텍스트와 도메인 네임을 이용한 메일 분류

김원화[0] 이일병
연세대학교 컴퓨터과학과
whkim@csai.yonsei.ac.kr[0], yblee@csai.yonsei.ac.kr

E-Mail Classification Using Text and Domain Name

Wonhwa Kim[0] Yillbyung Lee
Dept. Computer Science, Yonsei Univ.

요 약

정보화 시대에는 사람들의 모든 활동이 인터넷을 통해서 대부분 이루어진다. 이중에서 전자 메일이 차지하는 비중은 매우 크다. 고객 유치를 위한 기업들의 광고와 배움을 위한 강의, 자신의 관심 분야에 대한 정보 등을 전자 메일로 받아보게 되는 것이 더 많아 질것이다. 이러한 상황에서 사람들은 자신이 필요로 하는 메일과 필요로 하지 않는 메일을 분류하는데 많은 시간을 낭비한다.

사람들은 이러한 시간 낭비를 줄이기 위해서 메일 분류 시스템을 사용한다. 현재 사용되고 있는 메일 분류 시스템은 스팸 메일을 기준으로 하고 있다. 그러나 오분류되는 메일들이 있어 사용자가 스팸 메일을 다시 보는 경우가 있어 한계를 보인다.

본 논문에서는 사람들이 자신이 원하는 메일과 그렇지 않은 메일을 분류하기 위해서 1차 분류로 긍정어와 부정어를 이용하여 전자 메일을 분류하고 2차 분류로 도메인 네임을 이용하여 분류한다.

1. 서 론

정보화 시대에 인터넷은 필수품이라고 할 수 있다. 인터넷으로 정보 교환 및 쇼핑을 하는 것은 이제 일상이다. 인터넷으로 자신이 원하는 정보를 찾아서 얻을 수도 있고 뉴스 그룹을 이용할 수도 있다. 또, 인터넷 쇼핑을 할 때 쇼핑 정보를 각각의 쇼핑몰에서 직접 얻을 수 있고 쇼핑 정보 매거진을 받을 수도 있다.

인터넷에서 원하는 정보를 찾는 것이 현재와 같은 정보의 홍수 속에서는 시간과 노력의 낭비가 아닐 수 없다. 이러한 이유로 사람들은 뉴스 그룹이나 관련 매거진을 받아 보는 경우가 늘고 있다. 이제 인터넷의 기본은 정보의 천국이 아니라 원하는 정보를 전자 메일을 통해서 빠르게 얻는 것이다.

전자 메일의 가장 큰 특징은 기존의 일반 우편 제도와 달리 개별적인 우편을 발송함에 있어 아주 적은 이용료만으로 정보를 가장 신속하게 전달할 수 있는 것이다. 이러한 특징을 이용하여 기업에서는 고객 유치를 위해서 기업 광고를 전자 메일을 통해서 하고 있으며, 고객 관리를 위한 매거진을 전자 메일을 통해서 보내고 있다.

사용자가 받는 전자 메일에는 필요에 의해서 받는 메일과 무작위로 받는 메일이 있다. 기업의 광고 메일은 공정거래위원회 전자 상거래 보호법을 통해서 제목에 " (광고)" 를 명시하도록 되어있다.[1][2][3] 현재 기업에서 보내는 광고 메일은 제목에 " (광고)" 를 명시하여 발송하고 있어 손쉽게 분류할 수 있다. 그러나 개인이 발송하는 광고 메일은 (광고)라는 것을 명시하지 않고 있어서 분류가 어렵다.

이러한 메일 분류의 한계를 극복하기 위해서 전자 메일 분류에 대해서 많은 연구가 진행되어오고 있다.

2. 관련 연구

메일의 본문의 내용과 관련된 단어는 그 문서에서의 단어의 출현빈도와 관련이 있다고 생각한다.[4]

기존의 관련 연구들은 스팸 메일을 기본으로 헤더나 본문에서 단어를 추출하여 그 단어를 포함한 메일을 스팸 메일로 분류한다. 처음에는 메일의 헤더를 이용하거나, 본문에 단순한 스트링 매칭에 의한 방법을 주로 연구하였다. 현재처럼 대량의 스팸 메일에 시달리면서 여러 가지 방법들을 연구하게 되었다.

