시소러스 통합을 위한 개념기반 패싯 프레임워크 구축

Construction of the Concept-Based Faceted Framework for Thesaurus Integration

이 승 민(Seungmin Lee)*

〈 목 차 〉

초 록

각각의 시소러스가 지닌 고유한 특성 및 상이한 구축목적으로 인해서, 하나의 시소러스를 이용하여 특정 도메인의 정보를 조직하고 검색하는데 여러 가지 문제가 나타나고 있다. 시소러스의 통합은 이러한 문제를 해결하기 위한 한 가지 방법이 될 수 있다. 본 연구에서는 물리학 분야에서 널리 사용되고 있는 시소러스인 PACS와 PIRA를 선정하여 이들 시소러스를 개념적으로 통합할 수 있는 패싯 프레임워크를 구축하였다. 이들 두 시소러스를 연결하기 위해 물리학 분야 전체를 다룰 수 있는 계층구조를 구축하였고, 이 계층구조에 패싯분석을 적용함으로써 각각의 주제 사이의 관계를 설정할 수 있는 하나의 지식기반을 제공하고 있다. 또한, 이 개념기반 패싯 프레임워크를 통해 보다 효율적으로 정보를 조직하고 검색할 수 있는 접근방법을 제안하고 있다.

키워드: 시소러스, 시소러스 통합, 패싯분석, PACS, PIRA

ABSTRACT

Applying one specific thesaurus might cause several problems because each thesaurus has its own characteristics inherited from its construction process. Therefore, integration of thesauri can be an appropriate approach to overcome the difficulties. This current research selected physics as a domain and two thesauri in the domain: PACS and PIRA. By integrating these two heterogeneous thesauri, this research could construct a conceptual structure that covers the whole concepts related to physics. By constructing the conceptual structure with the use of facet analysis from integrated thesaurus, it provides knowledge base with hierarchical structure and clear relationships between concepts. It can be an alternate approach to effective and efficient information retrieval and knowledge discovery.

Keywords: Thesaurus, Thesaurus Integration, Facet, Facet Analysis, PACS, PIRA

* 중앙대학교 문헌정보학과 시간강사(seungminator@gmail.com)
• 접수일: 2010년 8월 27일 • 최초심사일: 2010년 9월 8일 • 최종심사일: 2010년 9월 27일

Ⅰ. 서 론

웹으로 대표되는 정보네트워크의 발전으로 인해, 도서관을 둘러싼 정보환경에서는 생성되는 정보의 양이 기하급수적으로 증가하고 있을 뿐 아니라 정보가 유통되는 속도 또한 엄청나게 빨라졌다. 또한 여러 가지 유형의 전자적 정보자원이 기존의 인쇄자료와 대등한 위치에 놓이거나 점차적으로 이를 대체하는 경향이 나타나고 있다. 이러한 환경의 변화에 따라, 전통적인 인쇄형태의 정보들은 전자적으로 가공, 처리, 전송될 수 있는 디지털 방식으로 생성되거나 아니면 전자적인 형태로 다시 출판되고 있다. 이러한 상황은 학술정보자원 뿐만 아니라 신문이나 잡지 등과 같은 일상적인 정보에도 적용되어 수많은 정보들이 웹 상의 페이지 형태로 제공되고 있다. 이와 같은 정보환경의 변화와 함께, 현재 도서관의 초점은 디지털 자원의 생성과 유지에서 디지털 컨텐츠의 조직과 활용을 촉진하는 인프라스트럭쳐(infrastructure) 및 이와 관련한 기술의 개발로 그 방향이 바뀌어 가고 있다. 또한 전통적인 모습의 도서관은 새로운 유형의 디지털 정보를 보다 효과적으로 다룰 수 있는 디지털 도서관으로 점차 변모하고 있다. 이러한 환경에서, 정보의 조직과 정보검색은 디지털 도서관 연구의 두 가지 주요 연구분야가 되고 있다.

정보조직은 지식을 일련의 개념 및 이들 개념 사이의 관계로 표현함으로써 정보를 조직하고 활용하는 메카니즘이다.[1] 전통적인 도서관 환경에서의 정보조직은 정적이고 고정되어 있는 주제 카테고리를 열거하여 물리적인 정보자원에 대한 접근을 제공하는 분류의 개념으로 사용되어 왔다. 이러한 방식은 인쇄물 형태의 정보자원에는 적합할지 모르나, 디지털 정보자원 또는 웹 상의 컨텐츠가 지닌 동적인 특성을 다루는 데에는 효과적이지 않다. 결과적으로, 이와 같은 전통적인 분류에 기반한 정보조직 방법은 동적이고 변화가 심한 새로운 유형의 정보자원을 조직하는데 있어서 많은 어려움을 초래하고 있다. 따라서, 웹과 같은 디지털 정보자원에 대해서는 기존의 분류를 통한 접근이 아닌 정보자원의 조직에 초점을 둔 접근방법이 필요하다. 이러한 접근방법은 이용자들의 다양한 정보요구에 대응할 수 있을 뿐만 아니라 정보의 동적인 특성을 처리할 수 있을 것이다.

이러한 문제를 해결하기 위해, 다양한 유형의 정보자원을 효과적이고 체계적으로 활용할 수 있는 시소러스와 같은 표준화 된 지식기반을 통한 정보의 조직의 중요성이 대두되어 왔다. 하지만, 기존의 분류체계와 마찬가지로, 현재의 시소러스 역시 미리 정해져 있는 고정된 구조를 채용하고 있기 때문에, 의미적으로 차이가 있는 주제나 개념을 조직하는데 있어서 여러 가지 한계를 보이고 있다. 또한, 동일한 도메인 내에서의 시소러스라 하더라도 같은 개념에 대해 상이한 용어를 적용하여 표현하고 있기 때문에, 각각의 주제 및 이들 사이의 관계를 명확하게 표현하는데 어려움을 겪고 있다.

1) Seungmin Lee and Elin K. Jacob, "Construction of a conceptual structure as a mediator between MARC and FRBR," *Proceedings of 2007 North American Symposium on Knowledge Organization(NASKO)*(June 14-15, 2007), Toronto, Canada.

이에 본 연구에서는, 기존의 정적이고 고정된 체계를 통한 정보조직이 아닌 각각의 주제가 지닌 개념에 기반하여 동적이고 융통성있는 특성을 지닌 프레임워크를 구축함으로써 기존의 제한된 시소러스를 서로 연결시켜 여러 가지 유형의 정보자원을 보다 효과적으로 조직하고 정보검색의 효율성을 극대화시킬 수 있는 방안에 대해서 논의하고자 한다.

