情報保護學會論文誌
第14券 第2號, 2004. 4

n-Gram 색인화와 Support Vector Machine을 사용한 스팸메일 필터링에 대한 연구

서 정 우,[a)†‡] 손 태 식,[a)] 서 정 택,[b)] 문 종 섭[a)]
고려대학교,[a)] 국가보안기술연구소,[b)]

A study on the Filtering of Spam E-mail using n-Gram indexing and Support Vector Machine

Jung-Woo Seo,[a)†‡] Tae-Shik Shon,[a)] Jung-Taek Seo,[b)] Jong-Sub Moon,[a)]
Korea University,[a)] National Security Research Institute,[b)]

요 약

인터넷 환경의 급속한 발전으로 인하여 이메일을 통한 메시지 교환은 급속히 증가하고 있다. 그러나 이메일의 편리성에도 불구하고 개인이나 기업에서는 스팸메일로 인한 시간과 비용의 낭비가 크게 증가하고 있다. 이러한 스팸메일에 대한 문제들을 해결하기 위하여 많은 방법들이 연구되고 있으며, 대표적인 방법으로 키워드를 이용한 패턴매칭이나 나이브 베이지안 방식과 같은 확률을 이용한 방법들이 있다.

본 논문에서는 기존의 연구에 대한 문제점을 보완하기 위하여 패턴 분류문제에 있어서 우수한 성능을 보이는 Support Vector Machine을 사용하여 정상적인 메일과 스팸메일을 분류하는 방안을 제시하였으며, 특히 n-Gram을 사용하여 생성된 색인어와 단어사전을 학습데이터 생성에 사용함으로서 효율적인 학습을 수행하도록 하였다. 결론에서는 제안된 방법에 대한 성능을 검증하기 위하여 기존의 연구 결과와 비교함으로서 제안된 방법의 성능을 검증하였다.

ABSTRACT

Because of a rapid growth of internet environment, it is also fast increasing to exchange message using e-mail. But, despite the convenience of e-mail, it is rising a currently big issue to waste their time and cost due to the spam mail in an individual or enterprise. Many kinds of solutions have been studied to solve harmful effects of spam mail. Such typical methods are as follows : pattern matching using the keyword with representative method and method using the probability like Naive Bayesian.

In this paper, we propose a classification method of spam mails from normal mails using Support Vector Machine, which has excellent performance in pattern classification problems, to compensate for the problems of existing research. Especially, the proposed method practices efficiently a learning procedure with a word dictionary including a generated index by the n-Gram. In the conclusion, we verified the proposed method through the accuracy comparison of spam mail separation between an existing research and proposed scheme.

Keywords: *Spam Mail Filtering, n-Gram Indexing, Support Vector Machine*

접수일: 2003년 11월 5일; 채택일: 2004년 3월 8일
* 본 연구는 대학 it연구센터 육성 지원 사업에 의해 수행되었습니다.
† 주저자, ‡ 교신저자 : korea002@korea.ac.kr

Ⅰ. 서 론

인터넷 사용자가 늘어나면서 이메일 사용자는 크게 증가하였다. 전화와 달리 이메일 시스템을 이용하면 수취인의 부재와 관계없이 메일을 보내거나 송신자가 원할 때에 메일을 보낼 수 있다. 특히, 같은 내용의 이메일을 많은 사람에게 동시에 보내는 경우 수신처를 복수로 지정하거나 그룹화 시켜서 동시에 발송할 수 있다. 하지만, 웹 페이지의 게시판이나 뉴스그룹에서 획득한 이메일주소 리스트를 이용하여 상업적인 내용이나 원하지 않는 이메일을 무차별적으로 발송하는 문제점이 있는데 이를 스팸메일 이라고 한다. IT 시장조사 전문기업인 IDC의 조사자료에 의하면 전세계적으로 연간 1조9천6백억 통의 스팸메일이 유통되고 있는 것으로 전망하고 있으며, 이는 전체 이메일 중 스팸메일이 차지하는 비중이 40%를 넘을 것으로 추정하고 있다. 이는 지난 2001년 8%에 불과했던 것이 최근 5배나 늘어난 수치이다. 결국, 스팸메일은 개인 및 기업에게 스팸메일 삭제에 엄청난 비용 및 시간의 부담을 준다.[1]

스팸메일 제거의 효율적인 방법은 스팸 메시지를 자동적으로 제거해주는 도구를 개발하는 것이다. 이와 같은 도구를 스팸필터라고 하며, 내용기반의 필터링의 경우는 메시지안에 특정 키워드 패턴들이 존재하는지를 검색하거나 발송 이데일주소와 스패머(Spammer)들의 블랙리스트를 비교함으로써 스팸메일을 필터링 한다. 그리고 전자문서 분류에 대한 연구에 많이 적용되는 나이브 베이지안 분류(Naive Bayesian Classification)의 경우는 문서내의 단어들을 대상으로 확률적인 방법을 적용하여 분류하기 때문에 특정 패턴을 따르지 않는 스팸메일을 걸러낼 수 있다.[2]

본 논문에서는 이진 분류 문제에 있어서 효율성과 정확성을 제공하고 있는 SVM(Support Vector Machine)을 사용하여 스팸메일 여부를 판정하였으며, 이진 패턴 분리의 정확성을 증대시키기 위하여 n-Gram을 적용한 색인어와 단어사전을 생성하였다. 생성된 단어사전은 이메일의 제목과 매칭하여 얻어진 이진값을 SVM의 입력 노드로 적용한다. SVM은 통계적 학습이론에 기반한 방법으로 경험적 위험을 최소화하는것이 아니라 구조적 위험을 최소화하는 이진 패턴 분리를 위한 알고리즘이다.[3,4,5]

