정보산업 분야 시소러스의 공학적 구축 방안

류법모*, 김재호*, 최기선*, 성원경**
*전자전산학과 전산학전공, 한국과학기술원
전문용어언어공학연구센터/언어자원은행
**한국과학기술정보연구원

*{pmryu,jjaeh}@world.kaist.ac.kr, kschoi@cs.kaist.ac.kr,
**wksung@kisti.re.kr

Toward IT Domain Thesaurus: An Engineering Approach

*Pum-Mo Ryu, *Jae-Ho Kim, *Key-Sun Choi, **Brian W.K. Sung
*Computer Science Division, KAIST KORTERM/BOLA
**Korea Institute of Science and Technology Information

요 약

이 논문은 공학적인 접근 방법에 기반한 단계적인 전문분야 시소러스 구축 방법을 제안한다. 시소러스 구축 과정은 용어 추출 단계, 용어 분류 단계, 계층 구조 구축 의 3 단계로 구성되고, 모든 단계에서 자동 처리와 전문가 검증 작업을 거친다. 추출된 용어를 미리 정해진 분류 체계에 따라 분리한 후 여러 개의 작은 시소러스를 구축하고, 마지막으로 전체 시소러스로 결합한다. 이 방법은 1) 시소러스를 구축하는 복잡도가 줄어들고, 2) 클래스 단위의 작은 시소러스가 다른 전문분야 시소러스에 쉽게 재사용될 수 있으며, 3) 각 클래스에 포함된 용어들의 분포를 쉽게 판단할 수 있는 장점이 있다. 제안한 방법을 이용하여 한국어 정보기술 분야 시소러스를 구축하였다. 시소러스 구축에 사용된 용어들은 정보기술 분야의 최근의 한국어 신문과 특허 문서에서 추출하였기 때문에 한국에서 만들어진 신조어를 포함한다. 구축된 시소러스는 81 개의 상위 레벨 클래스와 1,000개 이상의 용어로 구성된다.

1 서론

시소러스는 주어진 분야에서 통제된 용어들의 집합을 말하며, 용어들 사이의 동의어 관계 및 계층 관계가 명확하게 표시된다. 시소러스는 정보 저장 또는 검색 시스템에서 문서 검색의 효율을 높이고, 문서의 색인에서 일관성을 주기 위하여 사용된다 [1].

실용적이고 대규모의 시소러스 구축 과정에는 많은 어려움이 있다 [2]. 첫 째, 시소러스를 구축하고 유지하기 위하여 전문가의 수작업이 과도하게 많이 필요하고, 둘 째, 시소러스가 포함하고 있는 지식은 지속적으로 변화하고 발전하고 있기 때문에 시소러스가 출판됨과 동시에 최신의 지식이 아닌 낡은 지식을 표현하는 경향이 있으며, 셋 째, 시소러스가 응용 분야에서 효율적으로 사용되기 위해서는 해당 전문 분야 지식을 포함하여야 한다. 전문 분야는 각각 자신들만의 독립된 전문용어를 사용하기 때문에, 일반 분야 시소러스를 수정하지 않고 그대로 전문 응용 분야에 사용하는 것은 적절하지 않다. 따라서 본 연구에서는 위의 문제를 해결하기 위한 새로운 시소러스 구축 방법을 제안한다. 첫 째, 전문 분야 말뭉치에서 해당 분야 용어를 추출한다. 이 과정에서 전문 분야 말뭉치에서 추출한 대부분의 용어는 해당 분야의 전문 용어이기 때문에 위의 세 번째 문제점을 해결할 수 있다. 둘 째, 추출한 용어를 전문 분야 지식 분류 체계의 클래스 단위로 분류하고, 분류된 용어를 대상으로 여러 개의 작은 시소러스를 구축한다. 전문 분야 지식 분류 체계는 작은 시소러스들은 하나로 다시 묶어주는 역할을 한다. 이 방법은 시소러스를 분할-정복 (divide and conquer) 방법으로 구축하기 때문에 시소러스 구축의 복잡도를 줄일 수 있는 장점이 있다. 특히 다른 전문 분야 시소러스에서 이미 구축한 시소러스의 일부분을 쉽게 재사용할 수 있는 장점이 있다. 따라서 분야 전문적인 시소러스를 빨리 구축할 수 있기 때문에 위의 두 번째 문제점을 완화시킬 수 있다. 셋 째, 자동 구축과 지침에 의한 수동 검증 작업을 시소러스 구축 과정의 모든 단계에서 순차적으로 수행하였다. 자동 구축과정에서 사용된 용어 추출 시스템, 용어 분류 시스템, 그리고 관계 추출 시스템에 의하여 수작업 비용을 줄일 수 있고 전문가의 검증을 통하여 자동 처리 방법의 오류를 수정할 수 있다. Cimiano [10] 는 전문가의 개입이 배제된 자동

시소러스 및 온톨로지 구축 방법의 신뢰성에 많은 의문을 제시하고 있다. 따라서 본 연구에서 제시한 자동 처리와 전문가 검수의 혼합된 접근 방법은 실용적으로 시소러스를 구축할 수 있는 시작 단계로 볼 수 있다.

이 논문의 구성은 다음과 같다. 2장에서 제안한 방법의 개요를 설명하고, 3장에서는 자동 용어 추출과 검수 지침에 대하여, 4장에서는 자동 용어 분류와 전문가 검증 방법을, 5장에서는 자동 계층관계 추출 방법과 전문가 검증 방법을 설명한다. 마지막으로 6장에서 결론을 맺는다.

2 시소러스 구축 방법

본 연구에서 제시한 방법은 그림 1과 같이 용어 추출 및 디스크립터 검증 단계, 용어 분류 단계, 계층 관계 구축 단계의 3 단계로 구성된다.

