Proceedings of KFIS Autumn Conference 2006 Vol. 16, No. 2.

스팸 대응 시스템에서 특징 추출 방법 비교에 관한 연구

Comparison of Feature Selection Methods in Anti-Spam Systems

김종완[1], *김희재[1], 강신재[1], 황운호[1]

[1] 경북 경산시 대구대학교 컴퓨터 · IT공학부
E-mail: heejae0305@hanmail.net

요 약

본 논문에서는 스팸 대응 시스템의 특징 추출 방법들을 비교한다. 실험 결과는 퍼지추론 방법이 정보획득량, 카이제곱 통계량, 상호정보 방법에 비하여 정확률과 재현율의 결합 척도인 F-척도면에서 월등한 성능을 보여주지는 않는다. 하지만 제안된 퍼지추론 방법은 사용된 특징들의 수에 비례하여 성능이 증가하므로 좋은 특징 추출 방법으로 간주된다. 따라서 본 연구는 무수한 스팸 메일로 고통 받는 전자우편 사용자들을 위한 스팸 메일 필터링 시스템 개발에 도움을 줄 수 있다.

Key Words : 텍스트 분류, 특징 추출, 퍼지추론, 스팸 대응 시스템

1. 서론

인터넷의 급속한 보급에 따라서 적은 비용으로 메시지를 실시간으로 전달할 수 있는 편리성 때문에 오늘날 전자우편(email)은 사용자간 의사소통을 하는데 필수적인 통신수단이 되었다. 전자우편은 사용자에게 많은 편리성을 준 반면 매일 많은 양의 스팸 메일을 처리해야 하는 불편함도 주고 있다. 기본적으로 스팸 메일 필터링 문제는 문서분류의 특별한 한 형태로 볼 수 있기 때문에 여러 다양한 정보검색 기법과 기계학습 알고리즘들이 이 문제의 해결을 위해 사용되어져 왔다 [1-5]. 스팸 필터링은 스팸(spam)과 비스팸(non-spam) 일반 메일의 이진 문서 분류 문제에 해당한다. Sahami[1]는 나이브 베이지안 분류기(Naive Bayesian classifier)를 스팸 메일 필터링에 사용하였는데, 수작업으로 구축된 구(phrase) 정보와 발송자의 도메인 타입, 제목에서 기호 문자의 비율 등 다양한 비텍스트(non-textual) 정보를 도메인 속성으로 정의하여 사용하였다. 스팸 메일 필터링이나 메일의 자동분류에 관한 최근의 연구들을 대체적으로 살펴보면 TFIDF나 나이브 베이지안, 의사결정 트리와 같은 기존의 분류 알고리즘보다 Vapnik[2]가 고안한 SVM (Support Vector Machines)이 보다 나은 성능을 보이고 있음을 알 수 있다[3-4]. Yang[5]에서는 텍스트 정보와 발송자 이름, 발송자 소속 등과 같은 메타 데이터를 이용하여 스팸 메일을 구분하고자 하였는데, TFIDF보다 나이브 베이지안과 SVM이 훨씬 좋은 결과를 보임을 실험을 통해 입증하였다. 특히 메일의 헤더에서 추출한 속성을 SVM에 적용하였을 때 가장 좋은 결과를 보였다. 이는 SVM이 스팸 메일 필터링과 같은 이진 분류 문제(two-class problem)에 적합하기 때문이라고 볼 수 있다.

전자우편 문서를 잘 대표하는 특징 또는 용어를 선택하고 이들의 가중치를 부여하는 문제는 기존의 대표적인 선형 분류기인 Rocchio와 Widrow-Hoff 알고리즘들[6]이 학습 문서 집합을 대표하는 중심 벡터를 구성하는 것과 성격이 동일하다. 이들 알고리즘들은 용어의 가중치 산정시 발생 빈도수(TF)와 역문헌 빈도수(IDF)를 결합하는 방법을 취하고 있지만, 문서내 또는 문서 집합내 용어들 사이의 관련성을 용어의 가중치 계산에 반영하고 있지는 않다. 따라서 TF가 높은 용어는 높은 가중치를 가지게 되는데 대표 용어로서 실제 중요하지 않는 용어임에도 문서 내에 자주 발생만 되면 높은 가중치 값을 부여받을 수 있다는 단점을 지니고 있다. 이러한 문제를 해결하기 위해, 특정 용어의 중요도 계산에 사용되는 입력 정보(예: TF, DF, IDF)들은 정량적으로 정확히 해석될 수 없는 부정확하고 불확실한 특성을 내포하고 있으므로, 이러한 불확실성의 문제 해결에 효과적인 퍼지추론을 적용하여 후보 용어들의 가

중치를 계산하고 이 값들에 따라 선택 우선순위를 부여하는 방법도 있다[7].

