제36회 한국정보처리학회 추계학술발표대회 논문집 제18권 제2호 (2011. 11)

발신번호 특징 및 음절단위 기계학습을 통한 모바일 스팸 SMS 필터링 시스템

유환일, 채동규, 임을규
한양대학교 공과대학 컴퓨터공학부
e-mail: stockholm2@hanyang.ac.kr

A Mobile Spam SMS Filtering System using Machine learning about syllable and the features of caller ID

Hwan-il You, Chae-Dong Kyu, Eul-Gyu Im
Division of Computer Science & Engineering, Hanyang University

요 약

본 논문에서는 스팸 SMS 발신번호와 메시지 텍스트의 특징을 기계학습한 스팸 필터링 시스템을 논한다. 최근 변화하는 스팸SMS에 대한 적응력을 위해서, 각 트레이닝 셋의 수신 텍스트를 음절단위로 분석 할 것을 제안한다. 그리고 기존의 분류기는 성능이 미흡하거나 구현의 복잡성으로 인해 실제로 스펨 필터엔진으로 활용되지 않는 점을 극복하기 위해서 보다 단순한 분류기를 사용한다.

제안하는 시스템은 트레이닝 셋의 발신번호 및 수신 텍스트의 음절단위를 빈도수와 묶어 학습데이터를 구성하고, 테스트 셋을 스팸적·논스팸적으로 분석하여 스팸일 확률을 계산한다. 또한 Naive baysian를 바탕으로 한 경계값 기반 분류기를 통해, 타 분류기에 비해 구현 및 활용면에서 실용성이 높으면서도 성능이 뒤처지지 않는 시스템을 제안한다.

1. 서론

기존의 스팸 필터링 시스템과 관련된 수많은 논문은 전자메일에 대한 필터링 시스템을 제안하고 있다. 하지만 이는 스팸 SMS는 전자메일과는 다른 구성을 갖고 있고, 전자메일과는 다른 특징을 바탕으로 지능화 되고 있으며, 그 빈도가 증가함에도 스팸 SMS를 위한 필터링 시스템은 크게 부족하다.

또한 스팸SMS는 필터링 시스템을 우회하기 위해, 지능적이고 변칙적으로 발전하고 있다. 하지만 기존의 스팸 SMS 필터링 시스템은 인간에 의해 수동적으로 입력된 특정 문자열이 있거나, 특정 국번에서 발송된 SMS는 모두 필터링 해버리는 아주 기초적인 지원 중이다.[1] 이러한 방식으로는 변화하는 스팸 SMS 패턴에 제대로 대처하기 힘든 실정이다.

본 논문에서는 스팸 SMS의 특징을 알아내기 위한 다양한 실험을 거듭하였고 해당 결과를 논하였다. 이를 바탕으로 발신번호와 수신텍스트에서 스팸 SMS적 특징이 있다고 판단하고, 발신번호의 특징 및 수신 텍스트를 기계학습하여, 분류 알고리즘에 의해 필터링을 수행하는 스팸 SMS 필터링 시스템을 제안한다.

또한 기존 논문에서는 지능화되고 변칙적으로 수신되는 스팸 SMS에 적응하기 위해 시소러스(Thesaurus) 사전 등의 멥핑 테이블을 활용하는 방법을 제안하고 있다.[2] 하지만 이는 사전에 입력되어 있지 않거나, 학습되어 있지 않은 패턴에 대해서 성능이 떨어진다. 그래서 해결방안으로 수신 텍스트를 음절단위로 분석하는 방식을 제안하여, 별도의 멥핑 테이블 없이 필터링하는 방안을 제안한다.

논문의 구성은 다음과 같다 1절은 본 연구의 배경과 필요성을 설명한다. 2절에서는 제안하는 시스템의 개요를 서술하였다. 3절은 제안하는 시스템의 설계 및 구현을, 4절은 해당 시스템의 성능을 확인한다. 마지막으로 5절에서는 결론 및 향후과제를 언급하였다.

2. 관련 연구

기본 아이디어는 나이브 베이지안 분류자(Naive baysian classifier)에서 가져온다. 나이브 베이지안 분류자는 타 알고리즘에 비해 구현이 용이하여 많은 스팸 필터링 시스템에서 사용 중인 분류 알고리즘이다.[3] 해당 알고리즘을 사용하는 스팸필터는 “스팸으로 판단된 문서에서 자주 출현한 단어를 많이 포함한 문서는 스팸을 가능성이 높다.”[4]는 아이디어로 스팸지수를 계산한다.

