

# 휴대폰의 스팸문자메시지 판별 시스템

이성욱<sup>o</sup>  
충주대학교  
leesw@mail.cjnu.ac.kr

## A Spam Message Filter System for Mobile Environment

Songwook Lee<sup>o</sup>  
Chungju National University

### 요 약

휴대폰의 광범위한 보급으로 문자메시지의 사용이 급증하고 있다. 이와 동시에 사용자가 원하지 않는 광고성 스팸문자도 넘쳐나고 있다. 본 연구는 이러한 스팸문자메시지를 자동으로 판별하는 시스템을 개발하는 것이다. 우리는 기계학습방법인 지지벡터기계(Support Vector Machine)를 사용하여 시스템을 학습하였으며 자질의 선택은 카이제곱 통계량을 이용하였다. 실험결과 F1척도로 약 95.5%의 정확률을 얻었다.

주제어: 스팸문자메시지 차단, 지지벡터기계, SVM, 카이제곱 통계량

### 1. 서론

이동통신기기의 발전과 더불어 휴대폰이 널리 보급되었고, 음성통화와 문자메시지 송수신 기능은 휴대폰의 핵심 기능이다. 그러나 문자메시지를 기업의 상품 홍보와 쇼핑물 광고, 선거운동 등에 이용할 목적으로 사용하고 있어 각 개인은 하루에도 수많은 스팸 문자를 수신하고 있다. 스팸 문자 메시지 중에는 불법적인 광고내용도 있고 최근에는 휴대폰 문자메시지를 이용한 결제사기도 빈번하다. 스팸문자로 인한 사회적 문제가 심각해지자 수신자의 사전 동의없는 문자메시지의 발송을 금지하는 법까지 생겨났으나 스팸문자는 여전히 발송되고 있다.

본 연구에서는 이러한 휴대폰에 수신되는 스팸문자를 판별하는 시스템을 제안한다. 스팸문자 판별은 이진분류 문제이며, 우리는 이진분류기 중에서 좋은 성능이 보고된 지지벡터기계를 이용하였고 자질의 선택은 카이제곱 통계량을 이용하여 선택하였다. 다음 그림 1은 제안하는 시스템의 구조도이다.

제안하는 스팸문자 판별 시스템은 크게 두 단계로 나뉜다. 먼저 학습단계에서는 학습용 스팸문자 데이터로부터 지지벡터기계의 학습에 사용할 수 있는 벡터 자질을 추출하여야 한다. 학습용 문자메시지는 먼저 형태소 분석기에 의해 품사가 부착된다. 우리는 자질로 어휘/품사 쌍과 음절정보를 사용하는데 그 중 유용한 자질들을 카이제곱 통계량을 이용하여 선택한다. 선택된 각각의 자질은 벡터의 차원을 구성하는 축을 이루며 각 자질의 가중치가 해당 차원의 값이 된다. 이렇게 하나의 문자 메시지는 다차원 공간의 한 점이 되고, 모든 학습 데이터의 벡터가 모두 구성되면 이들로 지지벡터기계를 학습한다.

적용단계에서는 학습 때와 마찬가지로 분류용 스팸문자 데이터는 형태소 분석 단계와 자질 추출 단계를 거쳐 다차원 공간상의 한 점을 이루는 벡터가 되고 이를 학습된 지지벡터기계가 스팸 또는 정상 문자 메시지로 분류한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구를 살펴보고, 3장에서는 지지벡터기계의 학습에 사용된 자질과 좋은 자질을 선택하는 방법을 설명한다. 4장에서는 실험을 통해 제안된 방법의 성능을 보이고 5장에서 결론을 내린다.

### 2. 관련연구

스팸문자 판별 문제는 내용이 비교적 짧다는 점을 제외하면 스팸 메일 판별 문제와 매우 유사한 문제이다. 스팸 메일 판별과 관련된 연구에서 [1]은 카이제곱 통계량을 이용하여 자질을 선택하였으며 지지벡터기계를 이용하여 시스템을 학습하였고, 나머지 대부분의 연구는 베이저안 분류기를 기반으로 하고 있으며[2-6], 그 외, 마코프 랜덤 필드 (Markov Random Field) 모델[7]과 k-Nearest Neighbor(k-NN) 방법[8]을 이용한 연구 등이 있다.