본 연구에서는 성인 메일을 필터링하는데 정보획득량(information gain), 카이제곱 통계량(chi-squared test), 상호정보 (mutual information) 같은 기존의 대표적인 특징 추출 방법들에 비하여 약간 나은 성능을 보이는 퍼지추론에 의한 특징 추출 방법을 제안한다. 이를 보이기 위한 실험을 수행할 목적으로, 본 논문에서는 텍스트 정보(textual information)와 하이퍼링크(hyperlinks)에 기반한 2단계 스팸 메일 필터링에 관한 기존의 연구결과도 이용한다[8]. 제안된 시스템은 2가지 기본적인 아이디어를 기반으로 한다. 첫째, 높은 분류 판별력을 갖는 특징을 선택하기 위하여, 문서 분류에서 효과적인 방법[9]으로 알려진 정보획득량, 카이제곱 통계량, 상호정보 방법들과 퍼지추론 방법을 비교한다. 둘째, 스팸 메일은 2단계 필터링 방법으로 구분한다. 1단계에서는, 메일 발송자의 전자우편 주소, URL, 스팸 키워드 리스트가 적용된다. 다음 단계에서는 1단계에서 분류되지 않고 남은 전자우편들을 대상으로 전자우편 헤더와 본문뿐만 아니라 하이퍼링크를 참조하여 페치한 웹 페이지로부터 추출된 덜 중요한 정보를 사용하여 분류한다. 이러한 2단계 필터링은 필터를 1회 사용하는 1단계 시스템 보다 정확한 분류를 수행한다[8].

2. 특징 추출

최근 스팸 메일의 본문은 적은 텍스트 정보만 포함하므로 스팸 메일과 일반 메일을 구분하기 위한 충분한 힌트를 제공하지 않는다. 이 문제를 해결하려고 전자우편 본문에 포함된 하이퍼링크를 활용하여 전자우편 본문과 함께 페치된 웹 페이지로부터 추출된 모든 가능한 힌트를 활용하였다. 이들 힌트가 SVM 분류기를 구축하는데 사용된다. 우리는 스팸을 구분하기 위한 힌트를 명확한 정보와 덜 명확한 텍스트 정보의 2가지 종류로 구분하였다. 스팸 메일을 위한 명확한 정보에는 전자우편 주소와 URL 링크와 같은 발송자 정보와 "포르노", "광고" 같은 명확한 스팸 키워드 리스트가 있다. 전자우편이 스팸인지 아닌지를 구분하는 증거를 제공하는 많은 특별한 특징들이 있다. 예를 들어, 전자우편 본문의 개별 단어들, 발송자의 도메인 형태, 전자우편의 수신 시간, 메일의 제목안에 있는 비알파벳 문자 비율 등은 스팸 메일 필터링에서 덜 명확한 텍스트 정보를 나타낸다. 2단계 필터링 방법에서는, 명확한 정보를 사용하여 스팸 메일을 1차로 분류하고, 덜 명확한 정보를 가지고 남아있는 메일들을 2차로 분류한다.

덜 명확한 텍스트 정보로부터의 특징 추출은 예측에 가장 적합한 특징들을 발견하기 위하여 후보 특징 집합에 있는 모든 가능한 조합을 탐색하여 결정된다. 최적의 특징들을 발견하기 위해 고안된 방법들로는 문서 빈도수 임계(document frequency threshold), 정보획득량, 상호정보, 용어 강도(term strength), 카이제곱 통계량 등이 있다. 정보획득량은 기계학습 분야에서 자주 사용되는 기법으로 문서에서의 출현 빈도뿐만 아니라 출현하지 않은 빈도까지 고려해서 각 범주에서의 용어 정보량을 계산하는 방법이다. 정보획득량은 상호정보와는 달리 용어의 출현빈도를 고려한 상호정보의 평균값과 용어가 출현하지 않은 빈도의 상호정보 평균값의 합으로 계산된다. 카이제곱 통계량은 용어 t와 범주 c의 의존성을 측정하는 것으로 자유도 1인 카이제곱 분포와 비교될 수 있다. 카이제곱 통계량은 용어 t와 범주 c가 완전히 독립적이면 0의 값을 가진다. 상호정보는 연관 단어의 통계적 언어 모델링에서 일반적으로 사용되는 기준이다. Yang은 문서 분류에서 속성 선택 알고리즘들의 성능을 비교 평가하였는데, 상호정보를 제외한 나머지 속성선택 알고리즘들이 모두 비슷한 성능과 특징을 나타냄을 보였다[9].

