실세계의 FAQ 메일 자동분류를 위한 문서 특징추출 방법의 성능 비교*

홍진혁, 류중원, 조성배
연 세 대 학 교 컴 퓨 터 과 학 과
{hjinh, rjungwon}@candy.yonsei.ac.kr, sbcho@csai.yonsei.ac.kr

Comparison of Document Feature Extraction Methods for Automatic Classification of Real World FAQ Mails

Jinhyuk Hong° Jungwon Ryu Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

최근 문서 자동분류의 중요성이 널리 인식되어 다양한 연구가 진행되고 있다. 본 논문에서는 한글 문서의 효과적인 자동분류를 위한 다양한 특징추출 방법들을 구현하고 실제 질의메일에 대한 효율적인 특징추출 방법을 제시한다. 실험을 위해 문서 빈도(document frequency), 정보 이득(information gain), 상호 정보량(mutual information), x^2 통계, 적합성 점수(relevancy score), 교차비(odds ratio)과 단순화된 x^2등 7가지 특징추출 방법을 사용하였으며 463개의 실제 테스트 질의메일에 적용한 결과, x^2 방법이 74.7%의 인식률을 내어 성능이 가장 좋음을 알 수 있었다. 반면에 x^2와 함께 가장 자주 쓰이는 방법 중의 하나인 정보 이득은 인식률이 최대 40.6%밖에 되지 않았다.

1. 서론

컴퓨터와 PC통신의 보급과 인터넷의 확산으로 인해 많은 사람들이 통신 기반 서비스를 이용하게 됨에 따라, 서비스 제공 업체에는 사용자들로부터 많은 양의 질의 메일들이 오고 있다. 이러한 질의 메일들에 대하여 사람이 손수 분류하고 대답하기에는 한계가 있으며, 많은 인력의 낭비를 가져온다. 따라서 문서 자동분류의 중요성이 널리 인식되고 있으며, 최근 다양한 문서 자동분류 기법이 이용되어 그 성능을 입증받고 있다[1, 2, 3].

본 논문에서는 인공신경망 알고리즘을 이용한 효율적인 문서 자동분류 시스템을 위하여 다양한 문서 특징추출 기법의 성능을 비교분석하고, 가장 적절한 방법을 제시한다[2, 3].

2. 문서 분류시스템

문서 자동분류는 각종 분류기법을 이용하여 자동으로 새로운 문서를 미리 정의된 부류들로 나누는 것을 말한다. 그 과정은 그림 1에서와 같이 크게 전처리 단계, 특징 추출 단계, 문서 분류 단계로 구분된다. 사용자 질의 메일이 들어오면 어근화, 불용어제거 등의 전처리 단계를 거치게 된다. 이 때 메일에서 사용되는 주요 단어들로 구성된 색인어 사전을 통해 사전에 존재하는 단어들을 뽑아내는데, 본 논문에서는 미리 만들어 놓은 449개의 색인어를 가진

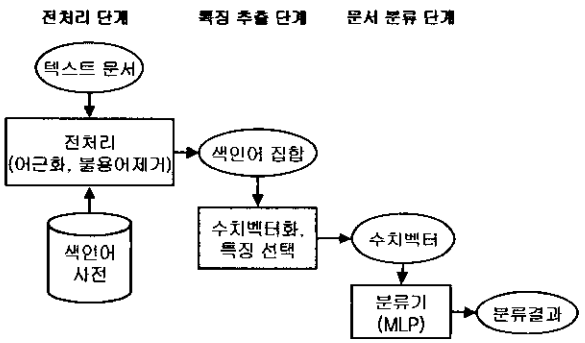

그림 1. 시스템 구성

사전을 사용하였다[3].

특징추출 단계에서는 이렇게 만들어진 키워드 집합을 수치화된 벡터로 표현한다. 이는 추출된 색인어가 문서에 나타나는 빈도수와 색인어가 나타난 전체 문서의 개수를 기반으로 계산되는 tf- idf값에 의하여 구해진다. tf- idf는 키워드 집합을 수치화된 벡터로 표현하는 모듈로 대부분의 정보검색 시스템이나 문서 분류시스템에서 사용되는 방법이다. 문서 d_i의 j번째 키워드 w_{ij}의 가중치는 다음과 같이 표현된다[1, 3].

$$w_{ij} = tf_{ij} \log\left(\frac{N}{df_j}\right)$$

* 이 논문은 (주)다음 소프트의 일부 지원에 의한 것임.