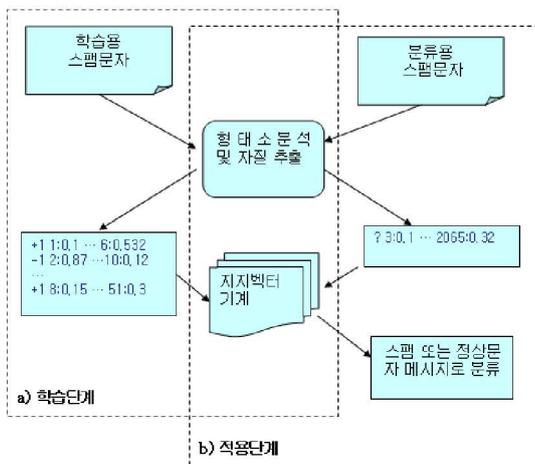


그림 1. 시스템 구조도

### 3. 자질과 카이제곱 통계량

우리는 수집된 휴대폰 문자 메시지들을 자체 개발한 형태소분석기를 이용하여 자동으로 품사를 부착하였으며, 품사가 부착된 어휘/품사 쌍을 기본 자질로 사용하였다. 휴대폰 문자 메시지의 특성상 한글맞춤법이 맞지 않는 단어의 출현이 많고 비교적 메시지의 길이가 짧은 특성 때문에 음절 n-그램 정보도 자질로 추가하였다(실험에서는 3-그램까지 사용). 따라서 가능한 자질의 종류는 수집된 문자 메시지에서 발견되는 모든 어휘/품사 쌍과 음절 n-그램이 되며, 많은 수의 자질이 나타나게 된다. 이 중에서 좋은 자질을 선택하기 위해 카이제곱 통계량을 이용해서 자질을 선택한다. 카이제곱 통계량을 계산하는 식은 다음과 같다[9].

$$\chi^2(f,s) = \frac{(A+B+C+D) \times (AD-BC)^2}{(A+B) \times (A+C) \times (B+D) \times (C+D)} \quad (1)$$

A는 스팸문자 s 중에 자질 f를 포함하고 있는 문서의 수이고, B는 범주 s 이외의 문서, 즉 정상문자에 속해 있는 문서 중에 자질 f를 포함하고 있는 문서의 수이다. 또한, C는 스팸문자 s에 속해 있는 문서 중에 자질 f를 포함하지 않는 문서의 수이며, D는 범주 s외의 문서 중에 자질 f를 가지고 있지 않는 문서의 수이다. 자질 f와 범주 s가 완전히 독립적이면 0의 값을 갖는다. 하나의 자질에 대해 카이제곱 통계량의 값을 결정하는 방법은 전체 범주에 대한 평균값을 사용하는 방법과 전체 범주에 대해 최대값을 사용하는 방법이 있을 수 있다. 우리는 이것을 이진 분류에 사용하므로 각 자질 당 하나의 카이제곱 값을 사용한다.

각각의 자질에 가중치를 부여하는 방법으로는 일반적으로 좋은 성능을 보이는 TF-IDF(Term Frequency-Inverse Document Frequency) 가중치 방법과 이진가중치 방법을 사용한다. 본 연구에 적용하기 위해 TF-IDF 값을 계산하는 경우, 용어(term)는 자질로, 문서(document)는 문자메시지로 범주(category)는 스팸문자와 정상문자로 간주하여 계산한다.

### 4. 실험 및 결과

본 연구에서는 지지벡터기계[10]의 학습을 위해 LIBSVM[11]을 이용하였고 선형 커널을 이용하여 학습하였다. 본 연구에 사용한 휴대폰 문자메시지 데이터는 본교 학생들이 수집한 것이며 스팸문자와 정상문자가 1:1의 비율로 구성된 1,160개의 문자메시지로 구성되어 있다. 4:1의 확률로 무작위 추출하여 287개의 데이터를 평가 데이터로 사용하고, 나머지를 학습 데이터로 사용한다.

다음 표 1은 자질의 종류에 따른 성능을 비교한 것이다. 음절 n-그램은 1-그램부터 3-그램까지 사용한 것이다. 표 1에서와 같이 음절 n-그램과 어휘/품사 자질을 같이 쓴 시스템의 성능이 더 좋았다. 제안 시스템을 후

대폰 등에 탑재하고자 할 때는 약간의 성능 손실을 감수 하더라도 음절 n-그램만을 자질로 사용하는 것이 좋을 것이다.

표 1. 자질의 종류에 따른 정확도

자질의 종류	정확도(%)	비고
음절 n-그램	92.0%	모든 자질
음절 n-그램, 어휘/품사	93.4%	

다음 표 2는 이진가중치와 TF-IDF가중치를 사용했을 때의 성능을 비교한 것이다. 본 실험에서는 TF-IDF 가중치보다 이진가중치를 사용하였을 때 더 좋은 결과를 가져왔다.

표 2. 가중치 표현에 따른 정확도

가중치	정확도(%)	비고
이진	93.4%	$\chi^2 > 4.5$
TF-IDF	86.8%	

표 3은 이진가중치로 자질을 표현하고, 자질의 개수를 카이제곱 값으로 제한했을 때의 정확도를 나타낸다. 카이제곱 값으로 자질의 개수를 제한하여 자질의 개수를 절반 이상 줄였을 때, 가장 좋은 결과를 보였으며 이는 카이제곱 값이 자질 선택에 유용하다는 것을 보여준다.

