확률 벡터를 사용한 전자 문서의 개념적 분류 기법*

조 완 섭 · 김 영 렬
(충북대학교 경영정보학과)
강 원 석
(안동대학교 컴퓨터교육과)
강 현 규
(한국전자통신연구원)

요 약

본 논문에서는 전자문서의 개념적 분류기법을 제안한다. 기존의 문서분류는 대부분 문서에 나타난 용어를 기반으로 분류하므로 개념적인 분류가 불가능하다. 제안된 기법에서는 한국어 시소러스를 사용하여 문서에 나타난 용어 뿐 아니라 용어의 상하위 개념을 기준으로 문서를 분류할 수 있다. 특히, 제안된 방법은 확률 벡터를 사용하는 방식으로써 점진적인 학습이 가능하다는 장점도 가진다.

1. 개 요

문서 분류(text categorization)란 문서들을 미리 정해진 주제어(subject) 혹은 범주(category)로 분류하는 작업이다[1,6]. 예를들어, 도서관에서는 신착 도서의 초록을 분석하여 특정 주제(들)로 분류하거나, 디지털 신문사에서는 기사를 분석하여 특정 주제(들)로 분류하는 것이 그 예이다. 폭주하는 문서를 주제어에 따라서 분류하여 컴퓨터에 저장함으로써 문서들의 용이한 관리와 신속한 검색의 효과를 얻는 것이 문서 분류의 목적이다.

문서 분류 기법은 크게 수동 방식(manual categorization) 과 자동 방식(automatic categorization) 방식으로 나누어지며, 최근에는 컴퓨터의 도움으로 자동 문서 분류에 관한 연구가 활발하다[6,7]. 문서를 사람이 분류하는 수동 방식의 경우, 문서에 나타난 구체적인 용어뿐 아니라 그로부터 유추되는 개념을 이용한 문서 분류가 가능하다는 장점이 있지만 시간과 비용이 많이 들며, 주관이 개입되므로 일관성 유지가 어렵다는 단점이 있다. 반면에 소프트웨어 프로그램을 이용하여 문서를 분류하는 자동 방식의 경

* 본 논문은 한국전자통신연구원의 97년도 위탁과제연구비의 지원으로 연구되었음.

우, 수동 방식의 문제점을 해결할 수 있지만 문서에 나타난 용어 만을 기준으로 문서가 분류되므로 개념적인 문서 분류가 어렵다는 단점을 가진다.

자동 문서 분류 기법은 규칙 기반 방식, 문서-문서 관련도를 이용한 방식, 확률 기반 방식으로 나누어진다[6,7]. 규칙기반 방식은 규칙의 작성이 어렵고 점진적 학습이 불가능한 방식이며, 문서-문서 관련도 기반 방식은 관련도 분석의 오버헤드가 확률 기반 방식에 비하여 크므로 본 연구에서는 이들 중 확률 기반의 문서 분류 기법을 채택한다. 확률 기반 문서 분류 기법은 문서와 범주를 용어들의 출현 빈도를 나타내는 용어-확률 벡터로 표시한 다음 문서-문서 관련도 측정 함수를 이용하여 관련 정도를 측정하여 문서를 분류하는 방식이다[3,4,6,7]. 이 방식은 특히, 시스템이 구축된 후 운용 중에 발생하는 분류 노하우를 시스템에 끊임없이 반영함으로써 점진적 학습이 가능하다는 장점을 가진다.

본 논문에서는 확률 기반의 문서 분류 방식에 시소러스 정보를 이용하여 개념적 문서 분류가 가능하도록 하는데 중점을 둔다. 시소러스는 용어와 그에 대한 상위/하위 개념들로 구성된 집합이며, 이를 이용하면 문서에 나타난 용어 뿐 아니라 그 용어의 상위/하위 개념을 기준으로 문서를 분류할 수 있게 된다. 이를 위하여 용어-확률벡터에 시소러스를 이용한 상위 개념들을 포함시켜 확장한 개념-확률벡터로 문서와 범주를 표현하고 이를 이용하여 관련도를 분석하는 방법을 제시한다.

논문의 구조는 다음과 같다. 제 2 장에서는 전자 문서의 분류 기법에 관한 관련 연구를 소개한다. 제 3 장에서는 기존 전자 문서 분류의 문제점을 제기한다. 제 4 장에서는 시소러스를 이용한 개념적 문서 분류 기법을 제안한다. 제 5 장에서는 결론을 맺는다.