Ⅱ. 이론적 배경

현재의 정보조직이나 정보검색이 당면한 여러 가지 문제를 해결할 수 있는 방법 가운데 하나로 시소러스와 같은 통제된 용어의 집합에 대한 필요성이 점차 증대되어왔다. 용어 및 해당 용어와 관련된 개념을 시맨틱 네트워크(semantic network)의 형태로 조직하는 시소러스는 폭발적으로 증가하는 정보를 검색하는데 있어서 중요한 도구로 여겨지고 있으며,[2] 특정 전문분야에서 사용되는 각각의 용어가 관련되어 있는 다른 용어와 맺게 되는 관계를 표현한 용어의 집합으로 정의된다.[3] 또 다른 측면에서 보면, 시소러스는 특정 분야에서 사용되는 용어 및 개념을 구조화시킨 체계로서, 용어 사이의 구조적인 관계를 표현해 주는 일련의 용어로 정의되고 있다.[4]

시소러스 내에 수록되어 있는 각각의 용어는 관련되어 있는 다른 용어와 여러 가지 관계를 맺음으로써 표현된다. 〈표 1〉에서 보는 바와 같이, 시소러스에서 이용되는 용어간의 기본적인 관계는 크게 등가관계, 계층관계, 연관관계의 세 가지로 나누어 볼 수 있다. 등가관계는 동의어나 유사동의어를 표현하는데 사용된다. 계층관계는 종속관계, 부분-전체(part-whole)관계, 사례(instance)관계 등을 모두 포함하며, 기호로는 BT/NT를 사용하여 나타낸다. 계층관계를 세분하여 종속관계는 BTG/NTG로, 부분-전체 관계는 BTP/NTP로, 사례관계는 BTI/NTI로 나타내기도 한다. 연관관계는 계층적이지는 않지만 등가관계도 아닌 밀접한 관계를 모두 일컫는 포괄적인 관계이며, RT/RT 기호로 표현된다.[5] 이러한 용어 사이의 구조적인 관계를 통해 시소러스는 지식의 조직, 분류를 위한 구조, 그리고 특정 도메인의 전반적인 체계 등을 제공해 준다.

---

2) Doerr Martin, "Semantic problems of thesaurus mapping," *Journal of Digital Information*, Vol.1, No.8(2001), 〈http://journals.tdl.org/jodi/article/viewArticle/31/32〉 [cited 2010. 8. 5].

3) 황순희, 정한민, 성원경, "패싯(facet)을 이용한 과학기술분야 시소러스 구축과 활용방안," 정보관리연구, 제37권, 제3호(2006. 9), p.65.

4) Robert M. Losee, "Decisions in thesaurus construction and use," *Information Processing & Management*, Vol.43, No.4(2007), p.1.

5) 이재윤, 김태수, "WordNet과 시소러스," 제11회 언어정보연찬회 발표논문집(1998. 2), 연세대학교.

〈표 1〉 시소러스에서 사용되는 의미적 관계

관계종류	기호
등가관계	USE/UF
계층관계(상위개념/하위개념)	BT/NT
종속관계	BTG/NTG
부분 - 전체 관계	BTP/NTP
사례관계	BTI/NTI
연관관계	RT/RT

* 출처: 이재윤, 김태수, "WordNet과 시소러스," 제11회 언어정보연찬회 발표논문집(1998), 연세대학교.

전통적으로 시소러스 구축에는 〈표 1〉에서 제시하고 있는 세 가지의 용어간 관계가 사용되었으나, 이 관계들만으로는 추론검색, 확장검색과 같은 정보환경의 변화를 충족시킬 수 없으며, 수록된 용어들 사이의 의미적 관계를 세분화하는데에도 한계를 보이고 있다.[6] 또한 이러한 시소러스의 구조는 전통적인 열거식 구조와 마찬가지로 용어를 조합하는 방식의 순서를 고정하여 미리 정해놓기 때문에, 일련의 정해진 접근점이나 관계형 구조 모두를 약화시키게 되며, 그로 인해서 이용자들의 다양한 정보요구에 부응하기가 어려워진다. 이러한 시소러스가 지닌 문제들은 대부분 시소러스의 고정적이고 정적인 구조로부터 비롯된다. 기존의 전통적인 분류체계와 마찬가지로, 시소러스 역시 미리 정해져 있는 고정된 구조를 채용하고 있기 때문에, 의미적으로 차이가 나는 용어나 개념을 표현하는데 있어서 한계에 부딪치고 있다.

시소러스가 지닌 또 다른 문제로는 시소러스 사이의 의미적인 차이를 들 수 있다. 시소러스는 특정 분야에서 사용되는 개념에 대해 광범위하게 인정받고 있는 용어로 구성된다. 이러한 용어에 대한 정의 및 개념은 해당 분야 전체에 걸쳐 공통적으로 적용되어야 하지만, 동일한 분야에서 사용되는 시소러스라 하더라도 같은 용어에 대한 의미를 서로 다르게 사용하는 경우가 많이 나타나고 있다. 또한 정보원 사이에서도 서로 다른 시소러스를 채택함으로써 특정 용어의 의미가 혼동되는 경우도 발생하고 있다. 이러한 의미적 차이는 시소러스 사이의 상호운용성을 저해하는 요인이 되며, 심지어 동일한 분야에서도 시소러스들이 고립되어 독립적으로 사용되게 함으로써 시소러스 활용의 효율성을 저해하게 된다.

이외에도, 하나의 도메인 내에서도 목적에 따라 여러 가지 다른 시소러스들이 존재하고 있으며, 각각의 시소러스가 지닌 이질적인 특성 및 상이한 적용범위로 인해서 여러 시소러스의 공존이 정보의 조직 및 검색에 도움이 되는 것이 아니라 오히려 혼선을 초래하는 경우가 발생하고 있다. 이러한 문제는 각 시소러스 구축 과정에서 비롯된 고유한 특징들에서 비롯된 것이다. 이와 같은 시소러

6) 황순희, 정한민, 성원경, "패싯(facet)을 이용한 과학기술분야 시소러스 구축과 활용방안," 정보관리연구, 제37권, 제3호(2006. 9), p.65.

스 사이의 이질성으로는 대표적으로 다음과 같은 것들을 들 수 있다.

1) 상이한 용어의 사용: 동일한 개념이 여러 가지 다른 용어를 사용해서 표현될 수 있다. 이 차이는 다른 언어를 사용할 때에도 발생할 수 있다.
2) 용어가 지닌 상이한 개념적 범위: 시소러스 구축의 목적과 구축 과정의 차이로 인해서, 특정 용어의 의미의 적용범위에 차이가 발생할 수 있다. 이는 시소러스의 계층구조에서 동일한 용어가 서로 다른 수준에 위치하여 상이한 의미로 사용되는 결과를 가져올 수 있다.
3) 용어들 사이의 불명확한 관계: 특정 용어가 지닌 의미적 범위의 상이함은 용어들 사이의 관계를 불명확하게 할 수 있다. 이러한 경우, 어떤 한 용어의 의미가 서로 다른 하나 이상의 용어들과 연결될 수 있다. 그러므로, 특정 용어의 의미적인 범위가 상이하게 설정될 경우 각각의 용어는 여러 용어들과 불명확한 관계를 맺게 될 수 있다.

이와 같은 시소러스가 지닌 문제점을 해결하기 위한 한 가지 방법으로, 상이한 시소러스를 통합하고자 하는 노력이 시도되어 왔다. 시소러스의 통합은 주제 전개에 융통성을 기대할 수 있으며, 일부 도입되어 있는 상관관계를 모든 주제에 일관성있게 적용함으로써 색인과 검색기법의 변화와 관리자 위주에서 이용자 위주로 바뀌고 있는 검색환경의 변화에 적응할 수 있는 새로운 환경에 맞는 용어관계의 확장방안을 제시해 준다. 이와 더불어, 시소러스와 기존의 분류체계를 통합한 통합적인 개념체계를 구축할 필요성도 제기되어 왔다.