단어에 가중치를 부여하고 단어의 출현 빈도를 통한 문서 분류 알고리즘을 사용하여 분류하는 방법이 있다. 이 방법은 스팸 메일과 비스팸 메일에서 단어의 출현 빈도 수를 이용한다. 저빈도의 단어에 대해서 신뢰할 수 있는 가중치 부여 기법인 카이제곱 통계량을 이용하여 단어의 가중치를 부여한다.[5] 단어의 빈도가 지나치게 낮은 영향력을 보충하고, 단어의 빈도가 높은 단어의 지나친 영향력을 낮추기 위해 로그 단어 빈도 가중치 공식을 사용한다.[6][7]

시소러스를 이용한 방법은 사용자 적합도를 판단하기 위해 사용자 관련 정보로부터 동적 시소러스를 구축한

다. 구축된 시소러스와의 비교를 통해 사용자에게 유용한 메일인지 아닌지를 결정하고, 사용자가 지정한 폴더 키워드를 중심으로 사용자 시소러스로부터 유전자 알고리즘을 이용해 추출한 키워드들과의 적합도 비교를 통해서 특정 폴더로 메일을 분류한다.[8][9][10]

본문에서 hyperlink를 이용한 방법으로 인터넷 주소의 특징을 추출한다. 예를 들어 본문의 내용에 www.girl.com/image.jpg 있다면 .(dot)사이에 있는 단어인 girl을 추출하여 메일을 분류하는데 사용한다. 이 방법 또한 단어로만 hyperlink를 사용하였다.[1]

그 밖에 본문 내용에 대한 의미 파악 방법 등이 연구되고 있다.

본 논문에서는 전자 메일을 1차 분류 방법으로 사용자가 선택한 긍정어와 부정어를 이용하여 분류해주고 2차 분류 방법으로 도메인 네임을 이용해서 스팸 메일을 분류하는 방법을 제시한다.

3. 긍정어와 부정어를 이용한 1차 분류

최근에는 교묘한 방법으로 스팸 메일을 보내기 때문에 광고 메일은 " 광고" 라는 단어를 사용해서 메일을 발송하게 하고 있으나 기대만큼 큰 효과를 보고 있지 않다.

스팸 메일에 자주 등장하는 단어를 포함하고 있는 메일이라고 해서 모두 사용자에게 불필요한 메일은 아니다. 사용자에게 꼭 필요한 메일임에도 불구하고 스팸 메일로 분류되어 사용자가 피해를 보는 경우가 종종 있다. 이로 인해서 사용자들은 스팸 메일로 분류되어진 메일들을 바로 삭제하지 않고 다시 보는 경우가 많다.

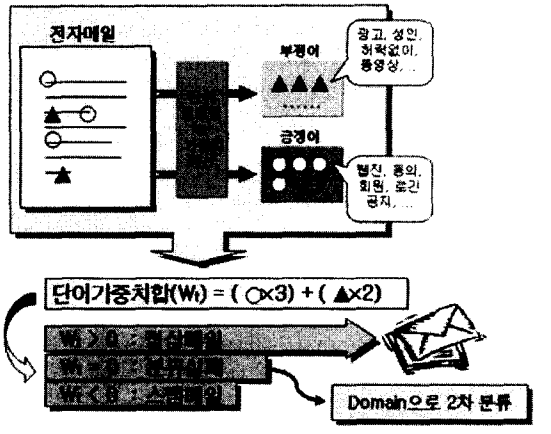
이런 문제점들을 보완하기 위한 방법으로 본 논문에서는 긍정어와 부정어를 사용하여 메일을 분류하는 방법을 제시한다.