첫 번째 단계에서는 도메인 말뭉치에서 용어를 자동으로 추출하고, 미리 만들어진 지침에 의하여 추출된 용어의 디스크립터 여부를 결정한다. 용어 추출 시스템은 정확한 도메인 용어를 추출하기 위하여 기존의 분야 전문용어 사전에 포함된 용어 리스트, 영어-한국어 음차 표기 정보, 용어 빈도수, 시간에 따른 용어 사용 빈도수 변화 (Term Temporal Salience Value: TTSV)와 같은 통계정보를 사용한다.

두 번째 단계에서는 디스크립터로 분류된 용어를 미리 정해진 전문 분야 분류 체계의 각 클래스로 분류한다. 이 단계에서 용어 분류 시스템은 각 용어에 가장 가능성 있는 한 개 이상의 클래스를 할당한다. 용어 분류의 목적은 시소러스 구축 문제를 단순화시키고, 구축된 시소러스를 쉽게 재사용할 수 있게 하기 위한 것이다. 문제의 단순화 측면에서는, 대규모 시소러스를 한 번에 구축하는 것보다 클래스 단위의 작은 규모의 시소러스를 여러 개 구축하는 방법은 한 개의 용어에 대하여 관계를 고려해 주어야 하는 용어의 개수가 줄어들기 때문에 문제의 복잡도가 낮아지는 장점이 있다. 시소러스의 재사용 측면에서는 클래스 단위의 소규모 시소러스는 관련이 있는 다른 분야의 시소러스에 쉽게 재사용될 수 있는 장점이 있다. 예를 들어 '전자상거래' 클래스는 '정보기술' 분야와 '경제' 분야에도 함께 포함되기 때문에, '정보기술' 분야 시소러스에서 구축한 '전자상거래' 클래스 시소러스는 '경제' 분야 시소러스에서도 쉽게 재사용될 수 있다.

세 번째 단계에서, 계층관계 추정 시스템은 용어들 사이의 가능한 계층 관계를 제시하고, 제시된 관계들을 분야 전문가가 검증한다. 이 과정은 두 번째 단계에서 분류한 클래스 단위로 진행된다. 관계 추출 시스템은 수직 관계 기반 방법, 정의문 패턴 기반 방법, 참조 시소러스 기반 방법 그리고 통계 기반 방법을 사용한다. 분야 전문가는 자동 추출된 관계를 검증하고, 검증된 계층 관계에 지침에서 정의한 관계의 종류를 부가한다.

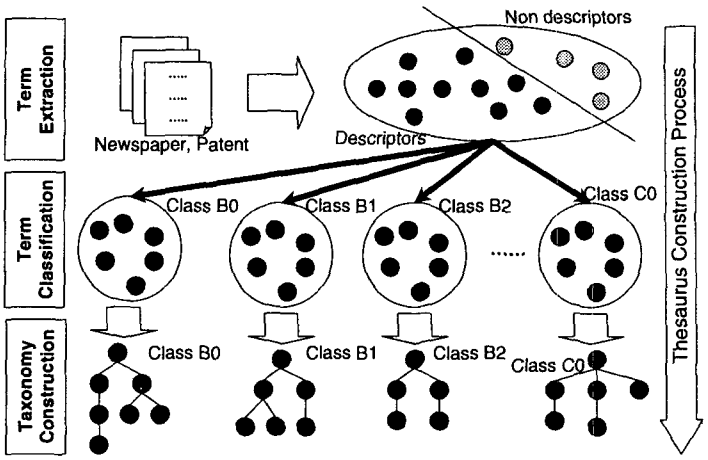

그림 1. 시소러스 구축 방법의 개요. 이 방법은 용어 추출 단계, 용어 분류 단계, 계층 관계 구축 단계로 구성된다.

3 용어 추출 및 검증

이 장에서는 정보기술 분야 한국어 시소러스를 구축하기 위한 첫 번째 절차를 설명한다. 용어를 추출하고, 스코프 노트 (scope note)를 용어 단위로 부가하며, 용어의 디스크립터 여부를 구분한다.

3.1 자동 용어 추출

이 절에서는 말뭉치에서 용어를 자동으로 추출하는 방법을 설명한다. 본 연구에서는 [3]에서 제안한 전문용어 추출기를 사용한다. 이 용어 추출기는 네 단계 과정으로 이루어진다. 첫 번째 단계에서는 클러스터링 기법에 의하여 사전간의 계층관계가 구축된다. 두 번째 단계에서는 명사구를 추출하며, 세 번째 단계에서는 추출된 명사구에 대하여 가중치를 부여한다. 사용한 가중치는 '사전에 의한 가중치', '통계에 의한 가중치', '음차표기를 이용한 가중치'이다. '사전에 의한 가중치 기법'은 해당 명사구가 나타난 전문용어 사전의 개수에 기반하여 가중치를 부여한다. 예를 들어 주어진 용어 후보 '멀티미디어 오브젝트' 는 기존의 사전 용어인 '오브젝트'에서 확장된 용어이기 때문에 이 후보에 높은 점수를 부여한다. '통계기법에 의한 가중치 기법'은 명사구의 출현빈도, 내포관계 등을 이용한다. 말뭉치에서 자주 등장하는 용어 후보는 해당 도메인의 대표적인 개념을 표현하기 때문에 높은 점수를 부여한다. '음차 표기를 이용한 가중치 기법'에 의해서 주어진 후보 명사구는 음차표기나 영어가 포함된 어절수에 의해 가중치가 정해진다. 한국어 전문용어는 외국어를 음차표기하거나 외국어를 그대로 사용하는 경우가 많기 때문에 영어-한국어 음차표기 정보를 이용하여 음차 표기된 용어 후보에 높은 점수를 부여한다. 네 번째 단계에서는 각각의 가중치 값을 하나의 값으로 통합하여 용어 후보를 순위화한 후 전문용어를 추출한다.

