의미의 상하위 정보를 이용한 웹문서 분류시스템

*강 원석°, **황도삼, ***최기선
*안동대학교 컴퓨터교육과, **영남대학교 컴퓨터공학과, ***한국과학기술원 전산학과
wskang@andong.ac.kr, dshwang@ynucc.yeungnam.ac.kr, kschoi@koterm.kaist.ac.kr

A Web-Document Categorization System Using the Hierarchical Information of the Concept ♣

*Wonseog Kang, **DoSam Hwang, and ***KeySun Choi
*Dept. of Computer Education, Andong National University
**Dept. of Computer Engineering, Yeungnam University, and
***Dept. of Computer Science, KAIST

요약

본 논문에서는 다양성을 가진 웹문서의 범주를 결정짓는 웹문서 분류 시스템을 설계, 구축한다. 웹문서는 일관된 형식과 내용이 없이 만들어지기 때문에 문서의 범주를 결정하는 시스템을 구축하기는 쉬운 일이 아니다. 제안한 웹문서 분류 시스템은 잡음 처리에 적합한 신경망 방식을 적용하여 다양한 내용의 웹문서의 범주를 결정짓는다. 본 시스템은 한국어 문장을 분석하는 한국어 형태소 해석기, 단어의 의미를 획득하는 개념 획득기, 단어의 사용된 의미를 고르는 애매성 해소기, 그리고 문서의 범주를 결정하는 신경망 범주 결정기로 구성된다. 본 시스템은 단어의 의미를 이용하여 문서를 표현하고 분석하는 개념 중심의 문서 분류 시스템이다.

1. 서론

정보 사회에서는 수많은 정보가 생성되고 소멸된다. 많고 다양한 정보 가운데 필요로 하는 정보를 찾기란 쉬운 일이 아니다. 문서 분류 시스템은 문서의 범주를 결정해주는 시스템으로 필요로 하는 정보 검색의 유용한 도구이다. 문서 분류 시스템에 대한 많은 연구[1,4,5,7]가 있으나 문서 분류 시스템의 대상이 되는 문서가 잡음이 많은 웹문서 분류 시스템에 대한 연구가 미흡하다.

웹문서는 웹문서 기술 양식에 따라 표현되나 그 내용은 천차만별이므로 웹문서의 범주를 결정하는 것은 어려운 일이다. 용어를 기반으로 한 문서 분류의 경우 용어 출현이 적은 웹문서의 분류는 더욱 어렵다. 본 논문에서는 이와 같은 웹 문서에 대해 문서의 범주를 결정하는 문서 분류 시스템을 설계, 구축한다. 다양한 분야의 문서를 효과적으로 분류하기 위해 학습 기능이 뛰어난 신경망 방식을 선택하였고 개념 중심의 방법을 적용하여 용어 중심의 방법이 지니는 문제점을 해결한다.

기존의 문서 분류는 규칙 방식, 확률벡터 방식, 관련도 비교 방식 등을 채택하였다[7]. 규칙 기반 방식은 문서의 정보의 패턴을 검사하는 규칙들로 문서를 분류하고자 하는 방법이다. 많은 시스템이 이 방법을 적용하나 웹문서 분류에는 적합하지 않다. 다양한 분야의 모든 문서에 대한 분류 규칙을 설계한다는 것은 많은 시간과 노력이 필요하다. 또한 어렵게 구축한 시스

♣ 본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

템을 확장하는 것은 더 어렵다. 확률 벡터 방식[1]은 문서에 나타나는 용어들의 출현 분포를 이용하여 문서를 분류하는 방식이다. 이 모델에서는 웹 문서 분류에서 얻어지는 용어들의 출현 분포도를 시스템에 추가, 반영하여 점진적 학습을 한다. 그렇지만 분류에서 얻어지는 분포도가 전체 분포도에 산술적으로 추가되기만 할 뿐 용어들의 상관관계에 대한 학습 효과가 부족하다. 문서-문서 관련도 기반 방식[5]은 이미 분류된 문서들과 새로운 문서 사이의 관련도에 의하여 새로운 문서를 분류하는 방법이다. 이 방식의 단점은 문서를 분류할 때 훈련 집합의 모든 문서에 대하여 문서-문서 관련도를 분석하므로 관련도 측정 오버헤드가 커다는 점이다.