정영미 등은 동일한 주제분석 도구라는 측면에서 시소러스와 분류표를 통합해야 하는 필요성을 제기하면서, 과학기술 분야를 포괄하는 분류표, 시소러스, 용어사전의 구축 및 이들의 효율적인 개발, 유지, 이용을 위해서 통합 개념체계를 구축하고자 하였다.[7] 손대형, 김태수는 분류와 시소러스의 통합을 위해 패싯화 된 시소러스를 구축하였다.[8] 이는 시소러스와 결합된 분류체계로서 전통적인 분류방식과 시소러스가 서로를 보완하여 문제점을 극복하기 위한 것이라고 주장하였다. 또한 분류와 함께 시소러스를 제공하는 것은 분류표로 제시할 수 있는 것보다 훨씬 많은 용어를 상세한 수준으로 제공할 수 있다고 하였다. Amann와 Fundulaki는 온톨로지와 시소러스의 통합에 대해 논의하면서, 상이한 두 개의 체계를 일관성있게 통합하는 시스템을 구축하였다.[9] 이 시스템은 시소러스에 수록된 용어를 온톨로지의 특화된 개념으로 보면서, 시소러스에 수록된 용어 사이에 존재하는 의미적 관계를 재정의함으로써 보다 세분화되고 명확한 체계를 구축하고자 하였다.

---

7) 정영미 등, "과학기술분야 통합 개념체계의 구축방안 연구," 정보관리학회지, 제19권, 제1호(2002. 3), pp.135-161.
8) 손대형, 김태수, "패싯분류체계를 이용한 시소러스 작성에 관한 연구," 한국정보관리학회 제5회학술대회논문집(1998. 8), pp.235-238.
9) Bernd Amann and Irini Fundulaki, "Integrating ontologies and thesauri to build RDF schemas," In *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries*(September 22-24, 2009), Paris, France, pp.234-253.

이와 같은 연구들은 시소러스를 통합하기 위해 각 시소러스에 수록된 용어들에 대한 일대일 매핑을 실행함으로써 상이한 시소러스를 하나의 새로운 시스템 안으로 병합하는 방식이 주를 이루고 있다. 하지만, 앞서 언급한 바와 같이, 동일한 의미라 하더라도 그 의미를 나타내는 용어의 레이블은 다르게 표현되고 있다. 예를 들면, 특정 저작의 저자를 의미하는 용어가 어떤 경우에는 'author'로 표현되지만 다른 경우에는 'creator'로 표현되기도 하며, 상황에 따라서는 'writer'나 이외의 다른 용어를 사용해서 표현되기도 한다. 이런 경우에, 용어의 레이블에 기반한 매핑은 같은 의미를 지닌 용어라 하더라도 그 용어를 둘러싸고 있는 상황을 반영하지 못하기 때문에 이들을 동일한 의미로 연결시키지 못하는 의미적인 문제점을 일으킬 뿐만 아니라 매핑 결과의 신뢰성을 저하시키는 결과를 가져오게 된다.

또한, 하나의 시소러스 안에 수록된 용어가 다른 시소러스에 수록된 용어와 의미적으로 매핑되지 않는 경우도 발생하게 된다. 이러한 매핑과정에서 누락된 용어들이 발생하는 이유는 통합되는 두 가지 시소러스의 구조적인 차이로 인한 것일 수도 있고 각각의 시소러스가 지닌 주제의 세분화 단계의 차이에서 비롯된 것일 수도 있다. 이와 같은 이유로 인해, 포괄적이면서도 정확한 용어의 매핑은 통합되는 시소러스가 동일한 구조와 동일한 정도의 세분화된 체계를 채용하고 있지 않는 한 거의 불가능한 것이다.[10]

이러한 문제점 이외에도, 시소러스를 통합하기 위한 기존의 접근방법은 통합되는 시소러스에 수록된 용어의 매핑을 통해서 새로운 용어간 관계를 설정하게 된다. 이 새로운 관계들이 모여서 통합되는 시소러스를 병합할 수 있는 하나의 구조적인 체계를 구축하게 되는데, 이는 결국 또 하나의 시소러스를 구축하게 되는 결과를 가져오고 있다. 이 새롭게 구축된 시소러스는 통합되는 시소러스의 특성과 체계를 그대로 이어받게 되기 때문에, 정적이고 고정적인 또 다른 시소러스를 생성하는 것에 지나지 않는다.

이와 같은 시소러스 통합의 어려움에 대해 Martin은 용어의 의미상의 차이와 상이한 계층구조, 용어들 사이의 상이한 의미적 관계로 인해서 서로 다른 지식기반에 속해 있는 용어를 조합하여 이를 통합하는 것은 의미적으로 불가능하다고 전제하면서, 용어의 레이블 간의 매핑을 통한 시소러스의 통합이 지닌 한계를 설명하고 있다.[11] 결국, 기존의 정적이고 고정된 분류체계에 기반을 둔 시스템을 구축하여 시소러스를 통합하려는 것은 상호운영이 어려운 또 다른 하나의 시소러스를 생성하는 것에 지나지 않을 것이다.

이에 본 연구에서는 여러 가지 유형의 정보자원을 효과적으로 조직하고 검색하는 한 가지 방법

10) John S. Kendall, "The use of metadata for the identification and retrieval of resources for K-12 education," In *Proceedings of the 2003 Dublin Core Conference: Suporting communities of discourses and practice - Metadata research and applications*(September 28-October 2, 2003), Seattle, Washington, United States.

11) Doerr Martin, "Semantic problems of thesaurus mapping," *Journal of Digital Information*, Vol.1, No.8(2001), 〈http://journals.tdl.org/jodi/article/viewArticle/31/32〉 [cited 2010. 8. 5].

으로 용어의 레이블이 아닌 각 용어가 지닌 근본적인 개념에 기반한 프레임워크를 제안하고자 한다. 이는 개념적인 프레임워크로 상이한 시소러스를 통합하는 것이 아닌 이들 시소러스들을 의미적으로 연결시켜 시소러스 사이의 상호운용성을 확보하는 개념적인 매개체로서의 기능을 하는 것이다.

Ⅲ. 특정 도메인에서의 시소러스 통합

일반적으로 시소러스는 수록되어 있는 용어 사이의 관계를 명확하고 구체적인 구조를 이용해서 표현해 준다. 따라서, 시소러스의 통합은 이들 용어 및 용어가 지닌 의미를 확인하고 계층구조를 구축하는 과정으로 고려될 수 있다.[12] 그렇기 때문에, 각 시소러스의 고유한 계층구조 및 용어 사이의 관계를 유지하면서 시소러스를 통합하는 것이 필요하다. 본 연구에서는 특정 도메인에서 각 시소러스가 수록하고 있는 용어의 개념에 기반하여 이들 용어 사이의 관계를 명확하게 설정함으로써 의미적으로 보다 명확한 시소러스의 통합에 중점을 두고 있다.

1. 시소러스 통합의 접근방법

시소러스가 지닌 개념에 기반하여 시소러스 통합을 위한 프레임워크를 구축하는 방법으로 본 연구에서는 상향식(bottom-up) 혹은 귀납적인 개념 클러스터링(clustering)의 과정을 반복적으로 실행하여 관련된 용어 및 개념을 확인하고 그룹화하는 방법을 채택하였다. 클러스터링을 통한 시소러스의 통합은 특정 도메인에서 일반적으로 받아들여지고 있는 용어 및 용어의 의미를 해당 도메인 내의 다른 용어들과 구체적인 관계를 지닌 패싯으로서 추출할 수 있도록 해준다.