본 논문의 구성은 다음과 같다. 2장에서는 스팸메일 분류를 위한 관련된 연구에 대하여 설명하고, 3장과 4장에서는 n-Gram 기반의 색인 방법 및 SVM의 개요 대하여 알아보며, 5장에서는 스팸메일 필터링 방안을 제안하고, 6장에서는 SVM을 통한 실험결과를 설명한다. 마지막으로 7장에서 결론 및 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

2.1 나이브 베이지안 분류(Naive Bayesian Classification)[2,6]

나이브 베이지안 기법은 확률 모델로서 이미 알고 있는 지식을 사전 지식으로 사용하여 학습 목표인 조건부 확률을 계산하는 베이지 정리에 기초를 두고 있다. 확률 모델에서 베이즈 정리(Bayes' theorem)와 전확률(Total probability)로부터 분류하려는 문서에 단어 W_1, W_2, W_3, ..., W_n이 출현할 경우를 사건 E, 문서가 범주 C_j에 분류되는 것을 C_j라고 하면, 문서가 범주에 분류될 확률은 다음과 같다.[2,6]

$$P(C_j|W_1,\ldots,W_n)=\frac{P(W_1,\ldots,W_n|C_j)\cdot P(C_j)}{P(W_1,\ldots,W_n)} \quad (1)$$

나이브 베이지안 분류에서는 한 문자가 특정 카테고리에 포함될 때 그 문서에서 나타나는 단어를 독립적이라고 가정하고, 범주의 할당도 상호 배타적이라고 가정한다. 이러한 가정에 의하여 공식을 표현하면 식 (2)와 같다.

$$P(C_j|W_1,\ldots,W_n)=\frac{P(C_j)\times\prod_i P(W_i|C_j)}{P(W_i)} \quad (2)$$

나이브 베이지안 분류는 속성값들이 주어진 목적값에 조건부 독립적이라는 가정을 기반으로 한다. 하지만 특정 단어를 선택할 확률보다 선행단어와 연관된 후위단어를 선택할 확률이 더 크다. 이러한 가정에도 불구하고 나이브 베이지안 분류는 효율적인 역할을 수행한다.

2.2 k-nearest-neighbor 알고리즘을 사용한 메모리 기반 분류(Memory-based classification)[2,7]

메모리 기반 방법들의 공통적인 특징은 메모리 구조안에 모든 학습 인스턴스를 저장하고 있으며, 직접적인 분류를 위하여 사용한다. 메모리 구조는 인스턴스안에 특성들에 의하여 다차원 공간으로 정의되며, 각 학습 인스턴스는 공간상에서 점과 같이 나타난다. 메모리 기반 분류 절차는 일반적으로 k-nearest-neighbor(k-nn) 알고리즘을 사용하는데, k-nn 알고리즘은 "Discriminatory Analysis: Non-parametric Discrimination: Consistency Properties"[11]란 논문에서 처음으로 발표되었으며, 다양한 클래스들 사이에서 유사점을 가지는 데이터들을 단순히 분리하는 역할을 한다. 이 방법은 분류할 인스턴스를 $a_i(x) = a_1(x), a_2(x), \ldots, a_n(x)$로 나타내고, 클래스별 훈련 인스턴스 $a_r(x_i)$의 거리를 식 (3)과 같이 계산한 다음, 분류대상 인스턴스 $a_r(x_i)$와 가장 가까운 k개의 훈련 인스턴스 $a_r(x_i)$를 선정한다.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \quad (3)$$

선정된 k개 중에서 가장 많은 개수의 인스턴스 예제가 소속된 클래스로 분류대상 인스턴스 $a_r(x_j)$가 분류된다. 식 (3)에서 r은 클래스의 종류이며, n은 클래스의 개수이다. k-nn 알고리즘은 일반적인 목적 함수(target function)를 학습하는 기계 학습 방법을 사용하는 것과는 다르게 예제들만을 색인하는 것으로 모든 학습 과정이 끝나며, 문서 분류 시에는 입력 문서와 유사한 k개의 예제들을 이용하여 문서의 범주를 할당한다.[7]

III. n-Gram 기반의 색인 방법 연구

3.1 Support Vector Machine 개요

SVM 이론에 따르면, 패턴 인식을 위한 전통적인 기법들이 경험적인 위험을 최소화하는데 n-Gram 색인 방법은 어절 단위 색인법에서의 복합 명사 띄어쓰기 문제를 완화할 수 있으며, 형태소 단위 해석에서와 같은 복잡한 문장 해석규칙이나 언어 정보의 개발을 요구하지 않는 색인 방법이다. 표 1에서는 n-Gram 방법을 이용한 색인 과정을 나타낸다[5].

표 1. 일반적인 n-Gram 기반의 색인 과정

단계 1 : 문서나 질의 내의 모든 어절들을 인식한다.
단계 2 : 불용어를 제거한다.
단계 3 : 각 어절에서 비색인 분절들을 절단한다.
단계 4 : 나머지 색인 분절을 n-Gram들로 분할하여 색인어로 선정한다.

n-Gram 기반의 색인 방법은 검색효과의 측면에서 다음과 같은 장점이 있다. 첫째, n-Gram 기반의 색인법은 어절 단위 색인법을 이용할 때의 절단 오류로 인한 파급 효과를 완화한다. 둘째, 복합 명사의 띄어쓰기 문제를 완화한다. 셋째, 철자 오류나 일관성 없는 외래어 표기 문제를 적절히 극복할 수 있다.

따라서, 이메일 제목에 포함된 특수문자나 공백들의 삭제로 인하여 재구성된 이메일 제목을 단어사전과 매칭하기 위하여 n-Gram 기반 색인 방법을 사용한다.

