나이브 베이지안 분류자와 메일 주소 유효성 검사를 이용한 스팸 메일 필터링 시스템

임정택, 김형준, 강승식
국민대학교 컴퓨터학부
arar5325@paran.com, dictions@nate.com, sskang@kookmin.ac.kr

Spam-Mail Filtering System by Using Naive Bayesian Classifier and Mail Address Validation Check

Jung-Taek Lim, Hyung-Joon Kim, Seung-Shik Kang
School of Computer Science, Kookmin University

요 약

본 논문에서는 가중치가 부여된 나이브 베이지안 분류자와 스팸 메일의 특성을 이용한 주소 유효성 검사를 결합하여 필터링하는 방식의 스팸 메일 필터링 시스템을 제안하였다. 주소 유효성 검사를 통해 스팸 메일을 효율적으로 필터링 할 수 있으며, 나이브 베이지안 분류자에 가중치를 부여함으로써 더욱 효과적인 분류를 할 수 있다. 또한, 각 요인의 중요도에 따라 다른 비중을 부여함으로써 메일의 특성을 고려한 필터링 환경을 구현하였다. 실험에서는 제안하는 요인들이 실제로 필터링 성능 향상에 어떤 영향을 미치는지 살펴보고 최적의 시스템 성능을 측정하였다.

1. 서론

전자 우편은 최소의 비용으로 실시간으로 배달된다는 편의성과 더불어 다수의 수신자에게 동시에 전송하는 특성으로 인해 오프라인 우편을 대체하는 수단으로 사용되고 있으며, 또한 데이터 전송이나 데이터 백업의 목적으로 활용되기도 한다. 그런데 매우 저렴한 비용으로 불특정 다수에게 광고물을 전송하는 마케팅 수단으로 사용되면서 상업적 광고성 메일이 무분별하게 발송되고 있다[1]. 이에 따라 메일 서비스 업체부터 개인 사용자에 이르기까지 많은 피해자가 속출하게 되었고, 그 피해 규모 또한 적지 않다.

스팸 메일이 심각한 사회 문제가 되는 문제점을 해결하는 방안으로 문서 분류에 이용되는 나이브 베이지안 분류자를 스팸 메일 필터링 시스템에 이용하는 방법이 시도되었다[2,3,4,5]. 나이브 베이지안 분류자(Naive Bayesian Classifier)는 문서 내의 키워드들을 기준으로 분류를 수행한다. 메일의 경우, 메일 헤더에서 얻을 수 있는 정보와 메일 본문에서 얻을 수 있는 정보를 최대한 활용하여 스팸 메일 분류에 특화된 분류자를 만들 수 있다. 본 논문에서는 받는 사람의 이메일 주소의 유효성을 검사하고, 제목과 본문에 대해 각각 나이브 베이지안 분류자를 적용하고 각 결과에 대해 가중치를 부여하는 스팸 메일 필터링 시스템을 구현하고, 가중치 지정에 따른 최적의 성능을 평가하였다.

본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 부분적인 지원을 받았음. 시스템 설계 및 구현에 국민대학교 컴퓨터학부 졸업 프로젝트 팀원(임연우, 김수연, 송성룡)이 기여하였음.

2. 스팸 메일 필터링 시스템

스팸 메일 필터링 시스템은 학습 모듈과 분류 모듈로 구분된다. 학습의 전제 조건으로, 먼저 학습 메일 문서들은 사용자에 의해 스팸과 정상 메일 셋으로 분류되어 있어야 한다. 학습 모듈은 스팸/정상 메일을 파싱하고 색인어 추출기를 이용하여 제목과 본문에 대해 각각 키워드를 추출한다. 추출된 키워드들은 {키워드, 스팸에 출현한 빈도수, 정상 메일에 출현한 빈도수} 쌍으로 저장 및 갱신되어 차후 나이브 베이지안 분류자의 데이터로 사용된다. 분류 모듈은 학습 모듈과 같은 방법으로 메일을 파싱한다. 이후, 파싱된 헤더에서 주소 유효성 검사를 시행하여 주소의 유효성을 검사한다.

2.1 주소 유효성 검사

스팸 메일의 특성을 찾기 위하여 데이터 집합을 분석하는 중 스팸 메일에 유독 메일 헤더 내에 받는 사람의 주소가 없거나 해당 사용자의 메일 주소가 아닌 경우가 많음을 발견할 수 있었다. 그 원인은 SMTP에게 전달하는 송신자와 수신자 정보와 메일 헤더 간의 관계가 없음으로 인한 것이다. 스팸 메일 발송자들은 이를 이용하여 메일 헤더를 일일이 수정하지 않고도 메일 주소가 다른 여러 사용자에게 같은 메시지를 발송할 수 있다.

