한글 문자 단위 서열 정렬을 통한 스팸 문자 필터링

임진수, 우균
부산대학교 컴퓨터공학과
e-mail:elphy@pusan.ac.kr

SPAM Filtering for short Message Using Korean Character Alignment

Jin-Su Lim, Gyun Woo
Dept of Computer Engineering, Pusan National University

요 약

휴대전화 사용이 늘어나면서 이를 노리는 광고 문자 또한 많아지고 있다. 이를 막기 위해 대부분의 휴대전화가 스팸 차단 기능을 제공하고 있다. 허나 현재 제공되고 있는 스팸 차단 기능은 발신 번호가 같거나 설정 문구가 같은 경우에만 막아주는 기초적인 기능뿐이다. 그리고 광고 문자를 보내는 쪽은 이러한 차단 기능을 염두에 두고 변칙적인 문구를 사용해서 보내는 경우도 많다. 본 논문에서는 한글을 문자 단위로 서열 정렬하여 광고 문자를 차단하는 방법을 제안한다. 제안한 방법은 사용자가 등록한 문구를 수신한 문구에 대해 서열 정렬하고 이 결과를 바탕으로 유사도를 비교하여 차단하고자 하는 문구를 지닌 스팸 문자를 최대한 차단할 수 있다.

1. 서론

근래 사용자들은 SMS를 더 즐겨 사용하고 있다. SK 텔레콤이 추산한 가입자의 SMS 대 음성 비율은 2006년 32.9%에서 2007년 39.9%, 2008년 47.3%로 계속 증가하는 추세였다. 이에 발맞춰 스팸 문자량 또한 크게 늘어났다. 한국인터넷진흥원의 조사에 따르면 2005년 39만 건에서 2009년 3560만 건으로 90배 이상 증가했음을 알 수 있다.

휴대전화 단말기는 통상적으로 문자가 왔을 때 바로 알아챌 수 있도록 설정을 한다. 이런 이유로 스팸 메시지 차단 실패는 사용자가 이를 바로 확인을 하게끔 만든다는 점에서 스팸 메일보다 큰 스트레스를 유발한다. 게다가 밤낮을 가리지 않기 때문에 이에 대한 문제를 토로하는 경우도 많다. 그렇기 스팸 메시지 차단 기능은 중요하다.

현재 휴대전화 단말기에서 제공하는 스팸 문자 차단 기능은 기대에 미치지 못한다. '광고' 라는 문구를 차단해도 '광*고' 와 같은 변칙 문구는 전혀 차단하지 못한다. 이 때문에 이런 변칙 문구를 모두 고려하여 '광*고', '광.고' '광/고' 등을 모두 등록해야하는 것이 현실이다. 사용자가 직접 수많은 경우를 예상해서 등록한다 하더라도 스팸 문자 발송자는 또 다른 변칙을 통해서 회피할 수 있다. 발신 전화번호를 이용한 차단도 있지만 여러 회선 또는 명의 대여 등을 통해 새로운 번호를 만들고 이를 이용하여 계속 발송하기에 이 또한 부족하다. 즉, 현재 휴대전화에서 제공되는 단순한 스팸 차단 시스템으로는 부족한 것이다.

2. 관련 연구

스팸 필터링은 전자메일 분야에서 많이 연구되었다. José와 Guillermo, Enrique, Francisco는 스팸 메일 필터링 방법을 다음과 같이 분류하였다[3].

- 화이트리스트와 블랙리스트:발송자를 확인하여 스팸 발송자가 아니라 판단하면 수신을 하거나 스팸 발송자라 판단하면 수신을 하지 않는 방법
- 발송료:발송을 할 때 마다 경제적으로 또는 다른 방법으로 발신에 대한 발송료를 매기는 방법
- 주소 관리:발송 주소가 임시로 또는 기계적으로 생성된 주소로 판단하면 모든 발송물을 버리는 방법
- 협업 필터링:수신자 중 누군가가 스팸이라고 판단했다면 모두 스팸이라고 판단하고 버리는 방법
- 전자 서명:전자 서명이 없는 모든 발송물을 스팸으로 분류하는 방법
- 내용에 기반을 둔 필터링:해당 단어가 포함된 발송물을 스팸으로 판단하는 방법

이를 스팸 필터링에 적용하기 위해 문서 분류 기법을 많이 사용한다. 분류 기법은 Naive Bayes, K-nearest neighbor, Support Vector Machine 등이 있다[1].