최초의 클러스터링에서는 각각의 용어가 지닌 일반적인 개념들을 확인하고, 가장 기본적이면서도 그 의미가 광범위한 개념들을 제공해 주는 보다 포괄적인 클러스터를 생성하게 된다. 여기에 속하는 개념들은 프레임워크의 전체적인 의미적 범위를 규정해 주며, 메인 카테고리를 형성하는 기본패싯(baseline facet)으로서의 기능을 하게 된다. 그 다음, 이들 기본패싯들은 보다 세밀한 의미를 지닌 하위의 패싯을 형성하기 위해 의미적으로 분석되고 결합된다. 이들 하위 패싯 사이의 결합 혹은 연결이 통합되는 시소러스의 의미적 상황과 의도된 목적에 맞게 적합한 상태로 유지된다면, 이와 같은 클러스터링을 통해서 개념 및 각 개념이 연결되는 용어 사이의 표현의 일관성을 확보할 수 있으며, 각각의 개념 사이에서 구조적인 일관성을 지닐 수 있게 된다.

12) Donna Harman, "How effective is suffixing?" *Journal of the American Society for Information Science*, Vol.42, No.1(1991), p.12.

2. 패싯과 패싯분석

넓은 의미에서 패싯은 개념적인 범주화이다. 세부적으로 보면, 패싯은 특정 주제를 구성하는 요소를 일반화시켜 표현하기 위해 일반적인 용어로 구성된 개념의 그룹이다. 각각의 패싯은 개념적으로 동일하다고 여겨지는 여러 가지 용어를 수록하게 된다.[13]

문헌정보학에서 패싯은 "명확하게 정의된, 상호배타적인, 그리고 전체적으로 완전한 특성을 지닌 클래스 또는 특정 주제"라고 정의된다.[14] 이와는 조금 다른 측면에서 보면, 패싯은 유사하거나 의미적으로 결합할 수 있는 카테고리로 구분되는 주제에서 얻어진 용어집 혹은 용어의 그룹을 구분짓는 것을 의미하기도 한다.[15] 이와 같이 특정 도메인에서 사용되는 용어를 구분하고 패싯을 생성하는 과정은 패싯분석(facet analysis)이라고 부른다. 이는 일련의 원칙과 가정에 기반해서 특정 주제분야를 패싯으로 분석하는 인식론적인 과정이며, 개념 사이의 의미적 관계를 명확하게 기술하는 지식구조로 귀결된다. 이 지식구조는 개념들을 조합할 수 있는 구문적인 규칙을 통해서 여러 가지 유형의 개념을 수용할 수 있는 프레임워크를 제공해 준다.[16]

패싯분석은 패싯의 분석과 종합의 두 가지 과정으로 나누어진다. 분석은 특정 주제를 근본적인 개념으로 나누는 과정이며, 종합은 이들 분석된 개념을 재조합하여 주제를 표현하는 문자열을 생성하거나 새로운 조합용어를 만드는 과정이다.[17] 이와 같은 분석과 종합의 과정을 거쳐서, 패싯분석은 특정 주제에 속해 있는 개념들 사이의 관계를 확인하고 표현해 주는 도구로서 사용된다. 또한, 이 개념들 간의 관계에 기반해서, 패싯분석은 새로운 주제를 수용할 수 있는 융통성 있는 구조, 즉 패싯구조를 구축하는 기반을 마련해 준다.

패싯구조는 서로 연관되어 있지 않거나 유사하지 않은 개념들을 서로 구분하고, 관련되어 있거나 유사한 개념들을 그룹화시킴으로써 개별단위적인 계층구조로 개념들을 조직할 수 있는 체계를 제공해 준다. 이러한 과정을 통해서 특정 도메인에서 사용되는 용어가 명확하게 구분되고 상호배타적인 패싯으로 구분될 수 있다.[18] 또한, 이 패싯구조는 특정 도메인의 정보자원이 수록하고 있는 컨텐츠를 적절하게 분류할 수 있도록 조정된 체계를 구축하는데 사용될 수 있다. 패싯구조를 구축하게 되면, 해당 도메인의 어플리케이션은 분류체계의 생성에만 국한되지는 않는다. 오히려, 패싯구조는 후조합 색인언어로써 정보의 조직에 적용될 수 있으며 또한 전조합 색인의 과정과 주제표

13) Shiyali Ramamrita Ranganathan, *Prolegomena to library classification*(New York : Asia Publishing House, 1967).
14) Amanda Maple, "Faceted access: A review of the literature," 〈http://www.music.indiana.edu/tech_s/mla/facacc.rev〉 [cited 2010. 8. 10].
15) Elaine Svenonius, *The intellectual foundation of information organization*(Cambridge : MIT Press, 2000).
16) P. S. G. Kumar, *Introduction to colon classification*(Nagpur : Dattsons, 1987).
17) Shiyali Ramamrita Ranganathan, *op. cit.*
18) Seungmin Lee, "Faceted framework for metadata interoperability," *Journal of the Korean Society for Information Management*, Vol.27, No.2(2010. 6), p.79.

목을 생성 및 개발하는데도 사용될 수 있다. 또한, 이용자가 생성한 패싯의 순서를 지원함으로써 개별 이용자들의 정보요구에 즉각적으로 대응할 수 있는 동적이고 유연성있는 클래스 구조를 제공하는데도 적용될 수 있다.[19] 이는 추출된 패싯을 통해 정보자원 사이에 새로운 관계를 형성할 뿐만 아니라 조직구조를 유연하게 재배치할 수 있도록 해주며, 따라서 전통적인 방식의 열거식 체계나 고정된 열거순서(citation order)를 사용하는 구조에 비해 보다 더 넓은 정보요구에 대응할 수 있게 된다.[20]

Ⅳ. 개념기반 패싯 프레임워크 구축 방안

1. 도메인 및 시소러스 선정

시소러스 통합의 첫 번째 단계는 도메인의 선정 및 해당 도메인 내에서 사용되고 있는 기존 시소러스를 검토하는 것이다. 의미적, 구조적으로 서로 이질적인 시소러스를 개념적으로 통합하기 위해 본 연구에서는 특정 도메인에서 널리 이용되고 있는 두 가지 시소러스를 선정하였다.

본 연구에서는 물리학 분야를 도메인으로 선정하였는데, 물리학 분야에는 여러 가지 다양한 시소러스가 사용되고 있으며, 시소러스 통합의 효율성을 평가하는데 충분할 정도로 많은 개념과 세부분야들이 존재하기 때문이다. 또한, 물리학의 주요 주제분야는 여러 가지의 관련된 세부 주제분야로 구분되고 있으며, 이 분야들 사이의 계층적인 관계가 명확하고 상호배타적으로 형성되어 있어 전체적인 도메인의 의미적 범위를 비교적 정확하게 파악할 수 있다는 장점이 있다.

도메인을 선택한 후, 다음 단계에서는 해당 도메인에서 사용되는 시소러스를 분석하여 선정하였다. 시소러스 선정의 기준으로는 물리학 분야에서 현재 사용되고 있으며, 명확한 계층구조를 지니고 있고, 물리학 분야에서 사용되는 개념을 광범위하게 수록하고 있으며, 수록된 용어들 사이의 관계가 확실하게 구축되어 있는 것 등을 기준으로 하였다. 이러한 기준에 근거하여, 각각의 시소러스에 수록되어 있는 용어들의 상호비교를 통해 물리학 분야에서 광범위하게 사용되고 있는 두 가지의 시소러스를 선택하였다: Physics and Astronomy Classification Scheme(PACS)와 Physics Instructional Resource Association(PIRA).

---

19) Barry M. Leiner, *The scope of the digital library: Dlib Working Group on Digital Library Metrics*, 1998, 〈http://www.dlib.org/metrics/public/papers/dig-lib-scope.html〉 [cited 2010. 8. 6].