스팸 메일 필터링 시스템은 이런 특징을 이용하여 사용자의 메일 주소와 메일 메시지 내의 헤더의 받는 사람 주소를 비교하여 스팸 메일을 효율적으로 필터링한다. 즉, 메일을 파싱하여 헤더를 추출하면 추출된 헤더에 대해 주소 유효성을 검사하고 유효하지 않은 경우 필터링함으로써 필터링 성능이 향상된다.

2.2 가중치가 부여된 나이브 베이지안 분류자

본 논문에서는 정보 검색과 문서 분류에서 사용되는 tf*idf를 각 단어의 가중치로 사용한 나이브 베이지안 분류자를 사용하여 필터링을 수행한다. tf*idf는 공통으로 포함된 단어에 대해서 그 단어들이 전체 문서에서 희귀한 단어일 경우 특징적인 키워드라고 추측, 상대적으로 높은 값으로 전환하기 위한 방법이다. 문서 D_i에서 단어 t_k가 나온 횟수를 tf_{ik}, 문서 전체의 수를 N, t_k를 포함하는 문서의 수를 n_k라고 했을 때, tf*idf를 이용한 t_k의 가중치 w_{ik}는 다음과 같다.

$$w_{ik} = tf_{ik} * \log(N/n_k) \quad (1)$$

이와 같이 tf*idf를 가중치로 부여함으로써, 분류 대상이 되는 메일의 특징적인 키워드에 높은 가중치를 부여하게 되어 필터링 성능을 향상시킨다.

2.3 제목-내용의 중요도 반영

스팸 메일 여부를 판단하기 위해 주소 유효성 검사 및 나이브 베이지안 분류자에 의해 메일 제목과 본문 내용에 대해 각각 스팸 메일 확률을 계산하는 방법을 적용한다. 그런데 통상적으로 스팸 메일의 제목에 주로 사용되는 특징적인 용어들이 발견되므로 제목과 본문 내용에 출현하는 용어의 중요도를 차별화함으로써 필터링 성능을 향상시킨다.

제목과 본문 내용의 상대적인 중요도를 차별화하여 비중을 다르게 반영하기 위하여 제목-내용 각각에 대해 계산된 스팸 확률값의 반영 비율을 다르게 적용한다. 제목과 본문의 반영 비율을 a, β라 할 때 $a + \beta = 1$ 이며, 구체적인 반영 비율은 최적화 실험을 통해 0.65, 0.35로 정하였다.

또한, 주소 유효성 검사에 의해 유효하지 않은 주소를 가진 메일은 모두 스팸으로 간주하는 방법도 있으나, 스팸이 아닌 경우도 있으므로 스팸 여부를 판단할 때 임계값을 차등 적용하도록 한다. 정상적인 메일의 임계값을 τ라 할 때, 주소 유효성 검사를 통과하지 못한 메일은 스팸이 가능성이 높으므로 임계치를 낮춰서 $\tau - a$ 로 적용한다. 상수 a값은 유효 주소인 경우 0이고, 그렇지 않을 때의 값은 최적화 실험을 통해 0.08로 정하였다.

최종적으로, 제목-내용의 중요도를 반영하고 주소 유효성 검사에 따라 임계치를 조절하여 스팸 메일 확률을 계산하는 식은 (2)와 같다.

$$a\frac{P(c_s|E_T)}{P(c_s|E_T)+P(c_n|E_T)} + \beta\frac{P(c_s|E_B)}{P(c_s|E_B)+P(c_n|E_B)} > \tau - a \quad (2)$$

전자 우편 E의 제목을 E_T, 본문의 내용을 E_B 로 분리하여 각각에 대한 스팸 확률을 계산할 때 E_T와 E_B의 스팸 확률은 식 (1)과 같이 스팸 범주에 속할 확률과 정상 메일 범주에 속할 확률을 각각 계산한 후에 스팸으로 분류될 비율로써 스팸인지를 판단한다.

3. 실험 및 성능 평가

학습 및 실험에 사용된 메일들은 수집된 실제 한/영 메일이며, 학습 메일 3,538개와 실험용 1,517개이다. 실험에 앞서 여러 요인들의 조합으로 분류 모델을 표 1과 같이 구성하였다.