긍정어과 부정어의 개념은 기존의 개념과는 다르다. 사용자가 필요로 하는 메일에 포함되어진 단어들로 사용자가 선택한 단어를 긍정어라고 하며, 사용자가 필요로 하지 않은 메일에 포함되어진 단어들로 사용자가 선택한 단어를 부정어라고 본 논문에서는 정한다.

정상 메일에는 사용자가 수신을 허락한 메일과 서신 메일 등을 포함한다. 광고 메일일지라도 사용자가 수신을 허락한 메일이면 정상 메일로 분류한다.

긍정어와 부정어는 각각의 긍정어와 부정어 파일로 저장되며 사용자가 필요에 의해서 갱신할 수 있다. 긍정어에는 +1이라는 단어 가중치를 부여하고 부정어에는 -1이라는 단어 가중치를 부여한다.

전자 메일의 헤더와 본문에서 긍정어 파일과 부정어 파일에 있는 단어와 동일한 단어를 찾아서 단어 가중치를 주고 단어 가중치의 합에 따라서 메일을 분류한다. 전자 메일의 단어 가중치 합(Wt)이 0보다 크면 정상 메일로 분류되며 0보다 작으면 스팸 메일로 분류한다.

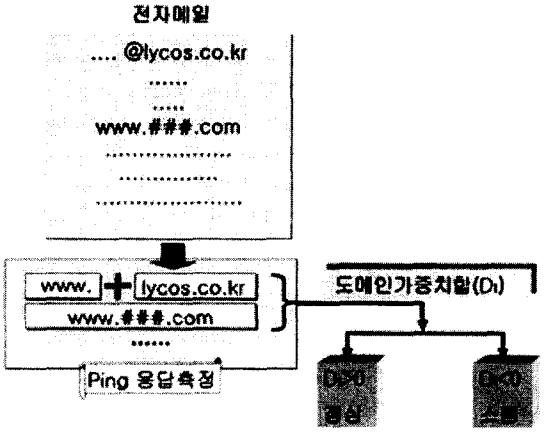


<그림1> 긍정어와 부정어를 이용한 1차 분류

4. 도메인 네임을 이용한 2차 분류

기업의 광고 메일과는 달리 개인의 무차별적 대량 광고 메일은 가상 도메인 네임을 사용하는 경우가 많다.

전자 메일에는 사용자의 주소록에 등록된 도메인 네임과 등록되지 않은 도메인 네임 중에 존재하는 도메인 네임과 존재하지 않는 도메인 네임이 있다. 등록된 도메인 네임과 등록되지 않았으나 존재하는 도메인 네임에는 +1이라는 도메인 가중치 값을 부여하고 존재하지 않는 도메인 네임에는 - 1이라는 도메인 가중치 값을 부여한다.



<그림2> 도메인 네임을 이용한 2차 분류

전자 메일의 헤더와 본문에서 도메인 네임들을 추출하여 도메인 네임 존재 여부를 알아보고 도메인 가중치 값을 부여한다. 도메인 가중치 값의 합에 따라서 메일을 분류한다. 전자 메일의 도메인 가중치 합(Dt)이 0보다 크면 정상 메일로 분류되며 0보다 작으면 스팸 메일로 분류된다.

5. 실험 및 결과

정상 메일과 스팸 메일로 구분해서 메일을 수집하였으며, 사용자가 수신을 허락한 광고 및 매거진은 스팸 메일로 구분하지 않았다. 사용자가 헤더와 본문에 대해서 선택한 긍정어와 부정어를 적용하여 1차 분류를 했고, 헤더와 본문에 있는 도메인 네임의 존재 여부를 사용하여 2차 분류를 하였다.

긍정어와 부정어는 설문 조사를 통해서 선호도가 높은 순서로 13개를 선택하여 긍정어와 부정어 파일을 생성하였다. 설문 조사로 만들어진 긍정어와 부정어에 대한 신뢰도를 위해서 전자 메일 309개를 사용하였고, 본 실험에서는 전자 메일 892개를 사용하였다.

