하이퍼텍스트 문서의 자동분류를 위한 워드넷 기반 특징 합병 기법

노준호*, 김한준*, 장재영**
*서울시립대학교 전자전기컴퓨터공학부
**한성대학교 컴퓨터공학과
e-mail:loece@uos.ac.kr, khj@uos.ac.kr, jychang@hansung.ac.kr

A WordNet-based Feature Merge Method for HyperText Classification

Jun-Ho Roh*, Han-Joon Kim*, Jae-Young Chang**
*School of Electrical and Computer Engineering, University of Seoul
**Department of Computer Engineering, Hansung University

요 약

본 논문은 하이퍼텍스트 문서의 자동분류 성능을 높이기 위한 새로운 접근법을 제시한다. 하이퍼텍스트 문서는 일반 문서와 달리 하이퍼링크로 서로 연결된 구조를 가진다. 이 하이퍼링크 정보는 대상문서와 연관도가 높은 정보를 가지고 있으며, 이러한 링크 정보로부터 특징을 보다 잘 선별하기 위해서는 보다 정밀한 접근법이 필요하다. 본 논문은 단어간 의미 유사도를 기반으로 하이퍼텍스트 링크 정보를 활용한 특징 가공기법을 제안한다. 제안 기법은 하이퍼링크 문서로부터 대상문서와 연관도가 높은 특징을 추출하기 위해 단어간 유사도 함수를 사용하며, 유사도 함수는 워드넷의 상/하위어 관계를 이용한다. 그리고 추출된 특징들 중 의미적으로 비슷한 개념의 특징들을 합병함으로써 의미적으로 보다 견고한 분류 모델을 구축한다. 제안 기법을 검증하기 위해 Web-KB 문서집합을 이용하여 실험을 수행하였고 실험 결과 기존 방법보다 우수한 성능을 보였다.

1. 서론

최근 웹 문서의 자동분류시스템의 성능을 향상시키기 위해 하이퍼텍스트 문서가 갖는 특성을 활용한 연구가 활발하다. 하이퍼텍스트는 반구조적 문서(Semi-structured document)로서 일반 문서와는 달리 태그(Tag) 정보와 하이퍼링크(Hyperlink) 정보를 가지고 있다. 이 중에서 하이퍼링크는 문서 간의 관계를 나타내는 중요한 정보가 담겨 있으며 하이퍼링크로 연결된 문서는 서로 밀접한 내용을 공유하고 있다.

하이퍼링크 정보를 이용하기 위해서는 링크 문서에서의 특징 추출이 선행되어야 하며, 이는 크게 두 가지 방식이 존재한다. 링크 문서들 중 대상 문서와 연관도가 높은 링크 문서를 선별하는 방식[1]과 모든 링크 문서 내의 특징 중 중요한 특징만을 선별하는 방식[2, 3]이 그 것이다. 이 들 모두 대상 문서와 연관도가 높은 용어만을 특징으로 사용한다. 예를 들어 [1]에서는 문서 간 유사도를 기반으로 신뢰도를 산정하여 링크 문서를 선별하고, 선별된 링크 문서와 대상 문서내의 공통된 용어들에 대해 가중치를 부여하는 방법을 제안하였다. [2]에서는 링크 문서들의 클래스 이름(Class label)만을 특징으로 사용하여 분류 정확[1]도를 향상 시켰고, [3]에서는 링크 문서내의 앵커 텍스트(Anchor text)와 앵커 텍스트 주변단어를 특징으로 추출할 것을 제안하였다. 그러나 이와 같은 방식은 정보손실이 발생하거나 대상문서와의 연관도가 떨어지는 특징을 추출하는 문제점이 존재한다.