본 연구에서는 위에서 설명한 정보 이외에 추가적으로 용어의 연도별 사용 경향을 설명하는 TTSV를 이용한다. TTSV는 [12]에서 제안한 '용어지배지수'와 유사한 개념으로 말뭉치에서 연도별로 사용 빈도수가 지속적으로 증가하면 높은 점수를 가진다. 세 개의 용어에 대한 연도별 출현 빈도수가 그림 2에 나타나있다. '통신 서비스'는 빈도수가 증가하고 있기 때문에 TTSV가 높고, 나머지 용어들은 빈도수가 점차 낮아지거나 변화하지 않고 있기 때문에 TTSV가 상대적으로 낮다. TTSV는 용어의 최근 사용 경향을 반영하기 때문에 TTSV 정보를 이용하여 추출된 용어집합에는 최근에 활발하게 사용되는 용어나 신조어가 포함된다.

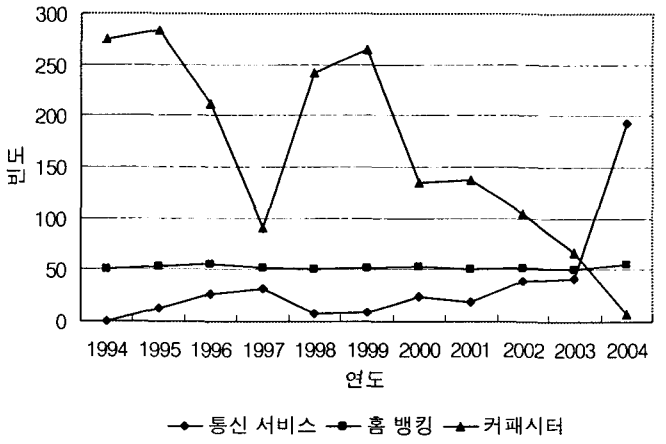

그림 2. '통신 서비스', '홈 뱅킹', '커패시터'에 대한 연도별 출현 빈도수

3.2 디스크립터 선택

디스크립터란 해당 분야의 개념을 표현하기 위하여 시소러스에서 선택된 용어를 말한다. 분야 전문가가 추출된 용어를 다음과 같은 기준을 이용하여 디스크립터인 용어와 그렇지 않은 용어로 구분한다. 같은 용어에 대하여 서로 다른 판단을 내리는 경우 다수 전문가의 판단 결과를 투표를 통하여 결정한다.

- 인명, 지명, 조직명 등을 나타내는 고유명사는 디스크립터에서 제외한다. 예를 들어 '마이크로소프트'는 조직명이기 때문에 디스크립터에 해당하지 않는다.
- 시간 또는 공간의 의미를 포함하는 용어는 디스크립터에서 제외한다. 예를 들어, '토종 리눅스' 에서 '토종'은 지역을 의미하기 때문에 디스크립터에 해당하지 않는다. '최신 컴퓨터' 에서 '최신'은 시간을 의미하기 때문에 디스크립터에 해당하지 않는다.
- 해당 분야의 전문 개념을 표현하지 않는 용어는 디스크립터에서 제외한다. 예를 들어 정보기술 분야 시소러스에서 '지르코늄' (*Zirconium*)은 정보기술 분야 용어가 아니기 때문에 디스크립터에서 제외한다.

스코프노트 (scope note)는 용어의 의미를 명확하게 전달하기 위하여 각 용어 별로 부가하는 설명문을 말한다. 본 연구에서는 기존 전문용어 사전의 정의문 또는 말뭉치에서 추출한 용례를 이용하여 전문가의 판단에 의하여 스코프노트를 구축한다.

3.3 실험 및 분석

본 연구에서는 IT분야 한국어 특허 및 신문기사에서 용어를 자동으로 추출하였다. 특허 문서와 신문기사는 1994년부터 2004년까지 각각 1,400만 어절과 1,500만 어절로 구성되어 있다. 총 765,468개의 단일 명사 또는 복합 명사로 구성된 용어 후보를 추출하였다. 추출한 용어 후보 중에서 상위 순위 3,688개 용어를 선택하였고, 이 용어들은 누적 빈도수로 볼 때 전체 용어 후보 빈도수의 10%를 차지한다. 선택된 용어를 세 명의 분야 전문가가 디스크립터와 그렇지 않는 용어로 구분하였다. 3,688개 용어 중에서 3,023개 용어(82.0%)가 디스크립터로 분류되었다. 디스크립터로 분류된 3,023개 용어 중에서 627개 용어(17.0%)는 기존의 IT분야 사전에 나타나지 않는 신조어였다. 예를 들어 '와이브로' (*WiBro*)는 새로운 무선 인터넷 서비스를 나타내는 신조어이다.

4 용어 분류

선택된 모든 디스크립터 용어를 대상으로 계층구조를 한번에 구축하는 작업은 복잡하고 시간이 많이 필요한 작업이다. 용어들이 의미적인 클래스 단위로 분류되어 있는 경우 클래스 단위의 작은 규모의 계층구조를 쉽게 만들 수 있으며, 클래스 단위의 계층구조는 다른 분야의 시소러스에 쉽게 공유될 수 있는 장점이 있다. 본 연구에서는 Inspec[1] 분류체계에 따라서 3장에서 선택한 디스크립터 용어를 클래스 단위로 자동 분류한 뒤 분류 결과를 검증한다. Inspec 분류 체계에서 상위 3개 레벨을 사용하며, 사용되는 분류 체계는 모두 81개의 클래스로 나뉘어 진다. 더 깊은 레벨의 클래스까지 사용하는 경우, 자료 희귀성 문제 때문에 정확한 분류가 어려워 진다.

[1] Inspec은 과학, 기술 분야 문서에 대한 접근 서비스를 제공하기 위한 IEE의 서지정보 시스템이다. 본 연구에서는 최상위 5개 분류 체계 중에서 '전기전자' (electrical engineering & electronics), '컴퓨터 및 제어' (computer & control), '정보 기술' (information technology) 분류를 이용한다.