II. 문서 분류의 관련 연구

본 절에서는 기존의 자동 문서 분류 기법과 그들의 특성과 장단점을 분석한다. 기존의 문서 분류 기법은 세가지로 구분할 수 있다[7]: 규칙기반 방식, 확률 기반 방식, 문서-분서 관련도 기반 방식. 여기서는 이들 방식을 차례로 살펴본다.

2.1 규칙 기반 방식

규칙 기반 방식[8]은 "if (패턴 혹은 용어) then 주제어" 형태의 규칙들로 구성되는

전문가 시스템(expert system)이다. 규칙들은 전문가가 문서를 분류하는데 사용하는 지식이다. 카네기 그룹에서 만든 CONSTURE 시스템이 이 부류에 속한다. 이 제품은 로이터 통신사의 기사 분류 시스템으로 사용되어 90 % 정도의 재현율(recall)과 정확률(precision)을 가지는 양호한 시스템으로 평가되었다. 그러나 단점으로서는 문서 분류 규칙을 사람이 작성하기 때문에 규칙 생성에 많은 시간이 소요되며, 일단 구축된 시스템에서 분류 규칙을 확장하는 것도 어렵다. 그리고, 문서 분류 시스템이 완성된 후 문서를 분류하는 과정에서 얻어지는 노하우를 시스템에 피드백시켜 점진적 학습을 하는 것도 불가능하다. 또한, 규칙에 나타나는 패턴도 문서의 작성에 사용된 자연어에 의존한다는 점도 단점으로 지적되고 있다.

2.2 확률 기반 방식

이 방식은 훈련 문서 집합으로부터 카테고리에 대한 자질 (features)을 추출한 다음, 베이지언 확률(Bayesian probability)을 이용하여 문서-카테고리 관련도를 측정하는 방식이다. 여기서, 베이지언 확률식은 문서 Dm 이 카테고리 Cj 에 할당될 확률인 조건부 확률식 P(Cj | Dm) 으로 정의된다. 이 조건부 확률을 계산하는 방법으로 M. E. Maron [2]과 D. D. Lewis[6] 등은 베이지안 확률을 이용하였다.

한편, S. K. M. Wong [4]은 문서와 카테고리를 그에 포함된 용어들의 출현 빈도를 나타내는 확률벡터(이하, 용어-확률벡터 라고 부름)로 표시한 후 관련성 측정 함수를 사용하여 관련도가 최대인 카테고리로 문서를 분류하는 확률벡터 모델을 제안하였다. 확률벡터 모델은 정보 검색 분야에서 제안된 벡터공간 모델[4]의 제한점인 용어간의 독립성 가정을 배제하기 위하여 제안되었다[3]. 확률벡터 모델에서는 카테고리의 표현도 하나의 문서 표현과 동일한 용어-확률 벡터이므로 훈련 문서 뿐 아니라 실제 문서 분류 과정에서 얻어지는 노하우를 카테고리의 자질로 계속 추가 (삭제)할 수 있으며, 따라서 시스템에 대한 점진적인 학습이 가능한 방식이다.

2.3 문서-문서 관련도 기반 방식

이미 분류된 문서들과 새로운 문서 사이의 관련도에 의하여 새로운 문서를 분류하는 방법이다. 즉, 새로운 문서와 각 훈련 문서와의 관련도를 측정한 후 가장 높은 관련도를 갖는 훈련 문서(들)가 속한 카테고리(들)로 새로운 문서를 분류한다. MBR(Memory

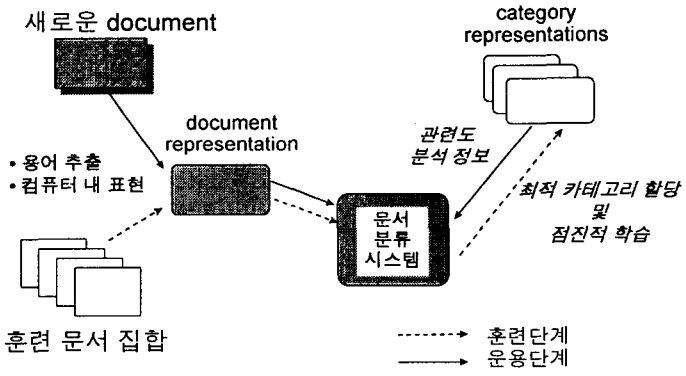
Based Reasoning) 모델[1]과 ExpNet[9] 모델이 이 방식에 속한다. 이 방식의 단점은 문서를 분류할 때 훈련 집합의 모든 문서에 대하여 문서-문서 관련도를 분석하므로 문서-카테고리의 관련성을 측정하는 확률-벡터 방식보다 관련도 측정 오버헤드가 커진다는 점이다.