퍼지추론에서는 각 용어의 TF(Term Frequency), DF(Document Frequency), IDF(Inverse Document Frequency)가 전처리된 전자우편으로부터 계산되고 정규화된다. 정규화된 NTF(Normalized TF), NDF, NIDF는 일반적인 삼각형 멤버쉽함수를 사용하여 퍼지화 한다. 입력 퍼지변수 NTF는 S(Small)과 L(Large)의 퍼지항을 갖고, NDF와 NIDF는 S(Small), M(Middle), L(Large)의 퍼지항을 갖는다. 출력 퍼지변수 TW(Term Weight)는 각 용어의 중요도를 6가지 언어 라벨 {Z(Zero), S(Small), M(Middle), L(Large), X(X Large), XX(XX Large)}로 표현한다. 표 1은 NTF 퍼지 입력값의 소속 정도에 따라 두 부분으로 나누어 규칙들을 표현하고 있다. NTF, NDF, NIDF 퍼지 입력값을 위의 결과로 생성된 18개의 추론 규칙별로 이들의 전건부의 소속 함수에 적용시킨다. 각각의 소속 정도가 구해지면 이들 중에서 최소값을 취한다. 그 결과 규칙별로 하나씩의 퍼지 값이 생성되며 이 퍼지 값들

을 퍼지 출력변수 TW에 따라 6개의 그룹으로 분류하고 그룹별로 해당 그룹에 속한 퍼지 값들 중 최대값을 취하여 총 6개의 퍼지 값들을 생성한다. 최종적으로 이들 6개의 퍼지 값들을 무게중심법으로 비퍼지화한 값이 해당 용어의 중요도 값으로 결정된다[10].

표 1. NTF 값에 따라 2그룹으로 구성된 퍼지 추론 규칙

	NIDF / NDF	S	M	L
NTF = S	S	Z	Z	S
	M	Z	M	L
	L	S	L	X
NTF = L	NIDF / NDF	S	M	L
	S	Z	S	M
	M	S	L	X
	L	M	X	XX

3. 실험

실험에 사용된 스팸 메일 말뭉치는 성인 스팸 메일 1,100개, 금융 스팸 메일 1,077개, 쇼핑 스팸 메일 397개로 총 2,574개이며, 비스팸 일반 메일은 2,218개로, 전체 말뭉치는 총 4,792개이다. 스팸 메일 필터링 시스템의 성능평가를 위해서 정보검색에서 일반적으로 많이 사용하고 있는 재현율과 정확률, F-척도를 사용하였다. 본 연구에서는 중요한 특징을 선택하기 위하여 Witten[11]이 개발한 WEKA (Waikato Environment for Knowledge Analysis) 패키지를 사용하였다. WEKA는 실세계 데이터 집합에 기계학습 기법을 적용하기 위해 개발된 도구이다. 본 실험에서 사용된 SVM 분류기도 WEKA에서 제공되며, SVM은 WEKA에서 제공된 기본 패러미터값을 그대로 사용하여 테스트하였다. 전자우편 문서 집합의 필터링 성능을 평가하기 위하여, 정보검색 분야에서 일반적으로 많이 사용되는 재현율(R: recall), 정확률(P: precision), 그리고 F-척도(F-measure)를 사용하였다. F-척도는 다음 수식과 같이 정의된다.

$$F = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall} = \frac{2PR}{P + R}$$

여기서 ß는 정확률에 대한 재현율의 상대적 비율로서, 두 척도의 가중치를 동일하게 주기 위해 실험에서는 1.0을 사용한다.

데이터 집합을 무작위로 선정하는데 따른 편차를 줄이는 객관적인 성능평가를 위하여 10층 교차 확인법(10-fold cross validation)을 사용하였다[11]. 이는 전체 전자우편 말뭉치를 균등하게 10등분한 다음, 9등분은 학습에 사용하고 나머지 1등분은 성능 테스트를 위해 사용하는 방법으로, 각 등분들이 한 번씩 테스트 용도로 사용되도록 10번 반복 실험을 한 후, 그 결과들을 평균 내는 방법이다. 표 2는 퍼지추론 및 기존의 정보획득량, 카이제곱 통계량, 상호정보를 성인 메일 대상으로 특징 추출하고 SVM 분류기에 적용한 실험 결과를 보여준다. 전처리 단계에서 불용어와 불필요한 단어들을 제거함으로써 약 7,600개의 형태소(morpheme)가 추출되었고, 이들 형태소들이 성인 메일 훈련에 사용될 후보 특징들이 된다. 본 연구에서는 7,600개의 후보 특징들 가운데, 정보획득량, 카이제곱 통계량, 상호정보 특징 추출 방법들은 WEKA를 사용하여, 퍼지추론 방법은 제안된 시스템에서 직접 200, 485, 838개의 특징들을 선택하였다. Yang[9]의 실험 결과와 마찬가지로 기존의 특징 추출 방법들의 성능에는 별 차이가 없음을 확인할 수 있었다.

