용어 가중치와 역범주 빈도에 의한 자동 문서 범주화*

국민대학교 컴퓨터학부, 첨단정보기술연구센터
이 경 찬† · 강 승 식

Automatic Text Categorization by Term Weighting and Inverted Category Frequency

Kyung-Chan Lee, Seung-Shik Kang
School of Computer Science, Kookmin University & AITrc, Seoul, Korea

요 약

문서의 확률을 이용하여 자동으로 문서를 분류하는 문서 범주화 기법의 대표적인 방법이 나이브 베이지언 확률 모델이다. 이 방법의 기본 형식은 출현 용어의 확률 계산 방법이다. 하지만 실제 문서 범주화 과정에서 출현하지 않는 용어들도 성능에 많은 영향을 줄 수 있으며, 출현 용어들에 대한 빈도 이외의 역범주 빈도나 용어 가중치를 적용하여 문서 범주화 시스템의 성능을 향상시킬 수 있다. 본 논문에서는 나이브 베이지언 확률 모델에 출현 용어와 출현하지 않는 용어들에 대한 smoothing 기법을 적용하여 실험하였다. 성능 평가를 위해 뉴스그룹 문서들을 이용하였으며, 역범주 빈도와 가중치를 적용했을 때 나이브 베이지언 확률 모델에 비해 약 7% 정도 성능 개선 효과가 있었다.

서 론(1절)

문서 범주화(text categorization)는 미리 정의된 범주에 문서를 할당하는 기법이다. 다량의 문서를 효과적으로 관리, 검색하는데 그 목적이 있다. 범주화 과정에는 학습 과정과 색인된 용어들에 대한 범주 할당 과정이 포함되는데 학습 과정은 범주를 표현할 수 있는 자질들을 선별하는 작업으로 표현될 수 있으며, 이는 벡터 공간으로 표현하고 선별된 자질들은 빈도로 표현된다. 범주 할당 과정은 단순 용어 출현 여부에 의한 방법부터 벡터 모델, 검색 모델, 확률 모델, 기계 학습 등 다양한 범주화 방법 등이 제시되었다.[1,3-5,7] 이 중 확률 모델은 초기에 나이브 베이지언 분류자(Naive Bayesian classifier)를 이용하여 문서가 범주에 포함될 확률을 계산하는 방법으로 많이 연구가 되었으며, 그 이후 계산 방식의 변형을 도입한 Bayesian networks 방식으로 연구 방법이 세분화되었다. 확률 모델의 기본 형식은 문서에 속한 용어가 범주에 속할 확률과 그렇지 않을 확률을 계산하여 가장 알맞은 범주를 찾아내는 방법으로 설명할 수 있다. 보통 문서가 범주에 속할 확률값의 표현은 문서에 출현한 용어의 빈도로 나타내는 것을 기본 원칙으로 하고 있으며, 출현하지 않을 확률은 일반적으로 범주의 낮은 값으로 표현하고 있다. 하지만 범주에 속하는 것과 속하지 않는 표현 값들에 대해서 실제로 범주화되는 것에 어느 정도 영향을 미치는지를 알아 볼 필요가 있으며, 아울러 기존의 빈도 표현 이외의 가중치를 부여하여 성능에 영향을 미치는지에 대한 부분도 연구하여 볼 필요가 있다. 본 논문에서는 이러한 부분에 대해서 기존 나이브 베이지언 모델과 비교하여 어느 정도 성능에 영향을 주는지 확인하는 실험을 한다.

본 논문의 내용은 다음과 같다. 2절에서는 기존 나이브 베이지언 모델에 대한 연구에 대해서 소개하고, 3절에서는 본 연구에서 제안한 방법들에 대한 설명을 하고, 4절에 실험 및 평가 결과를 기술한다.