4.1 K-NN을 이용한 자동 분류

본 연구의 자동 분류시스템은 k 최근접 이웃 (*k*-Nearest Neighborhood) 분류 방법 [4]을 이용하여 한 개의 용어에 대하여 최대 k개의 클래스를 제안한다. 용어와 클래스 사이의 유사도는 용어의 정보와 클래스의 정보를 벡터로 변환한 뒤, 코사인 유사도 계산 방식을 이용하여 계산한다. 아래는 용어 t 와 클래스 c 의 가능한 속성 벡터를 설명한다.

- 용어 t의 속성 벡터
  - V_{tw} : t를 구성하는 단어의 벡터
  - V_{td} : t의 정의문에 포함된 명사의 벡터
  - V_{tu} : t의 용례에 포함된 명사의 벡터

- 클래스 c의 속성 벡터
  - V_{cw} : c에 포함된 용어를 구성하는 단어의 벡터
  - V_{cd} : c에 포함된 용어의 정의문에 포함된 명사의 벡터
  - V_{cu} : c에 포함된 용어의 용례에 포함된 명사의 벡터

용어 t 와 클래스 c 사이의 유사도는 아래 식과 같이 계산한다. α, β, γ 는 각 유사도의 가중치 값이며, 반복된 실험에 의하여 찾아낸 최적값을 이용한다. 본 연구에서는 각각 0.6, 0.3, 0.1을 할당하였다.

$$Sim(t, c) = \alpha \cdot Sim(V_{tl}, V_{cl}) + \beta \cdot Sim(V_{td}, V_{cd}) + \gamma \cdot Sim(V_{tu}, V_{cu})$$

방법에서 제시하는 클래스의 개수 k는 샘플 테스트를 통하여 결정한다. 40개 용어에 대한 테스트에서 모든 용어에 자동으로 할당된 상위 클래스 12.98개 중에서 최소 1개의 정답 클래스를 찾을 수 있었기 때문에 k의 값을 13으로 결정하였다.

4.2 클래스 검증

자동 분류 시스템은 한 개의 용어에 대하여 13개의 후보 클래스를 제시한다. 수작업 검증 단계에서는 클래스 단위로 자동 분류 시스템에 의하여 할당된 용어의 포함여부를 결정한다. 따라서 자동 분류 결과를 클래스-용어 형식으로 변환한 다음, 전문가가 클래스에 할당된 용어를 순차적으로 해당 클래스와의 적합성 여부를 결정한다.

수작업 검증 단계에서 작업자들이 현재 고려하고 있는 용어와 클래스의 영역을 정확히 파악하는 것이 가장 중요하다. 작업자는 먼저 클래스 이름과 클래스에 포함된 용어를 보고 클래스의 의미 영역을 파악한다. 또한 작업자는 용어의 의미 영역을 파악할 때 1) 용어를 구성하는 단어, 특히 중심어를 보거나, 2) 용어의 정의문, 특히 정의문에서 상위 개념을 표현하는 핵심 단어를 참조하거나, 마지막으로 3) 말뭉치에서 추출한 용례를 참조한다. 작업자가 한 눈에 모든 정보를 참조할 수 있도록 클래스 이름, 포함된 용어, 용어의 정의문 및 용례를 한 개의 화면에 표현한다.

4.3 실험 및 분석

3장에서 디스크립터로 분류한 3,023개의 용어 중에서 2,470개 용어를 분류하였다. 나머지 533개 용어는 시스템이 제시한 13개의 클래스 중에서 적절한 클래스를 찾을 수 없었다. 실험에서 한 개의 용어에 평균 2.99개의 클래스를 할당하였다.

그림 3은 분류체계 두 번째 단계 클래스로 분류된 용어의 개수를 보여준다. 클래스 B (Electrical Engineering & Electronics)와 클래스 C (Computers & Control)로 분류된 용어의 개수가 클래스 D (Information Technology)로 분류된 용어의 개수보다 상대적으로 많다. 이 현상은 실험에서 사용한 말뭉치의 분류가 B와 C 클래스에 좀더 편중되었다는 사실을 설명한다.

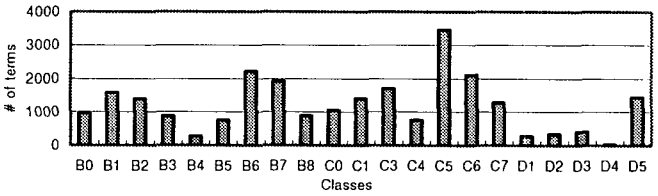

그림 3. Inspec 분류 체계의 두 번째 단계 클래스로 분류된 용어의 분포

5 계층 구조 구축

앞 단계에서 선정한 용어를 계층적으로 조직화시켜서 전문 용어 계층 구조를 구축한다. 계층 구조에 포함된 모든 용어는 한 개 이상의 다른 용어와 계층 관계를 가진다. 계층 관계에는 *is-a*, *part-of* 또는 다른 종류의 광의/협의 관계를 포함한다.

이 단계에서는 앞 단계에서 분류된 클래스에 포함된 용어들을 대상으로 클래스 단위의 여러 개의 계층 구조를 구축한다. 계층 구조 구축 과정은 현재의 계층 구조에 연속적으로 새로운 용어를 추가하는 과정을 반복한다. 계층 구조는 그림 4와 같이 초기에 비어 있는 상태에서 시작하여 반복적으로 새로운 용어를 추가하여 풍부한 구조를 가진다. 추가되는 용어는 용어의 전문성 값을 이용하여 정렬한다. 용어의 전문성은 주어진 분야에서 그 용어가 내포하는 분야 지식의 양을 정량화한 것이다 [5]. 분야 지식을 많이 포함할수록 높은 전문성 값을 가지고, 일반적인 지식을 표현하는 용어는 낮은 전문성 값을 가진다. 전문성이 높은 용어는 계층구조에서 하위 레벨에 위치하는 경향이 있고, 전문성이 낮은 용어는 상위 레벨에 위치하는 경향이 있다. 따라서 일반적인 용어부터 차례로 계층구조에 추가하면 계층구조는 상위 레벨부터 차례로 하위 레벨 방향으로 성장한다.