표 2. 속성선택 방법에 따른 실험결과 (성인 스팸 메일 SVM 분류기 대상)

특징 선택	특징 수	R	P	F
정보획득량	200	63.7%	84.5%	72.6%
	485	75.5%	91.7%	82.9%
	838	77.0%	91.1%	83.4%
카이제곱 통계량	200	49.3%	92.6%	64.3%
	485	72.8%	92.2%	81.4%
	838	76.5%	91.7%	83.4%
상호정보	200	65.5%	86.4%	74.5%
	485	73.7%	88.8%	80.6%
	838	76.3%	89.8%	82.5%
퍼지추론	200	68.7%	87.3%	76.9%
	485	76.5%	89.1%	82.3%
	838	79.3%	91.3%	84.9%

표 2에서 볼 수 있듯이, 퍼지추론 방법은 정보획득량, 카이제곱 통계량, 상호정보 방법에 비하여 평균적으로 2.3%, 6.5%, 2.8%의 F-척도를 향상시켰다. 비록 카이제곱 통계량이 다른 세 가지 방법들보다 약간 높은 정확률을 제시하지만, 재현율과 F-척도 면에서는 최악의 결과를 제공한다. 또한 우리가 퍼지추론 방법의 재현율 성능을 기존 방법들과 비교할 때 정

확률이나 F-척도의 성능에 비하여 약 2배 정도 향상시킴도 알 수 있다. 따라서 제안된 퍼지추론 방법은 사용된 특징의 수에 비례하여 점차적으로 성능이 향상되고, 선택된 특징의 수와 상관없이 안정된 특징 추출법이라고 여겨진다.

4. 결론

본 연구에서는 스팸 대응 시스템의 특징 추출 방법으로서 퍼지추론을 제안하였다. 우리는 성인 스팸 메일 분류 문제를 대상으로 특징 추출에 관한 비교 실험을 수행하였다. 비록 퍼지추론에 의한 괄목할 만한 성능 개선을 보여주진 않지만, 기존의 정보획득량, 카이제곱 통계량, 상호정보 기법에 비하여 재현율과 F-척도 면에서 나은 성능을 제시하였다. 정확률의 경우에는 카이제곱 통계량이 가장 좋은 성능을 보였지만, 이 방법은 사용된 특징수에 반비례하여 정확률이 감소하는 경향이 있었다. 또한, 카이제곱 통계량의 다른 평가척도들은 다른 방법들에 비하여 낮은 성능을 제시하였다. 따라서 제안된 방법은 신뢰할만한 특징 추출 방법으로 판단된다. 향후에는 높은 분별력을 갖는 특징을 보다 많이 찾아야 하고, 이를 통해서 스팸 대응 시스템의 필터링 성능을 향상시킬 필요가 있다.

References

[1] Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E., "A bayesian approach to filtering junk e-mail," In AAAI-98 Workshop on Learning for Text Categorization (1998) 55-62.

[2] Vapnik, V., The Nature of Statistical Learning Theory, Springer-Verlag, New York (1995)

[3] Drucker, H., Wu, D. and Vapnik, V., "Support Vector Machines for Spam Categorization," IEEE Trans. on Neural Networks, Vol.10(5) (1999) 1048-1054

[4] Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," ECML, Claire Nedellec and Celine Rouveirol (ed.) (1998)

[5] Yang, J., Chalasani, V., and Park, S., "Intelligent email categorization based on textual information and metadata," IEICE Transactions on Information and System, Vol.E86-D, No.7 (2003) 1280-1288.

[6] Lewis, D. D., Schapire, R. E., Callan, J. P., and Papka, R., "Training algorithms for linear text classifier," Proc. of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval (1996) 298-306.

[7] Kim, B. M., Li, Q., and Kim, J. W., "Extraction of User Preferences from a Few Positive Documents," Proceedings of The Sixth International Workshop on Information Retrieval with Asian Languages (2003) 124-131.

[8] 강신재, 김종완, "텍스트정보와 하이퍼링크에 기반한 지능형 스팸 메일 필터링," 한국퍼지및지능시스템학회 논문지, Vol.14, No.7, pp.895-901, 2004.

[9] Yang, Y, and Pedersen, J. P., "A comparative study on feature selection in text categorization," in Fourteenth International Conference on Machine Learning (1997) 412-420.

[10] Kim, J. W., Kim, H. J., Kang, S. J., and Kim, B. M., "Determination of Usenet News Groups by Fuzzy Inference and Kohonen Network," Lecture Notes in Artificial Intelligence, Vol.3157, Springer-Verlag (2004) 654-663.

[11] Witten, I. H. and Frank, E., Data Mining: Practical machine learning tools and Techniques, 2nd Ed., Morgan Kaufmann (2005).