일반적으로 자동문서분류 기법은 주로 기계학습(Machine learning) 기술을 사용하며, 기계학습 방법으로는 나이브 베이즈(Naïve bayes), 지지벡터기계(Support vector machine), 신경망(Neural network) 등이 있다. 이러한 알고리즘은 모두 문서를 용어집합(Bag of words)으로 표현하고 이를 기반으로 분류모델을 구축한다. 그러나 이러한 표현 방식은 모든 용어가 특징으로 사용되기 때문에 단어간 관계 정보는 무시되고 차원의 수는 커지게 된다. 이를 고려하기 위해서는 워드넷(WordNet)과 같은 어휘사전을 이용한 특징 가공을 수행해야 한다. [4]와 [5]에서는 각 특징마다 동의어나 상위어 등을 추가하는 특징 확장 방식을 사용하여 특징 집합의 질을 높였고, [6]과 [7]에서는 특징 집합 내에서 동의어인 특징들을 합병하는 방식을 제안하였다.

본 논문에서는 하이퍼텍스트 문서의 분류성능을 높이기 위해 기존 특징 추출 방법에서 워드넷 기반 단어간 연관도를 고려한 새로운 특징 추출 방법을 제안한다. 또한 추출된 특징 집합 내에서 연관도가 높은 특징들을 합병함으로써 의미적으로 보다 견고한 분류 모델을 구축한다. 최종적으로 대상문서는 링크 문서의 특징들로 표현되어지며, 특징 합병에 따라 문서를 나타내는 개념이 보다 분명해지게 된다.

* 교신저자 : 김한준 (khj@uos.ac.kr)

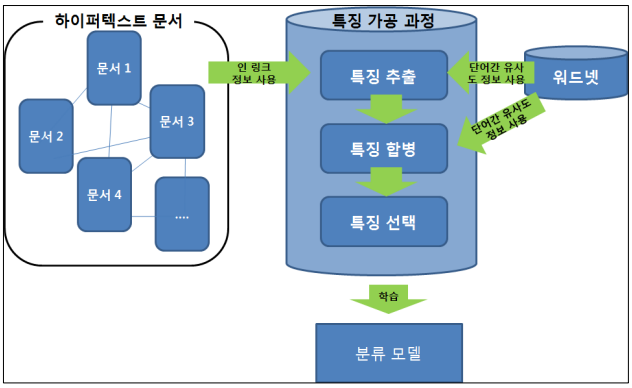
2. 워드넷 기반 하이퍼텍스트 문서 특징 가공

2.1 하이퍼 텍스트 문서의 특징 가공

문서 분류 시스템은 문서 집합으로부터 적합한 특징을 추출하고 양질의 특징을 선택하여 분류 모델을 구축하는 학습단계와 구축된 분류 모델을 적용하여 문서의 클래스를 예측하는 분류단계로 이루어진다. 일반적으로 문서분류의 정확도를 높이기 위해서는 학습단계에서의 특징 가공을 통한 최적의 특징 벡터 생성이 요구된다. 일반 문서 분류 모델의 경우 문서에 출현하는 특징만을 추출하여 특징 벡터로 사용하지만, 하이퍼텍스트 문서 분류 모델의 경우 링크 관계에 의한 정보로 인해 보다 정밀한 특징 가공 기법이 필요하다.

하이퍼링크는 대상문서로 들어오는 인링크(In-coming link)와 대상문서에서 나가는 아웃링크(Out-going link)로 나뉜다. [3]에서는 인링크 문서의 앵커 텍스트와 앵커 텍스트 주위 2-30 개의 단어를 사용하여 문서분류 정확도를 향상시켰다. 그러나 이와 같은 휴리스틱은 다소 임의적인 방법이므로 일반화하여 적용하기 어렵다.

본 논문에서는 의미 어휘사전인 워드넷을 이용하여 인링크 문서의 특징 중 대상문서와 연관이 높은 특징만을 추출할 것을 제안한다. 또한 추가적으로 주요 특징들의 가중치를 높이기 위해 인링크 문서로부터 추출된 특징들을 상위 개념으로 추상화하여 합병할 것을 제안한다. (그림 1)은 전체적인 하이퍼텍스트 문서 특징 가공 과정을 나타낸다.



