가중치 부여 휴리스틱을 이용한 개념 기반 문서분류기 TAXON의 개선

*강 원석[1], **강 현규, **김 영섬,
*안동대학교 컴퓨터공학교육과, **한국전자통신연구원

Improvement of A Concept-Based Text Categorization System(TAXON) Using Weight Determination Heuristic

*Wonseog Kang, **Hyunkui Kang, and **Youngsum Kim
*Dept. of Computer Engineering Education, Andong National University
**Electronics and Telecommunications Research Institute

요약

본 논문에서는 개념을 기반으로 문서의 분류를 하는 확률벡터 모델의 문서분류기 **TAXON**(Concept-based Text Categorization System)의 개선을 도모한다. TAXON은 한국어 문장을 분석하여 명사를 추출하고 명사의 개념을 시소러스 도구를 통해 획득한 후 이를 벡터화하여 주제와 입력 문서와의 관계성을 검사하는 문서 분류기이다. 본 논문은 문서 분류기 TAXON의 성능을 향상시키기 위하여 확률벡터 계산에 가중치 부여 휴리스틱을 도입한다. 그리고 시소러스 도구를 확장하여 문서 분류의 질을 높인다.

1. 서론[1]

문서 분류의 필요성에 의해 문서 분류에 대한 많은 연구[1,2,3,4,5,6,7]가 진행되고 있다. TAXON 시스템[11]은 그 중에 하나로서 용어 중심의 문서분류가 지니는 문제점을 개선한 시스템이다.

TAXON 시스템은 개념 기반의 문서 분류를 위해 시소러스 도구를 이용한다. 시소러스 도구는 용어가 가지고 있는 의미를 추출할 수 있게 한다. 즉, 시스템은 문서의 단어가 가진 의미를 시소러스 도구로 획득하여 문서의 주제를 결정한다. 앞으로 시소러스 도구에 의해 획득하는 의미를 개념이라고 표현한다. [11]에 사용한 시소러스 도구는 약 4000개의 단어에 대한 개념을 얻을 수 있는 것으로 확장이 필요하다. 본 논문에서는 15000개의 단어에 대한 개념을 얻을 수 있는 시소러스 도구로 확장하여 시스템의 성능을 개선한다.

[11]에서는 문장에 들어 있는 모든 명사에 대해 추출된 개념의 출현 빈도가 벡터에 표시된다. 그 벡터 값은 가중치가 반영되지 않았다. 하지만 아래와 같은 예를 보면 가중치를 반영해야 할 필요성이 있다.

(1) 서울 올림픽, 세계 기후, 가스 성운, 화학 반응

[11]에서는 복합 명사구 "서울 올림픽"에서 서울의 단어가 지니는 개념과 올림픽 단어가 지니는 개념이 대등한 비율로 벡터에 표현된다. 그런데 이 명사구에서 올림픽이라는 명사는 서울이라는 명사보다 문서의 주제에 더 많은 영향을 미친다고 볼 수 있다. 물론 이 명사구가 sports 영역이 아닌 다른 영역의 문서에 출현할 수도 있다. 그러나 일반적으로 이 명사구는 sports 영역에 해당하는 문서에서 빈번히 출현할 것이다. 따라서 본 논문에서는 올림픽 단어에 들어있는 sports 영역 개념에 대해 가중치를 부여하여 더 효과적인 문서 분류를 꾀한다.

(2) 소형 자동차, 외부 정형, 산성 식품

(2)와 같은 예에서는 앞 명사가 뒷 명사를 수식한다. 이 명사구의 주요 개념은 뒷 명사인 피수식 명사에 들어 있다[15]. 따라서 피수식 명사의 개념에 가중치를 부여한다. 본 논문에서는 확률벡터의 계산에 이와 같은 가중치 부여 휴리스틱을 적용하여 TAXON의 성능을 향상시키고자 한다.

TAXON 시스템은 단어의 개념을 획득하는 시소러스 도구와 입력 문장에서 단어를 분리하는 한국어 분석기, 개념의 가중치를 결정하는 가중치 결정기, 계산된 확률벡터로 문서의 주제를 찾는 관련도 분석기로 구성된다. 본 논문의 2장에 가중치 결정 휴리스틱을 기술하고, 3장은 문서분류기 **TAXON**의 구조에 대해 기술하고 4장은 실험과 결과를 분석하며 5장에서 결론을 맺는다.

2. 가중치 부여 휴리스틱

정보 검색 분야에서는 미등록어 문제와 recall의 향상을 위해 복합 명사를 분할한다[12,13]. 본 시스템에서도 미등록어 문

[1] 본 논문은 한국전자통신연구원의 98년도 위탁과제 연구비의 지원으로 연구되었음.