관련 연구(2절)

문서 범주화 모델에는 벡터의 내적값을 이용한 모델, k-

*본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.
†E-mail : ayin@cs.kookmin.ac.kr
E-mail : sskang@cs.kookmin.ac.kr

최근린법, 나이브 베이지언 모델, 결정 트리, SVM 등 다양한 방법들이 제시되었다. 이 중 나이브 베이지언 모델은 입력 문서가 범주에 속할 확률을 구해서 가장 확률이 높은 범주로 할당하는 방법이며, 이 확률 모델에는 두 가지 방식이 존재하는데 multi-variate model과 multinomial model이다. Multi-variate model은 입력 문서가 범주에 속할 확률을 용어의 빈도가 아닌 용어의 발생 여부에 의한 확률 계산 방법이고 multinomial model은 용어에 대한 빈도를 고려하여 확률 값을 계산하는 방식이다.[6] 본 연구에서는 빈도를 고려한 multinomial model 방식을 이용하였다.

문서 d가 범주 c에 속할 확률을 계산하는 나이브 베이지언 모델은 다음과 같이 기술된다.

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \approx P(c) \times \prod_t P(t|c)^{tf(t,d)} \quad (1)$$

이 식에서 P(t|c)의 확률은 범주 학습 문서에서 출현한 용어의 빈도수로 계산한다. 결국, 문서 범주화는 입력 문서에 대한 모든 용어에 대해서 각 범주 c에 속할 용어 t의 확률값을 계산하는 문제로 나타나게 된다.

$$P(t|c) = \frac{tf(t,c)}{\sum_i tf(t_i,c)} \quad , \text{if } tf(t,c) > 0 \quad (2)$$

오효정(2000)은 웹 문서의 자동 분류를 위해 나이브 베이지언 확률 모델을 이용하였는데, 범주 확률 이외에 하이퍼 링크된 문헌의 값에 의해서 성능을 검증하였으며, 고영중(2002)은 확률 계산시 idf(inverted document frequency)와 카이제곱 통계량, 그리고 문장의 중요도를 이용하여 나이브 베이지언 모델의 성능을 개선하는 방법을 제안하였다.[10,12]

용어 가중치와 역범주 빈도에 의한 문서 범주화(3절)

1. Smoothing 기법

나이브 베이지언 모델에서는 다음과 같은 문제점이 존재하는데 그 중 하나는 범주에 대한 용어의 확률 값이 0이 나오게 되는 경우이다. 즉 입력문서에 용어가 범주에 존재하지 않게 되는 것이 이에 해당한다. 이 문제를 해결하기 위해 존재하지 않는 용어의 확률 값을 매우 낮은 값으로 계산 하여주는 방법이 제시되었다. 이러한 기법을 smoothing 이라고도 하는데, Ponte(1998)는 언어 모델(language model)을 이용한 정보검색 확률 계산에서 smoothing 값을 cf(t)/cs로 정의하였다. cf(t)는 범주에서 출현한 용어 t의 빈도 합이고, cs는 범주의 총 빈도 값을 나타낸다. 언어 모델은 용어의 출현 확률 계산에 있어서 범주에 출현한 확률 값만을 계산하는 일반적인 방법과는 달리 범주에 출현하지 않은 범주내의 나머지 용어들의 확률값 또한 계산하는 방식이다.[7] 또한, Lavrenco(2001)는 이 조정 값의 계산에 있어서 smoothing 파라미터 λ=0.6을 주어 출현하지 않는 용어에 대한 값을 조정하였다.[8] 아래 식에서 P(t|G)는 범주에 용어 t가 생성될 확률이다.

$$P(t|c) = \lambda P(t|d) + (1-\lambda)P(t|G) \quad (3)$$

입력 문서에 출현한 용어가 범주에 존재하지 않을 경우 일반적인 계산 방법에 대해서 기술하였는데, 공통적인 부분은 범주에서 출현 할 가장 작은 값으로 계산하게 되며 그 값은 범주의 크기마다 값의 변화가 생긴다. 본 연구에서는 범주에 따른 유동적인 값으로 계산하지 않고 범주에 출현한 용어의 빈도 합만을 고려하여 고정된 값으로 조정하였다. 계산식은 다음과 같다.