(그림 1) 특징 가공 과정

2.2 특징 추출

2.2.1 인링크 문서에서의 특징 추출

링크 관계 문서 중 인링크 문서는 대상 문서를 나타내는 중요한 정보를 담고 있다. 인링크 문서에서 특징을 추출할 때는 전체 문서보다는 클래스 이름, 앵커 텍스트, 앵커 텍스트 주변단어 등을 사용한다. 하지만 이러한 특징들은 중요한 문제점을 가진다. 예를 들어 클래스 이름의 경우 전체 특징 개수가 많지 않고, 앵커 텍스트의 경우 한 문서의 내용을 간단한 단어로 추상화하였기 때문에 많은 정보손실이 발생한다. 앵커 텍스트 주변단어의 경우 대상 문서와 무관한 특징들이 다수 존재할 가능성이 높다. 그러므로 본 논문에서는 인링크 문서 내에서 앵커텍스트와 연관이 높은 용어들만을 추출할 것을 제안한다.

기본적으로 인링크 문서의 특징 추출은 아래와 같이 A, B의 방법을 사용한다[3]. 본 논문에서는 단어간 연관도를 고려한 특징 추출 방법인 C를 제안한다.

A. 앵커 텍스트 : 하이퍼링크 태그 안에 있는 앵커 텍스트를 특징으로 추출

B. 확장된 앵커 텍스트 : 앵커 텍스트 주변단어를 추출. 앵커 텍스트 전 후 20개 단어까지 특징으로 추출

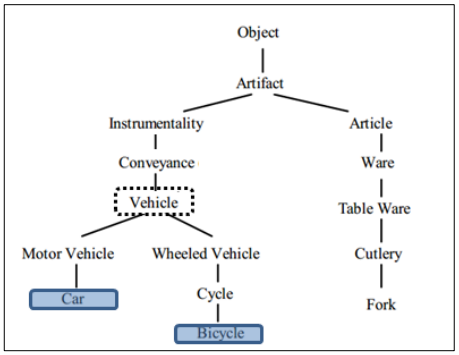
C. 워드넷 기반 확장된 앵커 텍스트 : 확장된 앵커 텍스트 단어 중 앵커텍스트와 연관도가 높은 단어를 특징으로 추출

제안 방법 C는 먼저 인링크 문서에서 앵커 텍스트와 앵커 텍스트 주변단어를 추출하고, 모든 앵커 텍스트와 주변 단어간 연관도를 계산한다. 그 후, 연관도가 일정 임계값(threshold) 이상인 특징들만을 추출한다. 여기서 임계값은 어느 정도의 유사도를 가지는 특징들을 추출할 것인지 정하는 상수이며, 이 값이 1 이면 동의어만을 특징으로 추출하게 된다. 이를 통해 대상문서와 연관도가 높은 특징들을 추출하면서도 대상문서와 무관한 특징들은 제거할 수 있다.

이러한 특징 추출을 하기 위해서는 각 단어들 간의 유사도를 계산하는 것이 필요한데, 이를 위해 워드넷을 이용한 유사도 함수가 요구된다. 유사도를 계산하기 위해 워드넷 관계중 상/하위어 관계를 활용한다.

2.2.2 워드넷 기반 단어간 유사도

워드넷은 영어 의미 어휘목록 관계정보를 담은 사전으로서 동의어집합(Synset) 단위로 이루어져 있다. 또한 각 동의어집합에 대한 상위어, 하위어, 등위어, 전체어 등의 의미 관계들을 제공한다. 단어간 연관도를 계산하기 위해 워드넷에서의 상/하위어 의미관계를 이용할 수 있다.



(그림 2) 워드넷에서의 계층 구조(상/하위어)

(그림 2)는 Car와 Bicycle 두 단어간의 상/하위어 의미 관계도를 나타낸다. 여기서 Vehicle은 두 단어의 공통된 상위어이고 이 노드를 통해 두 단어가 연결되어 있다. 때문에 단어간 거리는 연결된 노드를 지나는 최소거리로 정의할 수 있다[8]. 본 논문에서 단어간 연관도를 구하는 식은 다음과 같다[9].

$$\mathrm{sim}(a,f) = 1 - \frac{\min \mathrm{dist}(a,f)}{\min \mathrm{dist}(\mathrm{common\ parent}, \mathrm{root}) + \min \mathrm{dist}(a,f)}$$

