광고성 메일을 자동으로 구별해내는 Text Mining 기법 연구

이종호
서울대학교 인지과학 협동과정

Detecting spam mails using Text Mining Techniques

Jong-Ho Lea
Interdisciplinary Program in Cognitive Science, Seoul National University

요 약

광고성 메일이 개인 당 하루 평균 10통 내외로 오며, 그 제목만으로는 광고메일을 효율적으로 제거하기 어려운 현실이다. 이러한 어려움은 주로 광고 제목을 교묘히 인사말이나 답신처럼 변경하는 데에서 오는 것이며, 이처럼 제목으로 광고를 삭제할 수 없도록 은폐하는 노력은 계속될 추세이다. 그래서 제목을 통한 변화에 적응하면서, 제목 뿐만 아니라 내용에 대한 의미 파악을 자동으로 수행하여 스팸 메일을 차단하는 방법이 필요하다. 본 연구에서는 정상 메일과 스팸 메일의 범주화(classification) 방식으로 접근하였다. 이러한 범주화 방식에 대한 기준을 자동으로 알기 위해서는 사람처럼 문장 해독을 통한 의미파악이 필요하지만, 기계가 문장 해독을 통해서 의미파악을 하는 비용이 막대하므로, 의미파악을 단어수준 등에서 효율적으로 대신하는 text mining과 web contents mining 기법들에 대한 적용 및 비교 연구를 수행하였다. 약 500 통에 달하는 광고메일을 표본으로 하였으며, 정상적인 편지군(500 통)에 대해서 동일한 기법을 적용시켜 false alarm도 측정하였다. 비교 연구 결과에 의하면, 메일 패턴의 가변성이 너무 커서 wrapper generation 방법으로는 해결하기 힘들었고, association rule analysis와 link analysis 기법이 보다 우수한 것으로 평가되었다.

1. 서론

최근 들어 전자우편(e-mail)이 기업 홍보의 주요 수단이 됨에 따라, 개인들의 업무와 관련이 없는 광고우편(spam mail)의 처리가 개인들의 업무 효율을 위해 주요한 관심사가 되고 있다. 2001.12말 현재 국내 인터넷 이용자 수는 2,438만명으로 전체 인구의 56%에 해당하며, 이중 1,970만명이 전자우편으로 정보와 소식을 주고받고 있다[5]. 나라리서치에서 조사한 바에 따르면 개인들이 전자우편 중에서 스팸메일을 지우는 데 소비하는 시간은 연평균 44시간이며, e-mail 주소당 하루 평균 9.3통이며, 개인들이 보유하고 있는 메일 주소는 평균 4.83 개였다[1]. 이러한 개인 업무 효율의 저하가 국가적으로 연간 2조 6천억원이라는 추정도 있다[1].

심파일(http://file.simmani.com)에서 조사한 바에 의하면, 개인들이 스팸 메일(spam mail)을

처리하는 방법으로는, 응답자 2,908명 가운데 절반이 넘는 58.0%(1,688명)가 '무조건 삭제한다'고 답한 반면, '메일 프로그램의 필터링 사용'과 '메일마다 거부의사 표명'이 각각 14.8%(430명), 11.1%(322명)로 조사되어 주로 소극적인 대처를 하고 있는 것으로 분석됐다[2].

개인들이 이러한 소극적인 대처에 많이 의존하는 이유는, 제목만으로 광고를 식별하는 필터링 방식에 대처하기 위해 광고 발송자들이 광고 제목을 교묘히 위장하기 때문이며, 따라서 메일 프로그램에 의한 필터링 방식이 거의 무용지물이 되고 있는 현실이다[5]. 심파일(http://file.simmani.com)이 최근 네티즌 6,913명을 상대로 설문조사한 결과 '최근 가장 많이 보는 스팸메일의 제목'으로 응답자의 26.2%가 'Re:답변입니다'를 꼽았다. 이어 '오빠 나야'라는 제목이 19.7%로 뒤를 이었고 '당첨되셨습니다(12.5%)' '누구누구 동영상 공개(11.6%)도 많았다[3].

물론 공정거래위원회에서 2002년 7월부터 전자상거래보호법을 통한 스팸메일의 차단책으로, 수신거부 표시 의무화와 제목에 광고 표시 의무화, 메일주소 출처명시 등이 주요 골자인 해결책을 제시하고 있다[4][5]. 공정위가 제시한 불법 사례는 상업 메일이면서 '광고'라고 표시하지 않은 경우 [광*고], [광~고] 등과 같이 광고 메일임을 알아차리기 어렵게 표시한 경우 '답장' 'Re:질문' 등의 형태로 수신자를 속이는 경우 등이다[4].

