패싯(Facet)을 이용한 과학기술분야 시소러스 구축과 활용방안*

Building Thesaurus for Science & Technology Domain Using Facets and Its Application to Inference Services

황 순 희** · 정 한 민*** · 성 원 경****
Soonhee Hwang · Hanmin Jung · Won-Kyung Sung

차 례

초 록

본 논문은 시소러스 구축 시 직면하는 문제점과 구축방법을 비판적으로 검토하고, 여러 가지 구축 방법 중 직접구축 방법을 제안하였다. 또한, 확장검색의 효율성을 보장할 수 있는 시소러스 구축을 위해 의미적 준거인 개념패싯, 관계패싯 등의 도입과 활용을 도입하였으며, 이를 이용하여 구축한 과학기술분야 시소러스의 구축사례를 제시하였다. 특별히, 패싯이 반영된 시소러스에는 다양한 관점이 반영될 수 있으므로 검색의 효율성이 보장된다는 장점이 있고, 인접 과학기술분야에 응용될 수 있다. 일례로 본 시소러스는 과학기술 연구자들의 협업을 지원하기 위한 정보유통 서비스에 응용될 수 있으며, 향후 고도화된 지식 서비스에도 확장 응용될 수 있다.

키 워 드

지식기반, 지식표상, 온톨로지, 추론 서비스, 상세검색, 확장검색, 시소러스

* 본 논문은 한국과학기술정보연구원(KISTI), 부산대학교, (주)오름정보가 공동으로 개발, 구축하는 '범용 과학기술분야 전문용어 시소러스'의 구축 준거·절차 및 1차년도(2005년) 구축결과물을 연구대상으로 삼는다.
** 부산대학교 BK21 U-Port IT 산학공동 사업단 Post-Doc.(Pusan National University; Center for U-Port IT Research and Education, soonheehwang@pusan.ac.kr)
*** 한국과학기술정보연구원, 정보시스템연구팀, 선임연구원(Senior Researcher, Information System Division, KISTI, jhm@kisti.re.kr)
**** 한국과학기술정보연구원, 정보시스템연구팀 팀장(Director, Information System Division, KISTI, wksung@kisti.re.kr)
• 논문접수일자 : 2006년 8월 17일
• 게재확정일자 : 2006년 9월 12일

ABSTRACT

In this paper, we proposed one of the methods for building thesaurus in Science & Technology domain and investigated its applicability as an inference service based on ontology. There exist as many building methods for thesaurus as its role and function, and actually many thesauri capable of ensuring the accuracy and efficiency in information search are being built by many experts. After examining the previous studies related to the principles of building thesaurus and relevant concept "facet", we focused on its characteristics and applied it to building thesaurus. The facet is classified into 2 categories, conceptual facet and relational facet. The latter contains 3 subcategories: category relational facet, attribute relational facet and thematic relational facet. The thesaurus for Science & Technology domain using facets can be applied as a web-based inference service. As a result, the three types of inference service, COP(Communities of Practice), Researcher Tracing and Research Map are provided by means of ontology, and can be applied for the Query Expansion.

KEYWORDS

Knowledge Base, Knowledge Representation, Ontology, Inference Service, Advanced Search, Query Expansion, Thesaurus

1. 서 론

과학기술분야의 수많은 정보 자원들은 하루가 다르게 생산, 유통, 축적되고 있으며, 그 효과적인 활용을 위한 체계적이고 표준화된 지식기반(Knowledge Base) 자원의 중요성이 날로 강조되고 있다. 시소러스(thesaurus)는 기반지식 · 언어자원의 일종으로 최종 사용자의 활용 만족도를 높일 수 있도록, 보다 체계적이고 일관성 있는 방법론의 개발이 요구되는 자원이다. 본 논문은 범용 과학기술분야 전문용어 시소러스 구축의 배경과 방법론, 의미적 준거의 설정기준을 검토하고, 실제 시소러스 구축 결과물을 바탕으로 향후 추론 서비스로의 활용가능성을 진단하는 데 그 목적이 있다.

특정 전문분야에서 사용되는 빈도수가 높은 통제된 용어집합으로 정의되는 시소러스는 정보의 급증과 검색환경의 변화에 따른 기능변화가 절실히 요구되는 자원이다. 즉, 정보의 색인 및 확장검색(query expansion), 상세검색(advanced search) 또는 추론검색(inference search) 등의 검색 시스템에서 높은 효율성과 정확성을 보장할 수 있어야 한다. 시소러스는 전통적으로 문헌정보학과 전산학이 주축이 되

어 구축되어 왔으며, 전자는 여러 용어 간 관계 기술 및 구축에 명확한 의미적 준거를 고려하지 않았다(황순희, 윤애선 2005). 반면, 후자는 주로 통계적 기법에 입각한 자동구축에 비중을 두어왔다. 먼저 시소러스 구축 시 어려움은 다음 몇 가지로 요약될 수 있다.

첫째, 시소러스 구축 시 용어의 구축은 물론이고, 지식 및 정보의 증가 및 변화로 인한 향후 유지, 보수에 시간과 전문가의 수작업이 과도하게 요구된다(Brewster and Wilks 2004). 또한 자동구축 또는 기 구축된 어휘의미망을 이용한 간접구축의 경우라도 뚜렷한 정제 기준을 바탕으로, 지속적인 수정과 정제가 불가피하다.

둘째, 시소러스가 담고 있는 지식은 지속적으로 변화, 발전하므로 시소러스가 구축되는 시점에 시소러스에 포함된 지식은 동시대의 지식이 아닌 낡은 지식이 반영될 우려가 있다. 따라서, 한번 구축한 시소러스의 유지와 지속적인 수정이 가능하도록 초기부터 지능적으로 고안, 설계되어야 한다. 가령, 용어생명주기(Life Cycle of Terms)를 고려한 시소러스 구축은 구축 이후 시소러스의 효용성을 보장할 수 있는 방법이다(정한민 외 2005).

셋째, 자동구축 시 데이터의 산발성(data sparsity)이 문제가 된다. 통계기법에 의한 구축관련 선행연구들에 따르면 의미적 유사성을 지닌 용어들은 유사한 통사적 관계 속에 출현한다는 가정에서 출발하며, 용어 각각은 출현 가능한 문법적 문맥에 의해 그룹화 된다. 그런데, 시소러스의 구축대상이 되는 전문용어는 코퍼스상에서 상대적으로 낮은 빈도를 보이기 때문에 코퍼스를 이용하여 통계적으로 의미 있는 정보를 추출하기란 사실상 어렵다. 이 점은 자동구축에 의한 시소러스가 실제로 활용되는 예가 전혀 없다는 사실과 자동구축 자체가 실험적 수준을 벗어나기 어렵다는 한계를 뒷받침하는 부분이다.

넷째, 향후 구축되는 시소러스는 기 구축된 용어는 물론이고, 이를 기반으로 확장된 용어 및 외래어로부터 음차된 용어, 새로 생성된 신조어 등을 충분히 반영할 수 있도록 유연성 있는 설계가 요구된다. 또한, 추론 서비스 등으로의 상호 운용성을 염두에 두어야 한다.

본 연구는 특별히 시소러스의 자동구축방법이 갖는 문제점 해결을 위한 대안으로 의미적 준거를 이용한 직접구축(수동구축) 방법을 제안하고자 한다. 본 연구의 목적은 다음과 같다. 첫째, 급변하는 검색환경의 요구를 충족시키고 효율성이 보장된 시소러스의 구축을 위해 의미론적 관점에 입각한 구축준거를 제안하고, 이를 향후 시소러스 확장의 기준으로 삼고자 한다. 둘째, 본 연구를 통해 개발된 준거를 적용하여 구축된 과학기술분야 시소러스의 실제 구축사례를 통해, 향후 온톨로지 기반 추론 서비스로의 활용가능성을 살펴본다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 시소러스 구축방법과 기본원칙을 비판적으로 검토하고, 효율적인 시소러스 구축원칙 및 의미적 준거의 필요성을 언급한다. 3장에서

는 본 시소러스의 구축방법과 의미적 준거의 특성을 실제 구축사례를 통해 제시한다. 4장에서는 패싯(facet)을 적용한 시소러스의 장점들을 활용, 실제 과학기술분야에의 적용방향을 고찰한다. 5장에서는 본 연구의 제한점을 밝히며, 향후 연구방향을 제시한다.

2. 선행연구 및 문제점

정보의 급증으로 인한 정보검색 환경의 변화는 시소러스의 기능변화에도 영향을 미친다. 따라서, 이러한 요구에 부응할 수 있는 정교한 시소러스의 설계는 구축 이전에 고려되어야 할 사안이다. 2장은 시소러스 구축원칙과 관련된 선행연구를 검토한다. 이를 위해 구축방식에 따른 시소러스의 유형을 알아보고, 기존의 시소러스 구축과 문제점을 비판적으로 검토하며, 시소러스 구축의 의미적 준거로 개발, 확장될 수 있는 패싯의 필요성과 특성을 논의한다.