여기서 a는 앵커 텍스트를 나타내고, f는 앵커텍스트 주변단어를 나타낸다. min dist는 워드넷 상/하위어 관계에서의 최소 거리를 의미하고, common parent는 두 단어의 공통의 상위어를 의미한다. 마지막으로 root는 가장 최상위 개념의 단어를 의미한다. 이 식은 정규화된 식으로 두 단어가 동의어면 최소 거리가 0이 되어 연관도는 1의 값을 갖게 되며, 반대로 공통의 상위어가 없으면 최소 거리가 무한이 되어 연관도는 0의 값을 갖게 된다.

2.3 특징 합병

앞에서와 같이 생성된 특징 벡터는 대상 문서와 연결된 링크 문서의 용어이기 때문에 대상문서를 표현함에 있어서 중요한 정보가 된다. 하지만 단순히 단어의 빈도수를 벡터화 한 것이기 때문에 단어간 의미관계를 고려하지 못한다. 이로 인해 문서 내 중요한 개념을 갖는 특징일지라도 단어 빈도수가 낮을 경우 가중치가 낮게 책정될 수 있다. 이를 고려하기 위해 본 논문에서는 워드넷을 사용해 추출된 특징으로부터 서로 연관도가 높은 단어들을 합병함으로써 하나의 상위 개념으로 추상화할 것을 제안한다. 본 논문에서 제안하는 방식은 최근접 연관도 특징 합병 방식과 일정 연관도 내 모든 특징 합병 방식이다.

A. 최근접 연관도 특징 합병

문서에 존재하는 모든 특징 간 연관도를 계산하고 가장 높은 연관도를 가지는 2개의 특징을 임계값 이상이면 합병한다. 이 과정을 안정화될 때까지 반복 수행 한다. (그림 3)은 최근접 연관도 특징 합병 알고리즘을 보여준다.

B. 일정 연관도 내 모든 특징 합병

각 특징에 대해 문서에 존재하는 모든 특징 간 연관도를 계산한 후 일정 임계값 이상의 연관도를 가지는 특징들을 모두 합병한다. (그림 4)는 일정 연관도 내 모든 특징 합병 알고리즘을 보여준다.

```
입력 : 특징 벡터 {f1, f2, f3, ..., fN}
출력 : 특징 합병 후 개선된 특징 벡터 {f1, f2, ..., fM} (M ≤N)

    BEGIN
1   while( Max_sim > 임계값 ) {
2      for( i < 특징 벡터 크기) {
3         for( j < 특징 벡터 크기) {
4            sim[i][j] = sim(fi, fj)  /*유사도 계산*/
5         }
6      }
7      Max_sim = 0
8      for( i < 특징 벡터 크기) {
9         for( j < 특징 벡터 크기) {
10           if( Max_sim < sim[i][j] ) {
11              Max_sim = sim[i][j] /*Max_sim에 유사도
12                                    최대값 저장 */
13              Max_i = i  /*유사도가 최대일 때의 특징
14              Max_j = j  fi, fj를 Max_i,Max_j에 기억*/
15           }
16        }
17     }
18     if( Max_sim > 임계값){
19        tf(f_Max_i) = tf(f_Max_i) + tf(f_Max_j)
20  /* 두 특징의 단어 빈도수를 합하고, 하나의 특징집합
21  으로 f_Max_i, f_Max_j 합병 */
22     }
23  }
    END
```