새 용어 t_{new}를 계층 구조에 추가하는 작업은 기본적으로 아래의 4 단계로 구성된다.

1. 시스템은 t_{new}가 위치할 수 있는 가능한 위치의 후보를 기존의 계층 구조에서 찾아서 제시한다. 즉 시스템은 t_{new}의 가능한 상위어를 현재의 계층 구조에서 찾는다.
2. 전문가는 제시된 후보를 검증하여 한 개 이상의 상위어를 선택하여 t_{new}를 그 상위어의 하위어로 등록한다. 다음 용어를 t_{new}에 할당하고 단계 1을 반복한다.
3. 시스템이 제시한 후보 중에서 올바른 상위어를 찾지 못하는 경우, 전문가의 판단에 의하여 현재의 계층 구조에서 올바른 상위어를 찾고 t_{new}를 하위어로 등록한다. 다음 용어를 t_{new}에 할당하고 단계 1을 반복한다.
4. 시스템과 전문가 모두 t_{new}의 올바른 상위어를 찾지 못한 경우, t_{new}를 용어열의 가장 뒤에 보내고 다음 용어를 t_{new}에 할당하고 단계 1을 반복한다.

시스템의 상위어 추천 기능은 전문가의 수작업을 최소화시키고 계층관계를 일관성 있게 만드는 장점이 있다.

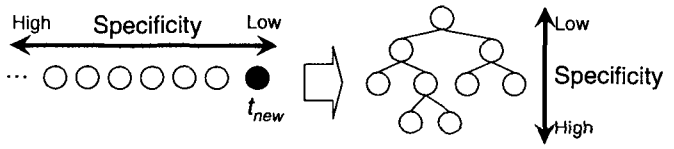

그림 4. 특정 클래스에 포함된 용어 중에서 전문성이 낮은 용어부터 차례로 클래스의 계층 구조에 등록한다.

5.1 계층 관계 자동 추출

이 절에서는 규칙 기반 방법, 참조 시소러스 기반 방법, 용어의 전문성과 유사도 기반 방법을 이용하여 용어의 계층 관계를 자동으로 추출하는 방법을 설명한다.

5.1.1 규칙과 참조 시소러스 기반 방법

이 절에서는 규칙과 참조 시소러스를 이용하여 계층 관계를 추정하는 방법을 설명한다. 전통적으로 이 방법들은 정확률은 높지만 재현율이 낮은 특징이 있다 [6].

- **수직관계 기반 방법**

분야 의존적인 개념을 용어로 표현할 때 많은 경우 기존의 용어에 수식어를 부가하여 만들어지는 경우가 많다 [11]. 예를 들어 '플래시 메모리'는 기존의 용어 '메모리'에 수식어 '플래시'를 부가하여 만들어졌고, 두 용어는 계층 관계를 가진다. 수직관계는 용어간 계층 관계를 표현하는 중요한 단서이다. 두 용어 t_1과 t_2, 에서 t_2가 t_1의 중심어와 일치하고 t_1이 t_2에 다른 용어나 형용사에 의하여 수식된 형태이면 두 용어 사이에는 *is-a*(t_1,t_2)관계가 성립한다 [6,7]. 이 방법은 분야에 독립적이고 비교적 정확한 관계를 추출하는 장점이 있다. 그러나 이 방법이 항상 올바른 *is-a* 관계만을 추출하는 것은 아니다. 예를 들어 두 용어 '배타적 논리합 게이트'와 '논리합 게이트' 사이에는 수직관계가 성립하지만 두 용어 사이에 *is-a*관계가 성립하지 않는다. 두 용어는 '게이트'의 일종으로 서로 다른 동작 특성을 갖는 동등한 수준의 용어이다.

- **정의문 패턴 기반 방법**

정의문은 과학 기술 문헌에서 특정한 개념, 동작, 객체 등을 설명하기 위하여 자주 나타나는 문장 패턴이다. 이 방법은 웹에서 추출한 정의문에서 정의문 패턴을 이용하여 계층 관계를 추출한다. 이 방법은 Hearst [8]가 제안한 어휘-구문 패턴 (lexico-syntactic pattern)을 이용하는 방법과 유사하지만 정의문 패턴을 중심적으로 이용하여 정확한 정확한 계층 관계를 추출할 수 있는 장점이 있다. 본 연구에서는 [9]에서 제시한 영어 정의문 패턴을 이용한다. 가장 대표적인 영어 정의문 패턴은 다음과 같다. 여기에서 *term*은 정의문 검색 대상 용어이고, *genus term*은 *term*을 정의할 때 핵심이 되는 용어이다. 이 경우 *is-a*(term, *genus term*)의 관계가 성립한다.

- a(n) *term* is a(an) *genus term* (which is) *verb*+ed ...
- a(n) *term* is a(n) *genus term* for *verb*+ing ...

예를 들어 '지지 벡터 기계'의 영어 대역어 '*support vector machine*'의 정의문을 검색하기 위하여 위의 정의문 패턴을 이용하여 'a *support vector machine* is a'를 웹 검색 시스템[2]에 질의어로 보낸다. 검색된 정의문 중 한 개는 다음과 같다.

- A *support vector machine* is a *supervised learning algorithm* developed over the past decade by Vapnik and others.

검색된 정의문에서 정의문 패턴을 적용하여 '*support vector machine*'의 핵심어 '*supervised learning algorithm*'를 추출하여 계층 관계 *is-a*('*support vector machine*', '*supervised learning algorithm*')를 만들 수 있다.

- **참조 시소러스 기반 방법**

계층 관계를 추출하기 위하여 WordNet[3]과 같은 기존의 시소러스를 이용할 수 있다. WordNet은 일반분야 시소러스이지만 다양한 분야의 전문 용어도 많이 포함하고 있다. 예를 들어 WordNet에서 '기호 논리'*(symbolic logic)*는 '부울 논리'*(Boolean logic)*의 상위어이다. 따라서 현재 등록하려는 용어가 '부울 논리'이고 기존의 계층구조에

[2] 정의문 검색을 위하여 Google (http://www.google.com)을 사용한다.