SMS에 비해 전자메일은 전달 용량 제한이 없어 내용이 다양하고 변칙적일 수 있다. 그래서 발송 주소나 제목을 이용하여 스팸 여부를 판단하려한 연구도 있다[2].

스팸 전자메일 필터링에 대한 연구를 토대로 스팸 문자 필터링에 대한 연구도 진행되었는데 국내에는 SVM 기반의 스팸 필터링 시스템 연구가 있다[4]. 제안한 시스템은 전처리 과정에서 특수 문자 제거와 자동 단어 띄어쓰기, 수사 어절 표준화, 불용어 제거를 수행하기도 한다.

위 6가지 방법을 활용한 연구 외에도 말뭉치를 기반으로 내용을 비교해서 스팸인지를 확인하는 방법이 있다[5]. 그리고 광고 문자의 문체가 구어체나 통상 많이 쓰이는 약어를 사용하지 않은 독특한 문체인 것에 착안하여 이를 기초로 스팸을 판단하는 연구도 있다[6].

살펴보면 대부분의 연구에서 내용에 기반을 둔 필터링을 시도하고 있는데, 이는 전자메일과는 달리 문자에는 제목과 발송자의 이름 같은 정보가 없기 때문이다. 내용에 기반을 둔 필터링이 잘 동작하려면 변형을 감지하기 위한 연구도 필요한데, Paul의 연구는 사람이 읽을 수 있지만 시스템 회피는 가능한 경우에 대해 설명하고 있다[7].

- 시각적 유사성 : 숫자 1과 알파벳 L의 소문자 l, 알파벳 I의 대소문자, 특수문자 |는 서로 바어도 인식할 수 있음
- 소리의 유사성 : 읽는 소리가 비슷한 경우 대체해도 인식할 수 있음

이 중 시각적 유사성을 이용한 변형 공격을 해결하기 위한 스팸메일 필터링 시스템이 제안된바 있으며[8], 형태 변화가 심하고 변형 공격 역시 빈번한 비속어에서도 이를 감지하기 위한 한글 범위 연구가 진행된 바 있다[9,10].

3. SMS의 내용에 기반을 둔 필터링

스팸 문자를 살펴보면 대부분이 인터넷 가입 및 대출 광고이다. 그리고 이 내용을 전달하기 위해 사용하는 단어는 그리 많은 편이 아니다. 한국인터넷진흥원 불법스팸대응센터의 스팸차단 금칙어 TOP10에서 2006년부터 2010년까지 많이 사용된 금칙어를 살펴볼 수 있다. 59개월간 10위까지 모은 자료로서 총계 590회에 달한다.

금칙어	횟수	비율	금칙어	횟수	비율
거부	50	8.47	자금	16	2.71
080	36	6.10	방문	15	2.54
고객	34	5.76	번호	14	2.37
당일	29	4.92	연체	13	2.20
상담	29	4.92	카드	13	2.20
최저	25	4.24	대리	12	2.03
무료	25	4.24	할인	12	2.03
현금	24	4.07	분할	11	1.86
대출	19	3.22	최대	11	1.86
인터넷	18	3.05	입금	10	1.69
총계	289	48.99	총계	127	21.52

표 1. 5년간 금칙어 순위권에 들어간 횟수

표 1의 5년간 금칙어 순위에 들어간 횟수를 살펴보면 같은 단어가 많이 나타나는 것을 알 수 있다. 20위까지의 단어는 총 416회인데, 이는 전체의 70.5%에 달한다.