2.1 구축방식에 따른 시소러스의 유형

시소러스는 구축방식에 따라 로제방식(Roget type) 시소러스, 워드넷(WordNet)/유로 워드넷(Euro WordNet) 시소러스, 말뭉치로부터의 용어 자동 추출·처리를 통한 자동구축에 의한 시소러스, 수동구축에 의한 시소러스 등 네 가지 유형으로 나뉠 수 있다. 유로 워드넷은 유럽공동체(EU)의 다국어 정보화 계획 프로젝트로 시작되었으며, 프린스턴 워드넷(Princeton WordNet) 1.5를 기반으로 하여 유럽 8개국어를 대상으로 구축된 다국어 어휘의 미망 구축 결과물이다.

첫째, 로제방식 시소러스에 속하는 'Roget 시소러스', 'Macquarie 시소러스' 등은 글을 쓸 때 특정 어휘의 동의어 및 보다 적합한 어휘 선택을 원하는 대중을 위해 고안된 시소러스이다. 로제방식 시소러스는 특정 포맷으로 구축되는데, 가령 'Roget 시소러스'는 의미분류에 기초한 총 6개의 클래스(Class)로 구성되어 있고, 각 클래스는 각각 부(Division), 과(Section) 등의 계층구조로 세분화된다. 각 계층은 고유한 표제정보를 지니며, 계층구조의 말단 노드에는 총 1,044개의 범주가 존재한다. 또한, 각 범주에는 품사별로 유의어 목록이 나열되어 있다(양재군, 배재학 2002). 한편, 'Macquarie 시소러스'에는 표제어의 사전적 정의가 없고, 표제어의 다의성이 기술되어 있어 시소러스 이용자는 표제어의 정의를 스스로 추론할 수 있도록 고안되었다.

둘째, 워드넷/유로 워드넷을 시소러스의 한 유형으로 간주할 수 있다. 이러한 입장은 워드넷의 기본 구성단위가 동의어집합(Synset)라는 점에서 워드넷은 사전보다는 시소러스와 유사한 구조를 지닌다는 점(Peters and Kilgarriff 2000)과 워드넷과 시소러스의 용어 기술을 위해 사용되는 '의미관계'가 차이가 없는 유사한 관계라는 점 때문이다(이재윤, 김태수 1998). 워드넷은 심리언어학 기반의 대용량

전자어휘 데이터베이스로 1985년 구축된 이래로 언어공학 연구에 폭넓게 활용되고 있고, 특히 언어자원(language resource)으로서의 그 유용성과 활용성이 다각도로 논의된 바 있다. 대표적인 활용사례로 자연언어처리(NLP) 분야의 의미분석에 대안제시, 다의어 또는 동형이의어의 중의성 해소(WSD)에 대안제시, 추론검색 또는 확장검색에의 활용 등을 들 수 있다. 워드넷이 NLP에 응용된 사례연구로 Buscaldi et al.(2006), Sinopalnikova(2004), Sundheim et al.(2006) 등이 있다.

셋째, 자동구축에 의한 시소러스는 구축 대상용어를 대용량 말뭉치로부터 자동추출·처리하여 구축되며, 시소러스의 자동생성이 주된 목표이다. 이 경우 용어의 자동추출, 디스크립터(descriptor: 해당 분야 개념을 표현하기 위해 시소러스상 사용된 용어) 선택, 용어의 자동분류, 계층관계의 자동추출 및 계층구조 설정 등이 모두 자동으로 이루어지도록 고안된다. 자동구축은 대개 사전의 정의문을 이용하여 상위어를 추출, 상향식(bottom up) 방식으로 구축되므로, 최종적으로 전문가의 검증이 반드시 수반되어야 한다.

수동구축에 의한 시소러스는 무엇보다 정보 및 색인 검색의 정확성을 보장할 수 있다. 따라서, 구축에 많은 비용과 시간이 소요됨에도 불구하고 대다수의 특정 전문분야(domain specific) 시소러스가 이 방식으로 구축된다. 구축기준은 용어 간 동의관계, 계층관계, 연관관계 등이며 분야에 따라 특정의 의미관계가 추가되기도 한다. 특별히, 자동구축에 의한 시소러스는 수동구축 시소러스에 비해 비용과 시간이 덜 요구되며, 애플리케이션에 따라 수동구축하는 것만큼의 정확도가 요구되지 않는 시소러스의 경우라면 '자동' 또는 '반자동(semi automatic)' 방식으로 구축될 수 있다. 반면, 수동구축에 의한 시소러스는 시소러스의 최종 사용 목적과 구축 전문가의 관점이 충분히 반영될 수 있는 장점이 있다.

2.2 관점에 따른 시소러스 구축

문헌정보학에서 구축하기 시작한 시소러스는 구조화된 특성 때문에 전산적 처리가 용이하므로, 문헌정보학과 전산학, 두 분야에 의해 구축되어 왔다. 본 절에서는 문헌정보학과 전산적 관점에서의 시소러스 구축방법을 검토하고 문제점을 지적한다. 1852년 처음 사용된 용어 '시소러스'는 '지식의 보고, 사전, 백과사전'을 의미하며, 그 당시 시소러스는 특정 용어를 이미 알고 있는 상태에서 의미내용을 찾는 일반사전과는 반대로, 뜻은 알고 있으나 그 개념에 해당하는 적당한 용어를 찾는 데 사용되는 용어집이었다(한상길 1999). 현대의 시소러스는 특정 분야에서 사용되는 사용빈도가 높은 전문용어 목록으로, 용어 각각은 서로 연관성 있는 다른 용어와 등가관계, 계층관계, 연관관계로 구현된다. 다음은 문헌정보학의 시소러스 구축에 이용되는 용어 간 기본적 의미관계로, 몇 가지 문제점을

지적할 수 있다(<표 1>).

첫째, 등가관계는 용어 간 동등관계로 색인과 검색에서 디스크립터(우선어)인 용어와 비디스크립터(비우선어) 용어와의 관계이다. 우선어와 비우선어 간 관계설정은 비우선어를 우선어로 연결시켜 색인과 검색의 효율성을 어느 정도 보장할 수는 있으나, 시소러스마다 등가관계의 기준이 명확하지 않으며 동의어 간 관계설정이 대부분 직관적으로 이루어지고 있다.

둘째, 계층관계는 우선어인 특정용어를 상·하위어 개념과 연결시켜 구축된 관계이다. 상·하위어를 중심으로 계층관계가 설정되며, 최상위어(top term)와 고립어(orphan term: 계층관계를 갖지 않는 용어)도 최종적으로 계층관계에 포함될 수 있다. 용어 간 계층관계는 속관계, 전체 부분관계, 사례관계로 분류되는데, 문제는 확실한 계층관계 구축을 위한 체계적 기준이 결여되어 있다는 점이다. 이 점을 보완하기 위해 본 연구에서는 향후 용어의 추가·확장을 염두에 두고, 용어 간 의미거리(semantic depth)를 충분히 반영하고자 하였으며, 계층구조의 조어적 기준과 의미적 기준, 상위어의 다중 할당이 허용되도록 고안되었다.

셋째, 연관관계는 등가관계 또는 계층관계에 속하지 않으면서, 개념적으로 밀접하게 관련된 용어 간 관계를 의미한다. 연관관계는 색인과 탐색에 이용될 가능성이 큰 대체 용어들인 관련어(RT: related term)를 의미한다. 그러나, 연관관계 설정의 일관성 있는 기준 역시

<표 1> 용어 간 기본적 의미관계

관계 유형	구현된 관계용어	하위 범주	예
등가관계 (equivalence relationship)	우선어-비우선어 (USE/UF: Used for)	-	1. 감염증상-감염증 2. 줄기세포-간세포
계층관계 (hierarchical relationship)	상·하위어 (BT: broader term/ NT: narrower term)	속관계 (NTG: narrow term generic)	1. 염색체 > 19번 염색체 2. 뼈 > 가슴뼈
		전체부분관계 (NTP: narrow term partitive)	1. 혈관 > 혈관벽 2. 컴퓨터 > 메인보드
		사례관계 (NTI: narrow term instance)	1. 네트워크 > KT망 2. 인공위성 > 무궁화 위성
연관관계 (associative relationship)	관련어 (related term)	-	1. 쥐-쥐약-사람 2. 병원-의사-환자

결여되어 있다.

이상과 같이 전통적으로 시소러스 구축에는 위의 세 가지 용어 간 관계가 동원되었으나, 이 관계들만으로는 추론검색·확장검색과 같은 급격한 정보환경 변화를 충족시킬 시소러스 및 의미관계 세분화를 만족시킬 수 없다. 따라서, 정보검색 환경의 변화에 따른 시소러스의 구축 원칙과 준거가 필요하게 된다.