[3] http://www.cogsci.princeton.edu/~wn

'기호 논리'가 있는 경우 '기호 논리'는 '부울 논리'의 상위어 후보가 된다.

5.1.2 용어의 전문성과 유사도 기반 방법

기존의 방법들은 비교적 정확률은 높지만 재현율이 낮은 단점이 있다. 따라서 재현율을 높이기 위하여 통계 기반 방식을 사용한다. 용어의 전문성은 용어간 계층 관계에서 필요조건으로 이용될 수 있다. 그림 5에서 두 용어 t_1과 t_2가 의미적으로 유사하면서 t_1의 전문성 값이 t_2 의 전문성 값보다 낮은 경우 t_1이 t_2의 상위어가 될 가능성이 높다. 이 가정을 기반으로 현재 계층 구조에 새롭게 추가되는 용어 t_{new}의 상위어 후보를 다음과 같은 단계를 거쳐서 추정한다.

1. 용어의 전문성을 이용하여 t_{new}의 기존의 용어 계층 구조에서 상위어 후보를 선택한다.
2. 단계 1에서 선택한 t_{new}의 상위어 후보를 용어간 유사도를 이용하여 순위화한 후 상위 n개를 선택한다.

그림 5에서 전문성 제약에 의하여 t_{new}의 가능한 상위어는 t_1, t_2, t_3, t_4 이다. t_{new} 의 전문성 값은 t_1, t_2, t_3, t_4 의 전문성 값보다 크다 . 상위어 후보는 t_{new}와의 유사도 계산을 통하여 정렬한다. 용어간 유사도는 두 용어의 의미 영역이 겹치는 부분을 정량화한 것이다. 용어 간 유사도는 용어 구성 성분 가정과 용어의 분포 가정을 기반으로 계산한다. 구성 단어를 많이 공유하는 두 용어는 의미적으로 유사하다고 할 수 있고, 비슷한 문맥을 가지는 두 용어도 유사하다고 할 수 있다. 유사도의 정도는 각 유사도 측정 방법에서 공유하는 비율을 이용하여 계산한다.

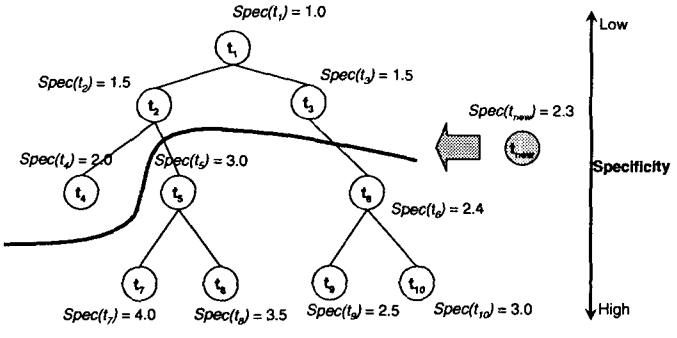

그림 5. 용어의 전문성을 이용하여 t_{new} 의 상위어 후보 선택

5.2 계층 관계 검증

시스템에서 제시하는 계층 관계는 전문가들의 검수 단계를 거친다. 그러나 전문가들의 전문 지식의 정도와 전문 분야에 따라서 서로 다른 검수 결과를 낼 수 있기 때문에, 검수 지침을 이용하여 일관성있게 검수할 수 있도록 유도한다. 검수지침은 Inspec 시소러스에서 제시한 다양한 계층 관계를 분석하여 작성하였다. 다음과 같은 단계를 통하여 시스템이 제시한 계층 관계를 검수한다.

1. 제시한 계층 관계에 포함된 두 용어 사이의 관계를 결정하는 핵심 특징을 결정한다. 정보기술 분야 용어간 관계에서 가능한 핵심 특징은 다음과 같다.
   - 객체 (A): 제시된 관계가 '객체'와 '객체+속성' 사이의 관계라고 판단되는 경우
   - 동작 (B): 제시된 관계가 '동작'과 '동작 + 속성' 사이의 관계라고 판단되는 경우
   - 속성 (C): 제시된 관계가 '속성'과 '객체 또는 동작과 관련된 속성' 사이의 관계라고 판단되는 경우
   - 기술 (D): 제시된 관계가 '기술'과 '기술 + 속성' 사이의 관계라고 판단되는 경우

   예를 들어 '네트워크←컴퓨터 네트워크' 관계는 '네트워크'라는 '객체'와 '네트워크' '객체'에 '컴퓨터' 라는 속성이 부가된 용어 사이의 관계이다. ←는 두 용어 사이의 상하위어 관계를 나타내고, 왼쪽 용어가 상위어이고, 오른쪽 용어가 하위어이다.

2. 두 용어 사이의 계층 관계를 결정하는 부가적인 '속성', '객체' 또는 '동작'을 결정한다. 즉 하위어에 추가된 특징을 결정한다. 예를 들어 '객체' (A)와 관련된 부가적인 '속성'의 종류는 다음과 같다.
   - 객체 ← 객체에 대한 *의미제약* (A01)
     예) 네트워크 ← *컴퓨터* 네트워크
   - 객체 ← 객체에 의한/대한 *행위* (A02)
     예) 네트워크 ← 네트워크 *관리*
   - 객체 ←객체의 *속성* (A03)
     예) 네트워크 ← 네트워크 *신뢰도*
   - 객체 ← 객체의 *인스턴스* (A04)
     예) 디지털 컴퓨터← *IBM 컴퓨터*
   - 객체 ← 객체의 *부분* (A05)
     예) 데이터베이스 관리 시스템 ← *데이터베이스 색인*
   - 객체 ← 객체의 *응용* (A06)
     예) 인터넷 ← *인터넷 전화*
3. 단계 1과 단계 2에 의하여 시스템이 제시한 관계를 검증할 수 없을 경우, 해당 관계를 제외하거나 새로운 검수 지침을 추가하여 해당 관계를 수용한다.