20) Shiyali Ramamrita Ranganathan, *Elements of library classification: Based on lectures delivered at the University of Bombay in December 1944*(Poona : N. K. Publishing, 1945).

가. PACS 시소러스

PACS는 물리학과 천문학 분야의 정보자원들을 분류하고 범주화하여, 해당 분야의 문헌을 효과적으로 검색하기 위한 계층적인 분류체계이다. PACS는 약 3,000여 개의 물리학 관련 주제분야를 대상으로 하고 있다. 이들 주제분야는 10개의 최상위 카테고리로 분류되고 다시 73개의 하위 카테고리로 구분된다. 즉, PACS는 십진법 구조를 차용하고 있으며 백진법 구조로 확장될 수 있는 개념적인 계층구조를 사용하고 있는데 이 구조는 다음 〈표 2〉와 같다.

〈표 2〉 PACS 첫 번째 수준에서의 카테고리

PACS 코드	코드 레이블
00	General
10	The physics of elementary particles and fields
20	Nuclear physics
30	Atomic and molecular physics
40	Electromagnetism, optics, acoustics, heat transfer, classical mechanics, and fluid mechanics
50	Physics of gases, plasmas, and electric discharges
60	Condensed matter: Structural, mechanical and thermal properties
70	Condensed matter: Electronic structure, electrical, magnetic, and optical properties
80	80. Interdisciplinary physics and related areas of science and technology
90	90. Geophysics, astronomy, and astrophysics

〈표 2〉에서 나타난 바와 같이, PACS의 첫 번째 수준에서의 카테고리는 총 10개의 넓은 의미의 주제분야로 구성되어 있으며, 십진법과 백진법의 구조를 혼합함으로써 계층적인 구조를 통해 관련된 주제들이 특정 관계를 형성하여 범주화될 수 있다.

이 구조는 크게 다섯 단계의 계층구조를 형성하는데, 하위 단계로 갈수록 보다 상세한 의미를 지닌 용어가 위치하게 된다. 각각의 카테고리와 용어에는 고유한 코드가 부여되며, 이들 코드는 주제어를 식별하는데 사용된다. PACS 코드는 계층구조의 세 번째, 네 번째, 다섯 번째 단계에서 사용되는데, 〈표 3〉에서 보는 바와 같이 각각의 PACS 코드는 여섯 개의 알파벳과 숫자가 혼합된 문자를 사용하여 해당 용어가 지닌 의미를 표시하고 있다.

〈표 3〉에서 나타난 바와 같이, PACS는 각각의 주제어에 부여하는 고유한 기호체계를 사용하고 있다. 이 코드는 두 자리의 숫자, 두 자리의 숫자, 두 자리의 문자로 구성되며, 이들 각 두 자리의 단위는 마침표(.)로 구분된다. 첫 두 자리의 숫자는 첫 번째 수준의 주제어인 메인 카테고리와 두 번째 수준의 주제어를 의미한다. 이후에 나오는 두 자리의 숫자는 세 번째 수준을 구성하는 세부적인 카테고리를 의미하며, 두 자리의 문자는 세부 카테고리를 보다 세부적으로 구성하는 주제분야를 표시한다. 두 자리의 문자 앞에 사용되는 '-'(혹은 '+') 기호는 해당 주제가 세분될 수 있는지 여부를 알려주는 기호로 사용되고 있다.

〈표 3〉 PACS의 계층구조 및 코드의 예[21)]

Hierarchy to 3rd level	Hierarchy to 4th level	Hierarchy to 5th level	Notes
00. General	30. Atomic and Molecular physics	90. Geophysics, Astronomy, and Astrophysics	Broadest category ; there are 10 such codes from 00 to 90, in increments of 10
04. General relativity and gravitation	32. Atomic properties and interactios with photons	91. Solid earth physics	More specific category ; up to 9 such codes under each 1st-level category
04.65.+e Supergravity	32.10.Hq Ionization potentials, electron affinities	91.25.-r Geomagnetism and paleomagnetism ; geoelectricity	Fairly specific category ; "-" or "+" as 5th character denotes resence or absence, respectively, of 4th level
		91.25.F- Rock and mineral magnetism	Most specific category found in most of PACS, "-" or a lowercase letter as the 6th character denotes presence or absence, respectively, of 5th level
		91.25.fd Environmental magnetism	Most specific category found in PACS, the 5th character is the same as for the 4th-level code, but lowercase

이와 같이, PACS는 뚜렷하게 구분된 계층구조 및 고유한 코드체계를 사용하여 물리학 분야에서 사용되는 주제어 사이의 관계를 명확하게 표현해 주고 있다. 하지만, PACS 체계 내에는 많은 부분에서 주제어가 사용되지 않은 공간이 나타나고 있다. 이 공간들은 앞으로의 PACS의 확장을 위한 것이며, 새로운 주제 또는 아직까지 코드가 할당되지 않은 주제에 사용될 수 있도록 마련된 것이다.

나. PIRA 시소러스

PIRA는 물리학 분야의 주제어 및 서지 등을 조직할 수 있는 프레임워크를 제공하기 위해 고안되었으며, 물리학 분야의 교과과정 자료를 위한 참고자료를 조직하기 위해 개발된 교육자료 분류체계이다. 따라서, PIRA에 수록된 주제어의 범위는 교육자료 혹은 학습을 위한 목적으로 제한되어 있다.

PIRA는 9,265개의 교육자료를 그 대상으로 하고 있으며, 이들 자료와 관련된 주제분야를 조직하기 위한 계층구조를 지니고 있다. 〈표 4〉에서 보는 바와 같이, PIRA는 9개의 메인 카테고리로 구성되어 있으며, 수록된 주제와 용어는 물리학 분야에서 사용되는 교육용 교재의 체계를 따르고 있다. 그렇기 때문에, 물리학 분야 전체를 다루고 있는 것은 아니며 일반적인 물리학 관련 자료에서 다루고 있는 일부 주제분야가 배제되어 있기도 하다.

21) AIP - Physics and Astronomy Classification Scheme®(PACS®) Home page, 〈http://www.aip.org/pacs/pacs2010/about.html〉 [cited 2010. 8. 6].

〈표 4〉 PIRA 메인 카테고리

코드	카테고리 레이블
1	Mechanics
2	Fluid mechanics
3	Oscillations and waves
4	Thermodynamics
5	Electricity and magnetism
6	Optics
7	Modern physics
8	Astronomy
9	Equipment

〈표 4〉에 나타난 바와 같이, PIRA는 9개의 메인 카테고리로 구성되어 있으며, 각각의 카테고리에는 고유한 코드가 부여되어 있다. 이들 메인 카테고리는 다시 61개의 하위 카테고리로 나누어지는데, 이는 PIRA의 계층구조를 형성하게 된다. 이 계층구조는 4단계로 이루어지며, 하위 단계로 내려갈수록 보다 세부적인 주제분야를 표현하게 된다. 각각의 주제분야는 PIRA가 다루고 있는 교육자료를 주제와의 관련성의 정도에 따라 수록하게 되는데, 각각의 교육자료에는 특정 코드가 부여되며 이 코드에 따라 각각의 자료가 PIRA 구조 내에 위치하게 된다. 이 코드는 PIRA가 대상으로 삼고 있는 교육자료를 식별하고 검색하는데 사용된다.

