있으며, 파괴된 문자열에 대한 복구 및 인식 과정 역시 존재하

지 않는다.

2.1.3 그 외 스팸 메일 필터링에 관한 연구

공미경, 이경순이 제안한 방법의 경우, 보내는 사람에 대한 블랙 리스트 및 화이트 리스트 기반의 메일 필터링을 우선적으로 실시한 후, 최대 엔트로피 모델을 통해 도출되는 스팸성 자질과 URL 자질 확률을 곱하여 스팸 메일을 분류하는 방법을 제안하였다[16]. 이 논문에서 스팸성 자질은 스팸메일에 빈번하게 나타나는 특징이나 스패머들이 인위적으로 메일에 삽입하는 패턴들을 숫자와 관련된 자질, 비정상적으로 변형된 자질, 강조의 목적으로 사용된 자질 등으로 나누어 검사한다. 또한, 비정상적으로 생성된 URL을 검사하여 스팸성 자질을 테스트 한다.

2.2 본 연구와 관련된 이론들

2.2.1 Levenshtein distance

Levenshtein distance는 edit distance라고도 하는데, 서로 다른 두 시퀀스 사이에 차이를 측정하는 척도이다. 특히 두 문자열 사이의 유사도를 측정하는 척도로 이용되며, 검색 엔진 및 입력된 단어 정정 방법 등에서 단어사이의 유사도를 계산하는 척도로 사용된다[7]. Levenshtein Distance는 하나의 문자열이 또 다른 문자열로 변환되는 과정을 생각할 때, 이루어 질 수 있는 연산은 삽입, 수정, 삭제 등이 있다. 각 연산이 수행될 때마다 총 연산 값에 1을 더해줌으로써, 하나의 문자열이 또 다른 문자열로 바뀔 때까지 이 값을 측정한다[7].

$$D(A,B) = \min[C(j) + I(j) + R(j)] \qquad (1)$$

식 (1)에서 C 와 I, R은 각각 수정과 삽입, 삭제를 의미하며, 문자열 A가 B로 변화될 때, 이들의 합이 최소가 되는 것을 두 문자열 사이의 최소 거리 혹은 유사도로 채택한다.

[표 1]은 Levenshtein distance의 예로 fire 가 found로 변환될 때, 위의 알고리즘을 적용한 결과를 나타낸 것이다. 우선 fire와 found의 “f”는 서로 같기 때문에 아무 연산도 하지 않으며, 총 연산 값도 그대로 0 이다. 그 다음 fire와 found의 두 번째 문자 “i”와 “o”는 다르기 때문에 문자 수정을 하며, 3번째, 4번째 문자도 다르기 때문에 두 문자 모두 수정한다. 그러나 5번째 문자의 경우 fire와 found의 문자는 각각 “”와 “d” 이므로, fire에 “d”를 삽입한다. 따라서 최종 연산 값은 수정 3, 삽입 1 이므로, 4가 된다. 이 과정을 [표 1]에서 진한 글씨로 표현하였다.

[표 1] Levenshtein Distance 의 예 : fire 와 found 의 유사도

		f	o	u	n	d
	0	1	2	3	4	5
f	1	**0**	1	2	3	4
i	2	1	**1**	2	3	4
r	3	2	2	**2**	3	4
e	4	3	3	3	**3**	**4**

2.2.2 Baysian Classifier

본 논문에서 사용한 베이지안 분류기는 확률 모델로 이미 알고 있는 지식을 사전 지식으로 사용하여 학습 목표인 조건부 확률을 계산하는 베이즈 정리에 기초를 두고 있는 매우 단순한 방법이다[8,10]. 식 (2)는 베이즈 룰을 나타낸 식이다.

$$P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)} \qquad (2)$$

식 (2)에 기반하여 관측된 특징 벡터 x가 주어질 경우, 그 특징 벡터가 속한 클래스를 결정하는 문제로 우도비 검증을 통해 베이지안 결정 규칙을 세운다.