5.3 실험과 분석

제안한 방법으로 1,042개의 용어로 구성된 정보기술 분야 계층관계를 구축하였다. 전체 용어 중에서 330개 (31.7%) 용어는 시스템이 제시한 상위어 중에서 선택하여 추가하였고, 712개 (68.3%) 용어는 전적으로 전문가의 판단에 의하여 계층 구조에 추가하였다 (표 1).

표 1. 시스템에서 제시한 계층 관계에서 결정한 계층관계 수와 전문가의 판단에만 의존하여 결정한 계층 관계의 수

수직 관계	정의문	워드넷	전문성/ 유사도	전문가	계
152	8	11	159	712	1,042

그림 7은 '객체'를 중심 특성으로 하는 관계 568개에 포함된 세분화된 계층 관계의 분포를 보여준다. 358개 (63.0%) 관계는 '객체'와 '객체에 대한 의미제약' 사이의 관계 즉 *is-a* 관계이다. 다음으로 많은 관계는 '객체'와 '객체의 부분' 사이의 관계 (*part-of*) 이다. (110개, 10.6%)

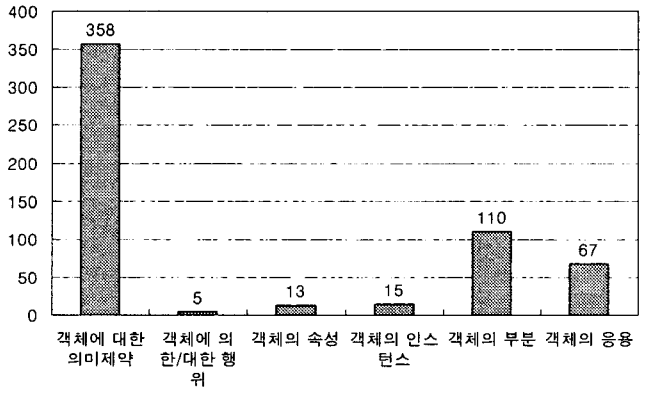

그림 7. 중심 특성이 '객체'인 관계들에 대한 세분화된 분포

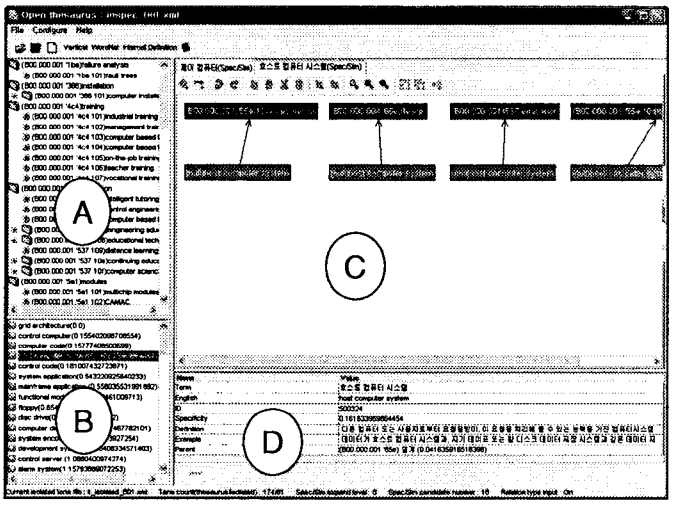

그림 8. 계층 관계 구축 지원도구

그림 8은 계층 관계 구축 지원 도구를 보여준다. 도구를 사용함으로써 작업자의 주관 및 오류를 최소화할 수 있다. 자동 계층 관계 추정 방법을 제공하며, 작업자가 직접 계층 관계를 설정할 수도 있다. 작업자가 용어의 의미를 정확하게 판단할 수 있도록 용어의 정의문, 용례 등을 볼 수 있는 기능이 있다. 그림에서 A는 구축중인 계층 구조를 나타내고, B는 A에 추가하려는 용어를 전문성 값에 의하여 정렬한 것이고, C는 B에서 선택된 용어에 대한 상위어 후보를 A에서 추정하여 보여주고 있다. 여러 개의 상위어 후보 중에서 한 개 또는 그 이상을 지정하면 선택된 관계는 A에 추가된다. D는 선택된 용어에 대한 정의문, 용례 등 판단을 위한 정보를 보여준다.

6. 결론

본 연구에서는 전문 분야 시소러스를 구축하기 위한 단계적인 분할-정복 방법을 제안하였다. 이 방법은 1) 계층 구조 구축의 복잡도가 낮아지고, 2) 각각의 클래스 단위의 시소러스는 다른 전문 분야 시소러스에 쉽게 재사용될 수 있고, 3) 전체 시소러스에서 각 클래스 별 용어 분포를 쉽게 파악할 수 있는 장점이 있다. 이 방법은 용어 추출 단계, 용어 분류 단계, 계층 관계 구축 단계의 3 단계로 구성된다. 모든 단계는 자동 처리와 수작업 검증 작업으로 구성되어서 수작업의 비용을 최소화하면서 자동 방법의 오류를 검증하는 체제로 구성된다. 본 연구에서 제안한 방법은 한국어 정보 기술 분야 시소러스 구축에 적용하였다. 한국어 특허 문서와 신문기사 말뭉치에서 용어를 추출하였기 때문에 한국에서 만들어진 다수의 신조어를 포함한다. 구축된 시소러스는 81개의 상위 클래스에 1,000개 이상의 정보기술 분야 용어를 포함한다.