PIRA 코드는 6자리의 알파벳과 숫자가 혼용된 형식을 지니고 있다. PIRA의 계층구조를 구성하는 카테고리에 이 코드가 부여되는데, PIRA 코드의 첫 번째 자리에 오는 숫자는 메인 카테고리를 의미한다. 두 번째에 위치하는 문자는 두 번째 수준의 주제분야를 나타내고, 세 번째와 네 번째에 위치하는 숫자는 해당 주제분야가 내포하고 있는 개념을 표현해 준다. 마지막에 위치하는 두 자리의 숫자는 PIRA가 대상으로 삼는 각각의 교육자료에 부여된다. 이와 같은 PIRA 코드는 다음 〈표 5〉와 같다.

〈표 5〉 PIRA 코드의 예(1D60.10 howitzer and tunnel)

1	Area	(mechanics)
D	Topic	(motion in two dimensions)
60	Concept	(projectile motion)
.10	Demonstration	(hotwitzer and tunnel)

〈표 5〉에서 보는 바와 같이, PIRA가 다루고 있는 각각의 교육자료에는 고유한 PIRA 코드가 부여되며, 상위 카테고리에 부여되는 코드는 관련된 교육자료를 포괄할 수 있는 의미적 범위를 규정해 주는 기능을 한다.

2. PACS 및 PIRA 시소러스 분석

시소러스를 선정한 후 다음 단계는, 선정된 시소러스에 대한 구조적, 의미적 분석이다. 이를 위해, 본 연구에서는 하향식(top-down) 접근방법을 사용하여 각 시소러스의 계층구조를 분석하고, 상향식(bottom-up) 접근방법을 사용하여 각 계층에서 주제를 표현하는 레이블로 사용된 모든 용어를 추출하였다.

각 시소러스의 계층구조와 용어를 분석한 결과, PACS는 5단계, PIRA는 4단계로 구성된 계층구조를 지니고 있다. 이들 시소러스는 계층구조의 체계에서 어느 정도의 차이를 보이고 있지만, 공통적으로 각 시소러스 내의 각 계층은 유사한 정도의 세부성을 지닌 개념들의 집합으로 이루어져 있다.

이에 반해, PACS와 PIRA 모두 카테고리별 하위 카테고리의 숫자에는 편차가 크게 나타나고 있다. PIRA의 경우, 메인 카테고리 중 '1 Mechanics'는 14개의 하위 카테고리를 지니고 있지만, '2 Fluid Mechanics', '8 Astronomy', '9 Equipment'는 세 개의 하위 카테고리만을 지니고 있다. PACS의 경우에는 그 편차가 더욱 심해서, '8000 Astronomy & Space Physics' 카테고리는 20개의 하위 카테고리를 지니고 있는 반면, '9000 Equipment'는 3개의 하위 카테고리만을 지니고 있다. 이러한 편차는 각 카테고리가 지닌 의미적 범위의 편차를 나타내고 있는데, 이들 카테고리의 의미적 체계는 시소러스의 구축 당시부터 미리 고정되어 있는 정적인 체계로서 기존의 전통적인 분류체계의 문제와 그 맥락을 같이 하고 있다.

이러한 시소러스 내의 의미적 편차 이외에도, 시소러스 사이의 의미적 편차 또한 나타나고 있다. PACS의 '8000 Astronomy & Space Physics'는 20개의 하위 카테고리를 지니고 있지만, PIRA의 '8 Astronomy' 카테고리는 3개의 하위 카테고리만을 지니고 있다. 이 두 가지 카테고리는 서로 다른 시소러스에 속해 있기는 하지만, 개념적으로 동일한 주제분야의 용어들을 수록하는 의미적으로 동일한 카테고리이다. 하지만, 이 동일한 개념을 세분하는 정도에는 큰 차이가 나고 있다. 이 세부적인 정도의 차이는 결과적으로 PIRA가 수록하고 있지 않은 용어들이 PACS에 수록되는 것으로 나타나고 있다. 이와 마찬가지로, PIRA에는 수록되어 있으나 PACS에는 수록되지 않은 카테고리 또한 나타나고 있다. 예를 들면, Planetary Astronomy라는 주제는 PIRA에는 수록되어 있으나(8A Planetary Astronomy) PACS에서는 이 주제분야를 다루고 있지 않다.

이러한 적용범위의 넓고 좁음의 차이와 수록하고 있는 용어 상의 차이는 시소러스를 구축하는 목적의 차이로부터 기인한 것이다. PACS는 물리학 분야 전반을 다루는 포괄적인 시소러스이지만, 높은 수준의 물리학 지식을 다루기 위해서 구축되었기 때문에 해당 도메인에 필요한 기본적인 개념들을 다루는데는 부족함이 있다. 반면, PIRA는 교육적인 목적으로 구축되었기 때문에 물리학 분야의 기본적인 개념들을 충분히 수록하고 있다. 반면, 교과과정을 뛰어넘는 높은 수준에서의 물리

학 개념들은 수록하지 않고 있다. 이러한 구축 목적의 차이는 두 시소러스에 수록된 전체 용어의 수에서도 큰 차이를 가져오고 있다. 〈표 6〉에서 보는 바와 같이, PACS와 PIRA 시소러스 사이에는 하위 카테고리에 수록된 주제분야로 세분될수록 포함된 용어의 수에 있어서 엄청난 차이가 나타나고 있다. 이러한 용어의 양적인 차이는 두 시소러스가 지닌 의미적인 범위에도 차이를 가져오게 되며, 이는 결국 동일한 분야에서 사용되는 시소러스임에도 불구하고 상당수의 용어가 PACS와 PIRA에 공통적으로 수록되지 않는 결과를 가져오게 된다.

〈표 6〉 PIRA와 PACS에서 사용된 용어의 수

	PACS	PIRA
메인 카테고리에 사용된 용어의 수(1st level)	10	9
하위 카테고리에 사용된 용어의 수(2nd level)	95	61
전체 용어의 수	6281	484

이러한 시소러스 구축목적에 따른 근본적인 차이로 인해서, 하나의 시소러스만으로는 물리학 분야 전체를 다루기에는 어려움이 있다. 따라서, 적용범위가 다른 PACS와 PIRA 시소러스를 통합하여 사용할 수 있다면, 이는 물리학 분야와 관련된 개념 전체를 다룰 수 있는 보다 효율적인 도구로 사용될 수 있을 것이다.

3. 개념 추출

선정된 시소러스의 구조를 분석하고 각 시소러스에 수록된 용어를 분석하고 난 후, 다음 단계는 이들 분석된 용어들로부터 개념을 추출하는 것이다. 이 단계에서의 개념추출은 선정된 시소러스에서 분석된 용어들 가운데 의미적으로 동일한 용어들을 확인하고, 이들이 지닌 공통적인 의미를 해당 용어가 지닌 근본적인 기본개념으로 추출하는 것이다.

하지만, 동일한 의미라 하더라도 각각의 시소러스에서는 상이한 용어를 사용하여 표현되는 경우가 많이 나타나고 있다. 예를 들면, 'Optics'라는 용어는 PACS와 PIRA 모두에서 공통적으로 사용되고 있으며, 그 표현과 의미가 모두 동일한 것으로 이는 의미적으로 정확하게 연결이 될 수 있다. 반면, PACS의 'Equations of state'와 PIRA의 'Change of state'는 동일한 의미를 지니고 있지만 그 의미를 표현하는 용어의 상이함으로 인해서 서로 다른 의미로 간주될 수 있으며, 이는 용어의 레이블에 기반한 시소러스의 통합에서는 의미적 모호성을 가져오는 결과를 초래하고 있다. 즉, 이러한 경우에는 동일한 의미라 하더라도 서로 연결되지 못하고 다른 개념으로 인식되거나 또는 의미적으로 차이가 있는 용어와 잘못 연결될 수도 있다.