식 (3)은 베이즈 정리에 기반하여, 우도비 검증 방법을 정리한 것이다[10].

$$\frac{P(x|\omega_1)}{P(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_2)}{P(\omega_1)} \qquad (3)$$

2.2.3 Support Vector Machine

Support Vector Machine(SVM)은 패턴을 고차원 특징 공간으로 사상시킬 수 있다는 점과 대역적으로 최적의 식별이 가능한 특징을 가진다. SVM은 각 클래스를

구분하는 최적 분리 경계면을 구하기 위해 분리 경계면과 가장 분리 경계면에 인접한 점(Support Vector)과의 거리를 최대화한다[8]. 이때 최적의 분리 경계면을 $f(x)=w^Tx+b=0$로 놓으면, Support Vector와 $f(x)$의 거리를 $\frac{2}{\|w\|}$로 나타낼 수 있다. SVM은 $\|w\|^2$를 최소화하여 분리 간격을 최대화하도록 하여 최적 분리면을 찾아낸다. 이 문제는 다음과 같은 최적화 문제가 된다[12].

$$\begin{aligned} &Minimize \quad \frac{1}{2}\|\omega\|^2 \\ &Subject\ to \quad d^i(\omega^Tx_i+b)\geq 1 \\ &\qquad\qquad \text{for } i=1,\cdots,N \end{aligned} \tag{4}$$

이 문제를 라그랑제(Legendra) 배수에 적용 시킨 뒤, 쌍대화 시키면 식 (5)와 같은 quadratic programming (QP) 문제가 된다[9-10].

$$\begin{aligned} &Maximum \quad W(\alpha)=\sum_{i=1}^{n}\alpha_i \\ &\qquad\qquad -\frac{1}{2}\sum_{i=1,j=1}^{n}\alpha_i\alpha_jy_iy_jx_i^Tx_j \\ &Subject\ to \quad \alpha_i\geq 0, \sum_{i=1}^{n}\alpha_iy_i=0 \end{aligned} \tag{5}$$

또한 ω는 식 (6)과 같이 구할 수 있다.

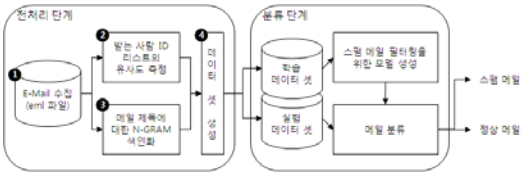
$$w=\sum_{i=1}^{n}\alpha_iy_ix_i \tag{6}$$

Ⅲ. 제안하는 방법

3.1 전체 구조

스팸 메일을 분류하기위해 제안하는 방법의 전체적인 구조는 [그림 1]과 같다. 실험의 단계는 크게 전처리와 분류, 두 단계로 나뉘며, 전처리 단계는 실험 데이터를 수집하여 최종의 데이터 셋을 생성하는 4개의 단계로 구분된다. 전처리 과정 중 2, 3번 단계에 해당하는 부분이 논문에서 제안하는 방법에 해당한다.

본 논문에서는 데이터마이닝에서 발생할 수 있는 프라이버시 침해 분쟁을 최소화하기 위해 메일의 헤더부



[그림 1] 본 논문의 실험에 대한 전반적인 구조도

분만을 사용하며, 그 중에서도 메일의 목적이 가장 많이 드러나는 메일 제목에 대해 N-GRAM 색인화를 사용하여, 분류를 위한 특징을 도출한다. 그러나 메일 제목이 없는 등, 메일 제목만으로 특징을 도출할 수 없는 상황이 발생할 수 있다. 따라서, 메일 제목에 대한 특징 이외에 받는 사람 리스트에 대한 유사도를 측정하여, 스팸 메일 필터링에 대한 추가적인 특징으로 사용한다. 받는 사람 리스트에 대한 유사도 측정을 하는 이유는 대부분의 스팸 메일 발송기가 하나의 타겟 서버를 대상으로 동시에 여러 계정에 발송하는 특징을 가지기 때문이다.

3.2 전처리 과정

이메일 데이터를 수집하여, 스팸 메일 필터링을 하기 위한 일련의 과정은 다음과 같다.