한편, 전산학분야의 시소러스 구축방법은 분포가설(distributional hypothesis)과 어휘 통사적 유형에 기반을 둔 분류학적 관계(taxonomic relations) 방법론에 입각하며, 통계기법을 활용한 자동구축이 주를 이룬다. 시소러스의 자동구축에 관한 국외의 연구로 Cimiano et al.(2004); Faure et al.(1998); Grefenstette(1994); Pereira et al.(1993) 등과 국내의 연구로 Ryu et al.(2006); 이창기; 이근배(1999); 임지희 외(2005) 등이 있다. Pereira et al.(1993)는 명사의 표지 미부착(unlabeled)의 계층구조 구축을 위해 하향식 클러스터링(top down clustering) 기법을 도입했으며, 엔트로피 기반(entropy based)의 평가방법을 제시했다. Grefenstette(1994)는 SEXTANT 체계를 이용한 시소러스 자동구축 방법을 도입했는데, 이 방법은 의미적 유사성을 지닌 용어들은 유사한 통사적 관계 속에 출현한다는 가정하에 출발하여, 각각의 용어는 출현가능한 문법적 문맥에 의해 그룹화 된다. 이 구축법은 실용적인 방법론이나, is a(일반화 관계) 또는 part of(전체 부분관계) 등의 특정관계를 구현할 수 없다는 단점이 지적된다. Faure and Nedellec(1998)은 유사한 문맥에 출현하는 명사들을 반복적 상향식 클러스터링(iterative bottom up clustering) 기법으로 구축하며, Cimiano et al.(2004)는 말뭉치로부터 추출한 개념의 계층구조의 자동구축방법을 제시했다. 형식적 개념분석(Formal Concept Analysis)에 기반을 둔 이 방법은 분포적 가설을 따랐으며, 용어의 문맥을 벡터로 구축하였다. 그러나, 분포적 가설에 의거한 자동구축은 근본적으로 앞서 언급한 데이터의 산발성을 해결하지 못한 취약점이 있다.

국내의 최근 연구로 류법모 외(2005); Ryu et al.(2006); 임지희 외(2005); [한의학 지식정보자원 시소러스] 등이 있다. 류법모 외(2005); Ryu, et al.(2006)은 〈분할 정복(divide and conquer)〉 구축법을 새로운 구축방법으로 제시하는데, 특정 분야 말뭉치로부터 용어의 자동추출, 추출용어의 전문분야 지식분류체계의 클래스 단위로 분류, 이렇게 분류된 용어를 대상으로 여러 개의 작은 시소러스들을 구축, 그리고 전 단계에서 얻은 작은 시소러스들을 하나로 다시 묶는 절차를 거친다. 본 구축법은 용어 추출, 용어 분류, 계층 구조 구축을 자동처리하여 시소러스 구축의 복잡도를 줄이고 기 구축된 시소러스를 재활용할 수 있다고 주장한다. 그러나, 이러한 방법으로 구축된 시소러스는 용어수가 1,000여 개 내외로 신뢰도의 문제, 본 시소러스에 이용된 분야분류체계인 Inspec와 한국어 용어와의 적합성 여부, 자

동구축이 갖는 계층구조 설정 시 발생하는 용어의 조어적, 의미적 문제, 검색의 성능보장에 관한 미검증 등의 여러 문제를 남기고 있는데, 이 점은 저자들도 지적한 부분이다(Ryu et al. 2006, 81).

임지희 외(2005)는 기 구축된 면역학전문용어사전, 의학용어사전, 표준국어대사전 등을 기반으로 면역학분야 시소러스의 간접구축 방법을 제안하고, 이렇게 구축된 시소러스를 이용하여 특정분야의 온톨로지 개발로의 확장성을 제시했다. 구축대상은 핵심용어와 관련용어(RT)를 포괄하며, 용어 간 의미관계 기술에 있어 어휘의미론적 관점에서 접근한다고 주장한다. 그러나, 실제 활용성은 미검증된 상태이다.

[한의학 지식정보자원 시소러스: http://jisik. kiom.re.kr/th/.]는 한의학 전체를 대상으로 하는 종합적인 데이터베이스 구축을 목표로 2004년부터 매년 3,000 디스크립터들을 대상으로 구축하고 있다. 구축하는 내용으로는 디스크립터 외에 범위주기(SN: scope note), 유의어, 관련어, 상위어, 하위어들이 있다. 그러나, 정영미 외(2002)나 기존의 시소러스들과 같이 관련어, 상위어, 하위어들에 대한 명확한 제약과 구축원칙을 제시하지 못하는 문제점을 지닌다. 특히, 관련어의 경우 약재명, 병명에 따라 필수적으로 포함되어야 하는 내용들이 일관성 없이 구축되어 있어 응용분야에서 개념들의 분류에 따라 자동 적용할 수 없다는 약점을 가진다.

2.3 패싯(facet) 도입의 필요성과 특성

앞 절의 관련 선행연구의 검토를 통해 시소러스 구축유형 및 방식의 장·단점을 주지하면서, 본 연구는 시소러스 수동구축 방식을 선택하고, 구축의 기본원칙으로 Nilsson et al.(2002)에서 제시된 다음 원칙을 따르기로 한다.

첫째, '주관적' 원칙으로, 이것은 동일개념에 대해 가질 수 있는 여러 관점을 수용하여 기술하는 것을 의미한다. 따라서, 개별용어 또는 관련 용어 간 의미관계를 명시적으로 세분화할 수 있게 된다. 이 점을 고려하여 본 연구에서는 개념패싯과 관계패싯을 개발, 시소러스 구축에 활용하고자 한다.

둘째, 용어의 '진화' 보장 원칙이다. 과학기술분야 용어는 과학기술발전의 속도에 비례하여 생성, 진화, 소멸함을 고려해야 한다. 기존의 과학기술분야 용어는 사전과 일반서적 등 출판에 상대적으로 긴 시간이 소요되는 검증된 문헌을 이용하여 선정, 구축되었다. 이 때문에, 시소러스 구축 이후 용어들에 출판물에 대한 용어들의 적용도(coverage)는 상당히 낮은 편이다(정한민 외 2005). 과학기술분야 용어는 용어생명주기(life cycle of terms)를 고려하여, 대용량 말뭉치에서 높은 적용도를 보이는 용어들을 선정하여 시소러스 구축의 대상으로 삼는 것이 시소러스 구축 완료 이후 효용성을 보장할 수 있다.

셋째, 용어의 '확장성'이 고려된 유연성 있는 구축이 요구된다. 용어 간 계층관계 구축은

조어적 기준과 의미적 기준에 의거하여 구축한다. 조어적 기준은 상·하위어간 형태적 유사성을 근거로 삼으며, 의미적 기준은 상·하위어 간 형태적 유사성은 없으나 의미적 상·하위어(hyper /hyponym)관계, 전체 부분어(holo /meronym)관계를 형성하는 경우이다. 상·하위어 간 계층관계 구축에 조어적, 의미적 기준을 고려하면 이미 구축된 계층관계에 새로운 용어를 추가, 확장하는 데 체계적인 일관성을 보장하게 된다.

특별히 본 연구에서는 용어 간 의미관계 명시를 위한 장치로 패싯개념을 도입, 이를 개발, 확장한다. 패싯은 S. R. Ranganathan이 도입한 '패싯분류(faceted analysis)' 라는 분석 합성적 방법론에 처음 출현한 개념이다(Grunenberg 2002). 시소러스 구축 시 대상 용어를 크게 대분류하고, 각 분류그룹 내에서 계층관계를 설정하게 되는데, 대분류의 기준이 되는 것이 바로 기본 패싯이다. Ranganathan은 패싯의 기본범주로 인간(personality), 사물(matter), 에너지(energy), 공간(space), 시간(time)의 다섯 가지를 제시했다. 여기서 패싯은 "어떤 부류(class)나 특정한 주제의, 분명하게 정의되고 상호 배타적이며 집단으로서 망라성을 띄는 양상들(aspects), 속성들(properties), 또는 특성들(characteristics)"로 정의될 수 있다(Taylor(1992). 재인용: Maple (1999)). 이처럼, 패싯은 용어 간 의미관계를 명시적으로 세분화할 메타 개념으로의 사용가능성을 보여주며, 이를 구축준거로 삼아 구축된 시소러스의 개발 및 확장은 용어 간 관계를 최대한 반영하여 효율적인 정보검색에 도움을 줄 것으로 예상된다. 그럼에도, 문헌정보학분야에서 정교한 패싯 개발과 범주화를 위한 체계적인 연구성과는 거의 전무한 상태이다. 패싯은 의미론과 전산언어학에서 개념과 어휘 기술에 이용하는 의미자질(semantic feature)과 공통점을 지닌다. 즉, 패싯과 의미자질은 분류 목적과 방식에서 유사성을 지닌다. 첫째, 용어(또는 개념)를 구성하고 대표하는 패싯과 의미자질은 이들을 규정하는 범주의 범위(category boundary)를 적절히 표상할 수 있는 장치이며 둘째, 패싯과 의미자질은 용어 각각 또는 용어 간 의미관계 기술을 위한 장치의 역할을 할 수 있다. 이러한 점은 의미자질에 착안한 몇몇 언어학적 접근 사례를 통해 확인될 수 있다(황순희, 윤애선 2005).

본 연구는 용어의 내재한 속성을 명시하기 위한 개념패싯(Category Facet)과 용어 간 관계를 명시하기 위한 관계패싯(Relational Facet)으로 나누고, 개발, 응용할 것을 제안한다. 관계패싯은 범주 관계패싯, 속성 키워드, 속성 관계패싯, 의미역 관계패싯으로 세분화된다.