이러한 문제를 해결하기 위하여, 본 연구에서는 용어의 표면적인 의미의 레이블이 아닌 용어가 지닌 근본적인 의미를 기반으로 각 시소러스에 수록된 용어를 분석하였다. 이 근본적인 의미는 상이한 용어로 표현되더라도 변하지 않으며, 용어가 다른 용어와 관계를 맺게 될 때 실질적으로 연결이 되는 것이다. 따라서, PACS와 PIRA에 수록된 용어를 분석하고 난 후 다음 단계는 이들 각각의 추출된 용어로부터 용어의 근본적인 의미인 핵심적인 개념을 추출하는 것이다. 또한, 이 단계에서는 각각의 용어가 시소러스의 계층구조 내에서 위치하고 있는 수준 또한 개념을 추출할 때 고려하였다. 시소러스가 지닌 계층구조는 각 시소러스가 수록하고 있는 용어의 의미적 범위에 영향을 미치게 되며 그로 인해서 동일한 의미를 지닌 용어라 하더라도 어느 수준에 위치하고 있는지에 따라 그 의미가 계층구조의 문맥에 따라 조금씩 변할 수 있기 때문이다.

각 시소러스에 수록된 용어로부터 기본적인 개념을 추출하기 위해서는 우선 동일한 의미를 지닌 용어들의 분석이 선행되어야 한다. 두 시소러스에서 공통적으로 사용되는 용어를 분석한 결과, 12개의 카테고리 용어가 PACS와 PIRA에서 의미적으로 동일하게 공통적으로 사용되고 있다. 이 12개의 용어들은 PACS와 PIRA의 메인 카테고리 및 두 번째 카테고리에서만 추출된 것이다. 왜냐하면 첫 번째와 두 번째 카테고리만이 각 시소러스에서 하위 카테고리를 지닌 의미적 카테고리로서의 기능을 하고 있으며, 세 번째와 네 번째(혹은 PACS에서의 다섯 번째) 카테고리는 특정적이고 세분적인 주제분야를 지칭하고 있기 때문이다. 또한, 이들 첫 번째 및 두 번째 카테고리는 각 시소러스의 전체적인 의미적인 범위를 규정해 주는 기능을 하고 있다. 이러한 이유로 공통적으로 사용되고 있는 용어를 추출할 때는 첫 번째와 두 번째 수준에서의 용어만을 대상으로 하였다. PACS와 PIRA 시소러스에서 공통적으로 사용되고 있는 카테고리 용어는 다음 〈표 7〉과 같다.

〈표 7〉 PIRA와 PACS 사이에 의미적으로 공통적으로 사용된 개념 비교

PACS	PIRA
mechanics(1)	general(1)
fluid mechanics(1)	fluid dynamics(2)
acoustics(2)	acoustics(2)
thermodynamics(1)	thermodynamics(1)
astronomy(1)	stellar system(2)
heat and the first law(2)	heat transfer(2)
change of state(2)	equations of state(2)
gas law(2)	physics of gases(2)
magnetic materials(2)	magnetic properties and materials(2)
optics(1)	optics(2)
nuclear physics(2)	nuclear physics(2)
equipment(1)	instruments(1)

* 비고: 괄호 안의 숫자는 각 시소러스 계층구조에서의 수준을 의미.

〈표 7〉에서 보는 바와 같이, PACS와 PIRA에서 사용된 카테고리는 대부분의 경우 주제를 나타내는 개별단위 용어의 조합으로 이루어져 있다. 이들 조합은 하나의 개념을 의미하기도 하지만, 반면 조합된 개념에 사용된 각각의 용어는 시소러스의 다른 계층에서는 다른 용어와 결합하여 전혀 다른 개념을 의미하기도 한다. 예를 들면, 'magnetic materials'와 'magnetic properties'는 모두 'magnetic'과 관련한 카테고리이지만, 이 'magnetic'이 'materials'과 연결되었을 경우와 'properties'와 연결되었을 경우에는 각각이 의미하는 범위가 달라지게 된다. 이 경우 'magnetic properties'는 'magnetic materials'를 포함하는 보다 넓은 범위의 의미를 지니게 된다.

이러한 용어 조합에 따른 의미의 변화를 최소화하고 관련된 용어를 수용할 수 있는 넓은 의미의 개념을 확인하기 위해, 본 연구에서는 조합된 용어들을 포괄할 수 있는 근본적인 기본개념을 추출하였다. 추출된 기본개념들은 넓은 의미를 지니고 있으며, 이 기본개념을 사용함으로써 조합되는 대부분의 개념들을 하위 개념으로 수록할 수 있는 확장성을 지니게 된다. 이들 추출된 기본개념은 〈표 8〉과 같다.

〈표 8〉 PACS와 PIRA로부터 추출된 기본개념

추출된 기본개념	PACS	PIRA
mechanics	mechanics(1)	general(1)
fluid	fluid mechanics(1)	fluid dynamics(2)
acoustics	acoustics(2)	acoustics(2)
thermodynamics	thermodynamics(1)	thermodynamics(1)
astronomy	astronomy(1)	stellar system(2)
heat	heat and the first law(2)	heat transfer(2)
state	change of state(2)	equations of state(2)
gas	gas law(2)	physics of gases(2)
magnetic	magnetic materials(2)	magnetic properties and materials(2)
optics	optics(1)	optics(2)
nuclear	nuclear physics(2)	nuclear physics(2)
equipment	equipment(1)	instruments(1)

* 비고: 괄호 안의 숫자는 각 시소러스 계층구조에서의 수준을 의미.

〈표 8〉에 나타난 바와 같이, PACS와 PIRA에서 공통적으로 사용되는 카테고리로부터 추출된 기본개념들은 해당 카테고리들을 포함할 수 있는 보다 포괄적인 의미를 지니고 있다. 따라서, 이들 추출된 기본개념들은 시소러스 내의 다른 용어들보다 상위에 위치하게 되며 메인 카테고리로서의 기능을 하는 기본패싯(baseline facet)으로서 사용된다. 이들 기본패싯은 PACS와 PIRA로부터 추출된 12개의 공통 개념으로부터 추출되었으며, 이들 공통적으로 사용되는 개념으로부터 핵심적인 개념만을 추출한 것이다. 따라서, 이들 기본패싯은 두 가지 시소러스에서 공통적으로 사용되고

있는 12개의 메인 카테고리에 비해 보다 포괄적인 의미를 지니고 있으며, 공통적으로 사용되고 있는 12개의 카테고리를 그 하위에 수록하는 최상위 카테고리로서의 기능을 하게 된다.

4. 계층구조의 구축

각 시소러스로부터 기본패싯을 추출한 다음 단계는, 각각의 시소러스에 수록된 용어를 관련된 기본패싯과 연결하는 것이다. 첫 번째 단계는 하나의 시소러스로부터 추출된 각각의 용어들을 그 용어가 지닌 근본적인 개념을 기준으로 관련된 기본패싯의 하위에 위치시키는 것이다. 이 클러스터링의 과정을 통해서, 추출된 용어들은 클러스터로 범주화되며 각각의 클러스터는 상호 관련된 개념들의 카테고리로서의 기능을 하게 된다. 이 과정은 기존의 메타데이터 매핑 등에서 사용되는 일대일(one-to-one) 매핑과는 달리 추출된 기본패싯을 기준으로 관련된 용어들을 개념적으로 서로 연결시키는 것이다. 또한, 이 개념적 클러스터링은 기본패싯과 관련된 개념들을 그 하위에 수록할 수 있도록 공간을 제공해 준다. 하지만, 기본패싯과 연결된 개념들 가운데 어떤 경우에는 PACS 시소러스의 낮은 계층에 위치한 용어가 PIRA 시소러스의 높은 계층에 위치한 용어와 하나의 패싯을 통해서 연결되며, 이와 반대의 경우도 종종 나타나고 있다. 이러한 경우에, 본 연구는 계층구조의 세 번째 수준 이하의 상이한 계층적 수준은 무시하고 해당 용어의 특정적인 의미는 보다 넓은 의미와 연결시켰다.

