과학기술 분야 시소러스 상에서의 패싯들

정한민°* 성원경* 박동인* 황순희**
*한국과학기술정보연구원 정보시스템부
*{jhm°, wksung, dipark}@kisti.re.kr, **soonheehwang@pusan.ac.kr

Facets on Thesaurus for Science and Technology Domain

Hanmin Jung°*, Won-Kyung Sung*, Dong-In Park*, and Soonhee Hwang**
*Information System Division, Korea Institute of Science and Technology Information
**Korean Language Processing Laboratory, Pusan National University

요 약

본 논문에서는 시소러스 상에서 개념간 세분화를 위한 의미적 기준인 개념 패싯 (Conceptual Facet)과 관계 패싯 (Relational Facet) 그룹을 사용하는데, 패싯이란 공통의 특성을 갖는 개념들을 함께 그룹화하고 용어간 관계를 구조화하기 위한 장치이다. 개념 패싯은 용어가 갖는 대표적인 의미속성, 범주를 의미하며, 용어 각각을 구별되는 의미장 (Semantic Field)에 분류하도록 한다. 관계 패싯은 상·하위 개념 간 의미관계를 표현한 메타 개념이다. 본 시소러스는 여러 전문가들의 다양한 관점을 반영하도록 설계되었다. 관점이란 주관적이며, 임의적이어서 개별 개념에 내재된 자질 또는 속성과는 구별되는 독립적 속성이다. 개념 패싯, 관계 패싯의 도입은 계층 관계, 동등 관계, 범주 관계 등과 더불어 용어 간 관계를 보다 구체적으로 명시함으로써 최종 사용자에게 검색의 효율성과 정확성을 제공할 수 있다.

1. 서 론

문헌정보학에서 패싯 개념을 도입한 이유는 급변하는 정보 검색 환경에서의 시소러스의 기능 확대와 형태 변화의 필요성이 대두되었기 때문이다. 이것은 기존의 시소러스가 갖는 지식 표상의 한계를 보완하려는 노력이며, 확장된 개념간 관계를 이용하면 개념들에 대한 보다 구체적인 표현이 가능해지기 때문이다. 일반적으로 하나의 상위 개념은 복수 개의 하위 개념들을 갖는데, 기존의 시소러스에는 상위 개념과 하위 개념들이 단순히 나열, 연결되는 수준에 그쳤다 [8]. 용어의 패싯 분석에 관한 국내외 연구의 공통점은 특정 전문 분야 용어를 인지적으로 추출·분석하여, 한정된 수의 개념 범주 (패싯)에 이들을 분류하고, 이를 바탕으로 특정 분야 시소러스에 대한 패싯 확장의 기준으로 삼는다는 점이다. [7]은 패싯을 "분명하게 정의되며, 상호 배타적이며, 또한 특정 부류 (Class) 또는 주제와 관련된 집합적인 관점 (Aspect), 속성 (Properties) 또는 특성 (Characteristics)"으로 정의하고 있다. 한편, 문헌정보학에 패싯이란 용어와 패싯 분류 (Faceted Analysis)를 최초로 도입하고, 일관성 있는 분석을 시도한 인도의 수학자이자 사서인, Ranganathan은 1933년 개발한 도서 분류 체계인 콜론 분류법 (Colon Classification)에 패싯과 콜론(:)을 처음 사용하였다. 그는 패싯과 관련된 접근법으로 분석-합성적 분류를, "분석 (Analysis)이란 특정 주제를, 그 주제를 구성하는 기본 개념으로 분류 (Breaking Down)하는 과정이며, 합성 (Synthesis)이란 개념들을 주제로 재합성 (Recombining)하는 과정으로, 분석과 합성은 모든 주제에 적용될 수 있고, 조직화될 수 있다"고 설명하고 있다 [5]. 그가 시도한 콜론 분류법에는 108개의 주요 분류와 10개의 일반 분류가 포함되는데, 이것들은 아라비아 숫자와 로마 문자를 혼합하여 표현되었다. 각 주요 분류들은 인간 (Personality), 사물 (Matter), 에너지 (Energy), 공간 (Space), 시간 (Time)의 5개 기본 패싯으로 구성된다. 그의 주요한 공헌은 이와 같은 기본적인 패싯이나 범주들을 명명했다는 점인데, 이것들은 모든 분야에 본질적인 것이어서 분류 체계에 반드시 이 내용들이 포함되어야 했다. 그렇지만, 콜론 분류법은 주제적 접근에 적용된 것으로 최근의 정보 검색에서의 확장 검색 용도로 사용할 때는 확장의 제약 조건으로서 재검토를 해야 한다. [3]은 실제 시소러스에 개념 패싯을 도입한 사례이다. 시소러스의 관련 개념들에 개념 패싯을 부여하여 관련 개념 관계를 좀더 명확히 하는 데 활용하고자 했다. 다만, 확장 검색의 대상인 상·하위 개념이 아닌 사용하기에 부담이 따르는 관련 개념에만 적용하였으며, 관계를 직접적으로 제약할 수 있는 관계 패싯을 고려하지 못하는 등 아쉬움이 남는다.