<표1> 긍정어와 부정어를 이용한 1차 분류

	정상 메일	스팸 메일
전체 전자 메일	406	486
분류 결과	265	457
분류 성공	260	437
잘못 분류	5	20
분류 실패	126	44

위의 실험 결과로 긍정어와 부정어를 이용한 1차 분류 재현율은 76.97%이고, 정확율은 96.61%이다. 잘못 분류된 메일은 사람의 실수로 잘못 분류된 것으로 나타났다. 분류에 실패한 메일들은 텍스트의 양이 아주 작아서 긍정어와 부정어의 개수가 동일하게 나온 메일들이며, 본문의 내용을 아스키 코드로 변환한 메일들이었다.

<표2> 도메인 네임을 이용한 2차 분류

	정상 메일	스팸 메일
전체 전자 메일	406	486
분류 결과	391	506
분류 성공	386	481
잘못 분류	5	20

위의 실험 결과로 도메인 네임을 이용한 2차 분류 재현율은 97.02%이고, 정확율은 97.36%이다. 잘못 분류된 메일은 사람의 실수로 잘못 분류된 것으로 나타났다.

<표3>에서는 2차 분류까지 실험한 최종 결과로 본 논문에서 제시한 방법은 재현율 97.02%와 정확율 97.36%의 성공율을 보인다. 잘못 분류된 것은 전자 메일을 수집할 때 잘못 분류한 것으로 나타났다.

<표3> 결과

	정상 메일	스팸 메일
전체 전자 메일	406	486
분류 결과	391	501
분류 성공	386	481
잘못 분류	5	20

본 논문에서 제시한 방법으로 사람이 잘못 분류하는 경우도 찾아 내었으며, 단어만을 가지고 사용해서 분류에 실패한 메일들이 도메인 네임을 사용하여 분류할 때는 분류가 되었음을 보여준다.

긍정어와 부정어를 이용한 1차 분류에서 본문의 내용이 아스키 코드로 변환되어있어 실패를 많이 했음을 보았다. 아스키 코드로 변환된 메일을 분류할 수 있는 방법이 향후 과제로 나타났다.

참고문헌

[1]이종호, " 광고성 메일을 자동으로 구별해내는 Text Mining 기법 연구" , 한국인지과학회, 춘계학술대회 논문집, 35-39, 2002.

[2]눈속임 광고메일 엄벌, (2002.4.23), 중앙일보, http://service.joins.com/asp/search_article.asp?aid=1726788&history=-2.

[3]공정거래위원회, (2002.4.23), " 스팸보도자료.hwp" , http://www.antispam.or.kr (공지사항).

[4]황도삼, 최기선, 김태석 공역, " 자연언어처리" , 홍릉과학출판사, 1999.

[5]한광록, 선복근, 한상태, 임기옥, " 인터넷 문서 자동분류 시스템 개발에 관한 연구" , 한국정보처리학회, 정보처리학회 논문지, 제7권, 제9호, pp2867-2875, 2000.

[6]고수정, 이정현, " Apriori 알고리즘에 의한 연관 단어 지식 베이스에 기반한 가중치가 부여된 베이지안 자동 문서 분류" , 멀티미디어학회, 멀티미디어학회 논문지, 제4권, 제2호, pp171-181, 2001.

[7]이재윤, 최보영, 정영미, " 문헌 자동 분류에서 용어 가중치 기법에 대한 연구" , 한국정보관리학회, 제7회 한국정보관리학회 학술대회 논문집, pp41-44, 2000.

[8]안희국, 노희영, " 동적 시소러스와 GA을 이용한 개별화된 E-Mail 분류시스템(PECS)" , 한국정보과학회, 한국정보과학회 춘계학술대회 논문집, 472-474, 2002.

[9]Michael J. Pazzani, " Representation of electronic mail filtering profiles" , Proceedings of the 2000 international congerence on Intelligent user interfaces, January 2000'

[10]조유근 외, " 알고리즘" , 이한출판사, 2000.