문서의 주제어별 가중치와 말뭉치를 이용한 한국어 문서의 자동 분류 : 베이지안 분류자

허준희*, 고수정*, 김태용**, 최준혁***, 이정현*
*인하대학교 대학원 전자계산공학과, **문경대학 컴퓨터정보과
***김포대학 컴퓨터계열
*E-mail : jjun2@nlsun.inha.ac.kr

An Automatic Classification of Korean Documents Using Weight for Keywords of Document and Corpus : Bayesian classifier

Jun-Hui Her*, Soo-Jeong Ko*, Tae-Yong Kim**, Jun-Hyeog Choi***, Jung-Hyun Lee*
*Dept. of Computer Science & Engineering, Inha University, Inchon, Korea
**Dept. of Computer Information, Munkyung College, Munkyung, Korea
***Dept. of Software Engineering, Kimpo College, Kimpo, Korea

요 약

문서 분류는 미리 정의된 두 개 또는 그 이상의 클래스에 새로 생성되는 객체들을 할당하는 방법이다. 문서의 자동 분류에 대한 연구는 오래 전부터 연구되어 왔지만 한국어에 대한 적용 및 연구는 다른 분야에 비해 아직까지 활발히 이루어지지 않고 있다.

본 논문에서는 문서를 자동으로 분류하기 위해 문서의 주제어에 가중치를 부여하고, 부족한 문서의 특징을 보충하기 위하여 말뭉치로부터 주제어들과의 상호정보에 의해 추출된 단어를 사용하여 문서를 표현한 후, 가중치를 부여한 문서의 주제어에 베이지안 분류자를 사용하여 문서분류를 수행한다. 실험은 한국어 정보검색 실험용 데이터 집합인 KTset95 문서 4,414개 중 1,300개의 문서를 학습 집합으로, 1,000개의 문서를 분류에 대한 검증 집합으로 사용하였다. 실험 결과, 순수 베이지안 확률을 사용한 기존의 방법보다 실험 집합과 검증 집합에서 각각 1.92%, 4.3% 향상된 분류 정확도를 얻었다.

1. 서 론

새로운 문서를 기존에 존재하는 클래스들에 할당하는 방법을 문서의 자동분류라고 한다[1]. 문서를 체계적으로 분류하고 관리하기 위한 문서 분류는 오래 전부터 사용되어 왔으나, 대부분의 분류 방법은 전문가의 수작업에 의해 진행되어 왔다. 그러나 수작업에 의한 분류는 전문가의 주관을 배제하기 힘들고 높은 비용에 비해 얻는 효과가 적다는 단점이 있다. 따라서, 컴퓨터에 의해 정해진 규칙에 의해 자동으로 문서를 분류하는 방법에 대한 연구들이 진행되어 오고 있다.

자동 문서분류에 대한 기존의 연구들로는 확률을 이용한 방법[2,5,10,11], 통계적인 기법을 이용한 방법, 벡터 유사도를 이용하는 방법[3], 엔트로피를 이용하는 방법[4] 등이 있다. 이들 중에서 확률을 이용하여 학습을 하고 문서를 분류하는 방법이 가장 많이 연구되었으며, 이 방법은 일반적인 문서 집합에 대해 가장 높은 분류효율을 나타내고 있다[1,5].

그러나 이러한 연구들의 대부분은 영어를 대상으로 연구된 결과이다. 따라서 각 나라마다의 언어적인 특성을 고려하지 않고 연구된 기존의 방법들을 한국어에 그대로 적용한다는 것은 많은 문제점이 있다. 한국어 문서에 대해 분류를 시도하기 위한 기존의 연구들로는 신경망과 k-NN을 이용하여 신문기사를 분류한 방법[6], 확률벡터와 교차엔트로피를 사용한 방법[7], 벡터테이블과 휴리스틱 자료를 이용한 방법[8], 역카테고리 빈도와 벡터 유사도를 사용한 방법[9] 등이 있다.

본 논문에서는 한국어의 특성에 맞게 문서들의 특징을 추출하고 자료의 희소성 문제를 해결하기 위해 말뭉치로부터 얻어진 주제어의 연관 단어를 사용한다. 그리고, 문서의 주제어별 가중치가 부여된 베이지안 분류자를 통해 자동으로 문서를 분류하는 방법을 제안한다.

