단어 가중치를 이용한 스팸메일 필터링

김호성 정경호 황도삼

영 남 대 학 교 컴 퓨 터 공 학 과
{hskim,khjung}@nlp.yu.ac.kr, dshwang@yu.ac.kr

A Filtering Method of Spam E-mails by Term Weighting

Ho-Seong Kim, Kyung-Ho Jung, Dosam Hwang
Department of Computer Engineering, Yeungnam University

요 약

현재, 전자메일을 정보전달의 수단이나 광고등의 목적으로 많이 이용하게 되면서, 메일 수신자는 원치 않는 상업적 광고, 불필요한 정보등의 스팸메일을 여과 없이 수신하게 되는 경우가 많아졌다. 이로 인하여 업무효율성 감퇴, 시간 낭비, 자원 낭비 등의 많은 문제점을 야기시키고 있다. 이러한 문제점을 해결하기 위한 기존의 메일 필터링 시스템들은 송신자의 주소나 도메인, 제목등의 메일 헤더정보만을 이용하거나, 사용자가 정의한 문장이 본문 내용에 나타날 때 필터링하는 방식들이 주류를 이루고 있다. 그러나 이러한 방식들은 메일의 내용에 대한 근본적인 필터링이 불가능하다.

본 논문에서는 메일의 내용을 파악하기 위해 메일의 내용을 대표할 수 있는 체언정보를 추출하여, 카이제곱 통계량 공식을 통해 단어 가중치를 부여하고, 이를 문서분류를 위한 로그 단어 빈도 가중치 공식에 적용하여 스팸메일을 필터링하는 방식을 제시한다.

본 논문에서 제안한 방법으로 실험한 결과, 스팸메일을 필터링하는데 84.61%의 재현율과 83.01%의 정확율을 얻을 수 있었다.

1.서론

정보화시대를 주도해 온 지능화된 통신망으로서, 인터넷은 초고속 정보시대를 지향하는 오늘날 전세계적으로 나날이 급성장을 하고 있다. 이러한 인터넷의 급성장에 가장 중요한 역할을 한 것이 전자메일이다. 전자메일은 기존의 일반 우편제도와 달리 개별적인 우편을 발송함에 있어 추가적인 요금이 부과되지 않고, 아주 적은 이용료만으로 대량의 정보를 가장 신속하게 전달할 수 있다. 이러한 전자메일의 장점을 이용하여 전자메일 마케팅이 성행하고 있다. 그러나, 수신자의 주소만 확보하면 수신자의 의도와는 상관 없이 메일 수신이 허용될 수 있으며, 이점을 악용하여 불특정 다수에게 발송되는 광고성 스팸메일에 의해 수신자는 많은 피해를 입고 있다. 한국 정보 보호원에서는 전자메일 서비스 제공업체중 55%정도가 스팸메일로 인하여 서비스가 정지되거나 지연된 적이 있었다고 보고했다. 유럽 연합(EU) 집행위원회에서는 인터넷 이용자들이 스팸메일을 봄으로써 발생되는 인터넷 접속료가 전세계적으로 연간 94억 달러에 이른다고 한다. 이처럼 스팸메일은 업무효율성 감퇴, 시간낭비, 자원낭비등의 많은 문제점을 야기시키고 있다. 이를 해결하기 위한 기존의 필터링 시스템들은 메일의 헤더를 이용하거나, 본문 내용의 단순한 스트링 매칭에 의한 필터링을 하고 있으나 효율적인 필터링이 되지 않고 있다.

그런데, 본문의 내용은 단어의 집합으로 근사하게 표현할 수 있으며, 특히 본문의 내용과 관련된 단어는 그 문서에서의 단어의 출현빈도와 관련이 있다[1]. 본 논문에서는 이 생각에 근거하여, 카이제곱 통계량 공식을 이용하여 단어에 가중치를 부여하고, 단어의 빈도에

[1] 본 논문은 KAIST AITrc를 통하여 과학재단의 지원을 받았음.

의한 영향력을 정규화하여 메일을 분류하기 위해 로그 단어 가중치를 이용함으로써 스팸메일을 분류하는 방법을 제시한다.