각 의미를 자세히 알아보자면, “사용자들이 스팸으로 등록한 트레이닝 셋에서(스팸으로 판단된 문서), 전체 트레이닝 셋에서 각 단어의 빈도수를 계산하는 훈련(Training) 과정을 통하여 스팸의 특성(Feature)으로 추출된 단어(자주 출현하는 단어)의 가중치들을 계산한 점수가 높을수록 스팸일 가능성이(많이 포함한 문서는 스팸일 가능성) 높다”[4]는 뜻이다.

본 논문에서는 수신텍스트의 단어 대신 음절단위로 분석한다. 직관적으로 봤을 때 음절 기반으로 Feature를 구성하면 기존의 연구들에 비해 False Positive가 늘어날 수 있다. 이와 같은 점을 해결하기 위해서 본 논문에서는 발신번호와 수신텍스트를 다각적으로 고려한다. 또한 스팸적, 논스팸적 Feature로 이중적으로 분석함으로써 False Positive 문제를 해결한다.

2.1. 수신 텍스트

수신 텍스트는 주로 광고를 목적으로 한다. 대출광고(47%), 일반광고(30%), 성인광고(23%) 순으로 구성되어 있다. 이에 일반광고는 신상품 광고, 이벤트 참여 광고 등이다.[5] 또한 텍스트의 크기는 80바이트(한글 40자)이내로 한정되어 있기 때문에, 어미나 조사 등이 생략된 경우가 많다.

이처럼 스팸 SMS의 한정된 크기와 목적 때문에, 텍스트로 집약적으로 구성되는 경우가 많고, 그래서 특정 음절의 출현 빈도가 높다.

또한 최근 스팸 SMS의 텍스트적 특징은 필터링 시스템을 통과하기 위해서 지능화되고 변칙적이다. 예를 들어 “대)출”, “대*출” 등과 같이 단어 사이에 특수문자를 삽입하기도 하고, “카쥐노”, “카지노”, “카즤노” 등 일부러 맞춤법을 비틀기도 한다. 하지만 의미 전달을 위해 모든 어휘의 음절이 변형되지 못하고, 특정 음절이 반복적으로 출현함을 확인할 수 있었다.

기존의 논문에서는 형태소 분석기를 통한 색인어 추출 후, 시소러스(Thesaurus) 사전을 통한 단어 표준화 과정을 거친다. 하지만 이는 시소러스 사전에 입력되지 않거나 학습되지 않은 패턴에 대해서 성능이 떨어진다.[2] 이러한 점을 보완하기 위해서 해당 논문에서는 단어 대신 음절단위로 분석하기로 하고 실험하였다.

음절단위 분석의 최적의 성능을 위해서, 테스트 셋의 SMS는 특수문자 및 숫자, 공백문자를 제외한 전처리 과정이 수행되었다.

2.2. 발신 번호

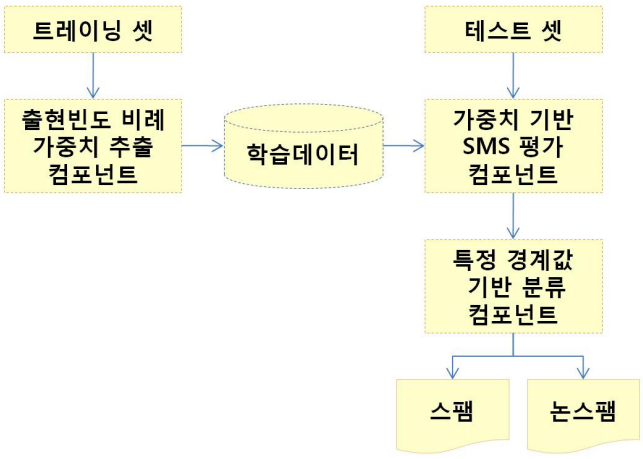
스팸 메시지(Spam SMS)의 발신번호의 주요 특징(Feature)을 알아보기 위한 실험을 수행하였다. 실험은 스팸 메세지 170개의 발신번호를 바탕으로 빈출하는 숫자의 시퀀스 및 주요 국번을 추출하였다.