하지만, 소비자를 기만하는 제목에 대한 기준이 다소 임의적인 측면이 있어서, 이런 유형의 제목 변형이 계속될 소지가 있다. 이러한 이유로 기존 전자우편에 대한 분류시스템이 제목과 내용의 일치성에 의존해왔던 점에서 벗어나[6], 전자우편 내용에 대한 분석이 더 강조되어야 할 필요가 있다.

본 연구에서는 전자우편의 내용을 단어 수준에서 파악할 수 있는 방안과 더불어, 전자우편의 내용을 간접적이면서도 분명하게 알 수 있는 hyperlink 구조분석을 병행하였다. 본 논문의 구성은 전자우편의 표상(representation)과 특성을 살펴보고, 스팸메일을 자동분류하는 방안과 그 처리 결과에 대해 살펴보는 순서를 택하였다.

2. 전자우편(e-mail)의 표상과 특질 선택 (feature selection) 방안

텍스트 마이닝(text mining)에서는 문서(document)를 "bag of words"로 보는 경향이 많다[6][7][8][10]. 이 "bag of words" 방식(단어들의 묶음으로서 문서를 보는 방식)은 문서가 갖고 있는 통사적 구조(syntactic structure)와 구문 순서 등을 무시하게 되므로 정보의 손실은 많으나 기계학습(machine learning)에 사용하기 편리한 표상(representation)인 feature vector로 변환하기 쉬운 측면 때문에 애용되고 있다[12][6][7][9].

전자우편은 from, to, subject, body 등으로 구별되는 구조가 있기는 하지만, 엄격하지 않고 생략이 가능한 "semi-structured data"이다[14]. 이러한 정보의 생략 가능성과, 광고의 내용을 보게 하려는 의도에서 제목 등을 기만하기 때문에, 문서 구분의 주요 단서 역할을 해왔던 제목(subject)에 의존한 전자우편 분류 방식(Wrapper 방식)이 어렵게 되었다[6][7][13].

제목은 그 길이가 보통 한 줄을 넘지 않으므로 길이가 일정한 feature vector로 변환하는 것이 쉬운 편이다. 제목에 등장하는 많은 단어들은 그 어휘를 제한하기 힘들기 때문에, 등장유무를 나타내는 Boolean 방식의 infinite attribute model로 표현하는 것은 적합하지 않다. 그래서, 다양하고 희귀한(sparse) 단어의 등장에 대해서도 처리할 수 있는 set-valued feature model(RIPPER 방식)이 적합하다[6][12][15].

앞서 이야기한 것처럼 제목을 통한 분류 방식이 스팸 메일의 기만제목 때문에 적용하기 어렵게 되었으므로, 내용(body)을 통한 분석방법이 필요한데, 제목 분석에서 사용하는 set-valued feature model 방식을 적용하기는 어려운 점이 있다. 그 이유는 우선, 편지의 길이가 제 각각이어서 길이를 표준화(normalize)하기 어렵고, 편지 내용은 그 형식이 일반 텍스트, HTML, 이미지 등으로 다양하기 때문이다. 그래서 본 연구에서는 분석하고자 하는 전자우편의 모델링을 다음과 같이 하였다.

type e-mail = title | body

type title = set(label * string)

type body = set(selected_string * anchor * form)

type selected_string = set(first40words, last20words)

전자우편(e-mail)에서 from 부분이 빠진 이유는 일반적인 광고들이 발신인을 여자 가명으로 보내는 경우가 대부분이기 때문이다(381통/500통).

Body 부분을 일반 text(string)로만 보기 어려운 이유는 대부분의 메일이 다양한 MIME 타입(encoding)을 갖고 있고, HTML 형식의 메일(form, table 등)이 많으며, 심지어는 광고임을 탐지하기 어렵도록 text 대신 이미지(jpg,gif)안에 광고 문안을 삽입(a href, img src)하는 사례도 많기 때문이다.

이러한 이유로, body 부분의 의미파악을 위해서 text도 이용하지만(RIPPER[6]), 광고의 대부분이 자신이 밝히고자 하는 내용을 별도의 hyperlink (form 부분의 action, a href 부분, img의 alt 부분)에서 표시하는 것에 착안해서, domain 주소일 경우에는 점(dot)을 기준으로 각 단어를 수집하여 다음 3장에서 나오는 hypernym 구조의 동의어들(synonym set; synset)과 일치하는 것이 있는지를 확인한다.