$$smoothing\ \alpha = \arg\min \frac{1}{cf(t)} \quad (4)$$

2. 용어 가중치 기법

용어 가중치는 국민대학교 형태소 분석기에 그 기능이 포함되어 있는데, 용어 빈도에 관련하여 자주 출현하는 용어들의 분포를 이용하고, 그 이외에 품사 유형 및 어절 위치 등을 고려하여 문서내의 용어 중요도를 계산한다.[10] 현재, 가중치의 범위는 0~1000 사이의 값으로 표현된다. 본 연구에서는 나이브 베이지언 확률 모델의 출현 빈도에 의한 계산법인 것을 고려하여 가중치의 값에 적절한 정규화 과정을 거쳐 실험을 하였으며, 식은 다음과 같다. 이 식에서 ω(t,d), ω(t,c)는 문서와 범주에서의 용어 가중치이다.

$$\omega_{t,d} = \omega(t,d) \times 0.1 \quad (5)$$

$$\omega_{t,c} = \omega(t,c) \times 0.1 \quad (6)$$

3. 범주화 방법

본 논문에서는 나이브 베이지언 모델에 기존 방법과 제안한 방법을 적용하여 성능을 비교해 본다. 실험 방법은 다음과 같다. 먼저 smoothing 기법으로 표현되는 범주내의 출현하지 않은 용어들에 대한 비교와 출현 용어에 대한 조정 값으로 알려진 ICF(Inverted Category Frequency)를 이용한 방법, 그리고 2.에서 제안한 용어 가중치 기법을 이용하여 용어의 표현에 있어서 기존 빈도만을 나타내던 방

법에 가중치를 추가 적용하였다.

우선 나이브 베이지언 모델에서는 용어의 출현 확률 P(t|c)의 값을 해당 범주 c의 모든 빈도에 용어 t 빈도의 확률 값으로 계산하며, 출현하지 않은 용어의 확률은 cf(t)/cs로 계산한다(TF-1).

$$(TF-1):\quad \begin{bmatrix} \prod_{a} P(t|c)^{tf(t,d)}, & \text{if } tf(t,c)>0 \\ cf(t)/cs & otherwise \end{bmatrix} \quad (7)$$

두 번째로 출현하지 않는 용어의 확률이 어느 정도 성능에 영향을 미치는지 확인하기 위하여 본 연구에서 제안한 방법을 이용하였다(TF-2). 이 방법은 범주 크기에 따른 유동적인 값보다 용어의 출현 범주 빈도를 이용한 고정된 값이 어느 정도 영향을 미치는지 알아보는 실험으로서 출현하지 않은 용어들에 대한 비중을 알아 볼 수 있는 실험이 될 수 있다.

$$(TF-2):\quad \begin{bmatrix} \prod_{a} P(t|c)^{tf(t,d)}, & \text{if } tf(t,c)>0 \\ & otherwise \end{bmatrix} \quad (8)$$

그리고 또 다른 고려 사항은 출현 용어의 확률값에 대한 것인데 나이브 베이지언 모델은 해당 범주에서의 빈도 확률만을 고려한다. 하지만 범주에 속하는 용어의 확률을 나타내려면 조금 더 변별력 있는 계산 값이 필요하다. 왜냐하면 동일한 용어도 여러 범주에 속할 수 있기 때문이다. 여기서 이에 대한 조정 값으로 용어 t에 대한 ICF값으로 계산하였다(TF-3).

$$(TF-3):\quad \begin{bmatrix} \prod_{a} P(t|c)^{tf(t,d)\times icf(t)}, & \text{if } tf(t,c)>0 \\ & otherwise \end{bmatrix} \quad (9)$$

그 이외에 용어 가중치를 이용하여 실험하였는데, 입력 문서와 학습 문서들에 대해서 각각 tf(ω,d), P(ω|c)로 가중치를 표현하였다.

$$tf(\omega,d) = tf(t,d)\times \omega_{t,d} \quad (10)$$

$$P(\omega|c) = \frac{tf(t,c)\times \omega_{t,c}}{\sum_{i} tf(t_i,c)\times \omega_{t_i,c}} \quad (11)$$

먼저 가중치 값을 학습 문서에는 기존 빈도만을 적용고 입력 문서에 대해서만 출현한 용어 빈도 표현에 가중치를 추가 적용하였다. 이 방법으로 입력 문서에서 출현한 용어에 대한 가중치의 적용만으로 어느 정도 범주 판별에 영향을 주는지의 여부를 확인할 수 있다(TW-1).