주요 빈도수를 보인 02, 010, 070, 15xx, 16xx, 080등의 번호는 모두 전국대표번호 및 인터넷번호, 무료번호 등으로 스팸메세지가 주로 발송되는 국번들이다. 이 중 02와 010을 제외한 국번은 웹투폰 방식의 대량 문자발송 서비스를 통해 스팸 SMS 발송이 증가하고 있는 국번이다.[6]

그래서 스팸 메세지 130개와 논스팸 메세지 130개를 발신번호의 주요 국번(070, 060, 080, 15xx, 16xx)만으로 필터링의 성능을 테스트 한 결과, 스팸 메세지 130개 중 77%가 발신번호 만으로도 필터링 되었다. 하지만 논스팸 메시지도 주요 국번(070, 060, 080, 15xx, 16xx)에서 수신되는 경향이 있기 때문에 기존처럼 특정 국번이라고 모두 필터링하기에는 False positive가 치명적이었다. 그래도 발신번호의 특징의 스팸과 논스팸 차이가 다소 명확하기 때문에 출현빈도에 의해 가중치를 주는 방식으로 False nagative를 어느 정도 안정화 시키며 필터링 시스템의 성능 및 신뢰도를 향상시킬 수 있다.

3. 제안하는 시스템의 설계 및 구현

본 절에서는 제안하는 새로운 스팸 메세지 필터링 시스템을 주요 컴포넌트 별로 나누어 소개한다. 해당 시스템은 크게 3개의 컴포넌트로 구성된다.



(그림 1) 주요 프로세스와 컴포넌트

3.1. 시스템 프로세스

제안하는 시스템의 주요 프로세스는 아래와 같다.

사전에 학습데이터를 구축하는 과정이 필요하다. 출현빈도 비례 가중치 추출 컴포넌트에서 스팸・논스팸 Training set을 음절단위 및 발신번호 특징별로 분석, 출현빈도 비례 가중치(학습데이터)를 추출(구축)한다.

위 학습데이터를 바탕으로 필터링 과정을 수행한다. 우선 Test set(신규 SMS)를 수신시, 전처리 과정으로 숫자, 특수문자, 공백문자를 제거한다.

둘째, 가중치 SMS 평가 컴포넌트에서 추출된 학습데이터(음절별・발신번호별 가중치)를 바탕으로, Test set의 각 SMS을 평가한다. 그리고 각각의 평가된 값을 바탕으로 계산식에 대입하여 ‘스팸일 확률(P)’을 구한다.

셋째, 특정 경계값 기반 분류 컴포넌트에서 ‘스팸일 확률(P)’을 최적 경계값 (T)와 비교하여 필터링 한다.

3.2. 출현빈도 비례 가중치 추출 컴포넌트

해당 컴포넌트는 학습(Training)과정을 수행한다. 2절에서 발신번호 및 수신 텍스트에서 확인한 특징(Feature)들을 바탕으로, 스팸・논스팸 트레이닝 셋에서 등장하는 음절 및 발신번호 특징에 대한 빈도수에 비례하는 가중치를 나타내는 해쉬 테이블(학습데이터)을 구성한다. 이러한 학습데이터(해쉬 테이블)는 스팸 발신번호 특징(a), 논스팸 발신번호 특징(b), 스팸 음절 특징(c), 논스팸 음절 특징(d) 등 4개로

구성된다.

<표 1> 각 특징별 학습데이터 구분

구 분	스팸SMS의 특징	논스팸SMS의 특징
발신번호	a	b
음절	c	d

3.3. 가중치 기반 SMS 평가 컴포넌트

출현빈도 비례 가중치 추출 컴포넌트로부터 구축된 4개의 학습데이터(각 특징과 가중치의 쌍)를 바탕으로, 4개의 영역에서 테스트 셋을 평가한다.

수신된 SMS(테스트 셋)의 발신번호 및 메시지 텍스트의 음절들을 각 a, b, c, d 테이블 포함여부에 따라서 가중치 (a_w , b_w , c_w , d_w) 혹은 0의 합으로 계산되고 아래의 계산식에 의해 '스팸일 확률(P)'로 평가된다.

$$P_{spam} = \frac{\sum a_w + \sum c_w}{\sum a_w + \sum b_w + \sum c_w + \sum d_w} * 100$$

음절 단위 및 발신번호 특징을 종합적으로 (스팸 가중치)/(스팸 가중치 + 논스팸 가중치)*100으로 계산된다.