여기서 tf_{ij}는 j번째 키워드의 문서 i에서의 가중치이며, df_j는 키워드 j가 전체 문서집합에서 나타나는 문서의 개수이다. N은 전체 문서의 개수를 말한다.

이렇게 얻어진 문서의 수치벡터는 색인어 사전에 존재하는 총 색인어의 개수인 449차원을 갖는다. 한편 벡터의 차원이 크면 학습의 속도가 늦어지고, 문서분류 성능 또한 나빠지게 되는데, 이러한 문제를 극복하기 위해 수치벡터의 차원을 적당한 크기로 축약하는 과정이 필요하다. 본 논문에서는 문서 빈도, 정보 이득, 상호 정보량, x^2통계, 적합성 점수, 교차비와 단순화된 x^2 등의 특징추출 방법을 사용하여 449개의 차원을 150개의 차원으로 축약하였다. 이렇게 축약된 수치벡터를 다중신경망을 이용한 분류기를 통하여 분류한다[3].

3. 문서 특징추출

3.1 문서 빈도

문서 빈도(*DF*, Document Frequency)란 전체 문서 집합 중에서 특정 색인어가 포함된 문서의 개수를 말한다. 이 방법은 발생이 드문 단어는 카테고리 분류에 별로 도움이 되지 못한다는 것을 전제한다. 문서 빈도를 이용한 방법은 특징벡터 축약의 가장 기본적인 방법이긴 하지만 단순한 색인어의 출현 빈도만으로 문서의 부류를 결정할 수 없는 경우가 많다. 학습 문서 중에서 각각의 단어에 대해 문서 빈도를 계산하고, 특정 임계치보다 낮은 문서 빈도를 가지는 단어들을 특징 공간에서 제거함으로써 특징 공간의 차원을 줄인다[1, 4].

본 논문에서는 각 부류별로 단어에 대하여 문서 빈도를 측정한 다음, 부류 중 가장 큰 값을 그 단어의 문서 빈도로 사용하였다.

3.2 정보 이득

정보 이득(*IG*, Information Gain)은 부류를 결정할 때, 문서에서 어떤 단어의 유무를 통해 단어와 부류간의 관계를 알아낸다. 전체 문서에서 부류들의 집합을 $\{c_i\}_{i=1}^{m}$이라고 할 때, 단어 t의 정보 이득은 일반적으로 다음과 같이 정의된다.

$$IG(t) = P(t_k|c_i)\log\frac{P(t_k|c_i)}{P(c_i)\cdot P(t_k)} + P(\overline{t_k}|c_i)\log\frac{P(\overline{t_k}|c_i)}{P(c_i)\cdot P(\overline{t_k})}$$

문서분류 문제에서 부류가 보통 여러 개로 되어 있기 때문에 이 정의를 사용한다. 본 논문에서는 학습 데이터에 대해 각 단어의 정보 이득을 계산하고, 예상하는 임계값보다 낮은 단어는 특징 공간에서 제거하였다[1, 4].

3.3 상호 정보량

상호 정보량(*MI*, Mutual Information)는 단어 연관성의 통계적 언어 모델링에서 보통 사용되는 방법이다. A는 부류 c에 속한 문서 중 키워드 t를 가진 문서의 개수, B는 부류 c에 속하지 않은 문서들 중 키워드 t가 발생하는 횟수, C는 부류 c의 문서들 중에서 키워드 t가 없는 문서의 개수, 그리고 N은 전체 문서수 라고 했을 때, 상호정보량은 다음과 같이 정의된다[1, 4].

$$MI(t,c) = \log\frac{P(t\wedge c)}{P(t)\times P(c)} = \log\frac{A\times N}{(A+C)\times(A+B)}$$

3.4 x^2 통계

x^2 통계는 키워드 t와 부류 c사이의 관계 정도를 구하여 문서분류에 있어서 키워드 t의 중요도를 구한다. A는 부류 c에 속한 문서 중 키워드 t를 가진 문서의 개수, B는 부류 c에 속하지 않은 문서들 중 키워드 t가 발생하는 횟수, C는 부류 c의 문서들 중에서 키워드 t가 없는 문서의 개수, 그리고 D는 부류 c에 속하지 않는 문서들에서 키워드 t가 발생하지 않는 횟수이다. N이 전체 문서수 라고 했을 때, 부류 c와 키워드 t간의 x^2값은 다음과 같이 정의된다.