3. 연구 범위와 내용

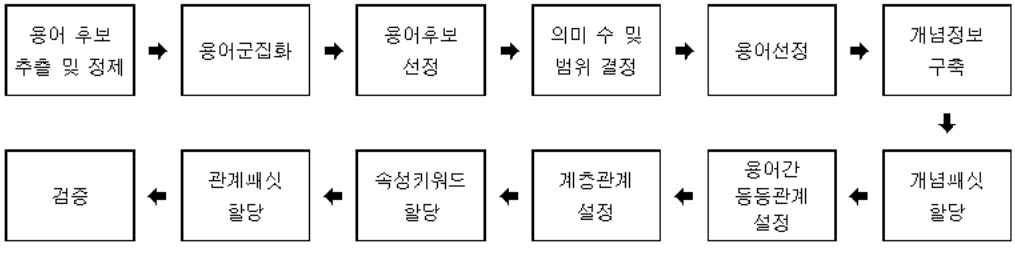
3장에서는 본 연구의 13개 과학기술분야(〈1. 전기전자, 2. 컴퓨터 공학, 3. 생물과학, 4. 기계공학, 5. 농수산, 6. 물리학, 7. 의약학, 8. 재료자원공학, 9. 토목건축공학, 10. 화학, 11. 화

학공학, 12. 수학, 13. 지구자원공학>) 시소러스 용어의 구성적 특성과 구축에 사용된 의미적 준거를 알아본다. 먼저 본 시소러스의 구축방법을 알아보고, 시소러스 용어의 구성적 특성을 감안한 언어학적 기준과 구축에 사용된 의미적 준거인 패싯들의 타당성과 실제 적용사례를 중심으로 논의한다.

3.1 시소러스 구축방법

본 시소러스 구축대상 용어는 대용량 말뭉치로부터 추출된 범용 과학기술분야 5만여 후보 용어 중, 1만5,000여 용어가 대상이다. 앞서 언급한 바와 같이 해당분야에서 가치가 증가하고 있는 전문용어를 우선시하고, 사용빈도가 높아지는 용어에 대해 가중치를 두어 순위화 한다. 본 연구에서 실행한 시소러스의 구축과정을 간략히 소개하면 다음과 같다. 시소러스 여러 방법으로 구축이 가능하며, 이에 관한 적합성 여부는 개별적인 연구과제이다. 시소러스 구축과정은 <그림 1>과 같이 요약된다.

먼저 특정 전문분야 말뭉치에서 등록할 용어후보를 추출하여, 주제적 또는 형태적으로 유관한 용어군을 만드는 용어의 군집화 단계를 거친다. 용어군에서 처리할 대상 용어를 선정하여, 다의성이 있는 용어와 그 의미를 파악한 후 그 수와 범위를 결정하게 된다. 이렇게 1차 정제된 용어를 시소러스 등록용어로 최종 선정한다. 개념정보 구축 단계부터는 실제 각각의 용어에 대한 세부 기술 단계로, 용어 각각에 대한 용례 · 범위주기 · 대응 외국어 등을 입력한다. 개념패싯 할당은 미리 선정된 개념패싯 집합에서 적절한 개념패싯을 선정된 용어 각각에 할당하는 단계이다. 동등관계 설정은 용어 간 우선어 비우선어 관계를 설정하는 단계로 USE/UF로 표현한다. 계층관계 설정은 용어 간 상 · 하위어를 표현하는 단계이다. 다음으로, 상 · 하위어 간 관계 표상을 위해 관계패싯 할당이 이루어지는데, 먼저 속성 키워드를 할당하며, 속성 관계패싯과 범주 관계패싯을 할당하며, 마지막으로 서술형 명사가 포함된 용어에 한하여 의미역 관계패싯을 할당한다. 끝으로 용어 전체에 대한 검증이 요구된다.



<그림 1> 시소러스 구축과정

3.2 전문용어의 특성을 고려한 언어학적 기준

앞 절에 소개된 구축방법과 병행하여, 전문용어의 구성적 특성을 고려한 언어학적 기준과 모델이 마련되어야 한다. 전문용어의 대부분은 복합어의 형태를 지니며, 단일용어 또는 복합용어의 여부는 용어 선정, 계층구조(hierarchical structure) 설정, 상·하위어 간 자질 상속(feature inheritance) 등의 문제와 밀접한 관계를 갖는다.

첫째, 어떤 용어를 전문용어로 판단할 것인가 기준이 필요하다. 본 연구에서는 다음의 경우를 전문용어로 인정하지 않는다. 동일한 의미를 지닌 용어의 형태적 반복(예: 'SARS 중증 급성 호흡기 증후군'), 동일하지 않은 개념의 반복(예: 'TV 오디오 에어컨'), 축약형태의 반복(예: '비타민 ADE'), 관련개념의 반복(예: '여성 호르몬', '에스트로겐') 등의 경우는 이를 전문용어로 선정하지 않는다. 또한 주관적 수식어구가 결합된 용어(예: '가해 아미노산', '최신식 시스템'), 상태·성질을 나타내는 형태소인 '형/용/성/식/적'이 추가된 용어(예: '최신형 컴퓨터', '신착계임'), 접속 조사 '와/과' 등이 연결된 용어(예: '밀리터리와 원적외선'), 관형격 조사 '의'가 포함된 용어(예: '표층수의 염분농도'), 출처나 원산지 표시 수식어가 포함된 용어(예: '중국산 소프트웨어'), '종류명+관련 고유명사'로 구성된 용어(예: '블랙홀 X선 관측위성 찬드라') 등은 전문용어로 인정하지 않는다. 반면, 수식어구가 추가된 경우지만 의미하는 범위나 지칭이 구체적이며, 전문용어로 굳어진 경우는 용어로 허용한다(예: 말기 간암, 악성 림프종, 왼쪽 마우스 버튼). 이 밖에 서비스나 기능을 나타내는 용어(예: 'MP3폰 재생제한'), PLO(Person Location Organization: 인명, 지명, 기관 및 단체명)관련 용어(예: '어플라이드 디지털'), PLO와 결합된 모델명, 브랜드(예: '삼성 하우젠') 등도 전문용어로 인정하지 않는다.

둘째, 시소러스 구축 시 용어 간 계층구조는 조어적(형태적) 유사성에 근거한 조어적 기준과 BT NT 간 의미적 기준에 근거한 계층구조 설정의 두 가지가 가능하다. 또한, 이들은 일반화, 전체부분, 사례의 세 가지 범주관계로 표현된다(본 논문의 범주관계 패싯). 가령, 용어의 조어·형태적 유사성만을 고려한 구축의 경우 BT '생태계'의 NT로 '자연생태계', '지구생태계', '갯벌생태계', '해조류생태계', '먹이생태계' 등이 동일한 층위(level)에 올 수 있지만, '생태계'의 NT들은 동일한 의미적 가중치(weight)를 갖지 않는다. 따라서, 이들 NT 간의 의미거리를 고려하여 계층관계를 재설정하여야 한다.

셋째, 원칙적으로 상위어의 자질은 하위어에 그대로 상속된다. '방사성 동위원소'의 의미자질은 직속 상위어인 '동위원소'에서 승계된다. 특정 용어는 두 개 이상의 직속 상위어에 의한 자질의 다중상속(multiple inheritance)이 가능하다. '무선 인터넷 접속'은 '인터넷 접

속'과 '무선접속'을, '심근'은 '심장'과 '근육'을 상위어로 갖는 다중상속의 경우이다. 논리적으로 상·하위어 관계에서 하위어는 상위어를 함의한다. 이것은 의미자질의 승계문제로, 상위어는 하위어에 비해 의미자질의 수가 적다. 언어 내적 관점으로는 대개 함의관계나 의미자질만을 상·하위어관계 기술에 동원하게 된다. 그러나 상위어가 모든 계층의 하위어를 포함하지 않는 경우도 가능하다. 가령, '저기압>열대성 저기압>허리케인'의 경우 모든 하위어는 상위어에 포함되지만 '생물>꽃>꽃술', '비행기>글라이더>행글라이더'의 자질승계는 부분적으로 충족된다.

3.3 의미적 준거와 특성

본 절에서는 용어 각각에 내재된 대표속성인 개념패싯의 특성과 종류를 기술하고, 용어간 의미관계를 표현하기 위한 장치인 관계패싯과 그 하위범주의 특성을 논의한다.

3.3.1 개념패싯(Conceptual Facet)

개념패싯은 용어가 갖는 대표적인 의미속성·범주를 의미한다. 본 연구에서는 최종 15개 개념패싯을 선정하였으며, 다음은 개념패싯명과 관련 패싯이 할당된 용어의 예이다. 개념패싯은 모든 용어에 비교적 쉽게 할당되나, 용어의 중의성 때문에 할당이 어려운 경우도 있다. 본 연구에서 설정한 15개 개념패싯은 연구관점, 연구자들 사이에 쉽게 합의를 보기 어려운 부분으로 이를 위해 본 연구는 다음의 절차를 거쳐 개념패싯을 설정하였다. 먼저 기존의 문헌정보학에서 설정한 20여 개 개념패싯을 전체 용어 중 15% 내외 용어에 할당해보고, 할당결과, 설정된 개념패싯의 잉여성 또는 불충분성 등을 고려하여 개념패싯을 재조정하였으며, 이와 동시에, 1차 선정된 개념패싯의 타당성 여부를 검증하기 위해, 어휘의미망의 대표적 사례인 프린스턴 워드넷에 사용된 상위개념(Upper Concepts)과 SUMO의 상위 온톨로지를 참고하였다. 즉, 상위개념을 따로 추출하여 이들로 해당 용어의 기술이 가능한 지 검토하고, 1차 선정된 개념패싯과 이를 비교하였다. 마지막으로, 용어 DB 자체의 실증적 분석 및 워드넷 상위개념을 이용하여 적은 수의 개념패싯으로 많은 수의 용어를 효율적으로 기술할 수 있도록 개념유형을 조정하였으며, 최종 15개 개념패싯을 선정하였다(<표 2>).