3.4. 특정 경계값 기반 분류 컴포넌트

테스트 셋 SMS의 스팸일 확률(P)이 계산이 되면 해당 컴포넌트에서 분류한다. 스팸일 확률(P)이 특정 경계값(T)를 기준으로 이상이면 스팸, 이하면 논스팸으로 분류・필터링하게 된다.

$$\text{스팸일 확률}(P) > \text{경계값}(T)$$

4. 제안하는 시스템의 성능

스팸 SMS는 일반 광고 및, 대출, 성인광고 등 의 텍스트를 지니면서 지능화된 자원을, 논스팸 SMS는 주로 모르는 번호로부터 수신될 가능성이 높은 공지사항 등의 텍스트로 구성되어 있는 자원을 수집·활용하였다.

수집된 스팸・논스팸 SMS는 학습데이터로 활용될 Training set과 성능을 측정하기 위한 Test set로 분류한다. Training set으로 스팸 메세지 130개와 논스팸 메세지 130개를 사용하였고, Test set으로는 스팸 메세지 120개와 논스팸 메세지 120개를 사용하였다.

최적의 성능을 위해서는 적절한 경계값(T)을 설정하는 것이 중요하다. 아래서는 최적의 경계값(T)를 찾는 실험을 하고 성능을 확인한다.

4.1. 성능 측정 기준

스팸 필터링 시스템의 성능 측정 기준은 보편적으로 활용되고 있는 False positive(논스팸 메세지를 스팸 메세지로 거르는 비율)와 False negative(스팸 메세지를 논스팸 메세지로 거르는 비율)이다.

<표 2> 실제 스팸메세지 여부와 분류 결과의 관계

	분류된 스팸메세지	분류된 논스팸메세지
실제 스팸메세지	w	y
실제 논스팸메세지	x	z

False positive는 실제 논스팸 메세지가 스팸메세지로 분류되는 경우로, 이 비율이 높아지면 수신되어야 할 논스팸 메세지가 필터링 될 확률이 커지게 된다. 이 점은 필터링시스템의 신뢰도에 큰 영향을 미치게 된다. False positive의 계산식은 아래와 같다.

$$False\ positive = \frac{x}{(x+z)}$$

False negative는 스팸 메세지를 논스팸 메세지로 분류한 경우로써, 필터링 시스템의 성능이 좋지 못함을 의미하게 된다. False negative의 계산식은 아래와 같다.

$$False\ negative = \frac{y}{(w+y)}$$

4.2. 성능 측정 결과

본 절에서는 제안한 스팸 필터링 시스템의 최적의 경계값과 성능을 확인한다. False positive와 False negative가 모두 안정적인 값을 갖는 최적 경계값(T)를 찾는 실험을 위해 경계값(T)를 0.5%단위로 증가시키며 실험하였다.

<표 3> 특정 경계값(T)에 따른 성능측정 결과 (단위 : %)

특정값(T)	False-positive	False-negative
48.0	10.0	2.5
48.5	8.3	2.5
49.0	8.3	3.3
49.5	5.8	5.0
50.0	5.0	5.8
50.5	**3.3**	**5.8**
51.0	**3.3**	**5.8**
51.5	**3.3**	**6.6**
52.0	**3.3**	**6.6**
52.5	**2.5**	**7.5**
53.0	**2.5**	**7.5**
53.5	1.6	10.8
54.0	1.6	12.5
54.5	1.6	13.3

그 결과 False positive와 False negative의 반비례적 관계를 고려하였을 때, 특정 범위에서 유의미한 결과를 확인할 수 있었다. 실험한 결과는 <표 3>과 같다. False

negative가 사용자에게 미치는 문제보다 False positive가 미치는 문제가 더 크기 때문에,[2,6] 두 값이 적정한 수준에서 낮은 상태에서도 False positive가 상대적으로 더 낮은 특정 경계값(T)를 유의미한 것으로 판단하였다. 그 결과 최적 경계값(T)가 50.5~53.0%에서 False positive가 2.5~3.3%, False negative가 5.8~7.5%의 뛰어난 성능을 보였다.