시소러스 구축 과정에서 각 단계별로 많은 용어들이 필터링되었다. 초기에 추출한 3,688개 용어 중에서 디스크립터로 선택된 용어가 3,023개이고, 이 중에서 분류 단계를 통과한 용어가 2,470개이며, 최종적으로 시소러스에 포함된 용어가 1,042개 이다. 용어들이 점차적으로 줄어든 이유는 첫 번째 단계에서 디스크립터로 분류하지 않아야 하는 용어를 두 번째, 세 번째 단계에서 추가로 필터링한 경우와, 분류 및 계층 관계 구축 단계에서 작업자가 용어의 의미를 정확하게 판단하지 못하여 분류를 하지 못했거나 계층 관계를 설정하지 못한 경우이다. 작업자의 전문 지식 부족 문제를 해결하기 위하여 전체 분류 체계에서 용어 분류 비율이 낮거나, 계층 관계의 설정 비율이 낮은 클래스에 대한 전문가를 추가로 활용하여야 한다. 정보 기술 분야는 전자, 전기, 컴퓨터, 정보 기술 등 광범위한 영역을 포함하기 때문에 소수의 전문가가 모든 분야를 관리하기는 어렵다.

시소러스 구축 과정의 모든 단계는 애매성을 많이 내재하고 있다. 따라서 향후 각 단계에서 객관적인 평가기준이 제시되어야 하고, 전체 과정을 반복하면서 점진적으로 시소러스의 오류를 수정하는 시소러스 구축 모델이 제안되어야 한다.

참고 문헌

[1] ANSI/NISO Z39.19-2003, Guidelines for the Construction, Format, and Management of Monolingual Thesauri, NISO Press, Bethesda, Mariland, U.S.A. (2003)

[2] B. Christopher, Y. Wilks, Ontologies, Taxonomies, Thesauri: Learning from Texts. In Proceedings The Use of Computational Linguistics in the Extraction of Keyword Information from Digital Library Content Workshop, Kings College, London, UK, (2004)

[3] J. Oh, K. Lee and K. Choi, Term Recognition Using Technical Dictionary Hierarchy, In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp 496-503, (2000)

[4] W. Hwang and K. Wen, Fast *k*NN classification algorithm based on partial distance search, Electronics Letters, Vo. 34, Issue 21, pp. 2062-2063, (1998)

[5] P. Ryu, K. Choi, Measuring the Specificity of Terms for Automatic Hierarchy Construction, In Proceedings of ECAI-2004 Workshop on Ontology Learning and Population, (2004)

[6] P. Cimiano, A. Pivk, L. Schmidt-Thieme, S. Staab, Learning Taxonomic Relations from Heterogeneous Evidence, In Proceedings on ECAI-2004 Workshop on Ontology Learning and Population, (2004)

[7] P. Velardi, P. Fabriani, and M. Missikoff, Using Text Processing Techinques to Automatically enrich a Domain Ontology, In Proceedings of the ACM International Conference on Formal Ontology in Information Systems, (2001)

[8] M.A. Hearst, Automatic Acquisition of Hyponyms from Large Text Corpora, In Proceedings of the 14th International Conference on Computational Linguistics, (1992)

[9] J. Pearson, Analysis of Definitions in Text, In Terms in Context (Studies in Corpus Linguistics), Vol. 1, John Benjamins Publishing Company, pp.89-104, (1998)

[10] P. Cimiano, A. Hotho, S. Staab, Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis, Accepted for JAIR - Journal of AI Research, 2005.

[11] ISO 704, Terminology work-Principles and methods, ISO 704:2000(E), (2000)

[12] 구희관, 정한민, 이병희, 성원경, 전문용어사전 구축을 위한 전문용어 추출 및 순위화, 제23회 한국정보처리학회 춘계학술발표대회 논문집, (2005)

부록

클래스 B61 (정보 및 통신 이론)에 포함된 계층 구조의 일부분

분류코드	용어	영어 대역어
533	패턴 인식	pattern recognition
533.107	특징 추출	feature extraction
533.107.a00	에지 검출	edge detection
8eb	신호 처리	signal processing
8eb.001	영상 신호 처리	video signal processing
8eb.002	영상 처리	image processing
8eb.002.002	영상 인식	image recognition
8eb.002.002.001	영상 정합	image matching
8eb.002.002.002	에지 검출	edge detection
8eb.002.002.003	지문 인식	fingerprint identification
8eb.002.003	컴퓨터 비전	computer vision
8eb.002.003.001	머신 비전	machine vision
8eb.002.004	영상 부호화	image coding
8eb.002.006	입체 영상 처리	stereo image processing
8eb.002.007	영상 개선	image enhancement
8eb.002.008	컴퓨터 단층 촬영	computerised tomography
8eb.002.009	영상 표현	image representation
8eb.002.00a	렌더링	rendering
8eb.002.00a.001	광선 추적법	ray tracing
8eb.002.00a.002	볼륨 렌더링	volume rendering
8eb.002.00b	화상 분석	image analysis
8eb.002.00c	영상 변환	image transformation
8eb.002.00d	세선화	thinning
8eb.004	의학 신호 처리	medical signal processing
8eb.005	데이터 압축	data compression
8eb.005.001	벡터 양자화	vector quantization
8eb.015	광 정보 처리	optical information processing
8eb.016	음향 신호 처리	acoustic signal processing
8eb.016.001	음향 합성	acoustic convolution
8eb.023	신호 검출	signal detection
8eb.023.001	차등 검파	differential detection
8eb.023.002	헤테로다인 검파	heterodyne detection
8eb.023.003	동기 검파	homodyne detection
8eb.02c	음성처리	speech processing
8eb.02c.001	음성 인식	speech recognition
8eb.02c.001.001	화자 인식	speaker recognition
8eb.02c.001.002	연속 음성 인식	continuous speech recognition
8eb.02c.001.003	화자 적응	speaker adaptation
8eb.02c.001.004	자동 음성 인식	automatic speech recognition
8eb.02c.003	음성 부호화	speech coding
8eb.02c.004	음성 압축	speech compression
8eb.02c.005	음성 합성	speech synthesis
8eb.02c.006	음성 분석	speech analysis

- 분류 코드는 용어 사이의 계층 구조를 표현한다.