IV. SVM 연구

전통적인 기법들이 경험적인 위험을 최소화하는데 기초한 반면, SVM(Support Vector Machine)은 구조적인 위험을 최소화하는 것에 기초하고 있다. 여기서 경험적 위험의 최소화는 훈련 집단의 수행도를 최적화하려는 노력을 말하고, 구조적 위험의 최소화는 고정되어 있지만 알려지지 않은 확률분포를 갖는 데이터에 대해 잘못 분류하는 확률을 최소화하는 것을 말한다.[4]

두 클래스에 속하는 학습 벡터의 집합을 선형적으로 분리 가능하도록 하는 문제를 생각해 보면, 가중치 벡터 ω와 바이어스 b로 구성되는 $(w_o^T \cdot x) + b_o = 0$의 초월면(hyperplane)을 가지도록 훈련 데이터 셋(training data set) $\{(x_i, d_i)\}_{i=1}^{N}$를 학습시키는 것을 나타내며, 여기서 x_i는 입력 패턴이고, d_i는 목표값이 된다. 초월면 $(w_o^T \cdot x) + b_o = 0$는 식 (4)의 조건을 만족하게 된다.

$$\exists w, b \quad s.t. \begin{cases} w^T x_i + b > 0 & \text{for} \quad d_i = +1 \\ w^T x_i + b < 0 & \text{for} \quad d_i = -1 \end{cases} \quad (4)$$

식 (4)에서 등호의 조건을 만족하는 입력패턴들 중에서 결정 표면(decision surface)에 가장 가까이 위치한 패턴들을 support vector라고 하며, 개념적으로 이 벡터들은 초월면에 가장 가까이 위치하여 분류하기가 어려운 벡터들이다. 따라서 분류를 위한 학습은 제약조건 식 (5)을 만족하는 최적의 초월면을 찾는 것이다. 이것은 제약조건을 가지는 최적화 문제로 훈련 데이터 셋 $\{(x_i, d_i)\}_{i=1}^{N}$이 주어질 때 최적의 초월면을 위한 최적의 파라미터 w와 b를 찾는 Quadratic 문제이다.

$$\begin{cases} Minimize \quad \Phi(w) = \frac{1}{2}\|w\|^2 \\ s.t. \quad d_i(w^T x_i + b) \geq 1 \quad \text{for } i = 1, \ldots, N \end{cases} \tag{5}$$

여기서 최적은 최대 마진(margin)을 가지는 것이며, 최대 마진 초월면은 최적으로 두 개의 클래스를 분리할 수 있는 초월면이다. 결국 최적의 선형 분리 경계면을 $g(x) = w_o^T \cdot x + b_o$로 놓으면, support vector와 g(x)의 거리를 $1/\|w\|$로 나타낼 수 있으며, 입력패턴을 최적으로 분류하는 초월면은 식 (6)과 같이 비용함수 $\Phi(w)$를 최소화한다.

$$\Phi(w) = \frac{1}{2}\|w\|^2 \tag{6}$$

식 (6)의 비용함수는 w의 블록함수이며, 제약조건 식 (5)는 w에 선형임을 확인할 수 있다. 지금까지 서술된 분류를 위한 SVM을 정리하면, 학습 패턴이 주어질 때 제약조건 식 (5)를 만족하는 가중치 벡터 w와 바이어스 b를 찾는 최적화 문제로 생각할 수 있으며, 이때 $\|w\|^2$을 최소화하여 분리 간격을 최대화하도록 하여 최적 분리면을 찾아낸다. 이 최적화 문제를 해결하기 위하여 라그랑제(Lagrange) 계수법을 이용하면 식 (7)과 같은 라그랑제 함수 $L(\chi, b, \alpha)$을 얻을 수 있다.

$$L(\chi, b, \alpha) = \frac{1}{2} w^T w + \sum_{i=1}^{n} a_i [1 - d_i(w^T x_i + b)], \quad a_i \geq 0, i = 1..n \tag{7}$$

식에서 a_i는 라그랑제 계수들이며, 최적화 문제에 대한 해는 χ와 b에 대해서는 최소화되며, $a_i \geq 0$에 대해서는 최대화되어야 한다. 따라서 χ와 b에 대한 $L(\chi, b, \alpha)$의 최소는 그 각각에 대한 미분으로 얻어질 수 있다.

$$\begin{aligned} \frac{\partial L}{\partial b} = 0 &\Leftrightarrow \sum_{i=1}^{n} a_i d_i = 0 \\ \frac{\partial L}{\partial w} = 0 &\Leftrightarrow w - \sum_{i=1}^{n} a_i d_i x_i = 0 \end{aligned} \tag{8}$$

식 (8)에서 a_i를 구하기 위해 기본 문제에 대한 라그랑제 함수 $L(\chi, b, \alpha)$를 이원문제(Dual problem)의 목적함수 Q(α)로 표현하면 식 (9)와 같이 나타낸다.

$$Q(\alpha) = \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j d_i d_j K\langle x_i, x_j \rangle$$
$$s.t. \quad a_j \geq 0, \ i = 1, \ldots, n \text{ and } \sum_{i=1}^{n} a_i d_i = 0$$

식 (9)의 목적함수는 일반적으로 Quadratic Programming 문제의 형태로 학습패턴의 항으로만 구성되며, 이때 커널함수는 $K\langle x_i, x_j \rangle = \langle \Phi(x_i), \Phi(x_j) \rangle$로 표현된다. 그러므로, 분류문제를 식 (9)의 이원문제로 생각하면, 이는 학습패턴 $\{(x_i, d_i)\}_{i=1}^{N}$이 주어질 때, 제약조건 $\sum_{i=1}^{n} a_i d_i = 0$와 $a_i \geq 0$ $(i = 1, 2, \ldots, n)$을 만족하는 목적함수 식 (9)를 최대화하는 라그랑제 계수 a_i를 찾는 것이다. 그러므로, Quadratic Programming 알고리즘에 따라 제약조건 식 (5)에서 목적함수 식 (9)를 최대로 하는 최적의 라그랑제 계수 a_i를 찾으면 최적의 가중치 벡터 w는 식 (8)에 의하여 계산될 수 있고, 최적의 바이어스 b는 support vector로부터 계산될 수 있다. 가중치 벡터와 바이어스에 대한 계산식은 식 (10)과 같이 나타낸다.