(예) www.girl.com ={www, girl, com}

Cohen[6], Provost[7]의 연구에 의하면, 사용자들이 며칠 동안만 오는 광고들을 별도의 폴더에 수집을 하면, 대략 20통의 메일이 될 것이다. 이렇게 모인 20통의 메일에 대해 기계학습을 하면, 그 광고의 경향성이 어느 정도 정확하게 파악된다고 하였다[6][7]. 이렇게 수집된 광고들을 가지고 decision tree 같은 rule을 학습할 수도 있겠으나[6], 사용자마다 광고에 대한 기준이 민감하게 다를 수 있으므로, 사용자가 개입할 수 있는 여지를 마련하는 것이 바람직할 것이다.

그래서 본 연구에서는 다음 3장에서 사용자들이 자신의 판단 기준을 개입시키면서, 기계학습으로 그 판단 기준의 관련도(relevance)를 알려주는 방식에 대해서 살펴보았다.

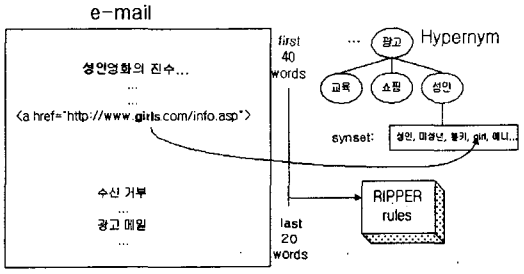

그림 1 RIPPER를 위한 특질 추출과 Hypernym 구조 예

3. 사용자의 판단 Hypernym 구조

사용자들이 판단 분류 깊이를 3 수준만 할 수 있도록 허용한 뒤에 해당 분류의 종단(leaf node)에 일반적인 분류(예: 광고-성인)를 두고, WordNet처럼 그 종단에 유사어들의 list(synset)를 두는 방식을 채택한다.즉, 사용자들이 판단 분류의 위계를 설정할 수 있을 뿐만 아니라, 이 위계에 속하는 특질(feature)을 직접 추가, 삭제할 수 있는 기능을 두도록 하는 것이 주안점이다.

그림 1에서 보듯이, "성인"이라는 종단(leaf)에 synset로서 "{출입금지, 미성년, 몰카, 애니메이션,…}"처럼 분류 기준어로서 사용자가 생각하는 단어들을 선택하거나, 추후에 삭제할 수 있다.

이러한 개인의 분류가 타당한지를 보여주는 기준으로서, 광고를 모으는 개인 폴더에 광고가 30통 이상 모였을 경우에는, 해당 synset 내의 단어끼리의 association rule 방식[15]을 도입해서 confidence와 support를 계산할 수 있을 것이다. 이렇게 계산된 confidence와 support를 이용해서 개인의 분류에 대한 지지 또는 수정 지침을 줄 수 있을 것이다. 가령, "애니메이션"이 비록 "성인" 분류에서 support가 많아도 동시에 "교육" 분류에서도 support가 많다면, 분류 기준어를 삭제할 수 있을 것이다.

그리고, 귀납추론(induction)에서 흔히 발생할 수 있는 Hempel's paradox(모든 까마귀는 까맣다는 것을 입증하기 위해 수많은 흰 운동화를 지지 근거로 삼는 역설)을 방지하기 위해, Kodratoff[11]가 제시한 세 번째 논항(argument) 방식을 채택하였다. 즉, 오직 같은 부모 노드(예: 새) 밑에 있는 다른 형제(sibling) 클래스(예: 까치 비둘기; 신발은 같은 형제 클래스가 아님)에 속한

문서들만을 골라서, 검토하고자 하는 분류 기준어(예: 까마귀)와의 association support를 조사해서 support가 적을수록 더 좋은 기준어가 되도록 측정하였다.

4. 실험 및 평가

4.1 데이터 구성 및 실험 방법

본 연구에서는 4명의 사용자가 1주일 동안 개인의 메일 계정 세 군데에서 온 광고메일(스팸메일)을 취합하여 500개의 메일 표집을 만들었다. 정상적인 업무 메일은 연구자가 6개월 동안 모아온 업무 관련 메일을 500개 모아서 대조군으로 하였다.

2장에서 살펴본 모델에서 추출한 특질(feature)들에 대해 RIPPER[6] 방식의 Wrapper[9][12]를 구성하여, 편지의 제목을 7단어만 고르고, 내용에서 html 태그와 anchor link등을 제외한 첫 40 단어와 끝 20 단어를 골라서, 이 단어들에 대해서 RIPPER 방식으로 rule을 생성하게 하였다(wrapper 방식).

두 번째 방법은 html의 form action 부분, img src 부분, a href 부분에 대해서 단어들을 추출하고, 이 단어들과 hypernym의 synset(사용자가 20 개의 메일에서 만든 hypernym과 synset)이 일치하는 상위 tag (분류 집단)를 선택하는 방식을 택하였다. 측정치는 일치하는 빈도 수로 하였으며, 역치는 3을 선택하여 겹치는 항목이 세 개 이상일 경우 그 분류집단에 속하는 것으로 평가하였다(link analysis 방식).