본 논문은 각 용어들의 상관 관계에 대한 학습 효과가 있고 관련도 분석 오버헤드가 적은 신경망 방식의 웹문서 분류시스템(Web-Document Categorization System: WeDCaS)을 설계, 구축한다. **WeDCaS** 는 신경망 시스템의 입력으로 개념-벡터[7]를 사용한다. 개념-벡터는 문서 단어의 개념을 획득하여 벡터로 표현한 것이다. 개념-벡터를 사용함으로 개념 중심의 문서 분류를 꾀할 수 있을 뿐 아니라 신경망 시스템의 처리 속도와 공간의 효율성을 얻을 수 있다. **WeDCaS** 는 문서의 문장에서 단어를 분리하는 한국어 형태소 해석기, 단어의 개념을 획득하는 개념 획득기, 단어의 사용된 의미를 선택하는 애매성 해소기, 문서의 범주를 결정하는 신경망 범주 결정기로 구성된다. 본 논문의 2 장에 웹문서 분류기 **WeDCaS** 에 대해 기술하고 3 장은 실험과 결과를 분석하고 4 장에서 결론을 맺는다.

2. 웹문서 분류 시스템

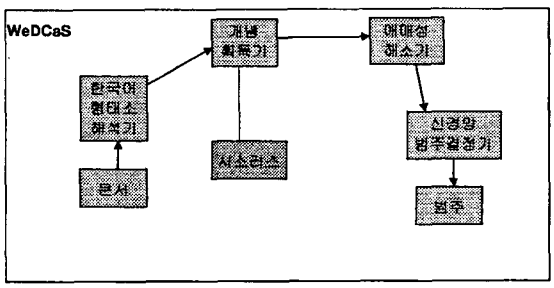

그림 1 웹문서 분류 시스템 **WeDCaS**

웹문서 분류시스템 **WeDCaS** 는 그림 1 과 같은 구조를 가진다. **WeDCaS** 는 한국어 형태소 해석기, 개념 획득기, 애매성 해소기, 신경망 범주 결정기로 구성된다. 한국어 형태소 해석기는 입력된 문서를 형태소 해석하여 명사 단어를 골라낸다. 개념 획득기는 선택된 단어가 가지고 있는 개념을 시소러스를 통해 획득한다. 애매성 해소기는 단어와 획득한 개념을 토대로 단어의 애매성을 해소한다. 그리고 신경망 범주 결정기는 선별된 개념들의 벡터값을 입력으로 문서의 범주를 결정한다.

2.1 한국어 형태소 해석기

문서 분류를 하기 위해서 먼저 한국어로 된 문서를 형태소 해석해야 한다. 형태소 해석은 접사 처리, 복합어 처리 등이 고려된 [7]의 한국어 형태소 해석기를 사용하였다. 이 형태소 해석기는 문서 분류에 중요한 역할을 하는 명사의 분석에 초점을 맞추었고 순수 명사 이외에 조용사 '하다'가 붙는 용언의 분석도 포함하였다.

2.2 개념 획득기

개념 획득기는 용어 중심의 해석방법에서 탈피하여 개념 중심의 해석 방법을 가능하게 하는 역할을 한다. 따라서 개념 중심의 문서 분류의 성능은 개념 획득기의 성능에 달려 있다. 좋은 개념 획득기를 개발한다는 것은 쉬운 일이 아니다. 본 논문에서는 10000 단어에 대해 개념을 획득할 수 있는 개념 획득기를 설계, 구축하였다. 이 도구는 [7]을 기초로 확장, 개선된 것이다.

개념 획득기는 시소러스를 사용한다. 시소러스는 문서 분류에 중요한 명사에 대해 개념이 정의되어 있다. 명사의 개념도 행위의 개념 등이 포함되므로 개념 분류는 크게 물체, 사건, 속성의 세 갈래로 구성된다. 그 분류는 [8]을 확장하여 설계하였다.

개념 획득기의 수행 결과의 예는 표 1 과 같다. 개념 획득기는 단어에 대해 해당하는 개념을 출력한다. 단어가 하나 이상의 개념을 가지면 개념 획득기는 단어가 가질 수 있는 모든 뜻에 대한 개념을 추출한다. 다음 단계의 애매성 해소기는 단어의 개념 가운데 사용된 의미의 개념을 선별한다.