본 시소러스에서는 이러한 점들을 보완하여 개념 자체, 그리고 상·하위 개념 간의 관계에 다양한 형태의 패싯들을 적용하여 필요에 따라 확장 검색의 제약으로 사용할 수 있는 기재를 마련하고자 한다.

2. 과 학 기 술 분 야 시 소 러 스

과학 기술 발전의 속도에 비해 지식 베이스의 구축은 아직도 국내에서는 그 예를 찾기가 힘들다 [3] [8]. 더욱이 이들은 시소러스의 기반이 되는 용어 선정을 위해 사전, 서적 등 출판에 걸리는 시간이 긴 검증된 문헌을 이용하였다. 그렇지만, [4] [6]에서 지적하듯이 현재 사용 중인 용어들에 대한 출판물의 적용도 (Coverage)는 상당히 낮다.

본 논문에서는 용어 생명 주기 (Life Cycle of Terms) 관점에서 선정된 용어들 [4] [6]과 최근 말뭉치에서 높은 적용도를 보이는 용어들을 섞어 사용함으로써 시소러스 구축 완료 시점 이후의 효용성을 고려한다.

기존에 구축된 시소러스들에서 개념에 대한 정보를 백과사전적인 접근에 의존하여 본말이 전도되는 현상과 가용한 용어 자원이 제약되는 현상이 나타났으나, 우리는 시소러스 상의 각 개념에 대한 정의를 구축자 간 의사소통이 원활히 이루어질 수 있는 수준으로 최소화한다. 즉, 개념을 대표하는 우선어 (Descriptor)/비우선어 (Non-descriptor)에 대한 예제들을 필수적으로 선택하고 범위 주기 (Scope Note)에 대해서는 동형이의어나 다의어적 성격을 가진 개념어들에 대해서만 부여한다.

관련 개념 (RT; Related Term)는 개념들 간 (Domain과 Range)의 관계를 표현하는 부분이지만 기존의 시소러스들은 Range의 명시에만 치중하고 어떤

관계 (Relation)인지에 대한 정보는 배제하거나 소홀히 하였다 [4] [8]. 우리는 이러한 부분이 일관성을 결여시킬 수 있는 점이라고 판단하여 관련 개념 구축을 본 시소러스에서는 제외한다.

기존 시소러스들이 많은 비판을 받는 이유 중의 하나는 다양한 관점을 반영하지 못한다는 것이다. 전문가들뿐만 아니라 해당 시소러스에 관심이 있는 사용자 입장에서 시소러스를 보는 관점은 모두 다를 수 밖에 없으며, 이들의 입장을 반영하고 일관성 있는 구축을 보장하기 위해서 우리는 개념 패싯 (CF; Conceptual Facet)과 관계 패싯 (RF; Relational Facet)을 도입한다. 관계 패싯은 다시 의미역 관계 패싯 (TRF; Thematic Role Relational Facet), 속성 관계 패싯 (ARF; Attribute Relational Facet), 그리고 범주 관계 패싯 (CRF; Category Relational Facet)[1]으로 나누어진다. 개념 패싯은 개념의 대표 의미 속성 분류 체계로서 의미 본질 보다는 활용적 측면을 고려한다. 예를 들어, "백신"은 컴퓨터 분야에서 사용되는 프로그램으로서의 의미와 의학에서 약의 통칭으로서의 의미가 있다. 전자의 경우에는 '내용'이라는 개념 패싯을, 후자의 경우에는 '물질•재료'라는 개념 패싯을 부여한다. 현재 '감각•감정,' '위치•공간,' '시간,' '생물' 등을 포함하여 15개를 정의하여 사용한다 (표 1 참조). 관계 패싯은 상위 개념 (BT; Broader Term)가 하위 개념 (NT; Narrower Term)을 바라보는 관점을 의미한다. 서술형 명사를 중심어로 가지는 "인터넷 접속," "원격 탐사"와 같은 용어들[2]에는 의미역 관계 패싯이 추가로 부여된다. 의미역 관계 패싯은 Fillmore로부터 발전한 동사의 의미적 수행에 필요한 의미역을 개선하여 '근원,' '대상,' '도구,' '목표' 등 10개를 정의하여 사용한다. 상위 개념과 하위 개념 간의 자질 상속을 고려하여 하위 개념에 붙는 변별적 자질을 키워드화하여 이를 속성 키워드 (AK; Attribute Keyword)로 명명한다. 속성 관계 패싯과 의미역 관계 패싯은 속성 키워드를 판단하여 부여된다. 속성 관계 패싯은 개념 관계 패싯 15개에 '사례'를 추가하여 사용한다. 상•하위 관계를 세분화하여 살펴보면, 일반화 관계 (IS-A), 부분전체 관계 (HAS_PART), 사례관계 (INTANCE-OF)로 나눌 수 있다. "메모리 → 플래시 메모리"는 일반화 관계, "심장 → 심근"은 전체부분 관계, "운영 체제 → MS Windows"는 사례 관계의 예이다. 우리는 이들을 범주 관계 패싯으로 정의하여 상•하위 관계의 속성으로서 부여한다.