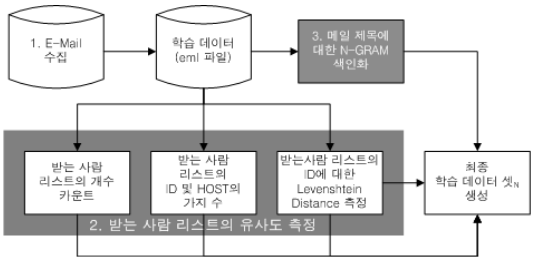
단계 1 : 스팸 메일 필터링을 하기 위해 정상 메일과 스팸 메일을 수집한다. 수집된 메일은 인터넷 메시지 포맷이다[17].

단계 2 : 수집된 메일에서 각 메일마다 받는 사람의 ID 리스트를 추출하여 받는 사람 리스트의 수, 받는 사람 ID의 수, 받는 사람 HOST의 수, Levenshtein distance 등을 추출하여, 각 메일 데이터에 대한 데이터 셋을 생성한다.

단계 3 : 수집된 메일의 모든 제목에 대해 N-GRAM 색인화를 실시한 뒤, 각 메일 데이터에서 색인화 된 특징의 발현 횟수를 카운트 하여, N-GRAM 색인화 특징 벡터에 대한 데이터 셋을 생성한다.

단계 4 : 동일한 학습 데이터에 대해 단계 2와 단계 3의 결과로 생성된 받는 사람 리스트의 유사도 측정값과 메일 제목에 대한 N-GRAM 색인화 값들을 병합 최종의 학습 데이터 셋을 생성한다. 또한 같은 특징 벡터에 대한

실험 데이터 셋도 생성하여 최종 분류 실험을 진행한다.



[그림 2] 전처리 단계 상세

[그림 2]는 앞서 설명한 전처리 단계를 좀 더 자세히 나타낸 것이다. [그림 2]에서 2. 받는 사람 리스트 유사도 측정은 단 한번만 실행하여, 학습 데이터 셋을 위한 유사도 평가 값을 도출한다. 그리고 N이 2라고 가정 후, 메일 제목에 대한 N-GRAM 색인화를 하여 유사도 평가 값과 병합 후 스팸 메일 분류 실험을 실행한다. N이 2일 때의 스팸 메일 분류 실험이 끝나면 N을 3으로 증가 시키고 메일 제목에 대한 N-GRAM 색인화를 다시 실행하여, 앞서 구한 받는 사람 리스트의 유사도 측정 결과와 병합 후 다시 스팸 메일 분류 실험을 진행한다. 이와 같은 단계를 반복해서 실행한다.

[표 2] 각 특징에 대한 설명 (총 5 + M 개)

특징	설명
분류	스팸 메일 혹은 정상 메일로 분류
받는 사람 리스트의 수	받는 사람 + 참조 + 숨은 참조의 리스트 수
받는 사람 ID 의 수	받는 사람 리스트의 ID 의 가지 수
받는 사람 HOST 의 수	받는 사람 리스트의 HOST 의 가지 수
받는 사람 ID 의 유사도	받는 사람 ID 리스트의 유사도 측정
메일 제목에 대한 N-GRAM 색인화	메일 제목에 대한 N-GRAM 색인화 결과로 생성된 특징

위의 전처리 단계를 실행하여 얻어지는 데이터 셋의 각 데이터는 [표 2]와 같은 특징 벡터를 가진다.

3.2.1 Levenshtein distance 를 통한 받는 사람 ID 리스트의 유사도 측정

이 단계는 [그림 1]에서 단계 2에 해당하며, [표 2]에서 설명한 특징 중, 받는 사람 ID의 유사도를 구하기 위해 Levenshtein distance를 이용하는 방법을 설명한다.