3.3.2 관계패싯(Relational Facet)

관계패싯은 용어 간 또는 상·하위어 간의 의미관계를 표현하는 메타 개념(Meta Concept)으로, 속성 관계패싯, 범주 관계패싯, 속성 키워드, 의미역 관계패싯의 네 가지 유형으로 세분될 수 있다. 기존의 시소러스가 용어간 의미관계를 충분히 반영하지 않고 상·하위어를 단순히 나열·연결하는 수준에 그친 반면, 관계패싯을 도입하면 용어 간 의미관계의 규명 및 세분화가 가능해 진다는 장점이 있다.

1) 속성 관계패싯(Attribute Relational

〈표 2〉 개념패싯의 종류와 용어

개념패싯명	개념패싯이 할당된 용어
감각/감정(feeling/sensation)	* 개념패싯 '감각 · 감정'은 (2005년 12월 말 현재) 구축분에서 할당된 예가 없으나, 시소러스 확장에 따라 필요한 개념으로 판단, 현재까지 유지되고 있다.
공간/위치(location/space)	게이트웨이, 광대역 통신망, 네트워크, 나선은하, 별자리
기기/장치/부속(equipment/apparatus)	고화질카메라, 양자컴퓨터, 인터넷전화, 지능형로봇, PC화면, 광케이블
상태/성질(property/attribute)	고해상도, 고혈압, 계수, 광년, 디지털
시간(time/age)	고생대, 중생대, 석탄기, 백악기, 트라이아스기 말기
생물(organism/living thing)	게임개발자, 게이머, 고지혈증환자, 네티즌, 기생충, 나비, 바이러스
분야/이론/방법(field/theory/method)	면역요법, 시험관 아기시술, 심폐소생술, 알고리듬
물질/재료(material)	간염백신, 칼륨, 고체연료, 글로불린, 에너지, 유해물질, 입자
조직(tissue/body/part)	뇌, 말초신경, 제대혈, 줄기세포, 말초신경계, 뇌혈관계
언어(language)	프로그래밍 언어, C 언어
단체(group/organization)	개인 커뮤니티, 유저 커뮤니티, 온라인 커뮤니티, 웹 커뮤니티
질병/증상(disease/symptom)	가려움증, 결막염, 관절염, 난청, 뇌중풍, 부인암, 여성 탈모, 요통, 전염병, 혈액암
내용(contents)	검색엔진, 에디터, 온라인게임, 온라인결제시스템, 3D게임, 토털설루션
행위(action)	데이터전송, 인간복제, 전자통신결제, 경주, 골수기증, 광고, 광대역접속
현상/사건(phenomenon/event)	고기압, 구름, 기류, 기온, 기후, 단층지진, 모래폭풍

Facet)

속성 관계패싯은 상 · 하위어 간 의미관계를 기술하는 방법으로 상위어에서 하위어를 본 관점을 뜻하는 '메타개념'이다. 관점이란 주관적이며, 임의적이어서 개별 어휘에 내재된 자질 또는 속성과는 확연히 구별되는 어휘 독립적인 속성이다. 속성 관계패싯은 개념패싯과 동기화하며, '사례' 관계를 하나 더 추가하여 총 16개 속성 관계패싯이 개발, 사용되었다.

2) 범주 관계패싯(Category Relational Facet)

범주 관계패싯은 어휘의미망 성립의 제1원칙인 상 · 하위어 간의 자질승계와 그 양상을 확인할 수 있는 장치이다. 또한, 용어 각각에 범주

〈표 3〉 범주 관계패싯의 종류와 용어

범주 관계 패싯명	세부설명	범주 관계패싯이 할당된 용어
전체부분관계 (whole-part relation)	1. 사물 전체와, 그 전체를 구성하는 부분을 중심으로 세분화함. 2. 일반적으로 '해당용어가 상위어의 부분/일부이다.' 라는 test를 통해 문장이 성립하면 이를 전체부분 관계로 간주함. 3. 본 연구 대상인 과학기술분야 시소러스는 그 적용범위를 '생체의 조직과 기관' 에 해당하는 것만으로 국한하여 사용함.	1. 귀(BT) 〉 내이(NT) 2. 귀 〉 중이 3. 내이 〉 반고리관
사례관계 (instance relation)	1. 상위어에 대해 하위어가 특정한 사례로 예시됨. 2. 하위어는 고유명사로 구현되며, PLO(인명, 기관 · 단체명, 지명, 제품명, 상품명)가 이에 해당함.	1. 네트워크 〉 KT망 2. 우주왕복선 〉 챌린저호 3. 달 착륙 우주선 〉 루나 9호
일반화관계 (is-a relation)	1. 일반화 관계는 위에서 정의한 전체부분관계, 사례관계 이외의 모든 관계를 이름. 2. 일반적으로 '해당용어는(은/이/가) 상위어의 일종이다.' 라는 test를 통해 의미가 성립하는 관계. 3. 특징적으로 전체부분 · 일반화 관계가 공통으로 성립하는 상 · 하위어는 일반화 관계로 판단함.	1. 감각수용체 〉 미각수용체 2. 악성종양 〉 간암 3. 금속 〉 비철금속

관계패싯을 할당하면, 용어 간 계층구조 설정의 타당성 여부와 계층구조 설정 시 사용된 기준이 용어 간 형태적 유사성에 입각한 조어적 기준에 의한 것인지, 또는 의미적 기준에 의한 구축인지 확인이 용이하다. 범주 관계패싯으로 다음 3가지를 설정하여, 이를 모든 상 · 하위 관계에 있는 용어에 할당한다(〈표 3〉).

3) 속성 키워드(Attribute Keyword)

상위어 '전화기' 와 그 하위어들 간 관계 기술방법을 보면, ① '전화기' 와 '휴대폰' 의 관계는 속성 관계패싯 '용도' 로, ② '휴대폰' 대신 동의어 '휴대 전화기' 가 되면 속성 관계패싯은 '행위' 로 바뀌며, ③ '휴대폰' 대신 '휴대용 전화기' 가 되면 속성 관계패싯은 '상태' 로 달라진다. 또한, 상위어가 '전화기' 대신에 동의어 '폰' 으로 바뀌면 그 하위어에 대한 속성 관계패싯 역시 달라진다. 즉, 용어의 어떤 동의어를 보느냐에 따라 관계패싯이 달라져 일관성을 유지할 수 없고, 검색의 기준에도 일관성이 결여되는 결과를 낳는다. '속성 키워드' 는 관계패싯의 일관성유지를 위해 도입한 동의어의 중심어로 정의될 수 있다. 이것은 일반적으로 시소

러스는 색인작성과 검색에의 효율성보장을 위해 특정 용어의 '대체용어 제시'를 기본원칙으로 삼기 때문이다. 즉, 용어 간의 유연성 있는 동의관계가 설정되고 있으며, 이를 우선어(descriptor: 대표표현)와 비우선어(non descriptor: 이형) 또는 USE/UF로 표현한다. 가령 '휴대폰'이 USE이면, 동의어로 간주되는 '셀룰러 폰', '핸드폰', '핸펀', '휴대전화기' 등이 UF에 해당될 수 있다.

위의 예에서 '전화기'와 '휴대폰'의 동의어군에 대해 관계패싯을 부여하는 중심어를 '휴대용'으로 결정하면, '휴대폰', {셀룰러 폰, 핸드폰, 핸펀, 휴대 전화기, 휴대용 전화기}를 모두 '휴대용' 측면에서 본다는 의미가 되며, 따라서 동의어를 가진 모든 용어를 하나의 관점에서 기술한다는 일관성을 유지할 수 있게