2. 기존의 스팸메일 필터링 시스템

현재, 제공되고 있는 대부분의 메일 필터링 시스템은 송신자의 주소, 도메인등의 메일 헤더의 정보를 가지고 수신허용 여부를 결정하여 필터링을 하고 있다. 서버 차원에서의 스팸메일에 대한 방화벽도 메일의 헤더를 이용하여 차단하고 있다. 이렇게 메일의 헤더만을 이용한 필터링 방법은 송신자의 주소나 도메인등 메일 헤더의 내용을 변경하여 송신할 경우 필터링이 되지 않는 문제점을 가지고 있다. 최근에는 메일의 본문 내용을 통한 필터링 시스템이 있으나, 이 시스템들은 메일 내용에 나타날 수 있는 문장을 데이터베이스에 저장해 둔 후, 본문의 문장과 단순 스트링 매칭을 통해 스팸메일로 분류하는 시스템들이 주류를 이루고 있다. 이렇게 단순한 스트링 매칭을 통한 필터링 방법은 스팸메일이 확실하더라도 사용자가 정의한 문장이 메일내에 존재하지 않으면 필터링이 되지 않는 한계를 가지게 된다. 그러므로, 기존의 방식으로는 근본적인 필터링이 되지 않는다. 그러나 메일 본문의 내용에 나타내는 단어를 추출하여, 스팸메일과의 관련도에 따라 단어에 대해 가중치를 부여하고 이를 이용하여 문서 분류 알고리즘에 적용 한다면 메일의 내용에 의한 필터링이 가능할 것이다.

이에 본 논문에서는 스팸메일을 필터링하기 위한 방법으로 단어와 문서의 범주를 통해 관여도를 계산하는 카이제곱 통계량 공식을 이용해 단어의 가중치를 부여하고, 단어의 출현 빈도를 통한 문서분류 알고리즘인 로그 단어 가중치 공식을 이용하여 필터링하는 방식을 제시한다.

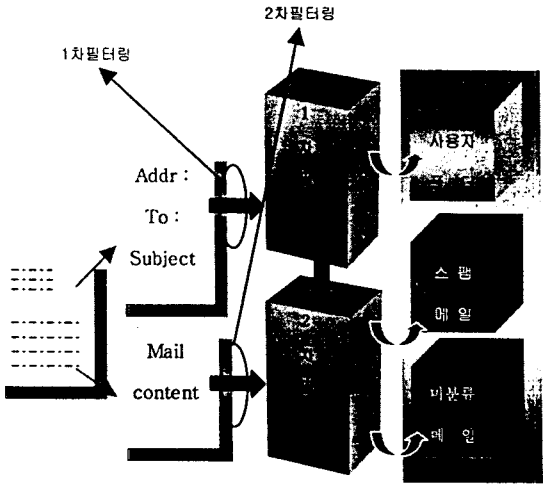
3. 스팸메일 필터링 알고리즘

본 논문에서 제시하는 전체적인 메일 필터링 시스템의 구조는 <그림 1>과 같다.

1차 필터링은 수신된 메일의 수신 거부, 받고자 하는 광고메일에 대한 수신허용등 사용자가 분류하고자 하는 메일을 헤더정보를 통해 필터링하는 것이다. 이는 헤더정보만으로 충분히 분류가 가능한 메일을 먼저 필터링함으로써, 본문 필터링 과정의 오버헤드를 줄이고, 2차 필터링에서 사용자가 수신을 원하는 광고메일을 스팸메일로 분류하는 오분류를 줄일 수 있다.

메일 본문을 통한 2차 필터링은 1차 필터링에서 분류되지 않은 메일들 중에서 스팸메일을 분류하는 필터링 과정이다. 이 과정에서, 단어에 대한 가중치를 부여 하여 구축하여 둔 스팸메일 사전과 문서분류 알고리즘을 적용하여 메일 필터링을 한다.

두 번의 필터링 과정을 거치게 되면, 스팸메일 분류할 때의 오분류를 막을 수 있고, 좀더 효율적인 필터링이 가능할 것이다.