제를 해결하고자 한국어 해석에서 복합명사를 분할한다 그리고 문서분류의 질을 향상시키기 위해 분할된 명사의 개념에 대해 가중치를 부여하여 확률벡터에 표현한다.

복합 명사 형태의 연구에 의하면 복합 명사의 96%가 명사+명사(58 4%), 명사+의+명사(22 6%), 명사+명사+명사(7 0%). 명사+의+명사+명사(3.1%), 명시+명사+의+명사(5.1%) 중의 하나라고 한다[12]. 본 논문의 가중치 부여 휴리스틱도 이 형태로 분석된 명사에 대해 다루었다 가중치 결정 과정은 아래의 단계로 구성된다 이 휴리스틱은 명사에 대한 의미 의존 관계[14,15]를 근거로 하여 정의되었다

복합 명사는 N1 + N2 로 표현한다. 위 5 가지 유형에서 N2 는 복합 명사의 마지막 명사를 뜻하고 N1 은 마지막 명사와 가장 근접한 명사를 의미한다. 본 논문에서는 N1 앞의 다른 명사에 대해서는 고려하지 않았다.

1 N1 이나 N2 가 domain 개념이나 그 하위 개념을 지니면 그 개념의 가중치를 +1 을 더한다. 각 개념에 대한 기본 가중치는 1 이다. 화학 반응과 같은 명사구가 이에 해당한다 화학이라는 단어의 개념에 chemistry 개념이 들어있고 반응의 단어는 event 개념에 속한다. 이 명사구에서 화학의 단어가 영역 정보를 가지고 있으므로 이에 대한 가중치를 부가한다

2 N1 이 feature 개념을 가지고 N2 가 thing 이나 event 의 하위 개념을 가지면 N2 가 가지고 있는 개념의 가중치를 +1 증가시켜 N2 의 중요성을 표현한다 소형 자동차와 같은 복합 명사는 수식어의 역할이 문서의 주제에 큰 영향을 미치지 못한다.

3. N2 가 feature 개념을 가지고 N1 이 thing 이나 event 의 하위개념을 가지면 N1 이 가지고 있는 개념의 가중치를 +1 증가시켜 N1 의 중요성을 표현한다.

4 N2 가 thing 의 개념을 가지고 N1 이 event 의 개념을 가지면 역시 N2 의 개념의 가중치를 +1 더한다

5 N2 가 event 의 개념을 가지고 N1 이 thing 의 개념을 가지면 역시 N2 의 개념의 가중치를 +1 더한다.

1 번의 경우 시소러스 도구를 통해 domain 개념이나 세부 개념을 얻는 경우 이는 문서의 주제에 큰 영향을 미칠수 있는 것이므로 이에 대한 가중치를 부여한 것이고 2,3.4,5 는 의존 관계에서 수식어보다 피수식어의 중요성을 인정하여 가중치를 더한 것이다. 계산된 가중치는 입력 문서의 개념-확률벡터에 표현된다 즉 입력 문서의 단어들에 대해 얻어진 개념들의 가중치를 모두 합하여 가중치 총합을 구하고, 그리고 각 개념에 대해 문서에서 나타난 가중치의 합을 구한 후 이를 가중치 총합으로 나눈 것이 각 개념에 대한 벡터 값이 된다

3. 개념 기반 문서분류기 TAXON

개념 기반 문서분류기 **TAXON** 은 그림 1 과 같은 구조를 가진다 **TAXON** 은 입력기, 한국어 분석기, 시소러스 도구, 가중치 결정기, 확률벡터 관련도 분석기, 출력기로 구성된다.

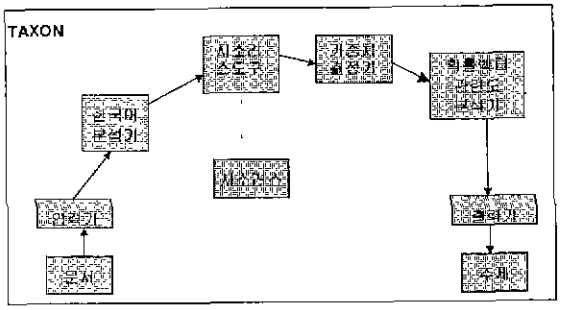

그림 1. TAXON 의 구조

입력기는 문서가 입력되면 그 입력의 결과를 한국어 분석기에 넘겨준다. 한국어 분석기는 입력된 문서를 형태소 해석하여 체언류를 분리한다 시소러스 도구는 한국어 분석기에서 넘어오는 단어에 대해 시소러스를 조사하여 단어가 가지고 있는 개념을 추출한다 가중치 결정기는 가중치 부여 휴리스틱에 의해 입력 명사들의 개념의 벡터값을 결정한다. 확률벡터 관련도 분석기는 입력 문서의 확률벡터와 주제의 확률벡터와의 관련도를 잎어 가장 근접하는 주제를 선택하여 **TAXON** 의 수행결과로 내어놓게 된다.