〈표 4〉 의미역 관계패싯의 종류와 용어

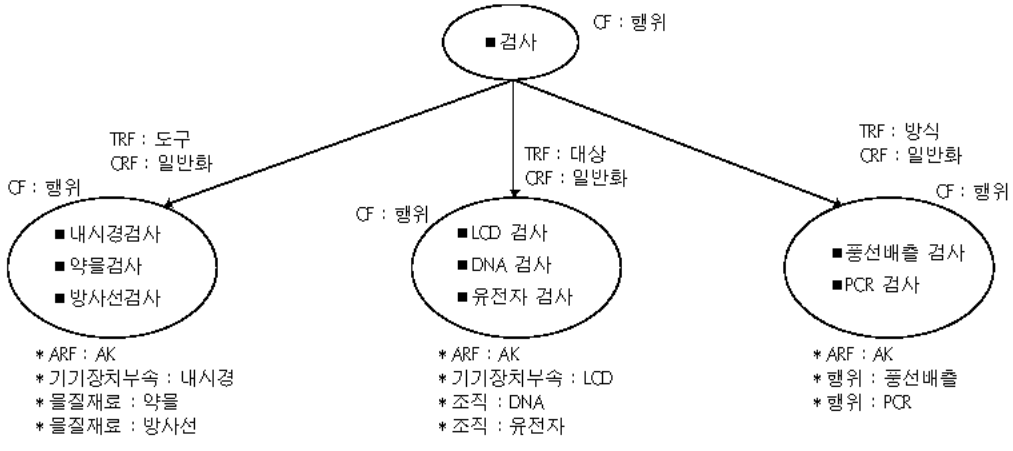
의미역 관계패싯명	세부설명	의미역 관계패싯이 할당된 용어
근원(source)	원인, 기원, 출처, 유래, 발단	1. 복사 > 태양복사 2. 감염 > 수혈감염
대상(object)	목적 대상	1. 검사 > 기형아 검사 2. 감염 > 광우병 감염
도구(instrument)	도구	1. 검사 > 방사선 검사 2. 농축 > 레이저 농축
시간(time)	시간성	1. 중계 > 실시간 중계 2. 스트리밍 > 실시간 스트리밍
목표(goal)	목적, 용도, 목표	1. 치료 > 재활치료 2. 수술 > 교정수술
장소(location)	공간, 위치, 장소, 지형, 지명, 지형, 천체	1. 경매 > 인터넷 경매 2. 네트워킹 > 홈 네트워킹
행위자(agent)	동작주, 행위의 주체	1. 폭발 > 태양폭발 2. 융합 > 핵융합
수혜자(beneficiary) * TRF '수혜자'는 '대상'과 중복할당의 경우가 많으며, '침식'의 두 가지 경우만 예시한다.	행위의 수혜자	1. 침식 > 모래침식 2. 침식 > 토양침식
방식(manner)	방법, 방식	1. 수술 > 이식수술 2. 운동 > 대류운동

된다.

4) 의미역 관계패싯(Thematic Relational Facet)

전문 용어를 구성하는 어휘(대부분 복합어)의 핵심어가 서술형 명사인 경우, 상·하위어 간 차이를 의미적으로 규명할 수 있도록 '의미적 논항관계'를 이용한 의미역 관계패싯의 도입이 필요하다. 전문용어와 직접적인 관계가 있는 명사의 경우도 동사와 마찬가지로 논항을 지닐 수 있다(이선웅 2004, 153). 가령, 행위성을 지닌 명사 '검사'는 검사의 행위주 또는 대상을 나타내는 어휘를 선택제약으로 요구하고 행위성이 없는 용어 '백신'은 무엇과 관련된 백신인 지(예: 컴퓨터 바이러스 백신, 조류독감 백신)를 명시할 수 있는 논항이 요구된다. 동사의 논항구조에 관해 세부적인 연구가 이루어진 반면, 명사의 논항에 관한 연구는 거의 전무한 상태이다. 이 점은 근본적으로 동사가 논항을 어떻게 실현시키느냐는 문장형식을 결정하지만, 명사가 논항을 어떻게 실현시키느냐는 명사구의 형식을 결정하는 것에 그치기 때문에 그 중요성이 덜 인식되었던 데 기인한다. 또한 의미역 관계패싯 할당은 기존의 의미역 관련 연구를 바탕으로 전산 언어학적 관점에서 활용한 연구(강신재, 박정혜 2003)를 참고할 수 있으나, 기존의 '행위주, 경험주, 수령주' 등의 구분과 구문관계에 따른 의미역 할당문제는 여전히 분명하지 않다.

대개, 전문용어는 기 구축 용어를 바탕으로 복합어형태의 신조어생성 및 이용으로 구축되는데, 이 경우 의미역 관계패싯은 용어 간 관계명시에 중요한 지표가 될 수 있다. 또한, 복합어형태의 전문용어는 특정 전문분야의 관점에 따라 다의적 의미해석이 가능하므로 용어를 분명하게 명시해야 한다. 이 경우 의미역 할당 역시 중복적으로 일어날 가능성이 크고, 그 판단 또한 용이하지 않다. 특정분야의 전문용어(technical terms)는 일상어휘(ordinary



〈그림 2〉 관계패싯을 이용한 BT '검사'와 NT들 간의 의미관계

terms)에 비해 원칙적으로 중의성 또는 '다의성을 전혀 지니지 않는 것이 특징'이라는 지적(최기선, 송영빈 2000, 111)과 달리, 실제로 상당수의 전문용어가 다의성을 지닌 것으로 확인된다. 전문용어는 지금까지 존재하지 않은 개념을 인위적으로 표현할 목적으로, 기존의 용어를 이용하여 복합어 구조로 재구성한 신어라는 특징과도 밀접한 관련이 있다. 따라서, 용어의 의미세분화(sense distinction)는 절실히 요구된다.

본 연구에서는 2개 이상의 의미역 관계패싯 할당을 허용하여, 이를 서술형 명사가 포함된 복합어 용어의 의미세분화 기준으로 적용한다. 2개 이상의 의미역 관계패싯 할당이 가능한 경우는 가령, '적외선 탐지기', '인터넷 검색' 등으로, '적외선을 이용하여 제 3의 대상을 탐지하는 기기' 또는 '적외선을 탐지하는 기기'의 의미로, '인터넷상의 검색' 또는 '인터넷을 이용한 검색' 등의 두 가지의 해석이 가능한 경우이다. 따라서, 의미적으로 큰 차이를 초래하지

〈표 5〉 패싯별 구축용어 통계

개념패싯	해당 용어 수	범주 관계패싯	해당 용어 수	의미역 관계패싯	해당 용어 수	속성 관계패싯	해당 용어 수
감각 · 감정	-	조어적(일반화)	11,697	도구	92	감각 · 감정	67
기기 · 장치 · 부속	3774	의미적(일반화)	2,263	대상	891	기기 · 장치 · 부속	878
내용	2706	의미적(사례)	74	수혜자	2	내용	1,074
단체	22	조어적(전체부분)	30	방식	271	단체	38
물질 · 재료	2620	의미적(전체부분)	9	장소	40	물질 · 재료	1,680
분야 · 이론 · 방법	348	조어적(사례)	289	시간	5	분야 · 이론 · 방법	234
상태 · 성질	847	총계	14,362	행위자	71	사례	414
생물	338			목표	48	상태 · 성질	5,813
시간	9			근원	15	생물	418
언어	18			총계	1,435	시간	106
위치 · 공간	752					언어	34
조직	896					위치 · 공간	640
질병 · 증상	1,290					조직	1,228
행위	1,782					질병 · 증상	752
현상 · 사건	124					행위	864
총계	15,526					현상 · 사건	76
						총계	14,316

않으나 다의적 해석이 가능한 경우는 복수 개의 의미역 관계패싯 할당을 허용한다. 또한, 의미역 관계패싯 '수혜자'의 경우, '수혜자' 이외의 의미역 관계패싯의 할당도 가능하다. 다음은 의미역 관계패싯의 종류와 예이다(<표 4>).

관계패싯을 이용한 BT '검사'와 NT들 간의 관계는 <그림 2>와 같다.

3.4. 구축결과

<표 5>는 의미적 준거를 적용하여 시소러스를 구축한 결과로 패싯별 구축용어 통계이다(2005년 12월 말 구축결과 기반).

개념패싯 중 가장 많이 할당된 것을 살펴보면 기기 · 장치 · 부속, 내용, 물질 · 재료의 순이다. 대표적 의미속성인 개념패싯으로 '기기 · 장치 · 부속'이 할당된 용어는 '인터넷전화', '지능형로봇', '광케이블' 등이며, 온 · 오프라인상 콘텐트에 해당되는 '내용'을 할당받은 용어는 '검색엔진', '온라인 게임', '에디터' 등이다. 한편, 개념패싯 '물질 · 재료'는 '칼륨', '고체연료', '입자' 등에 할당된다. BT NT 간 의미관계인 속성관계 패싯으로는 '상태 · 성질', '물질 · 재료', '조직'의 순으로 할당되었다. 또한, 의미역 관계패싯은 핵심어와 수식어 간의 관계가 '대상'으로 파악된 것이 가장 많았다.

4. 과학기술분야 시소러스의 활용방안

2장과 3장에서 살펴본 바와 같이, 패싯을 도입한 시소러스 구축은 다양한 관점을 반영하고, 일관성 있는 구축을 지원하는 등의 장점이 있다. 4장에서는 이러한 장점을 활용하여 과학기술 응용분야에 실제 적용한 시소러스 구축사례와 앞으로의 적용방향을 검토한다. 이를 위해 기 구축된 국내 시소러스들의 대표적 사례와 활용상의 제약점을 검토하고, 과학기술 연구자들의 협업을 지원하기 위한 정보유통 서비스상에서의 본 시소러스의 활용사례와 향후 개발될 지식 서비스상 패싯의 활용방안을 논의한다.

4.1 시소러스 구현사례와 문제점

국내의 과학기술분야 시소러스 구현사례로 원자력분야 용어를 대상으로 한 정영미 외(2002)가 있다. 본 시소러스는 특정 용어의 관련어(RT)에 대해 개념패싯을 도입하여, 시소러스를 온톨로지화 하기 위한 기반을 제공하였다는 데 의의가 있으나, 개념패싯이 관련어에만 도입되어 활용가능성에 대한 제약을 가져왔고, 관련어와의 관계에 대한 기본적인 제약을 포함하지 않아 시소러스 구축범위에서 디스크립터에 따른 비일관성이 수반된다는 지적을 받는다.