표 1. 필터링 성능 실험 모델

모델	설 명
F_N	기본적인 나이브 베이지안 분류자
F_A	주소 검사 사용 (NB 분류자 사용하지 않음)
F_NA	주소 검사 + NB
F_T	NB + tf/idf 가중치
F_AT	주소 검사 + NB + tf/idf 가중치
F_ATR	NB + tf/idf 가중치 + 제목-내용-주소 비중 조절

3.1 최적의 임계값 계산

실험 전체에서 사용할 최적 임계값은 기본 나이브 베이지안 방법 F_N의 F-measure측정을 통해 구하였다. 데이터는 학습 데이터를 사용하였고, 각각의 임계값 변화에 따른 스팸 메일 필터링 성능은 그림 1과 같다.

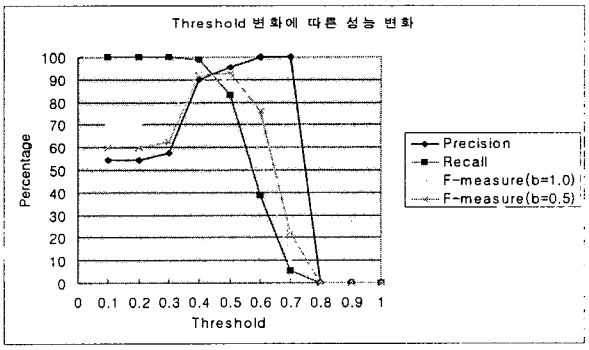

그림 1. 임계값 변화에 따른 필터링 성능

임계값을 0.1에서 1까지 변화시키면서 F-measure를 측정한 결과 b값을 1로 주었을 때에는 임계값이 0.4일 때 94.25%를 기록한 것이 최대이고, b값을 0.5로 주었을 때에는 임계값이 0.5일 때 92.64%를 기록한 것이 최대이다. 본 논문에서는 b값을 0.5로 취하였으므로 임계값 t는 0.5로 정하였다.

3.2 주소 검사에 따른 필터링 성능

이 절에서는 우선 주소 검사만으로 필터링한 성능을 측정하고, 실제로 나이브 베이지안을 사용할 때 주소 검사의 유무가 어느 정도의 성능 향상을 보이는지 측정

하였다. 주소 검사의 효용성만을 평가하기 위해 나이브 베이지안 분류자는 tf/idf 가중치를 사용하지 않았다.

주소 유효성 검사를 통해 주소가 잘못된 경우에 무조건 스팸으로 필터링해 보았다. 결과를 살펴보면, 가장 높은 정확률인 98.75%를 기록하였고 재현율 또한 76.98%를 기록하였다. 이를 통해 1.25%의 작은 정확률 하락을 통해 약 77%의 스팸 메일을 나이브 베이지안 분류자를 사용하지 않고 빠르고 확실하게 필터링할 수 있음을 확인하였다. 또한 혼합하여 사용할 경우 세 가지 중 가장 높은 성능을 보임으로써 주소 검사의 사용이 필터링 성능 향상에 도움이 됨을 확인하였다.

나이브 베이지안 분류자에 tf/idf를 적용하는 경우와 적용하지 않는 경우의 필터링 성능을 측정해 보았다. 효용성을 평가하기 위해 주소 유효성 검사는 하지 않았다. 나이브 베이지안 분류자에 tf/idf 가중치를 적용할 경우 적용하지 않은 경우보다 정확률과 재현율 모두 향상되었다. 또한 f-measure에서 2.62% 높은 성능을 보임으로써 tf/idf 가중치 적용이 필터링 성능 향상에 도움이 됨을 확인하였다.

3.3 제목-내용-주소 검사 비중 조절

제목-내용-주소 검사의 비중은 제목이 스팸일 확률에 부여되는 비중, 내용이 스팸일 확률에 부여되는 비중, 주소 체크를 통과하지 못하였을 경우에 부여되는 비중으로 총 3 가지이다. 먼저 제목-내용 비중의 최적값을 찾는 실험을 진행하였다. 두 값이 서로 상대적인 값을 갖도록 하기 위해 제목 비중과 내용 비중의 합이 1이 되도록 하였다. 제목 비중을 0.05 단위로 증가시키면서 필터링 성능의 변화를 관찰하였다. 또한, 임계값은 t를 1을 기준으로 하였을 때 0.5로 정하였으므로 그와 같은 0.5로 정하였다. 실험 결과, 정확도는 약간의 상승세를 보이다가 제목에 0.6 이상의 비중을 부여하면서 다시 하락세를 보였고, 재현율은 지속적인 상승을 보였다.