표 3. 자질의 개수에 따른 정확도 비교

자질의 개수	정확도(%)	비고
23,807	93.4	모든 자질
10,404	95.5	$\chi^2 > 1$
3,519	94.4	$\chi^2 > 2$
2,128	94.8	$\chi^2 > 3$

다음 표 4는 가장 좋은 성능을 보인 제안 시스템의 성능을 정확도(accuracy), 정확률(precision), 재현율(recall), Hm오류율, Sm오류율[12] 등으로 평가한 결과이다. Hm오류율과 Sm오류율은 다음과 같이 계산되며, 스팸메일 분류 등에서 사용하는 평가척도이다.

$$\begin{aligned} Hm(\%) &= \text{정상문자를 잘못 분류한 개수} / \text{정상문자의 수} * 100 \\ Sm(\%) &= \text{스팸문자를 잘못 분류한 개수} / \text{스팸문자의 수} * 100 \end{aligned}$$

일반적으로 정상문자를 잘못 분류한 Hm오류율이 Sm오류율보다 작은 시스템이 사용자에게 더 선호되는 시스템이며 그 이유는 정상문자를 스팸문자로 잘못 분류했을 때 사용자가 입는 피해가 더 크기 때문이다.

표 4. 제안 시스템의 성능

정확도	정확률	재현율	F1	Hm	Sm
95.5	95.5	95.4	95.5	3.2	5.9

표 4에서와 같이, 제안 시스템은 Hm오류율이 Sm오류율보다 더 낮은 결과를 보이며, 정확률과 재현율 모두 어느 한쪽에 치우치지 않은 결과를 보이고 있다.

시스템에서 발생한 대부분의 오류는 문자메시지에 특수문자가 포함된 경우와 한글맞춤법에 어긋난 단어가 포함된 문자메시지에서 발생했다. 이런 문제를 해결하기 위해서는 한글맞춤법을 수정하거나 특수 문자를 처리하는 전처리가 필요하다.

## 5. 결론

본 논문에서는 휴대폰의 스팸문자메시지를 판별하는 시스템을 제안하였다. 문자메시지의 어휘/품사 쌍과 음절 n-그램을 자질로 사용하였으며 카이제곱 통계량을 이용하여 유용한 자질을 선택하였다. 선택된 자질을 이진가중치로 표현한 후, 지지벡터기계를 학습하여 자동으로 스팸문자 메시지를 판별하는 시스템을 제안하였다. 실험에는 자체 수집한 데이터 집합을 이용하였으며, 실험 결과, F1 척도로 95.5%의 성능을 얻었다. 더 나은 성능을 위해서는 한글맞춤법 교정 등의 전처리 시스템이 필요하며, 더 많은 문자 메시지의 수집이 필요하다. 실험결과 형태소분석기를 이용하지 않고 n-그램 자질만을 이용하여도 형태소분석기를 사용했을 때와 크게 뒤떨어지지 않는 성능을 얻을 수 있었는데 이는 실제로 제안 시스템을 휴대폰 등에 탑재할 때에 형태소분석기도 탑재해야하는 부담을 줄일 수 있을 것이다.

## 참고문헌

[1] 이성욱, “카이제곱 통계량과 지지벡터기계를 이용한 스팸메일 필터”, 정보처리학회논문지, 제17-B권, 제3호, pp.249-254, 2010.

[2] Keselj, V., Milios, E., Tuttle, A., Wang, S., Zhang, R. “TREC 2005 Spam Track: Spam Filtering Using N-gram-based Techniques“, Proceedings of Text REtrieval Conference, 2005.

[3] 김현준, 정재은, 조근식, “가중치가 부여된 베이지안 분류자를 이용한 스팸 메일 필터링 시스템”, 정보과학회논문지, 31권 8호, 2004, pp.1092-1100.

[4] Segal, R. “IBM SpamGuru on the TREC 2005 Spam Track“, Proceedings of Text REtrieval Conference, 2005.

[5] Brakto, Al, Filipic, B., “Spam Filtering Using Character-Level Markov Models: Experiments for the TREC 2005 Spam Track“, Proceedings of Text REtrieval Conference, 2005.

[6] Breyer, L. A. “DBACL at the TREC 2005“, Proceedings of Text REtrieval Conference, 2005.

[7] Assis, F., Yerazunis, W., Siefkes, C., Chhabra, S., “CRM114 versus Mr. X: CRM114 Notes for the TREC 2005 Spam Track“, Proceedings of Text REtrieval Conference, 2005.

[8] Cao, W., An, A., Huang, X. “York University at TREC 2005: SPAM Track“, Proceedings of Text REtrieval Conference, 2005.

[9] Yang, Yiming and Jan O. Pedersen. A comparative study on Feature selection in text categorization. In proceedings of the 14th International conference on Machine Learning, 1997.

[10] V. Vapnik. The nature of statistical learning theory, Springer, NewYork, 1995.

[11] <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 2009.

[12] G. V. Cormack and T. R. Lynam. “TREC 2005 spam track overview,” The Fourteenth Text REtrieval Conference (TREC 2005) Proceedings, 2005.