반면, 본 시소러스에 도입, 활용한 개념패싯은 일관성 있는 확장이라는 측면과 구축된 시소러스의 오류를 최소화한다는 측면에서 중요한 의미를 지닌다. 단적인 예로, 본 시소러스를 구축하는 과정에서 개념패싯을 자동검증(상위

개념어와 하위 개념어는 개념패싯을 공유해야 한다는 제약을 적용)하지 않고 구축한 결과, 구축된 1만5,000여 개 용어에 대해 112개가 오류로 파악되어 약 0.75%의 오류율을 보인다. 시소러스의 크기가 커지고 분야가 다양해질수록 오류범위는 증가하게 되므로, 개념패싯과 같은 제약이 더욱 필요해진다. 또한, 관계패싯을 이용한 시소러스 구축은 국내외에서 본 연구가 유일하며, 4.3절의 향후 활용방안에서 그 의미와 역할을 살펴본다.

4.2 시맨틱 웹 기반 추론 서비스에서의 활용

현재와 같이 정보가 대량으로 무절제하게 생성, 유포되는 환경에서 지식 관리 시스템(KMS: Knowledge Management System) 또는 콘텐트 관리 시스템(CMS: Contents Management System) 등을 통한 체계화된 정보유통의 중요성은 크게 부각되고 있다. 그럼에도, 정보를 어떻게 지식화(Knowledgization)하고 검증할 것인가에 관한 연구는 정보유통 이외의 분야, 특히 시맨틱 웹 관련 분야에서만 독립적으로 이루어져 왔다. 이러한 연구영역 간의 분리는 결국 시맨틱 웹 기반 서비스들에서 정보의 생성, 저장, 관리, 검색, 유통의 동적 변화를 고려하지 못하게 만들었다. CAS, SEAL과 같은 대표적 서비스들도 배치작업을 통해 생성된 정보들만을 다룸으로써 동적 정보에 기반을 둔 정보유통 플랫폼에 바로 적용하기에는 어려움이 있다.

이에 본 연구의 응용분야로서 정보의 흐름을 관리하는 정보유통 플랫폼상에서 지식화와 추론 서비스를 포함하는 지식 서비스를 제공할 수 있도록 시맨틱 웹 기술을 접목하고자 시도하였다. 온톨로지를 통해 메타데이터를 지식화하고, 이들을 이용하여 〈연구자 네트워크(COP: Communities Of Practice)〉, 〈연구자 추적(Researcher Tracing)〉, 〈연구 맵(Research Map)〉의 세 가지 연구자 간 협업 지원 추론 서비스를 제공하게 되는데, 이들 서비스에서 연구주제에 대한 확장검색을 위해 본 연구에서 제안한 시소러스를 이용하였다.

한편, 국가과학기술 R&D(Research and Development) 기반정보의 핵심인 '과제', '출판물', '지적재산권' 등의 실 데이터 수집을 위해 한국과학기술정보연구원(KISTI)의 내부 성과물을 이용하였다. 또한, 이들 R&D 기반정보들 사이의 유발성과 연계성을 고려할 수 있도록 상호관계 정보를 데이터베이스 스키마를 분석하여 추출한다. 최종적으로, 국가과학기술 R&D 기반정보의 온톨로지 기반 추론 서비스 시나리오를 통해 과제중심의 추론 서비스와 연구성과 차원(출판물과 지적재산권 중심)의 추론 서비스를 충족시키는 구축 결과물을 도출하였다(〈표 6〉).

국가과학기술 R&D 기반정보 온톨로지(Science&Technology R&D Reference Information Ontology: 이하 RI 온톨로지)는, KISTI에서 관리하고 있는 국가과학기술 R&D 기반정보를 분석·활용하여 한국의 국가과학

〈표 6〉 내부 성과물 관리 데이터베이스의 주요 통계정보 예시

과제/출판물/지적재산권 Record 개수 - 과제 : 393개 - 출판물 : 1591개 - 지적재산권 : 325개	과제/출판물/지적재산권의 수량적 관계 분석 - 과제: 출판물, 과제: 지적재산권의 관계는 1: n - 과제 (A)/출판물 (B)/지적재산권 (C)의 집합 개수 ○ A ∩ B ∩ C - 33개 ○ A ∩ B - 1095개 (논문의 수), 148개 (과제의 수) ○ A ∩ C - 177개 (지적재산권의 수), 37개 (과제의 수)

기술 R&D 관련 정보를 과학기술 전문가를 비롯한 관리자, 일반 사용자 등에게 효율적으로 제공할 수 있도록 체계화시킨 OWL 기반의 온톨로지이다. RI 온톨로지는 시맨틱 웹 환경에서의 과학기술 기반정보에 대한 자료저장구조 및 관리체계를 반영하고, 과학기술 통합 정보 유통 서비스의 효율성을 고려한 온톨로지로, 시맨틱 웹 기술 기반의 추론 서비스에 활용된다. 본 연구에서는 온톨로지 기반 추론 기능을 실제 응용분야에 활용하기 위해 〈연구자 네트워크〉, 〈연구자 추적〉, 〈연구맵〉의 세 가지 추론 서비스 시나리오를 작성하고, 이 시나리오에 적합한 온톨로지 기반 추론 서비스 시스템을 구현하고자 시도하였다. 이는 RDQL(RDF Data Query Language) 기반의 질의 언어를 비롯한 추론규칙, 추론 엔진 등과 같은 추론 서비스 방법의 명확성을 위해 선행되어야 할 작업으로 판단된다.

4.3 개발된 패싯의 활용방안

앞에서 우리는 시소러스의 주된 활용이 정보검색 중 특히 '확장검색'이며, 그 효율성보장을 위해 BT NT 간 상하위 계층관계 및 다양한 패싯을 활용할 수 있음을 언급하였다. BT NT 간 계층관계는 설정 및 적용이 상대적으로 용이한 반면, 용어 간 RT는 체계적인 설계와 구축에 근거하지 않았기 때문에 적용에도 상당한 부담이 따랐다. RT는 온톨로지의 다양한 관계들을 반영할 수 있으나, 응용분야나 시나리오에 의존적이어서 단독으로 구축하는 경우 효용성이 떨어질 수밖에 없다. 본 논문에서 도입, 개발한 패싯은 기존의 RT와는 구분되며, 엄격히 정해진 범위 내에서 여러 종류로 기술되기 때문에 활용분야에 따라 필요한 패싯을 선택할 수 있는 장점이 있다. 패싯의 추후 활용방안은 다음과 같다.

첫째, 확장검색에서 '선택제약(Selectional Restriction)'으로 응용, 활용될 수 있다. 단순한 BT NT 계층 관계만을 사용한다면 확장검색을 하거나 하지 않거나 두 가지 중 하나를 선택해야만 한다. 반면, 선택제약으로 사용하는 경우에는 확장검색을 하더라도 특정한 NT들로만 확장을 할 수 있게 된다. 예를 들어, 질의응

답 시스템에서 "가속기를 형태로 구분하는 경우에 어떤 가속기들이 여기에 해당하는가?"라는 질문에 대해 '선형가속기', '원형가속기', '초대형 가속기' 등과 같이 속성 관계패싯이 '상태·성질'이며, 속성 키워드가 '~형'을 가지는 하위 개념어들로만 확장하여 정답을 찾을 수 있는 반면, "가속기가 다루는 물질에 따라 분류한다면 어떤 가속기들이 여기에 해당하는가?"라는 질문에 대해서 '방사광 가속기', '양성자 가속기', '이온 가속기' 등과 같이 속성 관계패싯이 '물질·재료'인 하위 개념어들로만 확장하여 정답을 찾을 수 있다.

둘째, 속성 키워드를 이용한 정답 제시에 활용할 수 있다. BT NT 계층 관계에서 상위 개념어들과 하위 개념어들 사이에 여러 관계패싯들과 속성 키워드를 가지는 본 시소러스 구조는 질의응답 시스템에서 정답추출 목적으로 활용할 수 있다. 예를 들어, '디스크'의 하위어들로 '만성디스크', '목디스크', '척추디스크', '퇴행성디스크' 가 있는 경우, "디스크가 올 수 있는 부위는 어디인가?"라는 질문의 질의 분석을 통해 부위를 신체 조직으로 해석하고 '조직'을 속성 관계패싯으로 가지는 하위 개념어들에서 속성 키워드를 추출한다면, '허리', '목', '척추' 등을 정답으로 얻을 수 있다. 기존 정보검색의 영역을 넘어서는 복잡한 질의를 처리해야 하는 응용분야에서는 패싯의 역할이 이처럼 더욱 중요해질 수 있다.