$$(TW-1):\quad \begin{bmatrix} \prod_{a} P(t|c)^{tf(\omega,d)\times icf(t)}, & \text{if } tf(t,c)>0 \\ & otherwise \end{bmatrix} \quad (13)$$

다음으로 학습되는 문서에 대해서도 벡터 표현 시에 용어 빈도에 가중치를 추가 적용함으로서 입력 문서와 학습 문서 모두에게 용어 가중치를 적용하는 방법으로 계산할 수 있다(TW-2).

$$(TW-2):\quad \begin{bmatrix} \prod_{a} P(\omega|c)^{tf(\omega,d)\times icf(t)}, & \text{if } tf(t,c)>0 \\ & otherwise \end{bmatrix} \quad (14)$$

실험 및 평가(4절)

1. 실험 데이터 및 성능 평가 방법

성능 평가를 위한 실험 집단은 뉴스그룹 문서를 이용하였다. 총 15개의 범주로 나누어져 있고 각 범주의 문서 수는 각각 다르다. 문서의 총수는 10,330 문서이며, 학습 집단은 7,224 문서, 평가 집단은 3,106 문서로 이루어져 있다.[12]

성능 평가를 위한 계산은 Table 1과 같이 나타낼 수 있다. 각 시스템과 실제 분류를 나누어 정확률(p)과 재현율(r)을 계산할 수 있다. 위의 표의 값으로 범주마다 정확률과 재현율은 각각 p=a/(a+b), r=a/(a+c)로 계산 할 수 있으며, 정확률과 재현율의 성능 평가 척도로 F1-measure 값을 사용하였다. 모든 범주의 통합 성능을 평가하기 위해 범주마다 문서 개수가 다른 점을 고려하여 micro-average 값을 사용하였다.

$$F_1(r,p) = \frac{2pr}{p+r}$$

2. 실험 결과

문서 범주화 실험은 학습 과정에서 추출된 총 94,573개의 용어들에 대해 카이제곱 통계량을 이용하여 용어를 선별하였다. 3절에서 설명한 5가지 방법에 대한 실험을 수행하였으며, 실험 결과는 Table 2와 같다. Table2에서 방법 TF-1과 TF-2는 smoothing값에 대해서 본 연구에서 제시한 고정값 α와 기존의 범주마다의 가장 낮은 값인지 대한 비교인데, 고정된 값의 성능이 평균 5%정도의 성능 향상을 보였다.

이는 우선 학습 문서들의 범주별 크기에 있어서 빈부의 차이가 심한 경우 범주별 값의 차이가 생기게 되는데 이를 고정값으로 제시하여 그 차이가 성능 향상에 도움이 되는

Table 1. 분류되는 경우의 수

실제 \ 시스템	옳음	옳지 않음
속함	a	b
속하지 않음	c	d

Table 2. 성능 평가

용어수 \ 방법	TF-1	TF-2	TF-3	TW-1	TW-2
5000	0.712	0.768	0.793	0.804	0.809
10000	0.714	0.780	0.812	0.820	0.823
20000	0.733	0.798	0.831	0.840	0.844
30000	0.746	0.816	0.844	0.847	0.853
40000	0.755	0.823	0.846	0.854	0.858
50000	0.767	0.833	0.857	0.859	0.866
60000	0.777	0.837	0.861	0.867	0.868
70000	0.791	0.848	0.863	0.862	0.869
80000	0.804	0.857	0.864	0.867	0.873

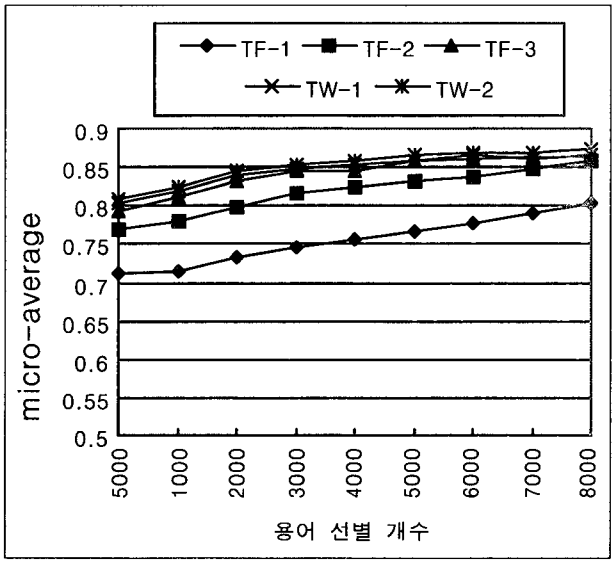

Fig. 1. 문서 범주화 실험 결과.