$$\begin{aligned} w &= \sum_{i=1}^{n} a_i d_i x_i \\ b &= -\frac{1}{2} w[x_r x_s] \end{aligned} \tag{10}$$

여기서 x_r과 x_s는 식 (11)의 조건을 만족하는 support vector들이다.

$$a_r, \ a_s > 0, \ d_r = 1, \ d_s = -1 \tag{11}$$

이때, SVM에 의한 분류식을 정리하면 식 (12)가 선형의 결정면을 가짐을 알 수 있다.

$$f(x) = sgn(w \cdot x + b) = sgn\left[\sum_{i=1}^{n} a_i d_i K(x_i \cdot x) + b\right] \quad (12)$$

여기서 sgn(·)의 ·이 양수이면 +1이고, 그렇지 않으면 -1을 갖는 함수이며, 식 (12)에서 정의된 커널함수는 다음과 같은 함수 중 선택될 수 있다.

- Dot kernel : k(x, y) = x * y
- Polynomial kernel : k(x, y) = (x * y + 1)d
- Radial kernel : k(x, y) = exp(-ϒ||x - y||2)
- Neural kernel : k(x, y) = tanh(ax * y + b)

하지만 선형으로 분류 가능하지 않는 문제에 대해서도 분류 가능하게 하는 일반화된 초월면을 구성하기 위해서 음수가 아닌 스칼라 변수 $\xi_i \geq 0$을 갖게 되는데, ξ_i는 잘못된 분류와 관계된 오차의 척도로 슬랙변수(slack variables)이다. 따라서 분류 불가능한 경우를 위한 슬랙변수 ξ_i를 포함하는 제약조건은 식 (5)를 식 (13)과 같이 변경함으로서 구할 수 있다.[8][9][12]

$$d_i(w \cdot x_i) + b \geq 1 - \xi_i, \quad i = 1, 2, \ldots, n \quad (13)$$

제약조건을 만족하는 가중치 벡터 w와 슬랙변수 ξ_i를 포함하는 비용함수 $\tau(w, \xi)$는 식 (14)와 같이 나타낼 수 있다.

$$\tau(w, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \quad (14)$$

이때 C는 학습 오차와 일반화 사이에 상관관계를 제어하는 양의 값을 갖는 파라미터이다. 본 논문에서 제안된 방법을 테스트하기 위하여 사용한 파라미터 C값으로 다양한 변수를 테스트함으로서 최적의 학습 오차를 갖는 값을 설정하였으며, SVM의 커널 함수로는 dot와 polynomial 그리고 RBF 커널 함수를 사용하였다.

V. 스팸메일 필터링 방안

기존의 내용기반 이메일 필터링의 경우 특정 단어의 패턴 매칭을 통해 필터링을 수행하지만, 스패머들에 의해 제목이나 본문의 내용이 변형된 형태의 스팸메일이 전송될 경우 이를 효과적으로 필터링하지 못할 수 있으며, 패턴의 수가 증가할수록 스팸메일을 필터링하기 위한 시간도 비례적으로 증가하는 문제점들이 존재한다. 이와 같은 문제점들을 해결하기 위해서 본 논문에서는 n-Gram을 사용하여 생성된 색인어와 단어사전의 매칭을 통하여 얻어진 데이터 셋을 SVM 분류기에 적용함으로서 스팸메일 필터링을 효율적으로 수행하도록 했다. n-Gram 색인화와 SVM을 사용한 스팸메일 필터링은 다음과 같은 과정을 수행한다.

5.1 스팸메일 필터링 전체 구성

본 논문에서 제안된 방법에 대한 실험을 위하여 다수의 사용자 이메일 계정으로부터 정상적인 메일(legitimate mail)과 스팸메일(spam mail)을 수집했고, 수집된 이메일에서 데이터 셋을 생성하기 위해 이메일의 제목부분만을 추출하여 텍스트 문서로 저장했다. 저장된 문서로부터 단어사전과 n-Gram 색인어를 생성하게 되는데, 단어사전의 경우는 이메일에 포함된 단어의 빈도수를 적용하여 수동으로 생성하고, 색인어 생성은 이메일의 제목에서 n-Gram을 적용하여 자동으로 색인어를 생성한다. 여기서 생성된 색인어와 단어사전을 매칭하여 학습 데이터 셋과 테스트 데이터 셋을 구성한다.

전처리 과정에서 생성된 학습 데이터 셋은 SVM 분류기의 학습과정을 수행하고, 테스트 데이터 셋은 스팸메일 여부를 테스트하기 위한 테스트 패턴으로 입력함으로서 정상메일 및 스팸메일 여부를 판정하게 된다.

n-Gram 색인어와 SVM을 사용한 스팸메일 필터링에 대한 전체 구성 모듈은 그림 1과 같다.