<그림1> 스팸메일 필터링 시스템 구조

3.1 카이제곱 통계량에 의한 스팸메일 사전 구축

문서를 분류하기 위해서는 구문분석, 의미분석이 선행되어야 정확한 분류가 가능하나, 이는 현재까지도 연구 진행 중인 분야이며 만족할 만한 성능에 이르지 못하고 있다[2]. 또한 전자메일은 일반문서와는 달리 분류를 위한 텍스트의 정보량이 많지 않다. 특히, 스팸메일의 경우는 대부분이 HTML의 형식을 취하고 있기 때문에 태그정보를 제거하면 메일 본문을 분석하기 위한 정보량은 아주 적다. 그러므로 메일 본문에서 형태소 분석 과정을 통해 추출된 체언의 단순한 출현 빈도만으로는 본문 내용을 분석하기가 어렵다. 따라서, 본 논문에서는 단어의 가중치를 부여함으로써, 적은 단어의 정보량으로도 메일을 분류할 수 있는 방법을 제시하고자 한다. 본 논문에서는 저빈도의 단어에 대해서 신뢰할 수 있는 가중치 부여기법인 카이제곱 통계량 기법을 이용한다[3]. 카이제곱 통계량은 일반적으로 통계분야에서 기대 도수와 관측도수의 차이가 유사한지를 판단하는 기준 또는 방법으로 사용되는 것으로 특정 단어와 문서 범주 간의 관계도를 측정하는데 사용할 수 있다[2,3,5]. 본 논문에서는 약 500개의 메일을 <표1>과 같이 스팸메일 범주와 비스팸메일 범주로 나누고, 각 메일에서 형태소 분석을 통해 추출된 결과 중 체언 정보만을 가지고 스팸메일에서 쓰이는 용어에 대한 가중치를 부여한 사전을 <표2>와 같이 구축한다.

<표1>단어와 스팸메일 범주간의 문서빈도 테이블

	스팸 메일 범 주	비스팸 메일 범 주
단어 t가 출현한 문서	A	B
단어 t가 출현 하지 않은 문서	C	D

<표2> 메일에 대한 사전의 예

단어	스팸메일에 나타난 횟수	비스팸메일에서 나타난 횟수	가중치
:	:	:	:
가격	48	25	8.55
가입	61	53	0.72
경매	34	0	37.16
정품	23	9	6.56
기념	82	51	9.99
긴급	15	24	0
뉴스	27	63	0
당첨	32	2	28.81
대박	10	0	10.22
돈	23	11	4.55
무료	187	98	22.30
이벤트	81	31	29.96
:	:	:	:

수신된 메일의 본문에 대해 태그 및 불용어 제거의 전처리 과정을 거쳐 형태소 분석 후 수신 메일에 대한 체언정보를 얻는다. 체언과 메일 사전을 이용하여 카이제곱 통계 방식을 통해 추출된 체언정보와 스팸메일의 관계도를 측정하여 이를 단어의 가중치로 부여 한다.

$$\chi^2(t,spam) = \frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \cdots(1)$$

위 식(1)은 체언정보 t와 스팸메일과의 관계도를 계산하는 공식이다. 이 식에서의 결과값이 단어 t의 가중치 값이 된다. 예를 들어 수신된 메일에서 추출된 체언정보가 "이벤트" 라면, 단어는 스팸메일 문서 250 개중에서 81개의 스팸메일 문서에서 한번이상 출현 했고, 비스팸메일 문서중에 31개의 문서에서 한번 이상 출현했다면,

N=500, A=81, B=31, C=169, D=219이고,

$\chi^2(이벤트,spam) = 29.96$ 가 되어, "이벤트" 는 29.96의 가중치값을 가진다. 다른 예로 "가입" 이라는 단어는

N=500, A=61, B=53, C=189, D=197이고,

$\chi^2(가입,spam) = 0.72$ 의 가중치값을 가지게 되므로, "이벤트" 라는 단어가 "가입" 이라는 단어보다 스팸메일 과 더욱 밀접한 관계를 가진다고 보는 것이다.