[그림 3]은 ID 리스트에 대한 Levenshtein distance의 적용을 설명하기 위해 임의로 작성한 ID 리스트의 예이다. 전자 메일에서 받는 사람, 참조, 숨은 참조 등에 나타나는 메일 주소의 순서대로 리스트의 ID 문자열이 입력되며, 이 리스트에서 ID들의 유사성을 측정하기 위한 식을 다음과 같이 정의하였다.

```
ID = { “randoll”, “hongildong”,
                “spyrr”, “spdfkdh” }
```

[그림 3] Levenshtein distance 적용을 위한 ID 리스트의 예

$$\begin{cases} \text{if } N<2,\ S(ID)=0 \\ \text{if } N\geq 2,\ S(ID)=\sum_{i=1}^{N-1} D(id_i, id_{i+1}) \end{cases} \quad (7)$$

[표 3] 그림 3의 리스트를 식 (7)에 적용한 결과

i	id_i	id_{i+1}	Levenshtein distance
1	randoll	hongildong	7
2	hongildong	spyrr	10
3	spyrr	spdfkdh	5
총 계 (측정된 ID 리스트의 유사도)			22

식 (7)에서 $S()$는 유사도를 도출하는 함수를 의미하며, $D(a,b)$는 a와 b 사이의 Levenshtein Distance를 의미한다. id_i는 [그림 3]의 리스트 내에서 i번째 ID 문자열을 의미한다. [표 3]은 [그림 3]과 같은 리스트가 있을 때, ID 리스트가 존재하는 순서대로 Levenshtein distance를 구하여 이들의 합으로 유사도를 표현한 것이다.

3.2.2 메일 제목에 대한 N-GRAM 색인화

이 단계는 [그림 1]에서 단계 3 및 단계 4에 해당하는 부분으로 메일 제목에 대한 N-GRAM 색인화를 위한

세부 단계는 다음과 같다.

단계 1 : 메일 제목에 대한 N-GRAM 색인화

단계 1.1 : 수집된 이메일 데이터 중 학습데이터로 사용할 데이터 셋에 포함된 모든 메일 제목을 병합한다. 이 때, 공백문자 ' '와 탭 문자, 개행문자는 삭제한다.

단계 1.2 : 병합된 메일 제목에 대해 N-GRAM 색인화를 실시한다.

단계 1.3 : N-GRAM의 윈도우 사이즈인 N 값을 2에서 목표하는 값 N 까지 단계 2에 대해 반복 수행한다.

단계 2 : 받는 사람 리스트에 대한 전처리 결과와 병합하여 수집된 메일 데이터에 대한 최종 데이터 셋 N-1개의 학습 및 실험 데이터 셋 쌍을 생성한다.

메일 제목을 병합할 때, 제목에서 공백문자 ' '와 탭 문자, 개행문자는 삭제하되 이외의 기호 및 특수문자는 삭제하지 않도록 한다. 제목에 대한 N-GRAM 색인화를 진행할 때, 바이트 단위를 특징 벡터를 생성한다. 다음은 각 N 값에 대하여 특징 벡터를 생성하는 예이다.

[표 4] 메일 제목에 대한 N-GRAM 색인화 결과

N	특징 1	특징 2	특징 3	특징 4	특징 i	특징 MN
2	“20”	“00”	“08”	“8X”	…	“구”
3	“200”	“008”	“08X”	“8년”	…	“X구”
⋮	⋮	⋮	⋮	⋮	⋮	⋮

[표 4]의 특징 벡터에서 “X"로 표시한 부분은 2바이트 문자인 한글 코드가 깨진 것을 표현한 것이다. N에 대해 생성된 각 특징 벡터는 받는 사람 리스트에 대한 전처리 결과와 교사학습을 위해 미리 결정된 범주 등과 병합하여, 최종 학습 및 실험 데이터 셋 쌍 N-1개를 생성한다.

Ⅳ. 실험 및 실험 결과

실험에 사용된 데이터는 10명이 국내외 대형 포탈 사이트와 교내 메일 서버에서 일주일 가량 수집한 것으로 [표 5]와 같으며, 총 2061개의 메일을 학습 데이터 셋 70%, 실험 데이터 셋 30%로 나누어 사용하였다.

뉴스레터에 등록된 메일과 스팸 메일은 받는 사람의 필요에 의해 메일링 리스트에 가입하여, 주기적으로 혹은 사용자 기호에 맞는 정보 제공을 목적으로 하기 때문에 사용자가 원치 않는 불특정 다수의 인물에게 보내지는 스팸 메일과 차이가 있어 정상 메일로 분류하였다.