3.1 시소러스 도구

시스템의 문서 분류의 성능은 단어의 개념을 제공하는 시소러스 도구에 달려있다 [11]의 시소러스 도구는 약 4000 단어에 대해 개념을 추출할 수 있는 것이다 본 논문에서는 이를 15000 단어에 대해 개념을 추출할 수 있는 시소러스 도구로 확장하였다.

시소러스의 개념은 150 개의 개념으로 이들 간의 상하위 관계가 정의되어 있다 시소러스는 일반적으로 명사류와 동사류, 형용사류, 부사류 등의 모든 개념이 정의되나 **TAXON** 의 시소러스는 문서 분류에 중요한 체언류에 대해 개념이 정의되었다 체언의 개념도 행위의 개념 등이 포함되므로 개념 분류는 크게 물체, 사건, 속성의 세 갈래로 구성된다 그 분류는 [11]에 있다

시소러스 도구의 단어 15000 개는 백과사전의 문서에서 엔트리를 설명할 때 자주 사용되는 용어 3309 개와 KAIST SET[8]과 ETRI SET[9]에서 사용 빈도수가 높은 단어 12000 여개를 선정하여 중복되는 단어를 제거한 것이다 주어진 단어에 대해 시소러스 도구의 수행 결과의 예는 표 1 과 같다. 시소러스 도구는 단어에 대해 해당하는 개념을 출력한다 단어가 하나 이상의 개념을 가지면 시소러스 도구는 단어가 가질 수 있는 모든 뜻에 대한 개념을 추출한다 본 논문에서는 중의성의 문제를 제외하였다

표 1 시소러스 도구의 실행의 예

입력 단어	시소러스 개념
가감	event change quantity-change
가감법	intellectual-thing abstract-thing thing means mathematics domain
가객	animal animate-thing art domain human music physical-thing thing
가건물	structure man-made-thing inanimate-thing physical-thing thing construction domain
가게	abstract-thing inanimate-thing man-made-thing organization physical-thing structure thing
가격	domain economics feature measurement society
가계	abstract-thing domain domestic industry organization thing

3.2 확률벡터 관련도 분석기

본 논문에서는 [11]에서 제안된 확률 벡터 모델을 그대로 사용하여 문서와 주제를 표현하고, 그들 간의 관련도를 측정한다. 관련도 분석은 다음 세 단계를 거쳐 이루어진다.

(1) 문서 D는 다음과 같은 개념-확률벡터로 표시한다.

$$D = (wd_1, wd_2, \ldots, wd_n)$$

여기서 wd_i는 문서 D에 나타난 i-번째 개념의 가중치(확률)를 나타낸다. 따라서, wd_i, i=1,2, , n는 0과 1 사이의 값이며, $\sum_i wd_i = 1$이 된다.

(2) 유사하게 범주 C도 다음과 같이 확률 벡터로 표시된다.

$$C = (wc_1, wc_2, \ldots, wc_n)$$

여기서 wc_i는 범주 C로 분류된 문서 집합에서 나타난 i-번째 개념 wc_i의 가중치를 나타낸다. 따라서, wc_i, i=1,2, .., n는 0과 1 사이의 값이며, $\sum_i wc_i = 1$이다.

(3) 문서 D=$(wd_1, wd_2, \ldots, wd_n)$와 범주 C= $(wc_1, wc_2, \ldots, wc_n)$의 관련도는 다음과 같이 정의되는 관련도 측정 함수 SIM(D, C)을 이용하여 측정한다.

$$SIM(D, C) = 1 - H[(D + C)/2] - [H(D) + H(C)]/2$$

단, 확률벡터 $P = (w_1, w_2, \ldots, w_n)$에 대하여 H(P)는 P의 불확실성 정도를 나타내는 엔트로피(entropy)로써 $H(P) = -\sum_i w_i \times \log_2 w_i$로 계산된다[3].

4. TAXON의 실험과 결과

TAXON은 계몽사 백과사전의 문서를 학습문서로 하였다. 학습문서는 17965 문서이고 검사 문서는 3593 문서이다. 검사 문서를 대상으로 시스템을 실험하여 표 2와 같은 결과를 얻었다.