세 번째 방법은 앞에서 살펴본 1, 2번의 방법을 합친, 동등한 가중치의 voting 방법으로 판단하게 하였다(voting 방식).

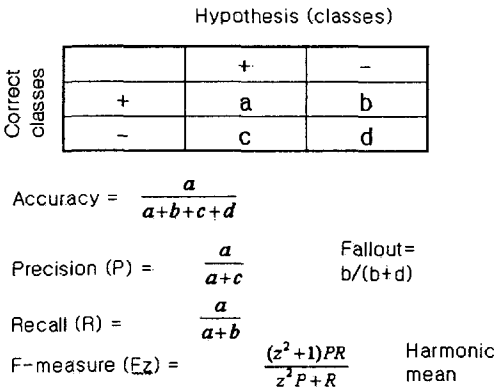

그림 2 편지 분류의 수행 측정치들

4.2 실험 결과

문서 분류의 정확도를 측정하는 방법은 그림 2에서 살펴보듯이 다양하다. 본 연구에서는 precision, recall, 그리고 F 값(z=1을 채택함. 즉, precision의 가중치를 recall과 동일하게 한 조화평균)을 앞서 4.1절에서 보았던 세 가지 연구 방법에 대하여 측정하였다.

	Pre-Cision	recall	F (z=1)	False alarm
wrapper	.858	.762	.404	.315
Link analysis	.564	.554	.279	.428
voting	.927	.884	.452	.07

연구 결과 wrapper 방식은, 전자 우편의 첫 40 단어와 끝 20 단어를 선택하는 범위의 제약성과, 광고 메일들이 핵심 내용을 직접적인 단어로 표현하지 않는 이유 때문에 정확도가 다소 나쁘고 false alarm이 높게 나타나는 것으로 보인다.

그래서 광고 메일의 본질적인 정보를 나타내는 link 정보를 이용하고, 아울러 기계가 자동으로 선택하기 어려운 특질(feature)에 대해 사용자가 정의한 hypernym과 synset을 이용한다면, 적은 수의 예(20통)를 가지고도 많은 수(500통에 대해 precision=.564)의 광고 메일에 대한 일반화가 가능한 것으로 보인다.

이러한 두 가지 방식이 결합한 voting 방식으로 정확도를 높이면서, false alarm도 줄이는 효과를 보일 수 있는 점이 발견되었다.

참고 문헌

[1] 스팸메일 손실 연2.36천억 (2002.4.30). 중앙일보. http://service.joins.com/asp/search_article.asp?aid=1730154&history=-4

[2] 네티즌 현혹 '스팸메일 제목들' (2002.4.15). 중앙일보. http://service.joins.com/asp/search_article.asp?aid=1722242

[3] Re:답변입니다 스팸메일 네티즌 현혹. (2002.4.17). 중앙일보. http://service.joins.com/asp/search_article.asp?aid=1723760&history=-2

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

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

[6] Cohen, W. W. (1996). "Learning rules that classify e-mail." *the AAAI Spring Symposium on Machine Learning in Information Access*, 18--25.

[7] Provost, J (1999). "Naive-Bayes vs. Rule-Learning in Classification of Email". http://www.cs.utexas.edu/users/jp/research/

[8] Sofus A. Macskassy & Haym Hirsh & Arunava Banerjee & Aynur A. Dayanik. (2001). "Using Text Classifiers for Numerical Classification", *IJCAI*, 885-890

[9] John G. & Kohavi R. (1997). "Wrappers for feature subset selection." *Artificial Intelligence*, v.97, 1-2, pp.273-324.

[10] Mladeni'c, D., Grobelnik, M., (1998). "Efficient text categorization", Text Mining workshop on the 10th European Conference on Machine Learning ECML98

[11] Kodratoff, Y. (1994). Induction and the Organization of Knowledge. In Michalski, R. & Tecuci, G. (Eds.), Machine Learning: A Multistrategy Approach (pp. 85-106). Morgan-Kaufmann.

[12] Sam Scott & Stan Matwin. (1999)."Feature engineering for text classification", Proc. 16th International Conf. on Machine Learning, pp.379--388, Morgan Kaufmann, San Francisco CA

[13] I. Muslea. (1998). Extraction patterns: from information extraction to wrapper generation. Technical report, ISI-USC.

[14] P. Buneman. (1997). "Semistructured data". Tutorial in Proceedings of the 16th ACM Symposium on Principles of Database Systems.

[15] J. S. Park, M. Chen, and P. S. Yu. (1995). "An effective hash based algorithm for mining association rules". In ACM SIGMOD Intl. Conf. Management of Data, May.