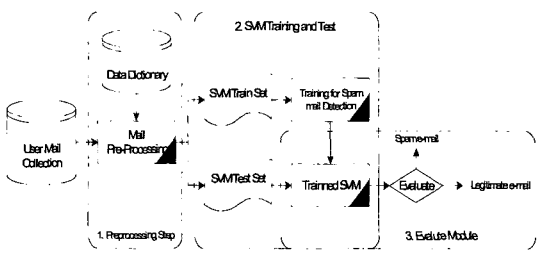


그림 1. 스팸메일 필터링 전체 구성도

5.2 스팸메일 필터링을 위한 전처리 과정

5.2.1 단어사전 생성

단어사전의 생성은 이메일의 제목에서 사용되는 단어의 빈도수를 조사하여 빈도수가 높은 단어 m개를 추출하여 생성한다. 생성된 단어사전에는 대출관련단어와 성인사이트 관련 단어들이 설정되어 있다. 단어사전의 색인은 SVM의 입력노드의 수와 일치하는 m개의 단어로 구성되어 있으며, n-Gram을 적용하여 생성된 색인어와 단어 사전에서 정의된 단어를 매칭 하는데 사용된다. 표 2는 생성된 단어사전의 예제를 보여준다.

표 2. 단어사전 생성 예제

사전의 색인	1	2	3	4	5	···	n
단어사전	성인 광고	몰카 자료	동영상	포르노	성방	···	샘플
사전의 색인	101	102	103	104	105	···	m
단어사전	카드 연체	허가 업체	연체금	서비스	연체	···	결제

5.2.2 n-Gram 색인어 생성

수집된 이메일에 대하여 학습 데이터 셋이나 테스트 데이터 셋을 생성하기 위하여 우선적으로 수집된 이메일에 n-Gram을 적용하여 색인어를 생성하게 되는데, 색인어 생성은 표 3과 같이 수집된 이메일을 사용하여 2Gram부터 4Gram까지 적용함으로서 색인어를 생성한다. 여기서 생성된 색인어와 표 2에서 생성한 단어사전을 매칭하여 학습 데이터 셋이나 테스트 데이터 셋을 구성하게 된다.

표 3. *n*-Gram 기반의 색인어 생성

수집된 이메일 제목 = {[광고]#카드증액을 통한 연체대금 대출받는 방법}
2-gram 적용 색인어 = {광고, 고카, 카드, 드증, 증액, 액을,....., 방법}
3-gram 적용 색인어 = {광고카, 고카드, 카드증, 드증액을,....., 는방법}
4-gram 적용 색인어 = {광고카드, 고카드증, 카드증액,....., 받는방법}

5.2.3 SVM 데이터 셋 생성

그림 2는 n-Gram을 적용하여 생성된 색인어와 단어사전을 적용하는 과정을 나타내고 있는데, 만약 n-Gram을 적용한 색인어와 단어사전의 단어가 일치하게되면 표 4의 예제와 같이 1(matching)의 값을 갖게되고, 단어사전에서 일치하는 단어를 찾지 못

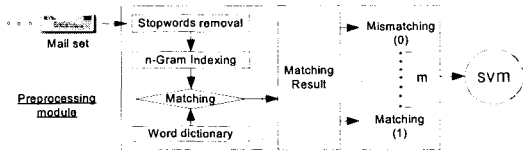


그림 2. n-Gram 색인화 및 SVM 적용을 위한 전처리 과정

표 4. 색인어와 단어사전의 매칭 결과

단어사전 인덱스	1	2	3	4	5	6	···	200	201
메일 셋(1)	1	0	1	1	0	0	···	1	0
메일 셋(2)	1	1	0	1	0	1	···	0	1
메일 셋(n-1)	·····								
메일 셋(n)	0	1	0	1	1	1	···	0	1

할 경우에는 0(mismatching)의 값을 가지게된다. 표 4는 n-Gram을 적용하여 생성된 색인어와 단어사전을 매칭하여 얻어진 결과값의 예제를 나타낸다.

표 4와 같은 매칭 결과값을 사용하여 학습 데이터 셋이나 테스트 데이터 셋을 구성한 후 SVM의 입력벡터로 사용하게 된다. 이때 SVM 분류기를 사용하여 학습과정을 수행하거나 테스트 데이터 셋을 사용하여 이메일에 포함된 스팸메일을 필터링한다.

VI. SVM을 통한 스팸메일 필터링 실험 및 결과

6.1 스팸메일 필터링 전체 구성

스팸메일 필터링 실험을 위하여 다수의 사용자 이메일 계정(아웃룩 및 상용 이메일)으로부터 데이터를 수집하였으며, 수집된 이메일은 5장에서 설명한 전처리과정을 통하여 SVM 적용을 위한 학습 데이터 셋과 테스트 데이터 셋을 구성하였다. 각 데이터 셋

은 정상메일과 스팸메일로 구성되며, 수집된 이메일 데이터에서 성인사이트와 대출관련 이메일을 스팸메일로 분류하고 나머지는 정상메일로 분류하여 구성하였다. 표 5에는 스팸메일 필터링을 위한 SVM 학습 데이터 집합을 나타내고 있으며, 학습 데이터 셋의 개수는 1000개로 구성되어 있다. 1000개의 학습 데이터 중 정상메일과 스팸메일은 각각 500개로 구성되었다.

표 5. SVM 학습을 위한 학습 데이터

DataSet / 메일종류	Training Set (총 1000개)
정상 메일	대출 및 성인 사이트 관련 이메일을 제외한 정상메일 (500개)
스팸 메일	대출 및 성인 사이트 관련 스팸메일 (500개)

학습 데이터에 의하여 학습된 결과가 정확하게 스팸여부를 판정할 수 있는지를 실험하기 위하여 테스트 데이터 셋을 구성해야 한다. 테스트 데이터 셋의 생성방법은 5장에서처럼 수집된 이메일을 단어사전과 매칭시켜 얻은 결과 값으로 구성되며, 정상메일 500개와 스팸메일 500개로 구성되었다.