III. 문제 정의

문서 분류의 과정은 그림 1과 같이 훈련 단계와 운용 단계로 구분하여 설명될 수 있다. 먼저 훈련 단계에서는 훈련 문서집합으로부터 문서를 가져와서 문서 분류 시스템의 처리에 적합한 형태로 표현한 문서 표현 (document representation) 을 생성하고, 문서 분류 시스템에 의해 적절한 개수의 카테고리를 생성한다. 훈련 문서 집합이란 카테고리를 준비하는데 사용되는 문서들의 집합이다. 예를들어, 신문사에서는 지난 1년치의 기사를 훈련 집합으로 하여 적절한 개수의 카테고리를 생성할 수 있다. 생성된 카테고리는 문서 분류 시스템 내에서 문서 분류에 적합하도록 표현되어야 하며, 이를 카테고리 표현(category representation) 이라고 한다. 일단, 훈련 문서 집합으로부터 카테고리 표현이 완성되면 새로운 문서에 대하여 기존의 카테고리중 어느것과 가장 관련도가 높은가를 측정하여 문서 분류를 할 수 있다.

문서 분류의 과정에서 다루어져야 할 중요한 작업은 다음 네가지로 요약된다.

1) 문서로부터 용어를 추출하는 작업
2) 훈련 문서에 대한 category 분류



<그림 1 > 문서 분류의 과정

3) 문서와 category의 문서 분류 시스템내 표현

4) 기존의 category와 새로운 문서와의 관련도 측정 및 분류

첫번째 작업은 형태소 분석기에서 담당하므로 여기서는 생략한다. 두번째 작업은 네번째 작업은 동일하다. 왜냐하면 훈련 문서의 경우 기존의 카테고리가 없는 상태에서 훈련 문서가 입력되므로 필요에 따라 적절한 카테고리를 하나씩 만들어가면서 문서를 분류한다는 차이는 있으나 기본적으로 네번째 작업과 동일하게 처리될 수 있기 때문이다. 세번째 작업은 관련도 측정에 적합하도록 문서와 카테고리를 문서 분류 시스템 내에 표현하는 방법을 의미한다. 예를들어, 이미지 형태로 스캔한 문서 자체는 관련도 측정에 적합하지 않으며, 대부분의 경우 문서 (카테고리)에 포함된 용어들의 벡터 형태로 문서를 표현한다.

본 논문에서는 첫번째 작업을 제외한 나머지 작업에 관하여 기술한다. 즉, (훈련) 문서를 문서 분류 시스템에서 사용하기에 적합하도록 표현하는 방법과 문서와 카테고리 사이의 관련도 측정 방법을 중심으로 살펴본다.

Ⅳ. 문서 분류 모델

여기서는 본 연구에서 사용하는 문서 분류 모델을 설명한 다음에 향후 구현될 시스템의 구조를 살펴본다. 특히, 시스템 내에서 개념적 문서 분류를 위하여 사용하고 있는 시소러스 정보의 이용 방안을 살펴본다.

4.1 확률 벡터 모델

본 연구에서는 참고문헌 [3]에서 제안된 확률 벡터 모델을 사용하여 문서와 카테고리를 표현하고, 그들 간의 관련도를 측정한다. 이 모델에서는 다음 세 단계를 거쳐서 문서를 분류한다.

(1) 문서 D는 다음과 같은 용어-확률벡터로 표시한다.

$$D = (wd1, wd2, \ldots, wdn)$$

여기서 wdi는 문서 D에 나타난 i-번째 용어의 가중치 (예를들어, 출현비율)를 나타낸. 따라서, wdi , i=1,2,..., n는 0과 1 사이의 값이며, $\sum_{i=1}^{n} wdi = 1$이 된다.

(2) 유사하게 카테고리 C도 다음과 같이 확률 벡터로 표시된다.

C = (wc1, wc2, ... , wcn)

여기서 wci는 카테고리 C로 분류된 문서 집합에서 나타난 i-번째 용어 wci 의 가중치 (예를들어, 출현비율)를 나타낸다. 따라서, wci , i=1,2,..., n는 0과 1 사이의 값이며, $\sum_{i=1}^{n}$ wci = 1이다.