표 2 시스템의 실험 결과

	학습 문서	비학습 문서
기존의 TAXON	69.6%(83.1%)	69.2%(84%)
가중치 결정기 추가	72.1%(83.7%)	71.8%(84.3%)
개선된 시소러스 도구 사용	78.2%(86.3%)	77.9%(85.4%)
개선된 시소러스 도구와 가중치 결정기 사용	78.5%(86.5%)	78.2%(85.6%)

표 2의 괄호 안의 성공률은 두번째 주제까지 포함한 경우의 성공률이다. 이 표에서 가중치 결정기를 사용한 경우 성공률이 69.2%에서 71.8%로 증가하였다. 그러나 개선된 시소러스 도구를 사용한 결과 성공률이 69.2%에서 77.9%로 그 변화가 뚜렷하였다. 이것은 개념 기반의 문서 분류의 성공률은 시소러스 도구에 달려있음을 의미한다. 개선된 시소러스 도구와 가중치 결정기를 둘다 사용한 경우 69.2%에서 78.2%의 성공률의 증가를 일었다. 앞으로 가중치 결정기의 개선을 위해 언어학적인 분석의 연구가 필요하다.

5. 결론

본 논문은 개념 기반 문서 분류기 **TAXON**의 개선을 위해 가중치 부여 휴리스틱을 적용하였고 그리고 개념을 추출하는 시소러스 도구를 확장하였다. 개선된 시소러스 도구와 가중치 결정기를 사용하여 시스템을 실험한 결과 시스템의 성능이 개선됨을 볼 수 있었다.

가중치 결정 휴리스틱은 명사들의 의미 관계를 이용하여 가중치를 결정한다. 이 휴리스틱의 적용은 시소러스 도구의 도움 없이는 불가능하다. 즉, 시소러스 도구는 가중치 결정 휴리스틱의 성공에도 주요한 역할을 담당한다.

확장된 시소러스 도구의 실험은 문서 분류기 TAXON의 성능 개선에 많은 도움을 주었음을 보여주었다. 따라서, 제안한 시소러스 도구는 문서 분류의 영역 뿐 아니라 정보검색 분야와 한국어 처리의 영역에 이용 가치가 있다. 가중치 결정 기술은 더 개선되어야 할 필요가 있다. 이를 위해서 시소러스 도구를 이용한 의미 애매성 해소 또는 의미 해석 등의 언어학적 분석이 선행되어야 할 것이다.

참 고 문 헌

[1] Masand. G Linoff, and D Waltz, "Classifying News Stories using Memory Based Reasoning," In *Proc Intl. Conf. on Research and Development in Information Retrieval, ACM SIGIR*, pages 59-65, 1992

[2] M E Maron, "Automatic Indexing An Experimental Inquiry." Journal of the ACM, 8:404-417, 1961.

[3] K M. Wong and Y. Y. Yao. "A Statistical Similarity Measure," In *Proc Intl. Conf on Research and Development in Information Retrieval, ACM SIGIR*, pages 3-12, 1987.

[4] K M. Wong and W Ziarko, V V Raghavan, and P C. N Wong. "On Extending the Vector Space Model for Boolean Query Processing," In *Proc. Intl Conf. on Research and Development in Information Retrieval, ACM SIGIR*, pages 175-185, 1986

[5] Salton, A Wong, and C S Yang, "A Vector Space Model for Automatic Indexing," CACM. 18(11) 613-620, 1975

[6] D D Lewis. Representation and Learning in Information Retrieval, Ph D Thesis, Computer Science Dept., Univ of Massachusetts, Amherest, 1992, MA 01003

[7] 권오욱, 확률벡터와 메타범주를 이용한 최적 문서 범주화 모델, 석사학위논문, 한국과학기술원 전산학과, 1995

[8] 한국과학기술원, 대한민국국어정보베이스 한국어 품사 부착 코퍼스, 한국과학기술원 & 자연어 정보처리 연구부, 1997

[9] 한국전자통신연구원 자연어처리연구실, ETRIKEMONG SET, 한국전자통신연구원, 1997

[10] 이현아, 이종혁, 이근배, "구문분석과 공기정보를 이용한 개념기반 명사구 색인 방법," 제 7 회 한글 및 한국어 정보처리학술대회 논문집, 1995

[11] 강원석, 강힌규, 김영섭, "개념기반 문서분류기 TAXON의 설계 및 구현," 한국정보과학회 '97 가을학술발표논문집(2), 24 권 2 호, 1997

[12] 님세진, 이지연, 신동욱, 최미옥, "복합명사의 통계적 처리에 대한 평가," 8 회 한글및한국어정보처리학술발표논문집, 1996.

[13] 박수준, 이현아, 장명길, 박제욱, 박동인, "효율적인 색인을 위한 복합 명사의 분해." 8 회 한글및한국이정보처리학술발표논문집, 1996

[14] 김광해, "{의}의 의미", 문법연구 제 5 집, 1984

[15] I.A Melcuk, Dependency Syntax Theory and Practice. State University of New York Press, 1988.