표 6. SVM 테스트를 위한 테스트 데이터

DataSet / 메일종류	Test Set (총 1000개)
정상 메일	대출 및 성인 사이트 관련 이메일을 제외한 정상메일 (500개)
스팸 메일	대출 및 성인 사이트 관련 스팸메일 (500개)

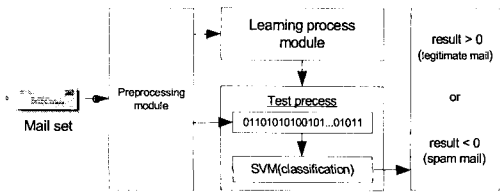


그림 3. SVM 학습을 통한 스팸메일 필터링 과정

본 논문에서 제안한 스팸메일 필터링을 위한 전체 구성은 그림 3과 같으며, 스팸메일 필터링을 위하여 mySVM 공개 라이브러리[10]를 사용하여 실험을 수행하였다. 7장에서는 mySVM 파라미터들에 따른 실험결과를 비교하였다.

6.2 스팸메일 필터링 실험 결과

제안된 방법을 테스트하기 위하여, 5장과 같은 스팸메일 필터링 시스템을 구현하였다. 구현된 시스템에 의하여 수집된 이메일에 대한 가장 효율적인 스팸메일 분류 방법을 살펴보고, 각 분류 방법에 대한 성능을 측정하였다. 테스트를 수행하기 위하여 실제로 2000개의 이메일을 수집하였으며, SVM 학습을 위한 학습 데이터 셋으로 1000개를 구성하였고, 테스트 데이터 셋으로 1000개의 이메일을 사용하였다. 여기서 정상메일과 스팸메일은 비율은 각각 50%의 비율로 구성하였다. 테스트 환경에서 사용한 커널의 종류는 dot와 polynomial 그리고 radial등 세 개의 커널함수를 사용하여 테스트를 수행하였다. 커널함수에 따른 스팸메일 필터링 성능은 다음과 같다.

6.2.1 dot 커널을 적용한 실험 결과

첫 번째 테스트의 커널함수로 dot를 사용하였고, SVM의 입력 노드로 단어사전과 일치한 m개의 노드를 사용하였다. 테스트 데이터 셋의 증가에 따른 스팸메일 필터링 비율은 그림 4와 같이 나타난다. 여기서 발견할 수 있는 중요한 특징은 테스트 셋이 500개를 넘어가면서 정상적인 스팸메일 필터링 비율은 감소하는 반면 오탐지 비율과 미탐지 비율은 증가하는 것을 확인할 수 있는데, 그 이유는 성인사이트를 광고하기 위한 이메일의 경우는 SVM을 통하여 학습된 결과와 상이한 데이터 셋이 정의되는 경우가 많기 때문이다.

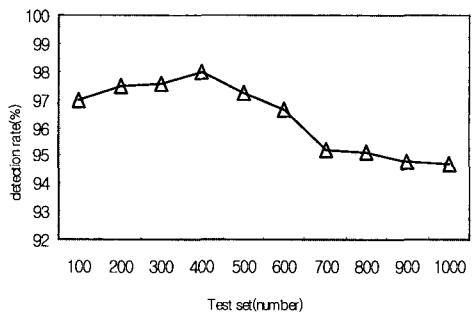


그림 4. dot 커널을 적용한 테스트 결과

6.2.2 polynomial 커널을 적용한 실험 결과

polynomial 커널의 경우는 파라미터 값을 포함하고 있으며, 식 (15)와 같이 파라미터 값을 조정하면서 테스트를 수행할 수 있다. 파라미터는 정수값을 가지며, 파라미터의 degree 값을 조정하면서 테스트를 수행하였다.

$$k(x,y) = (x*y+1)^d \quad (15)$$

polynomial 커널을 사용한 테스트에서는 dot 커널 방식과 비슷한 오탐지율과 미탐지율을 보였지만 스팸메일 필터링 비율은 좀더 우수한 테스트 결과를 확인할 수 있었으며, polynomial 커널을 사용한 전체적인 테스트 결과는 그림 5와 같다. 그림 5의 테스트 결과에서 확인할 수 있듯이 파라미터의 degree 값을 1로 설정한 경우가 3으로 설정한 경우보다 조금 향상된 결과를 확인할 수 있었으며, degree 값을 3 이상으로 설정하였을 경우는 오탐지율과 미탐지율의 증가로 결과값 도출이 불가능하였다.

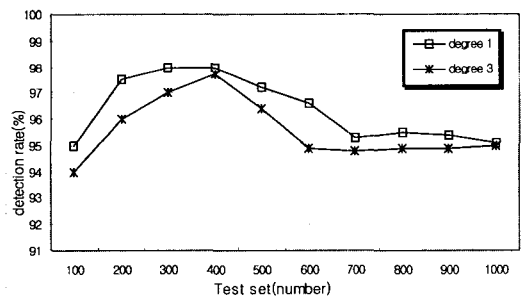


그림 5. dot 커널을 적용한 테스트 결과radial 커널을 사용한 방식의 스팸메일 필터링 결과는 그림 6과 같다. 테스트를 위한 gamma 값은 0.01과 0.1 그리고 0.5의 세 가지의 파라미터 값을 설정하였다.