표 1. 개념 획득기의 실행 예

입력 단어	단어 개념
가감	event change quantity-change
가감법	intellectual-thing abstract-thing thing means mathemathics 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
.....	.....

2.3 애매성 해소기

문서의 단어는 여러 가지 의미를 가진다. 그 문서에서 사용된 의미를 정확히 찾기 위해서는 문맥의 뜻을 분석하지 않으면 안 된다. 그러나 현재의 기술로 문맥 분석을 하기가 어렵다. 그래서 통계 정보, 공기 정보, 지식 정보 등을 사용한 애매성 해소의 방법이 나오고 있다[6].

[8]의 연구에서 가중치 결정기를 제안하였다. [8]에서는 단어가 가지고 있는 의미들을 비교하여 특정 단어의 가중치를 결정하는 것으로 단어의 의미를 선별하는 애매성 해소의 관점이 아니다. 본 논문에서는 한 단어가 가지고 있는 의미들 가운데 어떤 의미가 사용되는지를 선택하는 애매성 해소를 도모한 것이다. 이 단계의 첫 과정은 다음과 같다.

1. 단어의 개념들이 개념 트리의 일직선상에 있는가를 검사한다. 그렇다면 이 개념들을 한 집합으로 간주한다. 이것은 그 개념 이외에 다른 뜻이 없으므로 애매성이 존재하지 않는 것이다. 이 경우는 애매성 해소 시스템을 통과한다.
2. 단어의 개념들이 일직선상에 놓이지 않는다면 일직선상에 놓이는 개념들을 한 집합으로 간주하여 여러 개의 집합을 구분한다. 여러 집합이 생기면 집합 중 어느 집합에 대한 의미가 적합한지를 결정한다. 결정은 애매성 해소 시스템을 따른다.

애매성 해소 시스템은 신경망 시스템으로 입력으로는 다중 의미가 그대로 표현된 단어 개념 벡터, 위치 벡터, 문장개념 벡터가 들어오고 출력으로는 선별된 개념들이 나타난다. 입력의 문장 개념 벡터는 문장에서 나타난 단어들의 가능한 개념 모두가 표시된 것이고 위치 벡터는 애매성을 해소하고자 하는 단어가 문장에서 어느 위치에 있는가에 대한 정보이다. 위치 벡터는 구문 정보로서 지역 정보를 나타내고 있다. 이 벡터의 크기는 4로 하였다. 문장 개념 벡터는 문장이 가지고 있는 개념들을 모두 표현한 것으로 구문정보가 제외된 의미정보를 나타낸다. 이 벡터의 크기는 150으로 하였다. 출력 벡터는 선별된 단어의 개념 벡터로 입력 단어 개념 벡터에서 해당하는 의미만 표시된 것이다. 결국 시스템은 입력 단어의 의미들과 문장이 가지고 있는 의미들, 그리고 단어의 위치 정보 등을 이용하여 단어의 의미 선별을 한다. 시스템의 중간층의 크기는 200으로 하고 학습 알고리즘은 [10]의 역전파 알고리즘을 이용한다.

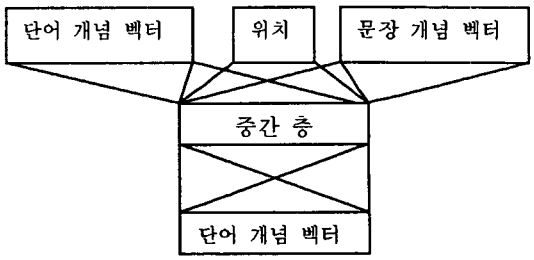

그림 2. 애매성 해소 신경망 시스템

2.4 신경망 범주 결정기

애매성 해소기에 의해 해소된 단어의 개념들은 다시 문서 개념-벡터로 표현되고 이 개념-벡터는 신경망 범주 결정기의 입력으로 사용된다. 학습 알고리즘은 애매성 해소기와 같이 [10]의 역전파 알고리즘을 이용한다. 문서 개념 벡터의 크기는 애매성 해소기와 같은 150으로 하였고 범주 벡터는 12와 76으로 하였다.