3. 개념 패싯과 관계 패싯

[9]에서 Nilsson은 Semantic Web 메타데이터 구조가 가져야 하는 조건들의 하나로 동일한 자원에 대한 다양한 관점을 지원하는 "subjective and non-authoritarian"을 제시했다. 다양한 관점은 응용 분야에 따라 필요한 정보를 선별적으로 선택할 수 있는 방안을 제공함으로써 무제한적인 확장을 막을 수 있다. 기존의 시소러스들은 상위 개념에 대한 하위 개념 그룹 내의 용어나 개념들이 서로 다른 특성을 가지고 있음에도 불구하고 이들을 구분할 수 있는 자질을 제공하지 못함으로 인해 하위 개념 확장에 대한 부담을 안을 수 밖에 없다. 본 논문에서는 이러한 난점을 개념 패싯과 관계 패싯들을 도입함으로써 해결하고자 한다.

3.1 개념 패싯

개념 패싯은 해당 용어가 갖는 대표적 의미 속성, 범주를 의미한다. 동일한 관계 패싯을 갖는 하위 개념 그룹 내의 용어들은 동일한 개념 패싯을 가져야 한다. 개념 패싯은 개념이 아닌 개념을 구성하는 Synset (USE/UF)의 모든 요소들에 할당된다. 이러한 사항들은 본 시소러스 구축 시의 가장 기본적인 원칙들 중 하나이다. 표 1은 범용 과학 기술 분야 용어들에 적용하기 위한 16개의 개념 패싯을 보여준다[3].

표 1. 개념 패싯

감각·감정	위치·공간	기기·장치·부속
상태·성질	시간	생물
분야·이론·방법	물질·재료	조직
언어	단체	질병·증상
내용	행위	현상·사건

3.2 관계 패싯

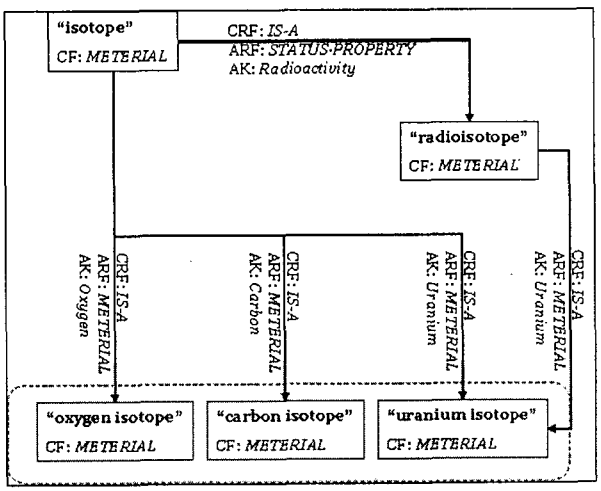

그림 1. 시소러스의 예

(1) 속성 관계 패싯: 속성 관계 패싯을 부여하는 기준은 상•하위 개념 간의 변별적 자질을 의미하는 속성 키워드이다. 속성 키워드는 관계 패싯들과 마찬가지로 BT-NT 관계에 모두 할당된다. 동일한 개념에 속하는 용어들은 모두 같은 속성 키워드에 의해 정의될 수 있어야 한다. 원칙적으로 BT-NT 관계에 하나의 속성 키워드가 할당되어야 하나, 상위 개념과 하위 개념 간의 의미적 거리가 먼 경우 (예. "전화기" → "카메라폰")에는 여러 개의 속성 키워드들 (예. "카메라," "휴대폰")이 할당될 수 있다. 이것은 상위 개념과 하위 개념 중간에 별도의 개념 (예. "휴대폰")이 도입될 필요가 있음을 의미한다. 속성 키워드는 개념 내의