어떠한 단어가 스팸메일에서 출현하는 빈도와 비스팸메일에서 출현하는 빈도를 이용하여, 카이제곱 통계량 공식에 의해 계산되어진 가중치 결과값은, 메일 본문에서 추출된 체언정보가 스팸메일에 관여하는 정도를 나타낸다. 만약 출현 단어가 스팸메일에서나 비스팸메일에서나 출현하는 빈도수가 유사하다면, 스팸메일을 분류하는 데 그 단어가 영향을 주어서는 안 된다. 그렇기 때문에 스팸메일에서나 비스팸메일에서 유사한 빈도수를 가진 단어에 대한 가중치는 낮게 된다.

3.2 로그단어 가중치를 이용한 분류

문서분류 알고리즘으로는 확률을 이용한 방법, 통계적인 기법을 이용한 방법, 벡터 유사도를 이용하는 방법, 엔트로피를 이용하는 방법등이 있다[5]. 이들 중 본 논문에서는 단어의 빈도가 지나치게 낮은 영향력을 보충하고, 단어의 빈도가 높은 단어의 지나친 영향력을 낮추기 위해 스팸메일을 분류하는 단계에서 로그단어 빈도 가중치공식을 적용한다[5,6]. 본 논문에서는 동일한 스팸메일 관련 단어라도 텍스트의 양에 따라 메일에서 차지하는 비중이 다르게 적용이 되어야 하기 때문에 메일 본문의 전체 체언의 수와 스팸메일 관련 체언의 수에 대한 비율로서 정규화를 한다.

본 논문에서는 메일의 본문을 형태소 분석과정을 통해 추출된 체언들과 카이제곱 통계량에 의해 가중치가 부여된 사전을 이용하여 가중치의 합과 수신된 메일 내에서 추출된 전체 체언의 수와 스팸메일 단어의 출현 빈도를 통하여 분류한다.

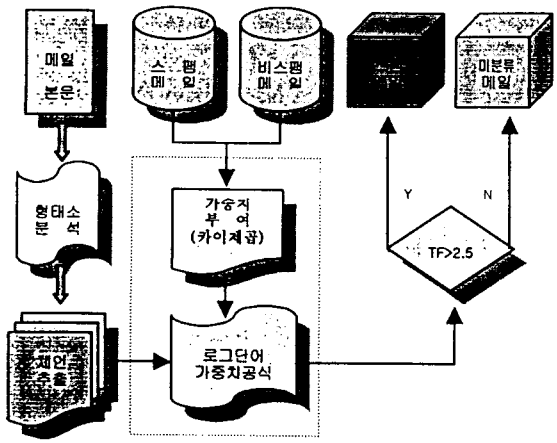
$$TF = 1 + \log(\text{용어가중치의합} \times \frac{\text{스팸관련체언수}}{\text{전체체언의수}}) \cdots(2)$$

식(2)에 의해서 계산된 결과값 TF가 임계치 이상이면 스팸메일로 분류가 된다.

3.3 2차 필터링 구성도

수신된 메일 중 1차 필터링에서 필터링되지 않은 메일은 2차 필터링을 거치게 된다. 먼저 메일의 본문을 형태소 분석하여 체언정보를 추출하고, 추출된 체언의 가중치를 사전에서 가져와 메일 본문의 단어 가중치의 합을 구한다. 이렇게 구해진 메일 본문의 단어 가중치의 합을 로그단어 가중치 공식으로 계산하여 결과값이 임계치 이상이 되면 스팸메일로 분류한다.

<그림 2>와 같이 두 번의 가중치 계산 공식을 거치면서 각 단어에 대해서 그리고 수신된 메일내의 스팸메일 관련 단어의 상대적 출현빈도를 가중치로 계산하므로 좀더 정확한 필터링이 가능하다.