신경망 범주 결정기는 미리 훈련 문서 집합과 범주 벡터로 학습을 하고 다음으로 새로운 문서가 입력되면 한국어 형태소 해석기, 개념 획득기, 애매성 해소기를 이용하여 문서의 개념-확률벡터로 표시한 후 이를 입력으로 문서의 범주를 결정한다.

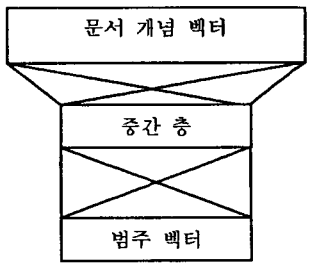

그림 3. 신경망 범주 결정기

3. 시스템의 실험과 결과

WeDCaS는 웹 문서를 대상으로 한다. 학습문서는 980문서이다. 따로 100문서를 검사 문서로 삼았다. 검사 문서를 검사한 결과 67.2%(73.1%)의 성공률을 얻었다. 이 성공률은 애매성 해소 단계가 포함된 시스템 전체의 성공률이다. 이 성공률은 아직 학습 문서가 적은 관계로 객관성이 부족하다. 앞으로 학습 문서의 수를 늘여 객관성 있는 결과를 얻을 것이다.

성공하지 못한 경우에 대해 분석한 결과 다음과 같은 사실을 알 수 있었다.

(1) 애매성 해소기로 인한 실패 : 단어가 가지고 있는 중의성을 해소하기 위하여 애매성 해소기를 이용한다. 그러나 애매성 해소기가 바른 결과를 내지 못한 경우 그 결과에 의해 올바른 범주 결정에 실패하였다. 이를 해결하기 위해서는 애매성 해소기의 성능 개선이 필요하다.

(2) 학습량의 부족 : 실패한 검사 문서의 유형에 대한 학습이 이루어지지 않아 시스템이 범주 결정에 성공하지 못하였다. 학습량을 늘임으로 이 문제를 해결할 수 있다.

(3) 개념 획득기의 부정확성 : 문서의 범주를 결정하기 위한 조건으로 개념 획득기는 단어의 개념을 획득하나 그 부정확성으로 실패가 발생하였다. 이를 해결하기 위해 개념 획득기의 개선이 필요하다.

(4) 개념 획득기의 비상세성 : 개념 획득기의 개발 여건상 단어의 개념을 구체적으로 표기할 수 있을 정도의 개념 분류가 부족하다. 이로 인한 실패가 일어났다.

4. 결론

본 논문은 웹문서 분류기 **WeDCaS**를 설계 및 구축하였다. 다양한 내용의 웹문서를 분류하기 위하여 학습 기능이 뛰어난 신경망 모델을 적용하였다. 시스템을 구축하기 위하여 상하위 정보의 시소러스를 구축하였고 이 시소러스를 이용한 개념 획득기를 구축, 개선하였다. 개념 획득기는 기존의 연구를 토대로 성능의 개선을 도모하였다.

그리고 문서 분류의 질을 향상하기 위하여 애매성 해소기를 포함하였다. 실험의 결과는 학습 문서의 수를 늘일 것과 애매성 해소기의 성능을 개선할 것을 제시하고 있다. 그리고 개념 획득기의 성능 개선을 요하고 있다. 차후 학습량의 증대와 애매성 해소기의 성능개선을 도모할 계획이다.

참 고 문 헌

[1] 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.

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

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

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

[5] 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.

[6] N. Ide and J. Veronis, "Word Sense Disambiguation: The State of the Art," Computational Linguistics, Vol. 24, No. 1, 1998.

[7] 강원석, 강현규, 김영섬, "개념기반 문서분류기 TAXON의 설계 및 구현," 한국정보과학회 '97가을학술발표논문집(2), 24권 2호, 1997.

[8] 강원석, 강현규, 김영섬, "가중치 부여 휴리스틱을 이용한 개념기반 문서분류기 TAXON의 개선," 한국정보과학회 '87가을학술발표논문집(2), 25권 2호, 1998.

[9] 강원석, 강현규, "시소러스 도구를 이용한 실시간 개념기반 문서분류 시스템," 한국정보과학회 논문지, 26권 1호, 1999.

[10] D.E.Rumelhart, G.E.Hinton, and R.J.Williams, "Learning Internal Representations by Error Propagation," *Parallel Distributed Processing*, Vol.1, 1986.