2. 자동 문서분류

베이지안 확률을 이용한 대부분의 연구는 naive Bayes 분류자라고 불리는 변형된 베이지안 분류법을 사용하였다. naive Bayes는 복잡한 베이지안 추정 방법을 단순화하여 분류에 적용하는 베이지안 규칙을 기반으로 한 분류법이다[2].

문서의 분류에 있어서 문서의 특징들은 문서에 포함되어 있는 단어들의 벡터집합으로 구성되는데, naive Bayes를 사용하는 분류법은 학습 데이터에서 발생하지 않은 단어는 무시하게 된다. 따라서 문서의 크기가 작거나 문서의 특징을 나타낼 단어가 부족해 문서를 적절히 표현할 수 없는 경우 문서를 오분류할 가능성이 크다. 본 논문에서는 말뭉치로부터 자동으로 생성된 단어군집을 이용하여 이러한 데이터의 희소성 문제를 해결하고자 시도하였다.

2.1 문서의 표현 및 특징 추출

문서를 분류하는데 있어서 문서의 특징이 될 수 있는 단어들을 추출하는 것은 중요하다. 또한 학습 집합에서의 특징 추출은 사전의 크기를 줄이는 데 유용하게 사용할 수 있다. 기존의 일반적인 학습 방법들은 학습 문서에 출현하는 모든 용어를 대상으로 사전을 구축하였기 때문에 분류에 영향을 미치지 않는 단어들에 대한 정보도 유지하게 된다. 이러한 사전 구성은 시스템의 저장 공간의 문제와 수행 속도, 계산의 복잡도 측면에서 비효율적이라고 할 수 있으며 문서 분류에 대한 오분류의 요인으로 작용할 수 있다.

본 논문에서는 각각의 문서에 대한 특징이 아닌 클래스 특징을 반영하기 위해 클래스 변수와 단어들간의 상호정보량을 이용해 특징을 추출하고 사전을 구성하게 되며, 이때 사용하는 식은 (식 1)과 같다[10].

$$I(C; W_t) = p(c, f_t)\log\left(\frac{p(c, f_t)}{p(c)p(f_t)}\right) \quad \text{(식 1)}$$

$$p(c, f_t) = \frac{\textit{frequency of } w_t \textit{ where class label is } c}{\textit{number of total words}}$$

$$p(c) = \frac{\textit{number of total words where class label is } c}{\textit{number of total words}}$$

$$p(f_t) = \frac{\textit{number of } w_t \textit{ occurence}}{\textit{number of total words}}$$

2.2 베이지안 추정치를 통한 학습

본 논문에서는 단어의 발생여부 만을 사용하는 방법이 아닌 단어의 출현빈도를 고려하는 다항 베이지안 학습법을 사용한다[11]. 학습 알고리즘으로는 형태소 분석을 통해 선택된 명사와 클래스 변수와의 상호정보 계산을 통해 추출된 특징을 이용하여 학습을 수행하는 [알고리즘 1]을 사용하였다.

```
Voc ← words through feature selection
Data ← training data set
for each class variable{
   docs_j ← documents which class variable is v_j
   P(v_j) ← |docs_j| / |Data|  //각 클래스의 출현확률
   V ← total words in docs_j
   n ← number of words in V
   /* 각 단어에 대한 추정치 계산 */
   for each w_k in Voc{
      n_k ← frequency of w_k in V
      P(w_k|v_j) ← (n_k + 1) / (n + |Voc|)
   }
}
```

[알고리즘 1] 특징이 추출된 사전을 이용한 베이지안 학습 알고리즘

2.3 문서 분류

분류될 문서가 주어지면 먼저 형태소 분석을 수행하여 주제어 후보들을 생성한 다음 (식 2)의 가중치 계산을 통하여 주제어를 추출한다[3].

$$W_i = TF \cdot IDF = TF_i(\log_2(n) - \log_2(DF_i) + 1) \quad \text{(식 2)}$$

만일 문서의 크기가 작아 문서의 특징을 나타내는 주제어들이 최소 특징 수보다 부족하면, 부족한 특징을 보충하기 위해 코퍼스로부터 단어 군집화를 통해 만들어진 유사단어 쌍을 이용하여 부족한 특징을 보충하여 데이터의 희소성을 해소할 수 있다.