이 중 제목에 0.65의 비중을 부여하였을 때 F-measure가 최고치인 96.39%의 결과를 보여 최적의 제목 비중을 0.65, 내용은 0.35로 정하였다. 다음 단계로 앞에서 최적의 제목과 내용의 비중을 이용하여 주소 검사 비중의 최적값을 구하기 위한 실험을 진행하였다. 주소 검사 비중은 0과 0.1 사이에서 0.01 단위로 변화를 주었다.

실험 결과 주소 검사에 대한 비중을 증가시킴에 따라 정확도는 조금씩 내려가고 재현율은 상승하였다. 그러나 0.09부터 정확도가 크게 떨어졌고, 재현율은 0.06부터 적은 폭으로 상승하는 결과를 보여 전체적인 F-measure는 최고치 이후로 계속 떨어지는 것을 확인할 수 있었다. 이로 인해 0.1 이상의 비중은 의미가 없음을 알 수 있었다. 이 중에서 주소 검사 비중을 0.08로 주었을 때 F-measure가 최고치인 97.18%의 성능을 보여 최적의 주소 검사 비중을 0.08로 정하였다. 이는 본 논문에서 제안하고 구현한 스팸 메일 필터링 시스템의 최적의 성능이기도 하다.

3.4 성능 평가

여러 가지 요인들을 모두 사용한 시스템의 필터링 성능을 측정하였다(표 2). 실험 결과, 모든 요인을 사용한 경우 정확률이 F_A와 F_T보다 약간 낮은 성능을 보이지만 재현율은 91.6%로 가장 높은 성능을 보임으로써 F-measure 측정에서 95.31%로 가장 높은 성능을 보였다. 이는 나이브 베이지안 분류자만 사용한 경우보다 정확률 2.1%, 재현율 11.57%, F-measure 4.24% 향상되었으며, 일반적으로 많이 쓰이고 있는 F_T보다 정확률 0.61%, 재현율 8.41%, F-measure 2.35% 향상되었다.

표 2. 스팸 메일 필터링 성능 비교

모델	스팸 ->스팸	정상 ->정상	스팸 ->정상	정상 ->스팸	정확률	재현율	F-measure
F_N	686	662	135	34	95.28%	83.56%	92.68%
F_A	632	654	189	8	98.75%	76.98%	93.46%
F_NA	743	654	78	42	94.65%	90.5%	93.79%
F_T	711	675	110	21	97.13%	86.6%	94.83%
F_AT	752	667	69	29	96.29%	91.6%	95.31%
F_ATR	780	678	41	18	97.74%	95.01%	97.18%

4. 결론

본 논문에서는 나이브 베이지안 분류자에 가중치를 부여하고, 주소 유효성을 검사하는 방법을 추가하며, 각 요인의 중요도에 따른 비중을 부여하여 필터링 성능을 향상시키는 방안을 제안하고 실제로 시스템을 구현하여 성능을 평가하였다. 실험을 통해 필터링 성능 향상 방안으로 제시한 방법들이 실제로 필터링 성능 향상에 기여함을 밝혔으며, 이를 적절히 통합한 시스템이 단순히 나이브 베이지안 분류자만을 사용하는 시스템에 비해 높은 성능을 보임을 확인하였다.

참고문헌

[1] 정보통신부, 정보통신망 이용 촉진 및 정보 보호 등에 관한 법률 시행령 제11조 "영리 목적의 광고성 전자 우편의 명시 방법", 2002.

[2] 조한철, 조근식, "나이브 베이지안 분류자와 메시지 규칙을 이용한 스팸 메일 필터링 시스템", 제29회 한국정보과학회 춘계학술대회, 제29권 1호, pp. 223-225, 2002.

[3] Sahami, M., Dumais, S., Heckerman, D. and Horvitz, E., "A Bayesian Approach to Filtering Junk E-Mail". Proceedings of the AAAI Workshop, pp.55-62, 1998.

[4] Diao, Y., Lu, H. and Wu, D., "A Comparative Study of Classification Based Personal E-mail Filtering", Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp.408-419, 2000.

[5] 박정선, 김창민, 김용기, "퍼지 관계곱을 이용한 내용기반 정크 메일 분류 모델", 정보과학회논문지: 소프트웨어 및 응용, 제29권 10호, pp.726- 735, 2002.

[6] 김현준, 정재은, 조근식, "가중치가 부여된 베이지안 분류자를 이용한 스팸 메일 필터링 시스템", 한국정보과학회 논문지:소프트웨어 및 응용, 제31권 8호, pp.1092-1100, 2004.