금칙어	횟수	비율
고객	11	10.0
당일	11	10.0
현금	9	8.18
최저	9	8.18
입금	9	8.18
인터넷	8	7.27
최대	8	7.27
팀장	7	6.36
상담	5	4.54
캐피탈	5	4.54
총계	82	74.54

표 2. 2010년 11개월간 금칙어 순위권에 들어간 횟수

표 2의 2010년의 금칙어를 살펴보면 총 23종류 단어가 나타났으며 상위 10건이 82회 나타나 74.5%에 달하는 것을 알 수 있다. 즉, 스팸 문자에 자주 쓰이는 단어가 있기 때문에 구문 필터링이 충분히 효과가 있을 것이라 예상할 수 있다. 이를 바탕으로 본 논문은 차단하고자 하는 구문을 포함한 문자를 완벽히 차단하는 것에 주안점을 두었다.

4. 한글 문자 단위 서열 정렬을 통한 필터링

위에서 보인 바와 같이 스팸 문자에서 사용하는 단어가 경향성이 있기 때문에 이를 필터링하면 즉시 높은 효과를 거둘 수 있다. 하지만 휴대전화 단말기에 제공되는 기능은 구문이 정확히 일치할 때만 차단하기 때문에 한계가 있다. 고객을 '고*객' 또는 '고/객'으로 변경하는 것은 기계적으로 적용 가능하기 때문에 쉽게 변형하고 발송할 수 있다. 그래서 이러한 변형을 감지하는 것은 중요하다.

관련 연구에서 알아본 학습을 통한 차단 시스템들은 변형이 진행되는 스팸 문자에 지속적으로 적용을 할 수 있다는 장점이 있다. 하지만 학습할 문자 메시지를 수집하기 힘든 환경에서는 사용하기 힘들다. 한국인터넷진흥원 추산 1인당 1일 스팸 문자 수신량은 0.5건 미만이다. 그리고 SVM 기반의 스팸 필터링 시스템에서 학습에 사용한 자료는 스팸 문자 100개, 비스팸 문자 200개이다. 해당 학습 효과에 도달하기 위해 스팸 문자 수집하기 위해 긴 시간이 소요될 수 있다는 것이다. 다른 통계 측정 방법을 사용해서 필요한 수집량을 줄일 수도 있겠지만 그럼에도 즉각 활용하기 힘들다는 단점이 있다. 스팸 메일을 처리할 때는 접속된 인터넷 환경을 사용해서 수많은 데이터를 얻을 수 있기에 학습을 통한 시스템을 구동하는데에 어려움이 없지만 스팸 문자를 처리하는 환경에서는 데이터 확보 측면에서 문제가 될 수 있다.

본 논문에서는 이런 변형을 감지하기위한 방법으로 생물정보학에서 유전자나 단백질을 분석하기 위해 사용하는

서열 정렬 기법을 사용한다. 서열 정렬은 서로 연관성이 높은 정도와 그 구간을 알아내는데 좋은 효과를 보여준다.

실험 데이터는 3개월간 수집한 스팸 SMS 86건이다. 이를 2010년 금칙어 누적 순위 10위까지에 오른 10개의 단어로 얼마나 차단을 해내는지 확인하는 것이 실험의 목적이다. 실험은 다음과 같이 진행된다. 먼저 일반 휴대전화에서 제공하는 기능인 문자열 비교로 차단 성공 횟수를 측정한다. 그리고 본 논문에서 제안한 방법인 한글 문자 단위의 서열 정렬을 통한 차단 성공 횟수와 이전의 방법을 비교한다. 이 실험을 통해 한글 문자 단위 서열 정렬 방법이 더 나은 차단 효과를 보이는지 알 수 있을 것이다. 실험의 서열 정렬을 처리하기 위해서 Freie 대학의 공개 라이브러리인 SeqAn을 사용하였다.

	차단 성공 횟수	비율
문자열 비교	63/86	73.3
서열 정렬	69/86	80.2

표 3. 문자열 비교 방법과 서열 정렬 방법의 스팸 차단 성능 비교

표 3을 보면 문자열 비교 방법보다 서열 정렬 방법이 차단에 더욱 효과적인 것을 알 수 있다. 실험을 위해 수집한 문자 86건 중 16건은 도박 및 성인물 스팸 문자이며 나머지가 대출 및 인터넷 스팸 문자이다. 차단 실험에 사용한 금칙어 10위까지 단어가 대출 및 인터넷 관련 단어이기 때문에 16건의 도박 및 성인물 스팸 문자에는 전혀 적용되지 못한 문제점이 있다. 본 논문에서 제안한 방법은 차단하려 하는 구문이 들어있는 경우 이를 차단할 수 있게끔 하는 것에 초점을 맞추고 있다. 이 점을 고려해보면, 위의 실험 결과는 다음과 같이 나타낼 수 있다.