문서의 특징추출로 각 주제어가 선택되었으면 (식 2)의 가중치를 고려하는 베이지안 분류자를 통해 확률 값이 가장 높은 클래스에 문서를 할당하게 된다. (식 3)은 문서에서 추출된 주제어들이 클래스 v_j에 포함될 확률의 곱을 나타낸다.

$$v_{NB} = \arg\max_{v_j \in V} p(v_j) \prod_{i \in I} W_i \, p(a_i|v_j) \quad \text{(식 3)}$$

여기서 $p(a_i|v_j) = p(w_1|v_j)p(w_2|v_j)\cdots p(w_i|v_j)$이고, W_i는 (식 2)와 같이 계산된다. 이때, DF_i는 전체 문서가 아닌 각 클래스의 문서들로부터 얻어진다. 본 논문에서 설계한 자동 분류 시스템을 정리하면 [알고리즘 2]와 같다.

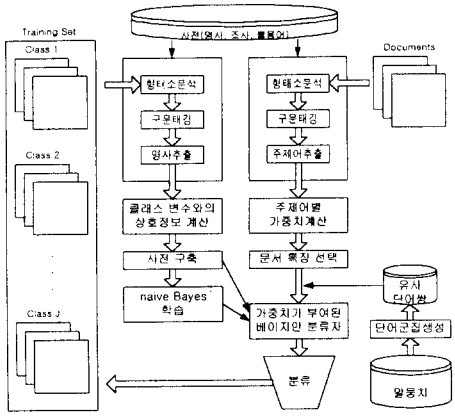
```
Num_words ← # of words in document;
w_k ← each word in document;
Num_class ← # of class variable;
z ← # of feature;
for(k=1; k≤Num_words){
    W_k ← Calc_Weight( w_k);
    Buf ← Save_Buf( W_k);
}
Sort(Buf); //TFIDF 값이 큰 순서로 정렬
Select(z); //상위 z개를 주제어로 선택
/* 주제어의 수가 부족하면 단어 군집으로부터 부족한
  특징을 보충 */
if(feature number < z)
    then(supplement keywords);
for(j=1; j≤Num_class){
    for(each keyword){
      Prob ← Calc_W( w_i) * Calc_P( w_i|v_j);
    }
    V ← Calc_P( v_j) * Prob;
    return Max(V); //최대확률 값
}
Assign(document); //V값이 최대인 클래스에 할당
```

[알고리즘 2] 주제어별 가중치가 부여된 분류자를 사용한 분류 알고리즘

3. 실험 및 결과

[그림 1]은 본 논문의 문서 자동 분류 시스템에 대한 구성도이다. 문서 분류는, 먼저 지식 베이스를 구축하기 위해 말뭉치로부터 단어 군집화를 수행하여 연관있는 단어의 집합(단어쌍의 형태)으로 구성한다. 학습 단계에서는 학습 데이터의 문서로부터 (식 1)을 통해 특징을 추출하여 분류에 사용될 사전을 구축한 후, 레이블 된 학습 데이터의 클래스로부터 베이지안 추정치를 계산하여 저장한다. 분류 단계에서는 새롭게 분류될 문서로부터 (식 2)를 사용해 주제어를 추출한다. 이렇게 선택된 주제어의 수가 역치 이하이면 군집화된 단어 쌍으로부터 얻는 연관 단어들을 특징으로 분류정보에 추가한다. 최종적으로 각 클래스에 대해 문서의 주제어별 가중치를 계산하고 (식 3)을 통해 계산된 값이 가장 높은 클래스에

문서를 할당한다.



[그림 1] 가중치 부여와 말뭉치를 이용한 문서분류 시스템의 구성도

실험을 위해 한국어 정보검색 시스템의 성능 평가용 데이터 집합인 KTset95 문서 4,414개 중 1,300개의 문서를 학습 집합으로, 1,000개의 문서를 검증집합으로 사용하여 실험을 수행하였다. 학습 집합의 클래스는 수작업으로 전산학 각 연구 분야의 14개 클래스로 분류하였다. KTset95 문서 중 정의된 클래스에 해당하지 않는 문서들은 사용하지 않았다. 학습 문서들로부터 78,156개의 용어들이 추출되었고, 이 중에서 중복된 명사들을 제거한 후 클래스 변수들과의 상호정보를 계산하여 3,750개의 단어들로 사전을 구성하였다.