[1] 기존에 시소러스에서 언급하는 NTG, NTP, NTI가 여기에 해당한다 [2]. 본 시소러스에서는 상위 개념에서 하위 개념으로의 관점만 유지하고 있으므로, BTG, BTP, BTI는 필요가 없으며, 관점이라는 성격으로 관계 패싯들을 다루고 있어서 하나의 관계 패싯으로 설정한다.

[2] 정의에 따르면, 서술형 명사는 '하다, 되다, 시키다, 당하다, 주다, 받다, 있다, 없다' 등의 기능 동사와 결합이 가능한 명사들이다. "게임," "메시징" 등과 같은 음차표기어 등에 대해서도 현재에 와서는 서술성을 가지는 단계에 있지만, 의미역 관계 패싯 부여에서는 이들을 제외한다. 또한, "가정용 전화"와 같은 수식 구조 형태의 용어도 제외한다. 이들을 일반화하면, 개념 패싯으로 '행위' 이외의 것을 가지는 용어들을 제외할 수 있다.

[3] WordNet 2.1에서 이들에 대한 평균 깊이는 약 4.5이다.

용어들이 동의어 관계인지 상•하위어 관계인지를 간접적으로 판단할 수 있는 기준으로서의 역할도 한다.

(2) 의미역 관계 패싯: 중심어가 서술형 명사인 경우 속성 키워드에 대해 술어 (중심어)가 요구하는 의미적인 논항을 따져 의미역 관계 패싯을 부여한다. 기존의 의미역 연구들과 달리 [1]에서는 전산 언어학적인 관점에서 의미역을 부여하고 있으나, '행위주, 경험주, 수령주'에 대한 구분을 포함하여 구문 관계에 따른 의미역 할당이 역시 용이하지 않다. 특히, 전문용어 영역에서 복합 명사에 의미역 관계 패싯을 할당하는 경우에는 그 어려움이 더하다. 예를 들어, "접속" → "네트워크 접속"의 경우에는 '목표, 장소, 도구'가, "경매" → "인터넷 경매"의 경우에는 '장소, 도구, 방식' 등이 성립가능하기 때문이다. 이러한 문제는 복합 명사가 가지는 의미역에서의 다의어적 특성에 기인하므로 이들을 명확히 판별하고 구분하는 것은 어려울 수 밖에 없다. 본 시소러스에서는 하나의 복합 명사가 가질 수 있는 의미역 관계 패싯 집합 내에 들어가는 의미역 관계 패싯이 하나 또는 그 이상인 경우를 허용한다. 이때, 각 의미역에 대해 문형 (예. AB로 조어를 이루는 용어에 대해 "A을 이용하여 B하다"가 성립하면 '도구' 의미역 관계 패싯을 부여한다.)을 기술하여 적용한다. 시소러스의 완성도가 높아질수록 의미역 관계 패싯 집합에 근접한 할당이 이루어질 것으로 기대하기 때문이다. 표 2는 10개의 의미역 관계 패싯을 보여주며, 그림 2는 속성 관계 패싯과 의미역 관계 패싯 차이를 설명하기 위해 "검사" 개념에 대한 하위 개념들을 보여준다.