	차단 성공 횟수	비율
문자열 비교	63/70	90.0
서열 정렬	69/70	98.6

표 4. 차단하고자 하는 단어를 포함한 문자에 대한 문자열 비교 방법과 서열 정렬 방법의 스팸 차단 성능 비교

표 4처럼 바꾸어 보면 한글 문자 단위 서열 정렬을 통한 스팸 문자 필터링은 단 하나의 문자만 차단에 실패했으며 그 외 모든 스팸 문자의 차단에 성공한 것을 알 수 있다. 차단하고자 하는 구문이 들어있는 대부분의 문자를 차단한 것이다. 차단에 실패한 문자는 '캐피탈'을 'ㅋ피탈'로 사용한 것으로 이러한 변칙에 대해서는 현재의 방법으로 대응할 수 없는 것도 확인할 수 있었다.

5. 결론 및 추후 연구 과제

정렬 방법을 적용하는 것만으로도 특수 문자를 첨가하는 스팸 차단 회피 방법을 쉽게 막을 수 있다. 정렬 방법은 현재 사용하고 있는 방법의 확장이기 때문에 휴대전화 단말기의 간단한 업데이트를 통해서 추가적인 효과를 거둘 수 있을 것이다. 게다가 간단한 연산만으로 결과를 추산할 수 있기 때문에 휴대전화에서 충분히 사용 가능하다.

추후 연구 과제는 형태가 비슷한 문자를 이용한 회피 및 일부 삽입 및 누락을 감지하는 것이다. 관련 연구에 제시된 방법은 2바이트 문자인 한글에서 사용하기 힘들다. 유사성에 의한 점수 매트릭스를 구성하는 것이 힘들기 때문이다. 이 문제를 해결한다면 휴대전화 단말기에서 사용 가능한 강력한 스팸 차단 기능을 구현할 수 있을 것이다.

참고문헌

[1] C,Lai, "An empirical stury of three machine learning methods for spam filtering", Knowledge Based System, Vol.20, No.3, pp.249-254, 2007.

[2] 공미경, 이경순, "스팸성 자질과 URL 자질을 이용한 최대엔트로피모델 기반 스팸메일 필터 시스템", 제 18회 한글 및 한국어 정보처리 학술대회, pp.213-219, 2006.

[3] José María Gómez Hidalgo, Guillermo Cajigas Bringas, Enrique Puertas Sánz, Francisco Carrero García, Content based SMS spam filtering, Proceedings of the 2006 ACM symposium on Document engineering, Oct. pp.10-13, 2006.

[4] 조인휘, 심혜택, "휴대폰 SMS를 위한 SVM 기반의 스팸 필터링 시스템" 제 34회 한국통신학회, pp.908-913, 2009.

[5] Gordon V. Cormack, José María Gómez Hidalgo , Enrique Puertas Sánz, Spam filtering for short messages, Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, Nov. pp.6-10, 2007.

[6] Dae-Neung Sohn, Jung-Tae Lee, Hae-Chang Rim, The contribution of stylistic information to content-based mobile spam filtering, Proceedings of the ACL-IJCNLP 2009, August, pp.4-4, 2009.

[7] Paul Gardner-Stephen, "A Biologically Inspired Method of SPAM Detection", International Workshop on Database and Expert Systems Application, pp.53-56, 2009.

[8] 이호섭, 조재익, 정만현, 문종섭, "비정상 문자 조합으로 구성된 스팸 메일의 탐지 방법", 제 18회 정보보호학회, pp.129-137, 2008.

[9] 박교현, 이지형, "SVM을 이용한 온라인게임 비속어 필터링 시스템", 제 33회 가을 학술발표논문집, 한국정보과학회, pp.260-263, 2006.

[10] 윤태진, 조환규, "한글 자소정렬을 이용한 온라인 욕설 필터링 시스템", 제 36회 가을 학술발표논문집, 한국정보과학회, pp.194-198, 2009.