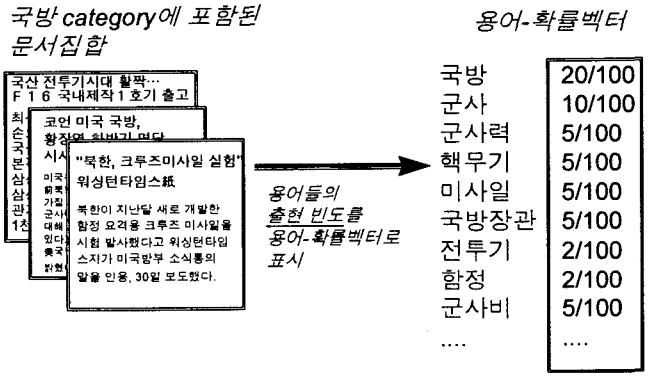
(3) 문서 D=(wd1, wd2, ..., wdn)와 카테고리 C= (wc1, wc2, ..., wcn)의 관련도는 다음과 같이 정의되는 관련도 측정 함수 SIM(D, C)을 이용하여 측정한다.

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

단, 확률벡터 P = (w1, w2, ... , wn) 에 대하여 H(P)는 P의 불확실성 정도를 나타내는 엔트로피(entropy) 로써 H(P) = $-\sum_{i}$ wi × log2 wi로 계산된다[3].

관련도 측정 함수 SIM(D, C)은 두 확률 벡터가 완전히 동일할 때 최대값 1을 가지며, 완전히 다르게 분포할 때 최소치 0를 가진다.

[예제 1] 그림 2는 카테고리에 대한 용어-확률벡터를 보여준다. 그림 2의 카테고리에서 용어의 총 출현 회수 (중복 포함)는 100이며, 용어 '국방'은 그 중 20회, 용어 '군사'는 10회 나타남을 보여주고 있다. 한편, 하나의 문서에 대한 용어-확률 벡터는 카테고리 내에 문서가 하나만 포함된 경우에 해당하므로 생략한다.



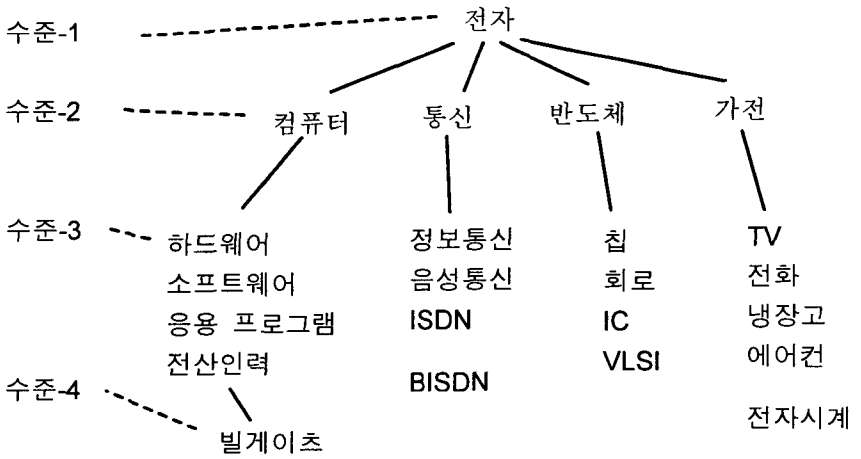
< 그림 2 > 카테고리 '국방'에 대한 용어-확률벡터

확률벡터를 이용한 문서 분류 방식에서는 점진적 학습이 가능하지만 개념적인 문서 분류는 불가능하다. 예를들어, 국방 카테고리에 새로운 (기존) 문서 D'이 할당(삭제)된

경우 시스템은 D'에 포함된 용어의 가중치를 국방 카테고리의 용어-확률 벡터에 첨가(삭제) 함으로써 점진적 학습을 실시한다. 이렇게 하면 분류된 문서가 많아질수록 카테고리를 나타내는 확률 벡터의 질(quality)이 더 좋아진다고 볼 수 있다. 그런데, 그림 2에서 보는 바와 같이 용어-확률 벡터가 문서들에 포함된 순수한 용어들의 출현빈도에 의존하므로 개념적인 문서 분류는 여전히 불가능하다.

4.2 시소러스 정보를 이용한 개념적 문서 분류

본 절에서는 확률-벡터를 사용한 문서 분류 방식에서 개념적 문서 분류가 가능하도록 하는 방안으로 시소러스 정보를 이용한 문서 분류 시스템을 제안한다. 시소러스란 주어진 용어에 대한 상하위 개념들의 집합으로 정의될 수 있다. 그림 3은 시소러스의 예를 보여준다.