[표 5] 실험에 사용된 데이터

메일 분류	학습 데이터	실험 데이터	총 계
정상 메일	1162	497	1659
스팸 메일	282	120	402
총 계	1444	617	2061

베이지안 분류기는 Weka 3.5.7[11]을 사용하였으며, LibSVM 2.8[18]을 사용하여 SVM 분류 실험을 진행하였다.

4.1 실험

스팸 메일 분류를 위한 전처리는 앞서 제안하는 방법과 같이 이루어지며, 최종적으로 도출된 학습 데이터 셋의 특징 벡터는 [표 2]와 같으며, 도출된 특징 벡터에 대한 실험 데이터의 특징을 도출한 결과로 얻어진 실험 데이터 셋과 학습 데이터 셋을 이용하여 스팸 메일 분류 실험을 진행한다.

스팸 메일 분류 실험을 위해서 베이지안 분류기와 SVM을 사용하였다. SVM 실험에서는 C-SVM을 사용하였으며, Radial kernel을 사용하였다. 또한, Gamma 값으로 0.0001에서 0.001까지 변경하며, 실험하였다.

4.2. 실험 결과

[표 6]은 본 논문에서 제안하는 방법에 의하여, 추출된 데이터를 사용한 베이지안 분류기를 사용한 결과와 해당 confusion matrix를 나타낸 것이다. 베이지안 결정이론을 적용하여 스팸 필터링을 실시한 결과 N이 2일 때, 스팸과 정상을 가장 정확하게 분류하였다. 그러나 스팸 메일 필터링에서는 스팸 메일을 정상 메일로 분류

[표 6] 베이지안 분류기를 통한 문류율 및 confusion matrix

N	분류율 (%)	정상 (%)	스팸 (%)	
2	**94.44**	97.0	3.0	**정상**
		16.5	83.5	**스팸**
3	93.18	97.8	2.2	정상
		26.1	73.9	스팸
4	93.03	98.6	1.4	정상
		30.0	70.0	스팸
5	91.90	99.0	1.0	정상
		37.5	62.5	스팸
6	90.28	99.2	0.8	정상
		46.7	53.3	스팸
7	**88.65**	**99.6**	**0.4**	**정상**
		56.7	**43.3**	**스팸**
8	87.03	99.6	0.4	정상
		65.0	35.0	스팸
9	86.22	99.6	0.4	정상
		69.2	30.8	스팸
10	84.93	99.6	0.4	정상
		75.8	24.2	스팸

[표 7] SVM을 통한 탐지율 및 confusion matrix (Radial Kernel, Gamma:0.008)

N	분류율 (%)	정상 (%)	스팸 (%)	
2	**94.76**	**100.0**	**0.0**	**정상**
		27.8	**72.2**	**스팸**
3	94.32	100.0	0.00	정상
		29.4	70.6	스팸
4	92.70	100.0	0.00	정상
		37.5	62.5	스팸
5	91.41	100.0	0.00	정상
		44.2	55.8	스팸
6	89.95	100.0	0.00	정상
		51.7	48.3	스팸
7	87.84	100.0	0.00	정상
		62.5	37.5	스팸
8	87.20	100.0	0.00	정상
		65.8	34.2	스팸
9	85.90	100.0	0.00	정상
		72.5	27.5	스팸
10	85.58	100.0	0.00	정상
		74.2	25.8	스팸

하는 것 보다, 정상 메일을 스팸 메일로 분류하는 것이 매우 치명적으로 작용하게 된다. [표 6]에서 N이 2일 때 False- Positive(FP)가 3% 정도 나타났으며, N이 7일 때, FP가 0.4%로 나타났다. 그러나 FP가 최소화 되도록 하기 위해서 N = 7을 선택한다면, N=2일 때 보다, 정상적인 분류율은 보다 약 6% 감소하고, False- Negative는 40.2%의 증가를 피할 수 없게 된다.

[표 7]은 추출된 동일한 데이터 셋을 이용하여 SVM 분류기에 적용한 결과와 confusion matrix를 나타낸 것이다. 이때 사용한 커널 함수는 Radial 함수이며, Gamma 값으로 0.008을 설정하였다. SVM 결과는 [표 6]의 베이지안 분류기에 대한 실험과는 다른 결과를 보여준다. N이 2일 때, 스팸 메일 분류율이 가장 높은 94.76%로 나타났으며, FP 및 FN역시 가장 작게 나타난다.