<그림 2> 2차 필터링 구성도

4. 실험 및 결과

본 논문에서 제안한 방법에서 단어의 가중치 계산을 위해 250개의 스팸메일과 250개의 비스팸메일을 가지고 메일 사전을 구축했다. 스팸메일로 구분된 메일들은 경품정보, 상품정보, 주식정보, 사이트 소개등과 관련된 메일들로 구성되어 있고, 비스팸 메일은 정치, 경제, 사회, 문화의 뉴스레터들과 일반적인 메일로 구성되어 있다. 형태소 분석을 위한 도구로서 KAIST의 KTS (Korean Tagging System)을 사용하였으며, 스팸메일에서 추출한 8,258개의 단어정보와 비스팸메일에서 추출한 11,246개의 단어정보를 가지고 출현한 단어들의 가중치를 계산하여 스팸메일 사전을 구축했다. 사전 구축에 사용하지 않은 새로운 52개의 스팸메일과 35개의 비스팸메일을 테스트 메일로 실험했을 때, 임계치가 2.5인 경우에 결과는 <표3>과 같이 스팸메일 44개와 비스팸메일 9개가 스팸메일로 분류되고, 나머지가 비스팸메일로 분류되었다.

<표3>임계치 2.5일때 필터링 결과

	스팸메일	비스팸메일
메 일 수	52	35
필터링후	53	34
분류성공	44	26
오 분 류	9	8

위 실험 결과, 논문에서 제시한 메일 필터링 시스템은 재현율 84.61%와 정확율 83.01%의 성공율을 보였다. 실험에서 나타난 오분류된 메일의 대부분은 HTML형식의 메일들이며, 이러한 메일들을 분류하기에는 텍스트의 양이 아주 적기 때문 발생하였다.

5. 결론 및 향후 발전 과제

본 논문에서는 스팸메일을 필터링 하기위해 문서분류 알고리즘을 스팸메일 필터링 시스템에 적용하여 기존의 스팸메일 필터링 시스템보다 좀더 근본적이고 정확한 필터링 방법을 제시한다. 이는 메일 본문의 내용을 분석하여 필터링하기 때문에 메일의 제목이나 송신자의 주소등과 같은 메일헤더 정보만으로 필터링 할 수 없는 스팸메일도 필터링 할 수 있다. 그러나 좀더 개선된 필터링을 위해서는 스팸메일과 비스팸메일의 분류기준을 명확히 하여 사전을 구축해야 하며, 정확한 형태소 분석을 통해 본문의 체언정보를 정확하게 추출해야 할 것이다. 또한 단순한 단어의 출현 빈도에 따라 가중치를 부여하기 보다는 단어의 의미적 자질에 대해 가중치를 부여한다면 좀더 정확한 필터링이 가능할 것이다. 또한 대부분의 메일들이 텍스트의 양이 적고 이미지의 양이 많은 HTML 형식으로 구성되어 있으므로 단어 정보와 함께 메일 내의 이미지 정보도 필터링에 이용한다면 적은 양의 텍스트로 필터링을 할 때 나타나는 오분류등의 문제점을 해결하고 좀더 명확한 분류가 가능할 것이다.

6. 참고 문헌

[1]. 황도삼, 최기선, 김태석 공역, "자연언어처리" 홍릉과학 출판사, 1999.
[2]. 신진섭, 이창훈, "단어의 연관성을 이용한 문서의 자동분류" ,한국정보처리학회,정보처리논문지,제6권 제9호, pp.2422-2430, 1999.
[3]. 한광록,선복근,한상태,임기옥, "인터넷 문서 자동분류 시스템 개발에 관한 연구" ,한국 정보처리학회 논문지 제7권 제9호, pp.2867-2875, 2000.
[4]. 김상범, "범주간의 상호관계를 고려한 자동 문서 범주화의 개선" , 고려대 석사학위 논문, 1999.
[5]. 고수정, 이정현, "Apriori알고리즘에 의한 연관 단어 지식 베이스에 기반한 가중치가 부여된 베이지안 자동 문서 분류" ,멀티미디어학회 논문지, 제4권, 제2호, pp.171-181, 2001.
[6]. 이재윤, 최보영, 정영미, "문헌 자동분류에서 용어가중치 기법에 대한 연구" ,제7회 한국정보관리학회 학술대회 논문집, pp.41-44, 2000.
[7]. 강원석, 황도삼, 최기선, "의미의 상하위 정보를 이용한 웹문서 분류 시스템" ,제11회 한글 및 한국어 정보처리, pp.36-39, 1999.
[8]. 김진상, 신양규, "베이지안 학습을 이용한 문서의 자동분류" ,정보과학회논문지,Vol.11,No.1,pp.19-30, 2000.