6.2.3 radial 커널을 적용한 실험 결과

마지막으로 적용한 커널 함수로 radial를 사용하였으며, 식 (16)에서와 같이 커널 함수가 정의된다. 파라미터는 실수값을 가지며, 파라미터의 gamma 값을 조정하면서 테스트를 수행하였다.

$$k(x,y) = \exp(-\gamma\|x-y\|^2) \quad (16)$$

radial 커널을 이용한 테스트에서도 dot나 polynomial 커널방식을 사용한 테스트 결과와 비슷한

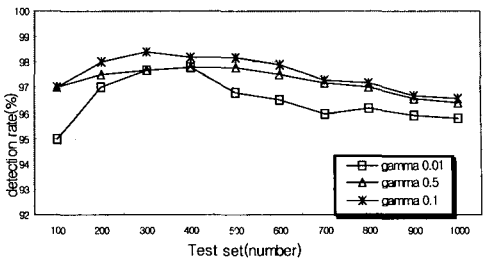


그림 6. radial 커널을 적용한 테스트 결과

오탐지율과 미탐지율 보였지만, 전체적인 스팸메일 필터링 결과에서는 가장 우수한 성능을 보였다.

radial 커널을 사용한 방식의 중요한 특징은 gamma 값을 0.5로 설정하였을 경우에 미탐지율이 가장 우수하였지만 상대적으로 오탐지율이 증가하는 문제점이 있었다. 또한 gamma 값을 0.01로 설정하였을 경우에 스팸메일 탐지율은 dot나 polynomial 커널 방식을 사용한 스팸메일 탐지율과 비슷한 결과를 얻었지만 다른 gamma 값과 비교하여 약간 떨어지는 성능을 보였다. 각 커널함수와 파라미터에 따른 테스트 결과는 표 7과 같이 나타난다.

스팸메일 분류 성능 평가는 정확도(spam precision)와 재현율(spam recall)로 나타냈다. 결과적으로, 표 7에서와 같이 radial 커널을 적용한 방식이 dot나 polynomial 커널을 사용한 방법보다 스

표 7. n-Gram 기반의 색인어 생성

Kernel definition	feature	parameter	Test set (1000개)			
			FP	FN	spam recall	spam precision
dot kernel	201	N/A	1.5	3.8	94.7	96.8
polynomial kernel	201	degree1	1.5	3.4	95.1	96.8
		degree3	1.4	3.6	95.0	97.0
radial kernel	201	gamma0.5	2.3	1.1	96.6	97.9
		gamma0.1	2.0	1.1	96.9	97.9
		gamma0.01	2.0	2.1	95.9	95.8

* spam precision(%) = #actual spammail / #classified spammail
spam recall(%) = #actual spammail / #total spammail[7]
FP(%) = false positive, FN(%) = false negative.

팸메일 필터링 테스트 결과에서 우수한 성능을 보였으며, 특히 gamma를 0.1 설정했을 경우에 가장 우수한 스팸메일 필터링 성능을 보였다.

6.3 스팸메일 필터링 성능 비교

본 논문에서 제안한 방법과 기존의 스팸메일 필터링 방법들과의 성능을 비교하기 위하여 기존에 발표된 논문[6][13]의 결론 부분을 인용하였으며, 스팸메일 필터링을 위한 테스트 필터는 나이브 베이지안 방식과 키워드 패턴을 사용한 방법이 사용되었다. 논문[13]에서 제안한 방법은 성능 측정을 위하여 사용자 계정으로부터 2815개의 이메일이 수집하였으며, 정상메일과 스팸메일의 비율을 80%와 20%로 구성하여 스팸메일 필터링을 진행하였다.

표 8. 스팸메일 필터링 성능 비교를 위한 분류 결과

Filter used	parameter	false positive (%)	false negative (%)	spam recall (%)	spam precision (%)
SVM	gamma0.1	2.0	1.1	96.9	97.9
keyword patterns	N/A	N/A	N/A	53.01	95.15
Naive Bayesian		5.0	8.0	95.8	93.0

제안된 방법과 기존 방식의 성능 평가 방법은 스팸메일 필터링 정확도(spam precision)와 스패메일 재현율(spam recall)을 사용하였데, n-Gram 색인어와 SVM을 적용한 방식이 스팸메일 필터링 정확도면에서 키워드 패턴을 사용한 방식보다 좋은 성능을 보였으며, 나이브 베이지안 방식과는 비슷한 결과를 확인하였다. 또한 스팸메일 재현율에서는 나이브 베이지안과 키워드 패턴을 사용한 방법들보다 우수한 필터링 성능을 보였고, 제안된 방법을 이용한 스팸메일 필터링 적용 커널함수에 따라 약간의 차이는 있지만, 평균 약 2초 정도의 시간이 소요되었다. 위와 같이 n-Gram 색인어와 SVM을 사용한 스팸메일 필터링을 수행하였을 경우 스팸메일 필터링 수행에 있어서 우수한 성능을 확인할 수 있었으며, 스팸메일 재현율 에서도 기존 방법들보다 향상된 결과를 확인하였다.

VII. 결론

본 논문에서는 최근 들어 큰 사회적 문제로 대두되고 있는 스팸메일 필터링을 위한 연구를 수행하였으며, 또한 n-Gram 색인어와 SVM(Support Vector Machine)을 사용한 스팸메일 필터링 방안을 제안하였다. 제안된 방법에 대한 테스트를 위해서 다수의 사용자 계정으로부터 수집된 이메일을 사용하여 테스트를 진행하였으며, n-Gram에 의하여 생성된 색인어와 단어사전에 의하여 생성된 데이터 셋을 SVM 분류기에 적용함으로서 스팸메일을 필터링 하였다. SVM 분류기에 사용된 커널 함수로는 dot, polynomial, radial 방식을 사용하였으며, 각 커널 방식에 대한 테스트 결과는 표 7에서 설명하였다.

SVM 분류기의 커널함수로 radial 방식을 사용하였을 경우 가장 우수한 성능을 보였으며, 나이브 베이지안 방식이나 메모리 기반 방식과 같은 기존 연구들의 비교에서도 스팸메일 필터링 성능의 향상을 보였다(표 8 참조). 하지만 성인광고의 경우 SVM 분류기에 의하여 학습된 결과값을 벗어나는 벡터들의 비율이 증가함으로서 필터링 성능이 떨어지는 문제점이 있는데, 이에 대한 해결책으로 이메일에 포함된 성인사이트 URL 주소를 패턴 매칭 방식으로 검출함으로서 해결할 수 있을 것이다.