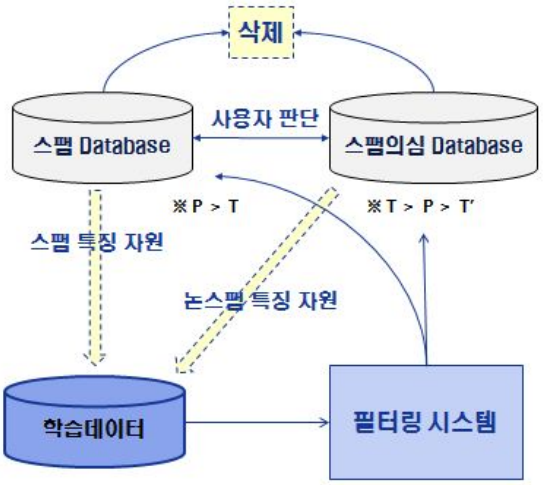
스팸 메일을 대상으로 한 논문이 많지만, 스팸 SMS을 위한 필터링 시스템은 그 수가 매우 적다. 또한 그 중에는 형태소 분석기로 분석하고 SVM 분류 알고리즘을 활용한 논문이 존재하지만[2], 트레이닝 및 테스트 셋이 다르고, 분석 범위(수신 텍스트, 수신 텍스트+발신 번호)가 다르기 때문에, 실험결과를 단적으로 비교하기에 무리가 있다. 그래도 굳이 성능을 비교하자면, 본 논문에서 False positive와 False negative가 모두 향상된 성능을 확인 할 수 있다.

5. 결론 및 향후과제

기존에 이론적으로 잘 정립되어 있고, 분류 성능이 좋은 것으로 알려져 있는 Naive Bayes Classifier, SVM, CRF 등의 분류기를 사용하지 않고 별도로 경계값 기반의 분류기를 제안한 이유는 다음과 같다.

일단 Naive Bayes Classifier의 성능이 우려할 만한 수준으로 스팸 필터엔진으로 사용하기에 많이 미흡하다. 하지만 또 다른 기존의 분류기에 비해 Naive Bayes 방법이 많이 사용되고 있는 이유는 구현이 쉽기 때문이다.[3] 그래서 본 논문에서는 단순하고 구현이 쉽지만 성능이 부족하지 않은 방법을 통해서, 실제 스팸 필터엔진으로써 실용성을 높고 대중적으로 활용될 수 있는 시스템을 위해 경계값 기반의 분류기법을 제안했다.

제안한 방법은 기존의 형태소 분석기나 기타 색인어 추출 방법에 비해 분석 과정이 단순하고 비용적인 측면에서 효율적이다. 이러한 장점은 어플리케이션 단(Level)에서 누적된 Database를 바탕으로 개인 스팸SMS패턴에 따른 맞춤형 필터 구축을 가능하게 해준다.



(그림 2) 개인 맞춤형 스팸SMS 필터 구조

(그림 2)는 개인 맞춤형 필터를 생성하는 구조이다. 해당 구조는 어플리케이션 단(Level)에서 필터링 시스템에 의해 분류된 스팸·스팸의심 SMS 데이터베이스를 활용한다.

스팸일 확률(P)가 최적 경계값(T) 이상인 SMS는 스팸 Database에 저장되고, 최적 경계값(T)보다는 작지만 스팸일 확률(P)가 적정 수준 이상(T′)인 SMS는 스팸의심 Database에 저장된다. 이와 같이 분류된 자원이 누적되면 사용자가 오분류된 SMS를 재분류하고 각각을 (그림 2)와 같이 스팸·논스팸 자원을 추가 학습하면 개인별 스팸 SMS 패턴에 맞춤형 스팸 필터(학습데이터)가 구축된다.

하지만 해당 방식이 성능 개선에 효과적인지는 추가적인 실험이 필요할 것으로 보인다.

참고문헌

[1] 방송통신위원회·한국인터넷진흥원, 이용자를_위한_불법스팸방지_안내서, KISA 안내·해설 2011 -3호, 2011. 1.

[2] 조인휘·심혜택, 휴대폰 SMS를 위한 SVM 기반의 스팸 필터링 시스템(A SVM-based Spam Filtering System for SMS), 韓國通信學會論文誌, Vol.34 No.9, 2009

[3] 김동건·조진남·장도석, 스팸메일 포착을 위한 데이터마이닝 기법 비교(Comparison of data mining techniques for detecting spam mails), 同德女子大學校 情報科學硏究所 정보과학연구, Vol.8 No.-, 2004

[4] 김우현, 나이브 베이지안 알고리즘을 이용한 스팸 필터링, NHN

[5] 방송통신위원회·한국정보보호진흥원, 2008스팸방지 가이드라인, 2008. 9.

[6] 방송통신위원회, '11년도 스팸방지 종합대책, 2011.1