그러나 두 실험 결과 모두 분류율이 최상일 때, 스팸 메일을 정상 메일로 분류하는 False-Negative(FN)가 15%~30% 사이로 나타나는 것을 확인할 수 있다. 잘못 분류된 스팸 메일 대부분은 정상 메일로 분류한 메일링 리스트에 등록된 쇼핑 및 뉴스, 정보 관련 메일과 유사한 제목을 가지고 있었다.

Ⅴ. 결 론

본 논문에서는 최근 들어 점점 심각한 보안 문제로 부각되고 있는 스팸 메일 필터링을 위한 연구를 수행하였다. 스팸 메일이 한 번에 여러 사람에게 전송된다는 특성을 적용하기 위해 받는 사람 ID 리스트에 대한 Levenshtein distance을 통해 ID의 유사도를 측정하였다. 또한 문법적으로 파괴된 문장이 가독 상에서 해석될 수 있고, 스팸 메일들이 이러한 방법을 통해 기존의 스팸 메일 필터링 알고리즘을 회피하기 때문에, 이를 해결하기 위해 단어 사전 및 어휘 분석을 제외한 모든 메일 제목에 대해 N-GRAM 색인화를 수행하여, 스팸메일 필터링에 적용하는 방법을 제안하였다.

서로 다른 성격을 가지는 분류기에 대해 미치는 영향을 알아보기 위해 베이지안 분류기와 SVM 분류기에 대해 실험하였다. 실험 결과로 베이지안 분류기에 대한 결과로 N이 최소일 때 분류율이 가장 높게 나타났다. 그러나 스팸 메일에서 치명적으로 작용할 수 있는 요소인 FP가 분류율이 최고일 때, FP도 최소로 나타나지는 않았다. 신뢰할 수 있는 스팸 메일 필터링 시스템을 위해서는 이 값이 최소가 되도록 해야하기 때문에, FP가 최소로 나타나는 N이 7일 때의 데이터 셋을 이용한다면, 분류율 감소 및 FN 증가는 피할 수 없게된다. 반면,

SVM을 사용하였을 때는 N에 관계없이 FP가 전혀 나타나지 않았으며, N이 가장 작을 때 가장 높은 분류율을 가지는 것을 알 수 있다.

그러나 베이지안 및 SVM을 사용한 분류 실험 모두 최대의 분류율을 가질 때, FN이 15% ~ 30% 사이로 나타나는 것을 확인하였다. 본 논문의 실험에서는 사용자가 메일링 리스트에 등록하여 발송된 쇼핑 및 뉴스, 기타 정보 메일을 정상 메일로 분류하여 학습 데이터를 생성하였는데, 실험을 통해 정상 메일로 잘못 분류된 스팸메일의 대부분이 메일링 리스트에 의해 발송된 메일과 비슷한 메일 제목을 가지고 있었다. 이런 메일에 대한 분류는 White List 기반의 분류나 서로 다른 특징을 가지는 특징을 추출 및 생성해내는 방법이 필요하다.

본 논문에서 제안한 제목에 대한 N-GRAM 색인화의 결과로 매우 많은 개수의 특징 변수를 얻게 된다. 본 논문에서 사용한 데이터 셋의 N-GRAM 색인화 결과로 최소 5,000 여 개에서 최대 58,000 여 개의 특징을 도출하였다. 따라서 최종의 데이터 셋을 도출하고, 분류하는데 많은 시간과 비용이 소비된다. 향후에는 N-GRAM의 색인화 결과로 얻어지는 특징 변수를 최소화하면서 분류율을 최대화 할 수 있는 방안에 대한 연구가 필요하다.

참고문헌

[1] A. Bratko, “FIGHTING SPAM WITH DATA COMPRESSION MODELS”, Virus bulletin, http://www.virusbtn.com, pp. s1-s4, Mar 2006.

[2] G. V. Cormark, “Email Spam Filtering: A Systematic Review”, Foundations and Trends in Information Retrieval, 1(4), pp. 335-455, 2008.