표 2. 의미역 관계 패싯

근원	대상	도구
시간	목표	장소
행위자	수혜자	방식
동반자		

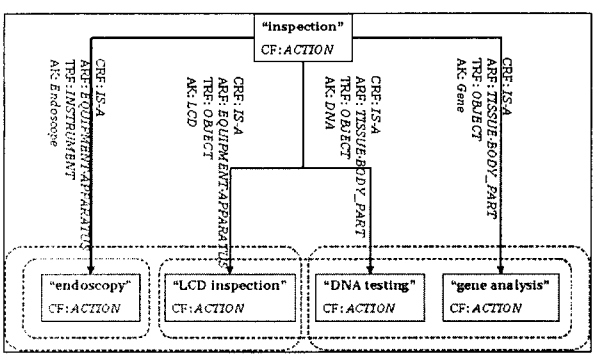

그림 2. 속성 관계 패싯과 의미역 관계 패싯의 예

(3) 범주 관계 패싯: 모든 상·하위 관계에는 일반화, 전체부분, 사례의 세가지 범주 관계 패싯 중 하나를 할당한다. 일반화 관계는 할당에 어려움이 없지만, 전체부분과 사례는 혼동되는 현상이 발생한다. 예를 들어, "별자리 → 전갈 자리"의 경우에 별자리의 부분으로 "전갈 자리"를 볼 수도 있으며, 별자리의 종류로도 볼 수 있다. "심장 → 심근"의 경우에는 부분으로 밖에 해석될 수 없다. 이러한 사례들을 정리하여 본 시소러스에서는 "~의 일종"과 "~의 부분"을 모두 가질 수 있는 상·하위 관계에는 일반화 관계를 할당하며, "~의 부분"만을 가지는 경우에 한해서 전체부분을 할당한다.

4. 시소러스와 패싯의 구축 및 활용

현재 약 6,000여 용어들에 대해 패싯들을 부여하고 있으며, 이들은 다음의 측면에서 그 활용 방안을 찾을 수 있다.

(a) 확장 검색에서의 제약: 모든 확장 검색에서 상·하위 관계만을 이용하여 확장하는 것보다는 사용자 질의에 따라 관계 패싯을 제약하여 확장 검색하는 것이 정확율 향상에 도움이 된다.

(b) 동형이의어에 대한 용이한 구분: 동형이의어의 구분을 범위 주기나 정의문 등의 비정형적인 표현에만 의존하는 경우에는 기계적 처리가 힘든 경우가 생기므로, 이를 보완하는 목적으로 개념 패싯을 활용할 수 있다.

(c) 정보 검색에서의 효율적 질의 생성: 동일 개념 패싯에 대해서는 'OR'를, 상이한 개념 패싯에 대해서는 'AND'를 적용하는 방식으로 자연어 질의어의 불리언 질의로의 변환을 가능하게 한다.

(d) 의미적 유사도 반영: 동일한 하위 개념 그룹일지라도 관계 패싯에 따라 세부적으로 그룹화될 수 있으며, 이 경우에 그룹 내의 의미적 거리가 그룹 간의 의미적 거리보다 가까워 진다.

(e) 시소러스 확장에서의 일관성 보장: 관계 거리가 멀어지더라도 동일한 관계 패싯으로 유지되는 Path는 일관성을 유지 측면에서 유리하다.

5. 결 론

본 논문은 공통의 특성을 갖는 개념들을 그룹화하고 이들을 구조화 개념 패싯과 관계 패싯 그룹을 시소러스에 도입하였다. 다양한 패싯의 도입은 확장 검색에서 다양한 제약을 가할 수 있는 방안이 되며, 시소러스를 온톨로지 수준으로 확장하기 위한 발판이 된다. 현재 범용 과학 기술 분야 용어들을 대상으로 이러한 설계를 반영하고 있으며, 차후 다양한 과학 기술 분야로 세분화하여 시소러스를 구축할 예정이다.

참 고 문 헌

[1] 강신재, 박정혜, 대규모 말뭉치와 전산 언어 사전을 이용한 의미역 결정 규칙의 구축, 한국정보처리학회 논문지 B, 10 (2), 2003.

[2] 이재윤, 김태수, WordNet과 시소러스, 언어 정보의 탐구 1, 1999.

[3] 정영미, 김명옥, 이재윤, 한승희, 유재복, *과학기술 분야 통합 개념체계의 구축 방안 연구*, 정보관리학회지 19 (1), 2002.

[4] 정한민, 구희관, 이병희, 성원경, 효율적인 자원 운영을 위한 전문용어 생명주기 관리 연구, *Korea Computer Congress*, 2005.

[5] L. Grunenberg, Facet Analysis: Using Faceted Classification Techniques to Organize Site Content and Structure, *Proceedings of the ASIS&T*, 2002.

[6] H. Jung, H. Koo, B. Lee, and W. Sung, Toward Managing the Life Cycle of Terms Using Term Dominance Trend, *Proceedings of Pacific Association of Computational Linguistics*, 2005.

[7] A. Maple, Faceted Access: A Review of the Literature, http://www.music.indiana.edu/tech_s/mla/facacc.rev, 1999.

[8] http://jisik.kiom.re.kr/th/

[9] http://wwwconf.ecs.soton.ac.uk/archive/00000221/01/index.html