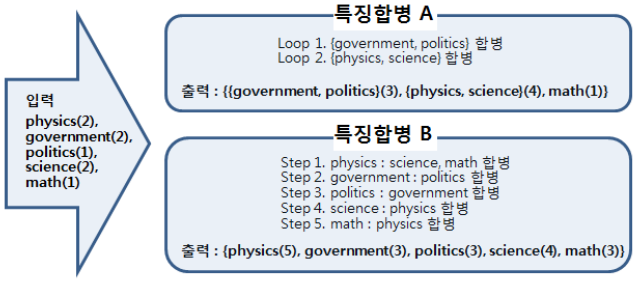
(그림 3) 최근접 연관도 특징 합병 알고리즘

```
입력 : 특징 벡터 {f1, f2, f3, ..., fN}
출력 : 특징 합병 후 개선된 특징 벡터 {f1, f2, ..., fN}

    BEGIN
1      for( i < 특징 벡터 크기) {
2         for( j < 특징 벡터 크기) {
3            sim[i][j] = sim(fi, fj)  /*유사도 계산*/
4         }
5      }
6
7      for( i < 특징 벡터 크기) {
8         for( j < 특징 벡터 크기) {
9            if( sim[i][j] > 임계값 ) {
10              tf(fi) += tf(fj)
11  /* fi와 유사도가 높은 모든 fj의 단어빈도수를 합함 */
12           }
13        }
14     }
    END
```

(그림 4) 일정 연관도 내 모든 특징 합병 알고리즘

(그림 5)는 임의의 특징 벡터가 입력될 때에 각 특징 합병 방식에서의 출력 결과 예시이다. 여기서 괄호 안에 숫자는 단어 빈도수를 의미한다.



(그림 5) 특징 합병 예시

일련의 과정을 거쳐 하이퍼텍스트 문서의 링크 관계를 이용한 특징 벡터를 생성하였다. 하지만 이 특징 벡터는 차원의 수가 많고 특정 특징들이 클래스마다 고루 분포되어 있을 수 있다. 이러한 특성은 분류모델의 악영향을 미치며 이를 해결하기 위해서 일반적으로 특징 선택 기법을 사용한다. 본 논문에서는 χ^2-statistics을 이용한 특징 선택을 수행하였다.

3. 성능분석

3.1 실험방법

본 논문에서 제안한 기법의 효율성을 검증하기 위해서 Web-KB 문서집합을 이용한 실험을 수행하였다. 이 문서집합은 여러 대학교의 웹 페이지를 7개의 클래스로 구분된 8,282개의 웹페이지로 구성되어져 있다. 본 논문에서는 서로 링크 관계가 존재하는 데이터가 필요하기 때문에 이 중 Wisconsin 대학의 문서집합을 대상으로 실험을 실시하였고, 클래스 간에 문서의 분포가 불균형한 문제로 인해 문서 수가 10개가 되지 않는 클래스의 문서는 실험에서 제외하였다. 이로 인해 4개의 클래스로 구분된 1,224개의 문서를 사용하여 실험을 수행하였고, 문서분류 알고리즘은 MALLET 시스템[10]에서 제공하는 나이브 베이지안 알고

리즘을 사용하였다. 실험은 기존 특징 추출 방법과 제안된 특징 추출 방법의 문서분류 정확도를 비교하였고, 추가로 워드넷 기반 특징 합병의 효능을 알아보기 위해 2가지 특징 합병 방식의 문서분류 정확도를 비교하였다.

3.2 실험결과

<표 1>은 특징 추출 방법에 따른 분류 정확도와 특징 합병 전후의 분류 정확도를 보여주고 있다. 비합병 중 제안 방법인 워드넷 기반 확장된 앵커 텍스트를 사용하였을 때가 기존 방법을 사용했을 때보다 분류 정확도가 모두 높았고, 약 0.2-1.0% 정도의 정확도 향상을 보였다. 특이한 점은 기존 방법에서 앵커 텍스트를 사용하였을 때가 확장된 앵커 텍스트를 사용할 때보다 분류 정확도가 높다는 점이다. 이는 웹 문서의 특성상 앵커 텍스트 주변 단어들에서 대상 문서와 연관도가 떨어지는 단어들이 많이 존재한다는 것을 반증한다. 하지만 제안 방법을 사용하였을 때는 오히려 정확도가 올라갔으며, 이는 링크 문서들이 분류에 도움이 되는 정보를 가지고 있다는 것을 의미한다.