$$x^2(t|c) = \frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

그리고 다음과 같이 학습 데이터에서 각 단어와 각 부류에 대해 통계를 계산한다. 각 단어마다 계산된 x^2값들 중 최대값을 가지는 것이 그 단어의 최종 x^2값으로 결정된다.

$$x^2_{max}(t) = \max_{i=1}^{m}\{x^2(t|c_i)\}$$

x^2의 계산은 이차식이며, 상호 정보량나 정보 이득과 비슷하다. x^2와 상호 정보사이의 주된 차이점은 x^2가 표준화된 값이라는 것이고, 따라서 x^2값은 같은 부류를 갖는 단어들간에 계산이 가능하다. x^2통계값은 특징추출에 있어서 뛰어난 성능을 보인다. 하지만 x^2 통계값은 저빈도 단어들에 대해서는 신뢰할 수 없다고 알려져있다[1, 4].

3.5 기타 방법

기타 다른 방법으로는 적합성 점수(*RS*, Relevancy Score), 교차비(*OR*, Odds Ratio)과 단순화된 x^2(sx^2, simplified x^2)의 세 가지를 이용하였다[1].

$$RS(t_k|c_i) = \log\frac{P(t_k|c_i)+d}{P(\overline{t_k}|\overline{c_i})+d}$$

$$OR(t_k|c_i) = \frac{P(t_k|c_i)\cdot(1-P(t_k|\overline{c_i}))}{(1-P(t_k|c_i))\cdot P(t_k|\overline{c_i})}$$

$$sx^2(t_k|c_i) = P(t_k|c_i)\cdot P(\overline{t_k}|\overline{c_i})\cdot P(t_k|\overline{c_i})\cdot P(\overline{t_k}|c_i)$$

위의 수식에서 $P(t_k|c_i)$는 임의의 문서가 부류 c_i에 속하고 단어 t_k를 가질 확률을 말한다. 또한 RS에서 d값은 감폭변수(damping factor)이다.

4. 실험 결과

4.1 실험환경

본 논문에서 소개한 특징추출 방법에 따른 문서분류기의 성능을 확인하기 위하여 약 한달간 수집된 한메일넷 사용자 질의 2,204개를 대상으로 실험하였다. 질의 집합 중 임의로 선택한 1,718개의 문서들을 학습데이터로, 463개의 문서를 성능 평가를 위한 실험 데이터로 사용하였다. 표 1에서처럼 질의 메일은 20개의 응답되어져야 할 부류와 질의의 특성상 직접 분류가 필요없이 운영자에게 포워딩해야할 부류로 총 21개의 부류가 존재한다.

표 1. 응답 종류에 따른 분류

	부류 개수	질의 개수
응답되어야할 질의	20	1,475(66.9%)
운영자에게 포워딩해야할 질의	46	729(33.1%)

본 논문에서는 총 21개의 부류를 가진 질의 메일 분류를 위한 인식기로 다중 신경망을 사용하였다. 신경망은 패턴인식 문제에서 널리 쓰이고 있는 분류기로 일반화 능력이 높으며 안정적인 학습 알고리즘을 가지고 있다[3].

4.2 결과분석

성능은 패턴인식 지표로 사용되는 인식률로 평가하였다. 여기에서 분류성공 질의 메일개수는 질의 메일이 응답되어져야 할 부류로 제대로 분류된 것과 포워딩부류로 분류해야 할 질의 메일을 성공적으로 포워딩 부류로 분류한 것을 말한다.