[3] H. Lee, A.Y. Ng, “Spam deobfuscation using a Hidden Markov Model”, Proceedings of the Second Conference on Email and Anti-Spam (CEAS05), July 2005.

[4] I. Cid, L. R. Janerio, J. R. Méndez, D. Glez- Peña, F. Fdez-Riverola, “The Impact of Noise in Spam Filtering: A Case Study”, Advances in Data Mining. Medical Applications, E- Commerece, Marketing, and Theoretical Aspects, 8th Industrial Conference (ICDM 2008), Springer-verleg, LNCS 5077, pp. 228-241, 2008.

[5] S. Cucerzan, E. Brill, “Spelling correction as an iterative process that exploits the collective knowledge of web users”, In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 293-300, 2004.

[6] V. Freschi, A. Seraghiti, A. Boliolo, “Filtering Obfuscated Email Spam by means of Phonetic String Matching”, 28th European Conference on IR Research (ECIR 2006), Springer-verleg, LNCS 3936, pp. 505-509, 2006.

[7] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals.”, Soviet Physics Doklady, 10(8), pp. 707-710, 1966.

[8] S. Theodoridis, K. Koutroumbas, Pattern recognition 3/E, Academic press, pp. 13-116, 2006.

[9] V. Kumar, M. Steinbach, P. N. Tan, Introduction to Data Mining, Addison-Wesley, 2005.

[10] R. O. Duda, D. G. Stork, P. E. Hart, Pattern Classification 2/E, Wiley-Interscience, 2000.

[11] I. H. Witten, E. Frank, Data Mining: Practical machine learning tools and techniques 2/E, Morgan Kaufmann, 2005.

[12] 한학용, 패턴인식 개론 : MATLAB 실습을 통한 입체적 학습, 한빛미디어, 2005.

[13] 김현준, 정재은, 조근식, “가중치가 부여된 베이지안 분류자를 이용한 스팸 메일 필터링 시스템”, 한국정보과학회논문지 : 소프트웨어 및 응용, 31(8), pp. 1092-1100, 2004.

[14] 강승식, “메일 주소 유효성과 제목-내용 가중치 기법에 의한 스팸 메일 필터링”, 한국멀티미디어학회논문지, 9(2), pp. 255-263, 2006.

[15] 서정우, 손태식, 서정택, 문종섭, “n-Gram 색인화와 Support Vector Machine을 사용한 스팸메일 필터링에 대한 연구”, 정보보호학회논문지, 14(2), pp. 23-33, 2004.

[16] 공미경, 이경순, “스팸성 자질과 URL 자질의 공동 학습을 이용한 최대 엔트로피 기반 스팸메일 필터 시스템”, 한국정보처리학회논문지 (B), 15B(1), pp. 61-68, 2008.

[17] P. Resnick, “Internet Message Format”, RFC

Editor, 2001.
[18] C. C. Chang, C. J. Lin, “LibSVM - A Library for Support Vector Machines”, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

<著 者 紹 介>

이 호 섭 (Ho-Sub Lee) 학생회원
2006년 2월 : 동국대학교 컴퓨터학과 학사
2006년 3월~현재 : 고려대학교 정보경영공학전문대학원 석사과정
<관심분야> 패턴인식, 시스템 보안, 네트워크 보안

조 재 익 (Jae-Ik Cho) 학생회원
2005년 2월 : 동국대학교 컴퓨터학과 학사
2008년 2월 : 고려대학교 정보경영공학전문대학원 석사
2008년 3월~현재 : 고려대학교 정보경영공학전문대학원 박사과정
<관심분야> 네트워크 모델링, 패턴인식

정 만 현 (Man-Hyun Jung) 학생회원
2006년 2월 : 동국대학교 컴퓨터학과 학사
2006년 3월~현재 : 고려대학교 정보경영공학전문대학원 석사과정
<관심분야> 패턴인식, 시스템 보안, 네트워크 보안

문 종 섭 (Jong-Sub Moon) 정회원
1981년 2월~1985년 : 금성 통신 연구소 연구원
1991년 : Illinois Institute of technology 졸업(전산학 박사)
1993년~현재 : 고려대학교 전자 및 정보공학부 교수
<관심분야> 생체인식, 침입탐지, 운영체제