지 알아보는 실험이다. 이 결과로 알 수 있는 사실은 용어의 범주 출현 확률만큼이나 출현하지 않은 용어에 대한 비중이 그 만큼 크다는 것을 나타낸다고 볼 수 있으며, 더 나은 smoothing 기법에 대한 또 다른 연구를 시도해 볼 가치가 있음을 알 수 있다.

두 번째로 방법 TF-2와 TF-3는 단순 범주의 용어 확률보다는 출현한 범주의 개수를 고려한 방법이 좀 더 용어의 변별력을 높일 수 있다는 기존 연구를 확인하여 주는 것으로 생각할 수 있다. 마지막으로 빈도이외에 용어의 가중치 값에 의한 비교에서도 기존 빈도에 의한 확률보다 조금 나은 성능을 보였다. 결과적으로 나이브 베이지언 확률 모델 TF-1 보다 용어 가중치를 적용한 TW-2 방법이 약 7% 정도의 성능 개선 효과가 있었다.

결론 및 향후 과제(5절)

본 연구에서는 나이브 베이지언 방식에서 출현하는 용어 만큼이나 출현하지 않은 용어의 비중이 크다는 것을 smoothing 기법의 비교 실험을 통해 알 수 있었으며, ICF 값이 용어의 변별력을 높일 수 있다는 것을 확인 할 수 있었다. 그리고 빈도 이외의 용어 가중치 기법을 적용하여 좀 더 나은 성능 개선 효과가 있었다. 앞으로 용어 선별 기법 등, 다른 실험들에 대해서도 연구해 볼 수 있으며 다양한 문서들에 대한 실험도 필요할 것이다.

REFERENCES

1) Yang Y, Xin Liu(1999) : *"A Re-examination of Text Categorization Methods", Proc. of Conference on Research and Development in Information Retrieval (SIGIR 99), pp42-49*
2) Yang Y, Pedersen JP(1997) : *"A Comparative Study on Feature Selection in Text Categorization", In Jr. D H. Fisher (Ed.), the 14th International Conference on Machine Learning, pp412-420*
3) Sebastiani F(1999) : *"Machine Learning in Automated Text Categorization", Technical Report IEI-B4-31-1999, Istituto di Elaborazione dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT*
4) Vapnik V(1995) : *"The Nature of Statistical Learning Theory", Springer-Verlag*
5) Yang Y(1999) : *"An Evaluation of Statistical Approaches to Text Categorization", Journal of Information Retrieval, Vol 1, No. 1/2, pp67-88*
6) Andrew McCallum, Kamal Nigam(1998) : *"A Comparison of Event Models for Naive Bayes Text Classification", AAAI-98 Workshop on Learning for Text Categorization*
7) Ponte J(1998) : *"A Language Modeling Approach to Information Retrieval", Proceedings on the 21st annual international ACM SIGIR conference, pages 275-281*
8) Lavrenko V, Croft WB(2001) : *"Relevance-Based Language Models", Proceedings on ACM SIGIR01, pp120-127*
9) Dumais S, Platt J, Heckerman D, Sahami M(1998) : *"Inductive Learning Algorithms and Representations fot Text Categorization", Proceedings of the 7th International Conference on Information and Knowledge Management, pp148-155*
10) Hyo-Jung Oh(2000) : *"A Practical hypertext categorization method using links and incrementally available class", Proceedings of the 23rd annual international ACM SIGIR conference, pp264-271*
11) 강승식, 이하규, 손소현, 홍기채, 문병주(2001) : "조사 유형 및 복합명사 인식에 의한 용어 가중치 부여 기법", 한국 정보과학회 가을 학술발표논문집, *Vol.28, No.2, pp196- 198*
12) 고영중, 박진우, 서정연(2002) : "문장 중요도를 이용한 자 도 문서 범주화", 봄 정보과학회논문집, *Vol.29, No.6, pp417-423*