학습집합과 검증집합의 문서에 대하여 본 논문에서 설계한 시스템에 의해 분류한 결과는 [표 1]과 같다. 분류 정확도는 바르게 분류된 문서 수를 전체 문서수로 나누어 백분율로 나타내었다.

[표 1] 분류 실험 결과

		naive Bayes	제안된 시스템
학습집합	오분류된 문서 수	70	45
	분류 정확도(%)	94.62	96.54
검증집합	오분류된 문서 수	120	77
	분류 정확도(%)	88.00	92.30

본 논문에서 제안한 시스템을 사용하는 경우 기존의 naive Bayes 분류자를 사용했을 때보다 분류 정확도가 실험집단에서 1.92%, 검증집단에서 4.3% 향상되었다.

제안된 시스템의 오분류에 대한 내용을 살펴보면, 문서에 자주 등장하여 높은 가중치를 가지는 단어가 해당하는 문서의 내용과 연관성이 적은 단어이거나, 문서의 내용이 둘 이상의 주제에 대해 다루고 있는 경우가 오분류된 문서의 95% 정도를 차지하고 있다. 기존의 naive Bayes를 이용한 분류실험의 경우 오분류된 문서의 90% 이상이 문서의 크기가 50어절 미만인 경우였다. 이는 문서의 특징을 나타내는 단어의 수가 부족하기 때문이다. 본 논문에서는 이러한 50어절 미만의 문서에 대해서도 연관 단어 군집을 이용해 해당되는 클래스로의 분류를 시도하였다.

4. 결 론

본 논문에서는 부족한 문서들의 특징을 보충하기 위해 말뭉치로부터 상호 정보에 의해 군집화된 연관 단어들을 사용하고, 각 문서별로 추출된 주제어들에 가중치를 부여하여 문서들을 분류하였다. 기존의 naive Bayes를 사용한 분류는 문서에 출현한 모든 단어에 대해서 추정치를 계산하고 이를 바탕으로 분류를 수행하였기 때문에 문서의 특징들을 정확히 반영하기 어렵고, 많은 잡음들의 영향으로 문서를 오분류하게 된다. 본 논문은 분류될 문서의 특징을 나타내는 주제어들과의 연관 단어 군집을 통하여 이러한 문제를 해결하고자 시도하였다. 실험 결과, 기존의 방법보다 학습집합에서 1.92%, 검증집합에서 4.3% 향상된 분류정확도를 얻을 수 있었다.

향후, 말뭉치로부터 좀 더 정확한 중의성이 해결된 단어 군집화와 자연언어 처리 기법을 도입해 문장의 문맥 파악이 가능하다면 더욱 정확하게 문서를 해당 클래스에 분류할 수 있을 것으로 기대한다.

참 고 문 헌

[1] Yonghong Li and Anil K. Jain, "Classification of Text Documents," Proceedings of the 14th International Conference on Pattern Recognition, Vol.2, pp.1295-1297, 1998.

[2] David D. Lewis, "Naive (Bayes) at forty: The Independence Assumption in Information Retrieval," In European Conference on Machine Learning, 1998.

[3] W. Frakes and R. Baeza-Yates, *Information Retrieval*, Prentice Hall, 1992.

[4] 정영미, *정보검색론*, 구미무역 출판부, 1993.

[5] Hein Ragas and Cornelis H.A. Koster, "Four text classification algorithms compared on a Dutch corpus," Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.369-370, 1998.

[6] 조태호, "신경망 또는 k-NN에 의한 신문 기사 분류와 그의 성능 비교," 한국정보과학회 가을 학술발표논문집, Vol. 25, No. 2, pp.363-365, 1998.

[7] 한미성, 송영훈, 송점동, 이정현, "확률 벡터간의 교차 엔트로피 계산을 이용한 자동 문서 분류 시스템," 한국정보처리학회 추계 학술발표논문집, Vol.4, No.2, pp.625-630, 1997.

[8] 최동시, 정경택, "카테고리와 키워드의 밀접성 정보에 의한 문서 자동 분류 시스템 설계 및 구현," 한국정보과학회 가을 학술발표논문집, Vol.22, No.2, pp.639-642, 1995.

[9] 조광제, 김준태, "역 카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동 분류," 한국정보과학회 봄 학술발표논문집, Vol.24, No.1, pp.507-510, 1997.

[10] Andrew McCallum and Kamal Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," AAAI-98 Workshop on Learning for Text Categorization, 1998.

[11] Tom M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.