셋째, 패싯은 서로 다른 분야에서 구축한 시소러스들을 병합하기 위한 기제로 활용될 수 있다. 기존 시소러스들은 상위 개념어와 하위 개념어 간의 관계 형성방식이나 근거가 시소러스에 선언되어 있지 않기 때문에, 같은 개념어를 서로 다른 시소러스들이 공유하고 있을지라도 그 하위 개념어들과의 의미적 거리나 계층 관계 설정 범위가 다를 수밖에 없다. 시소러스들이 각 분야에서 독립적으로 구축되고 있는 현 상황에서 ISO와 같은 표준조차 어떻게 관계를 설정해야 하는 지와 어떤 제약을 가해야 하는지를 제시하지 못한다. 다양한 패싯들을 포함하는 서로 다른 분야에서 만들어진 시소러스들을 통합하는 것은 제약들을 매칭하는 과정을 통해 보다 원활히 이루어질 수 있다. 가령, 전기·전자분야의 '검사', '반도체 검사', 'LCD 검사'와 의학분야의 '검사', '내시경 검사', '뇌파 검사'를 병합할 때, 속성 키워드에 의해 의미적 거리가 이미 검증되어 있으므로 '검사'의 하위 개념어로서 '반도체 검사', 'LCD 검사', '내시경 검사', '뇌파 검사' 등을 둘 수 있다. 또한, 다른 분야의 'LCD 검사', '내시경 검사'가 속성 관계패싯에 의해 '기기·장치·부속'의 동일한 속성을 갖는다는 사실도 쉽게 확인할 수 있다. 이처럼 분야 간 시소러스 병합은 기반 지식자원의 일관성 있는 확장 측면에서도 효율성을 보장받을 수 있다. 향후 본 연구를 통해 얻은 패싯 기반 시소러스는 위에 제시된 다양한 활용가능성을 염두에 두며, 지속적인 확장이 이루어질 전망이다.

5. 결론 및 향후 연구

본 연구에서는 시소러스 구축 시 직면하는 문제점과 구축방법을 비판적으로 검토하고, 여러 구축방법의 하나인 수동구축 방법을 제안하였다. 또한, 확장검색의 효율성을 보장할 수 있는 시소러스 구축을 위해 의미적 준거인 개념패싯, 관계패싯 등의 도입과 활용을 제안하였으며, 이를 적용하여 구축한 과학기술분야 시소러스의 사례를 제시하였다. 패싯이 반영된 시소러스에는 다양한 관점이 반영되어 검색의 효율성을 보장한다는 장점이 있으며, 인접 과학기술분야에 응용될 수 있다. 일례로 본 시소러스는 과학기술 연구자들의 협업을 지원하기 위한 정보유통 서비스에 응용될 수 있으며, 향후 고도화된 지식 서비스에도 확장 응용될 수 있다.

본 연구에서 제안한 패싯의 장점은 다음으로 요약된다. 첫째, 확장검색에서 '선택제약'으로 활용될 수 있으며, 특히 질의응답 시스템에서 보다 적절한 정답을 제시하는 데 기여할 수 있다. 둘째, 패싯의 한 가지 유형으로 확장 개발된 속성 키워드는 BT NT 간 속성 관계패싯의 효과적 추출에 변수로 작용하여, 검색의 효율성보장에 도움이 된다. 셋째, 패싯은 서로 다른 분야에서, 각기 다른 방식에 의해 구축된 시소러스들을 하나로 병합하는 데 중요한 기제로 활용될 가능성이 있다. 이것은 기존의 시소러스에는 BT NT 간의 관계 형성방식이나 근거가 체계적, 일관적으로 선언되지 않았다는 점과 직결된다. 따라서, 동일한 개념의 동일한 용어가 상이한 시소러스에서 구축된 경우, 이들 용어에 동일한 속성이 공유되어 있음을 규명할 방법이 없는데, 패싯이 도입되면 이 부분의 투명성이 밝혀질 가능성이 있다.

본 연구는 다음의 추후 연구과제를 남기고 있다. 첫째, 현재 구축된 1만5,000여 용어는 상·하위어 간 조어적(형태적) 유사성에 의한 기준이 우선적으로 적용 구축되었는데, 의미적 기준이 먼저 적용된 구축의 경우 문제점이 검토되지 않았다. 이것은 용어 또는 개념이 지닌 의미자질의 승계문제와 직결된다. 현재까지는 과학기술 13개 분야의 대용량 말뭉치로부터 용어가 추출되었으므로, 부정합의 문제는 발생하지 않는 것으로 확인되었다. 둘째, 본 연구에서 제안, 활용한 메타 개념인 패싯의 지속적인 개발이 수반되어야 하며, 이를 기반으로 구축된 시소러스는 인접분야 온톨로지 기반 추론 서비스에 효과적으로 활용될 수 있을 것이다.

참고문헌

한국과학기술정보연구원. 2005. 범용과학기술분야 전문용어에 대한 계층적 개념망/어휘망, 15,000여 구축용어, 서울: 한국과학기술정보연구원(KISTI).

강신재, 박정혜. 2003. 대규모 말뭉치와 전산언어 사전을 이용한 의미역 결정 규칙의 구축. 『한국정보처리학회 논문지』, 10(2B): 219 228.

류법모, 김재호, 최기선. 2005. 정보산업 분야 시소러스의 공학적 구축 방안, 『제17회

한글 및 한국정보처리학술대회 논문집』, 13 20.

양재군, 배재학. 2002. 온톨로지 정보를 이용한 범주 재편성: Roget 시소러스의 경우. 『제17회 한국정보처리학회 춘계학술발표대회 논문집』, 9(1): 515 518.

이선웅. 2004. 명사구 사전의 통사 정보 기술에 대하여. 『한국사전학』, 4: 153 184.

이재윤, 김태수. 1998. WordNet과 시소러스. 『언어정보개발연구』, 203 237.

이창기, 이근배. 1999. WordNet을 이용한 한국어 시소러스 자동구축, 『제11회 한글 및 한국정보처리학술대회 논문집』, 156 161.

임지희, 최호섭, 배영준, 옥철영, 최성필, 성원경, 박동인. 2005. 면역학 시소러스 및 온톨로지 구축. 『제17회 한글 및 한국정보처리학술대회 논문집』, 21 27.

정영미, 김명옥, 이재윤, 한승희, 유재복. 2002. 과학기술 분야 통합 개념체계의 구축 방안 연구, 『한국정보관리학회지』, 1013 0799, 19(1): 135 161.

정한민, 구희관, 이병희, 성원경. 2005. 효율적인 자원 운영을 위한 전문용어 생명주기 관리 연구. 『한국컴퓨터종합학술대회』, 32(1B): 457 459.

최기선, 송영빈. 2000. 『전문용어연구 2』, [Korterm: 전문용어언어공학연구센터].

한상길. 1999. 『시소러스 용어관계의 확장에 관한 연구』. 박사학위 논문, 중앙대학교 대학원, 문헌정보학과

황순희, 윤애선. 2005. 의미적 준거의 세목화를 고려한 과학기술 분야 시소러스 구축. 『한국어의미학』, 18: 99 124.

Brewster, Ch., Wilks, Y. 2004. "Ontologies, Taxonomies, Thesauri: Learning from Texts." in *Proceedings of the Use of Computational Linguistics in the Extraction of Keyword Information from Digital Library Content Workshop*. [cited 2006. 5]. 〈http://www.kcl.ac.uk/humanities/cch/ake/final/redist/pdf/brewster.pdf〉.

Buscaldi, D., Rosso, P., Arnal, E.S. 2006. "WordNet as a Geographical Information Resource." in *Global WordNet Conference 2006 Proceedings*, Jeju Island, Korea, 37 42.

Cimiano, P., Hotho, A., Staab, S. 2004. "Clustering Ontologies from Text." in *Proceedings of the Conference on Lexical Resources and Evaluation(LREC)*, Artipol, Portuga, 1721 1724.

Faure, D., Nedellec, C. 1998. "A Corpus-Based Conceptual Clustering Method for Verb Frames and Ontology." in *Proceedings of the*

LREC Workshop on Adapting lexical and corpus resources to sublanguages and applications, Granada, Spain, 1 8.

Grefenstette, G. 1994. *Explorations in Automatic Thesaurus Discovery*. Dordrech: Kluwer.

Grunenberg, L. 2002. "Facet Analysis: Using Faceted Classification Techniques to Organize Site Content and Structure." in *Proceedings of the ASIS&T*. [cited 2006.5]. 〈http://www.willpowerinfo.co.uk/thesbibl.htm〉.

Maple, A.(1999), "Faceted Access: A Review of the Literature." [cited 2006.5]. 〈http://www.music.indiana.edu/tech_s/mla/facacc.rev〉.

Nilsson, M., Palmér, M., Naeve, A. 2002. "Semantic Web Metatdata for e Learning Some Architectural Guidelines." [cited 2006.5]. 〈http://www2002.org/CDROM/alternate/744/#mini:lomrdf〉.

Pereira, F., Tishby, N., Lee, L. 1993. "Distributional Clustering of English Words." *in Proceedings of the 31st ACL*, 183 190.

Peters, W., Kilgarriff, A. 2000. "Discovering Semantic Regularity in Lexical Resources." *International Journal of Lexicography* 13 (4): 287 312. [cited 2006.5]. 〈http://www.lexmasterclass.com/people/Publications/2000 PetersKilg IJL SemRegularity.pdf〉.

Ryu, P.M. Kim, J.H., Nam, J., Huang, J.X, Shin, S. 2006. "Toward Domain Specific Thesaurus Construction: Divide and Conquer Method." in *Global Wordnet Conference 2006 Proceedings*, Jeju Island, Korea, 69 83.

Sinopalnikova, A. 2004. " Word Association Thesaurus As a Resource for Building WordNet." in *Global Wordnet Conference 2004 Proceedings*, Brno, Czech Republic, 199 205.

Sundheim, B.M., Mardis, S., Burger, J. 2006. "Gazetteer Linkage to WordNet." in *Global WordNet Conference 2006 Proceedings*, Jeju Island, Korea, 103 104.