비합병과 특징 합병 A, B를 적용하였을 때의 성능을 비교하면 특징 합병을 하였을 때의 정확도가 비합병 일 때의 정확도보다 높은 것을 알 수 있다. 특히 특징 합병 B를 적용하였을 때의 분류 정확도가 가장 높게 향상 됐다. 마지막으로 워드넷 기반 확장된 앵커 텍스트를 사용하고 특징 합병 B를 수행할 때 78.94%로 가장 높은 정확도를 보였다. 이는 2가지 제안 방법인 특징 추출과 특징 합병이 모두 긍정적으로 분류 시스템의 성능을 향상시켰다는 점에서 고무적이다.

<표1> 분류 정확도(%)

	특징추출방법	비합병	특징 합병 A	특징 합병 B
기존 방법	앵커 텍스트	77.12	78.26	78.14
	확장된 앵커 텍스트	77.01	77.16	77.89
제안 방법	워드넷 기반 확장된 앵커 텍스트	77.26	78.82	**78.94**

4. 결론

웹 문서에는 실제 많은 연관 정보 및 의미 정보를 가지고 있지만 컴퓨터는 이러한 정보를 해석할 수 없다. 이 정보를 컴퓨터가 처리할 수 있도록 가공하면 문서분류에 큰 도움이 될 수 있다.

본 논문에서는 하이퍼텍스트 문서의 자동분류 정확도를 높이기 위해 링크 정보와 단어 의미 정보를 사용하였다. 링크 정보는 인링크 문서내의 용어들 중 앵커 텍스트와 유사도가 높은 특징들을 사용하였으며, 단어 의미 정보를 알기 위해 유사도가 높은 특징들을 대상으로 특징 합병을 수행하였다. 이를 위해 기본적으로 워드넷을 이용하여 단어간 유사도를 계산하였다. 또한 마지막으로 분류 모델에 도움을 주는 특징들을 가려내기 위해 특징선택을 수행하였다.

향후 연구 과제는 링크 문서를 선별하는 유사도 함수를 고안하여 대상 문서와 연관도가 보다 높은 특징만을 추출하는 것이다. 또한 단어들의 의미 중의성 해소를 통해 특징 합병이 더욱 정밀하게 이루어지게 하는 것이다.

5. 감사의 글

이 논문은 2010년 정보(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것이며 (과제번호: NRF-2010-0025212), 또한 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임.(과제번호: NRF-2011-0022445).

참고문헌

[1] 오효정, 맹성현, "점진적으로 계산되는 분류정보와 링크정보를 이용한 하이퍼텍스트 문서 분류 방법", 정보과학회논문지, 제29권, 제7·8호, 2002, pp.498-509.

[2] S. Chakrabarti, B. Dom and P. Indyk, "Enhanced hypertext categorization using hyperlinks", Proceedings of the ACM SIGMOD International Conference, 1998, pp.307 - 318.

[3] H. Utard and J. Fürnkranz, "Link-Local Features for Hypertext Classification", Semantics, Web and Mining: Joint International Workshops, EWMF/KDO, LNCS, Vol.4289, 2005, pp.51 - 64.

[4] T. Mansuy and R. Hilderman, "Evaluating WordNet Features in Text Classification Models", Proceedings of the 19th International Florida Artificial Intelligence Research Symposium, 2006, pp.568-573.

[5] S. Scott and S. Matwin, "Feature engineering for text classification", Proceedings of 16th International Conference on Machine Learning, 1999, pp.379-388.

[6] Z. Elberrichi, A. Rahmoun and M. A. Bentaalah, "Using WordNet for Text Categorization", The International Arab Journal of Information Technology, Vol.5, No.1, 2008, pp.16-24.

[7] Z. Lu, Y. Liu, S. Zhao and X. Chen, "Study on Feature Selection and Weighting Based on Synonym Merge in Text Categorization", 2nd International Conference on Future Networks, 2010, pp.105-109.

[8] J. Jiang and D. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy", Proceedings on International Conference on Research in Computational Linguistics, 1997, pp.19-33.

[9] RiTa.WordNet, A WordNet library for Java/Processing, http://rednoise.org/rita/wordnet/documentation/index.htm

[10] MALLET, MAchine Learning for LanguagE Toolkit, http://mallet.cs.umass.edu/