$$\text{인식률} = \frac{\text{분류성공 문서개수}}{\text{전체 문서수}}$$

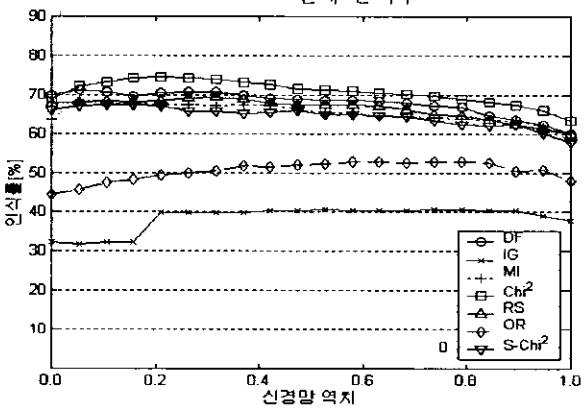

그림 2. 특징 선택 방법별 성능 평가

그림 2를 보면 각 특징추출 방법별 성능을 비교할 수 있다. 질의 메일의 부류를 판단할 때 그 신뢰정도가 신경망 역치 이상이 될 때에만 부류를 판별하고, 그 이하일 때에는 신경망에서 기각하도록 한다. 따라서 신경망 역치 값이 커짐에 따라서 인식률이 전반적으로 떨어지는 모습을 보이고 있다. 이것은 신경망에서 기각되는 질의의 수가 늘어나는 것을 의미한다.

여기에서 인식률이 가장 큰 것은 x^2이고, 문서 빈도, 교차비, 단순화된 x^2 등이 그 다음으로 좋은 성능을 보였다. 문서 빈도는 보통 특징추출에 부수적으로 사용되지만 본 실험에서는 부류 중 최대 문서 빈도값을 사용함으로써 좋은 성능을 얻을 수 있었다. 이는 각 부류에서 많이 나온 단어들이 분류에 있어서 유용하다는 것을 의미한다. 반면에 정보 이득 방법은 비교적 많은 곳에서 사용되고 있기는 하지만 한메일넷 질의 분류에 대해서는 좋지 않은 성능을 보였다. 정보 이득이 사실상 더 이상의 정보가 가치가 없어도 특성값이 크다면 값이 작은 것보다 많은 정보를 뽑아 사용한다는 특성에 기인한다. 표 2를 살펴보면 같은 임계치에서 x^2 통계가 다른 방법에 비해 인식률이 높음을 알 수 있다.

표 2. 실험 결과

특징추출방법	분석지표	임계치(0.1)	임계치(0.5)	임계치(0.9)
문서 빈도	인식률	0.706	0.686	0.622
	기각률	0.129	0.209	0.332
정보 이득	인식률	0.321	0.406	0.390
	기각률	0.069	0.578	0.604
상호 정보량	인식률	0.676	0.654	0.617
	기각률	0.118	0.233	0.319
x^2 통계	인식률	0.732	0.712	0.658
	기각률	0.090	0.190	0.295
적합성 점수	인식률	0.475	0.522	0.507
	기각률	0.116	0.285	0.421
교차비	인식률	0.682	0.671	0.611
	기각률	0.138	0.231	0.345
단순화된 x^2	인식률	0.671	0.650	0.602
	기각률	0.177	0.263	0.360

5. 결론

본 논문에서는 7가지 특징추출 방법을 이용하여 한메일넷의 사용자 질의를 분류하는 시스템의 성능을 확인해 보았다. 실험 결과는 x^2 통계와 문서 빈도가 한메일넷의 사용자 질의 분류에 대해서 다른 방법들보다 나은 성능을 나타내었다. 앞으로는 특징추출 방법에 따라서 분류시 가중치를 주거나 특징추출 방법을 결합함으로써 문서분류기의 성능을 향상시킬 수 있도록 하는 연구가 필요하며, TREC이나 REUTER같은 문서집합을 사용하여 특징추출 방법간의 성능을 객관적으로 검증할 필요가 있다.

참고문헌

[1] F. Sebastiani, "Machine learning in automated text categorisation: a survey," *Technical Report IEI-B4-31-1999*, Istituto di Elaborazione dell'Informazione, C. N. R., Pisa, 1999.

[2] S. M. Weiss, C. Apte, F. J. Damerau, D. E. Johnson, F. J. Oles, T. Goetz, and T. Hampp, "Maximizing text-mining performance," *IEEE Intelligent Systems*, p.63~69, 1999.

[3] 이지행, 조성배, "다중 신경망을 이용한 한메일넷 질의 자동분류 시스템," 제27회 춘계학술발표회 논문집, p.232~234, 한국정보과학회, 2000.

[4] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," *Proc. of 14th Int. Conf. on Machine Learning*, p.412~420, 1997.