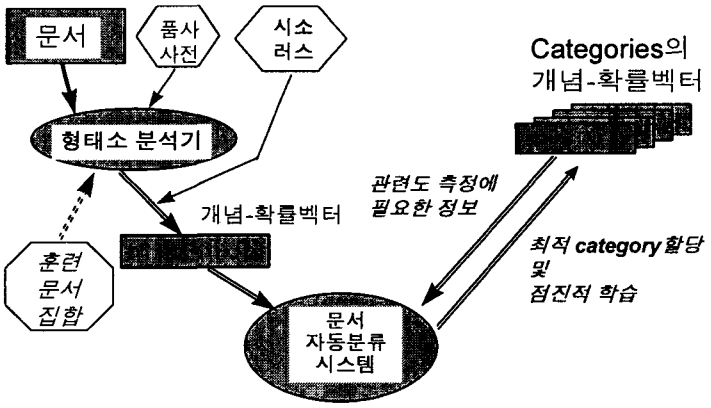
<그림 3> 시소러스의 예

제안된 문서분류기법의 핵심 아이디어는 문서에 대한 용어-확률벡터 대신에 시소러스 정보를 이용하여 용어-확률벡터에 포함된 용어들을 상위의 개념들로 대치한 '개념-확률벡터'를 사용한다는 점이다. 즉, 개념-확률벡터는 용어-확률벡터를 상위의 개념으로 재표현한 것이다. 예를들어, '빌게이츠'가 나타난 문서는 그림 3의 시소러스를 참고

하여 그 문서가 '전산인력'이나 '컴퓨터' 혹은 '전자'를 포함하도록 개념-확률벡터를 생성한다. 다음으로 확장된 개념-확률벡터에 대하여 제 3.1 절에서 소개한 관련도 측정 함수 SIM( )을 사용하여 관련도를 측정한 후, 문서를 최적의 카테고리로 분류하게 된다. 따라서, '빌게이츠'라는 용어가 나타난 문서도 시소러스를 이용하여 '전자'나 '컴퓨터' 분야로 분류될 수 있으므로 문서의 개념적 분류가 가능하게 된다.

4.3 문서 분류시스템의 구조

제안된 문서 분류 시스템의 구조는 그림 4와 같다. 먼저, 훈련 문서 집합으로부터 문서 분류 시스템을 이용하여 카테고리 집합 (개념-확률벡터들로 구성됨)을 생성한다. 다음으로 새로운 문서가 입력되면 시소러스를 이용하여 개념-확률벡터로 표시한 후 기존의 카테고리와 관련도를 분석한다. 그리고, 관련도가 최대인 카테고리로 새로운 문서를 분류하고, 새로운 문서에 포함된 용어(개념)의 가중치를 카테고리의 개념-확률 벡터에 반영하여 조정한다.



< 그림 4 > 문서 분류 시스템의 구조

4.4 문서 분류 시스템의 구현

제안된 문서 분류 시스템의 프로토타입을 다음의 환경에서 구현하였다. 먼저, 훈련 문서 집합으로 21258 개로 구성된 계몽사 백과사전의 문서를 사용하였다. 그리고, 백과사전과 연관된 681개의 단어에 대하여 한국어 시소러스를 구성하였다. 현재 구현된 문

서분류 시스템의 정확도는 약 70% 정도이며, 앞으로 시소러스의 확장과 더욱 정확한 한국어 시소러스의 사용을 통하여 정확도가 더욱 향상될 것으로 예상한다. 주의할 점은, 정확도와는 별개의 문제로서 개념을 기반으로 한 문서 분류가 가능해진다는 장점을 가진다는 점이다. 즉,문서에 나타난 '빌게이츠'라는 용어로부터 그 문서를 '전산인력'이나 '컴퓨터' 분야의 문서로 분류할 수 있도록 시도하였다는 점이다.

V. 결 론

본 논문에서는 한국어 시소러스를 이용한 전자 문서의 개념적 분류 기법을 제안하였다. 기존의 문서 분류 기법들은 문서에 나타난 용어들의 관련성 기반으로 문서를 분류함으로써 개념적인 문서 분류가 불가능하였으나 제안된 방법은 한국어 시소러스를 이용하여 이러한 문제점을 해결하였다. 특히, 제안된 방법은 확률 벡터를 사용함으로써 점진적인 학습 효과를 얻을 수 있으며, 이는 문서 분류 작업을 오랫동안 진행함에 따라서 더 정교한 분류가 가능해질 수 있음을 의미한다.

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] S. 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] S. 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] G. 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] P. J. Hayes, "Intellignet High-Volume Text Processing Using Shallow, Domain-Specific Technique," In Paul S. Jacobs, editor, Text-Based Intelligent Systems: Currnet Research and Practice in Information Extraction and Retrieval, pages 227-241, Hillsdale, New Jersey, 1992.

[9] Y. Yang, "Expert Network: Effective and Efficient Learning from Human Decision in Text Categorization and Retrieval," In Proc. Intl. Conf. on Research and Development in Information Retrieval, ACM SIGIR, pages 13-22, 1994.