이번 연구에서는 가장 대표적인 스팸메일 분류인 성인광고와 대출에 관련된 광고에 한정하여 스팸메일 필터링을 수행하였지만, 향후에는 다양한 형태의 스팸메일을 탐지하는 것이 필요하다.

참 고 문 헌

[1] http://kr.fujitsu.com/webzine/dream/special _report/20030708_specialreport/special_0307.html

[2] Androutsopoulos. I, Koutsias. J, Konstantinos V. Chandrinos, Constantine D. Spyropoulos, "An Experimental Comparison of Naive Bayesian and Keyword-Based Anti- Spam Filtering with Personal E-mail Messages" 23rd ACM International Conference on Research and Development in Information Retrieval, pp. 160-167, 2000.

[3] Campbell. C and Cristianini. N,

"Simple Learning Algorithms for Training Support Vector Machines", Technical report, University of Bristol, 1998.

[4] Cristianini N., Shawetaylor. J, "An Introduction to Support Vector Machines", Cambridge University, 2000.

[5] 이준호, 안정수, 박현주, 김명호, "한글 문서의 효과적인 검색을 위한 n-Gram 기반의 색인 방법", 정보관리학회지, pp. 47-63, July 1996.

[6] Ion. A, Georgios. P, Vangelis. K, Georgios. S, Constantine. D, "Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach", PKDD 2000, pp. 1-13, Sep. 2000.

[7] 임희석, 남기춘, "경험적 정보를 이용한 k-nn 기반 한국어 문서 분류기의 개선", 컴퓨터교육학회지 논문지, Vol. 5, No. 3, pp. 37-44, 2002.

[8] Takuya. I, and Shigeo. A, "Fuzzy Support Vector Machines for Pattern Classification", 2001.

[9] Cortes. C and Vapnik. V, "Support Vector Networks", Machine Learning, pp. 273-297, Sep. 1998.

[10] Joachmims. T, "mySVM - a Support Vector Machine", Univerity Dortmund.

[11] Fix, E. and Hodges, J.L., "Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties," Report No. 4, Project No. 21-49-004, 1959.

[12] 조용현, "모멘트를 이용한 Support Vector Machines의 학습성능 개선", 한국정보처리학회 논문지, Vol. 7, No. 5, pp. 1446-1455, May 2000.

[13] Mehran. S, Susan. D, David. H, Eric. H, "A Bayesian Approach to Filtering Junk E-Mail", In AAAI-98 Workshop on Learning for Text Categorization, 1998.

[14] Xavier. C, and Lluis. M, "Boosting Trees for Anti-Spam Email Filtering", 4th International Conference on Recent Advances in Natural Language Processing, pp. 58-64, 2001.

[15] William W. Cohen, "Learning Rules that Classify E-Mail", AAAAI Spring Symposium: Machine Learning in Information Access, pp. 124-143, March 1996.

[16] Pontil. M, and Verri. A, "Properties of Support Vector Machines", A.I. Memo No. 1612: CBCL paper No. 152, Massachusetts Institute of Technology, Cambridge, 1997.

[17] Yanlei. D, Hongjun. L, and Dekai. W, "A Comparative Study of Classification Based Personal E-mail Filtering", 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'00), pp. 408-419, 2000.

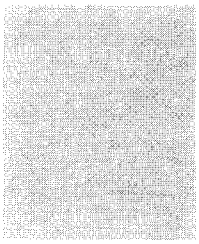
[18] Joachmims. T, "Text Categorization with Support Vector Machine: Learning with Many Relevant Features", European Conference on Machine Learning, pp. 137-142, 1998.

[19] R¨uping. S, "MySVM-Manual", University of Dortmund, Lehrstuhl Informatik VIII, 2000.

〈著者紹介〉

서 정 우(Jung-Woo Seo)
2002년 2월: 호남대학교 정보통신 공학부 졸업 (공학사)
2003년 3월~2004년 2월: 고려대학교 정보보호기술연구센터 연구원
2004년 2월: 고려대학교 정보보호대학원 졸업(공학석사)
2004년 2월~현재: (주)삼성전자
〈관심분야〉 시스템/네트워크 보안, 생체인식, 신경망, 영상처리

손 태 식 (Tae-Shik Shon)
2000년 2월: 아주대학교 정보 및 컴퓨터 공학부(학사)
2002년 2월: 아주대학교 정보통신전문대학원 정보통신공학과(석사)
2004년 2월: 고려대학교 정보보호대학원 정보보호학과(박사 수료)
2003년 9월~ 12월: 서경대학교 정보통신공학과 강사
2003년 5월~ 12월: 한국정보보호교육센터 강사, ICU 부설 정보통신교육원 강사
2002년 8월~현재: 고려대학교 정보보호기술연구센터 연구원
2004년 2월~현재: Dept. of Computer Science, University of Minnesota 객원연구원
〈관심분야〉 네트워크 보안, 패턴인식, 신경망, 리눅스 보안



서 정 택(Jung-Taek Seo)
1999년 2월 충주대학교 컴퓨터공학과 졸업(공학사)
2001년 2월 아주대학교 대학원 컴퓨터공학과 졸업(공학석사)
2000년 11월 ~ 현재 ETRI 부설 국가보안기술연구소 선임연구원
〈관심분야〉 정보전, 시스템/네트워크 보안, 취약점 분석·평가

문 종 섭(Jong-sub Moon)
1981년 2월: 서울대학교 계산통계학과 학사
1983년 2월: 서울대학교 계산통계학과 석사
1992년 2월: Illinois Institute of Technology 박사
1993년~현재 고려대학교 전자 및 정보공학부 교수
고려대학교 정보보호대학원 겸임 교수
〈관심분야〉 IDS, 신경망, 생체인식, 운영체제