이 개념적인 연결 과정의 결과, 첫 번째 및 두 번째 수준에서의 카테고리에 사용된 22개의 용어가 기본패싯을 통해 정확하게 연결되고 있고, 전체적으로 136개의 용어가 개념적으로 포섭-수록 관계를 통해 연결되었다. 연결된 136개의 용어 가운데, PACS에 수록된 75개의 용어는 하위 카테고리를 가지고 있다. 예를 들면, '초전도체(superconductivity)'라는 용어는 PACS에서 11개의 하위 카테고리를 지닌 용어로 사용되고 있다. 이 용어는 PIRA에도 수록되어 있어 이는 두 가지 시소러스 사이에서 개념적으로 정확하게 연결이 된다. 하지만, PIRA에서의 '초전도체'는 하위 카테고리가 없는 특정적인 용어로 사용되고 있다. 이러한 경우 PACS에서 사용된 '초전도체'의 하위에 위치한 용어들은 기존의 일대일 매핑에서는 누락된 용어(missing terms)로 여겨져 왔으며, 전체적인 매핑의 신뢰성을 저해한다는 잠재적인 단점을 지니고 있다. 기존의 매핑방법이 지닌 이러한 문제점을 해결하기 위해, 본 연구에서는 PACS에서의 '초전도체' 및 그 하위의 11개 카테고리 모두가 PIRA의 '초전도체'와 연결되는 것으로 간주하였으며, PACS에 수록된 '초전도체'의 하위 카테고리들은 PIRA의 초전도체라는 개념을 상위 카테고리로 삼는 것으로 간주하였다. 이와 같은 접근방법은 매핑이 아닌 용어간의 개념적인 연결을 시도하여 이 용어를 상위 카테고리에 위치하는 넓은 개념으로 간주함으로써 상이한 체계를 지닌 두 시소러스 사이의 계층구조의 수준의 차이를 상쇄하고 두 시소러스 사이의 상호운용성을 확보하고 있다.

이와 같은 접근방법을 통해, 상호연결된 용어 가운데 기본패싯의 하위 카테고리로 사용될 수 있는 개념들을 분석하였다. 이들 분석된 개념들은 경우에 따라서는 상호 관련되어 있는 보다 세부적인 개념들을 수록할 수 있는 공간을 제공하게 되며, 관련된 세부 개념이 없을 경우에는 하나의 특정적인 주제어로서의 기능을 하게 된다. 따라서, 본 연구에서는 이들 분석된 개념들을 앞으로의 계층구조의 확장에 따라 패싯으로서의 기능을 할 수 있는 잠재적 패싯(potential facet)으로 구분하여 분석하였다. 이 결과, 두 시소러스 사이에서 공통적으로 사용되는 기본패싯과 각 기본패싯에 속한 하위의 잠재적 패싯들을 선정하였다(Appendix A 참조).

이렇게 선정된 기본패싯과 그 하위의 잠재적인 패싯을 통해서 PACS와 PIRA를 개념적으로 연결시킬 수 있는 계층구조가 구축되었다. 이 계층구조 내에서 기본패싯은 최상위에 위치하는 메인 카테고리의 의미를 지니며, 잠재적 패싯들은 메인 카테고리 하위에서 관련된 세부적인 개념들을 수록하는 하위 카테고리로서의 기능을 한다. 이들 하위 카테고리는, 경우에 따라서, 보다 세부적인 개념들을 수록하는 카테고리로서의 기능을 하거나 아니면 특정적인 주제어로서의 기능을 할 수도 있다. 이 계층구조는 개념적인 구조로서 PACS와 PIRA에 수록된 용어들을 수록하는 하나의 프레임워크를 형성하게 된다. 이 프레임워크는 PACS와 PIRA에 수록된 용어들을 직접 매핑하는 것이 아닌 용어가 지닌 개념들을 프레임워크를 통해서 개념적으로 상호 연결시켜 주는 것이다. 따라서, 이 프레임워크는 하나의 새로운 구조가 아닌 PACS와 PIRA 사이의 중개자로서의 기능을 하게 된다.

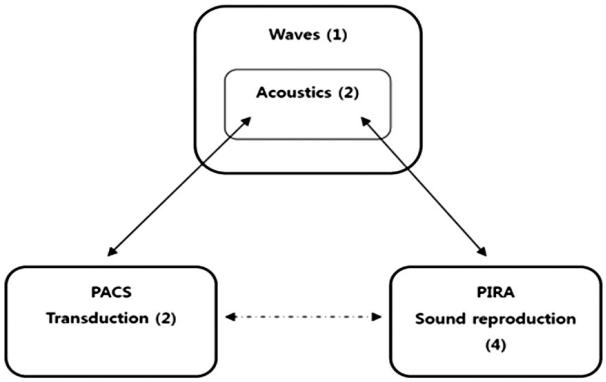
이 프레임워크의 최상위에는 기본패싯에서 추출된 메인 카테고리가 위치하며, 이는 프레임워크의 전체적인 개념적 범위를 규정해 준다. 메인 카테고리의 하위에는 관련된 잠재적 패싯들이 위치하여 세부적인 개념들을 수록할 수 있는 공간을 제공해 준다.

〈표 9〉 프레임워크에 수록된 카테고리 및 용어의 수

기본패싯(메인 카테고리)의 수	12
잠재적 패싯(하위 카테고리)의 수	95
패싯에 수록된 세부 용어	3,406

〈표 9〉에서 보는 바와 같이, 이 프레임워크의 구조는 해당 도메인의 주제를 표현해 내기에 적합한 수의 기본패싯과 잠재적 패싯들을 수록하고 있다. 패싯의 수가 너무 많아지면 전체적인 구조가 복잡하게 되는 결과를 초래할 수 있으며, 프레임워크를 구축하는데 적용하기 어려울 수 있기 때문에 적정 수의 패싯을 유지하는 것이 중요하다. 또한, 기본패싯과 잠재적 패싯 사이의 관계는 메인 카테고리와 하위 카테고리로 이루어지는 계층구조를 형성하게 되어 시소러스에 수록된 용어들 사이의 관계를 표현해 주게 된다. 이 프레임워크의 구조는 최상위 카테고리와 하위 카테고리, 그리고

하위 카테고리에 수록되는 세부 용어로 이루어진 세 단계의 계층을 형성하는데, 이는 물리학 분야에서 사용되는 수많은 주제분야를 표현하기에는 다소 부족하다고 여겨진다. 하지만, 이 프레임워크는 상이한 시소러스를 병합하는 것이 아닌 개념적으로 연결시키기 위한 체계이기 때문에, 이 프레임워크의 계층구조를 통해서 각 시소러스에 수록된 세부 용어를 그 근본적인 개념에 기반해서 상호 연결시킬 수 있으며, 이의 예는 다음 〈그림 1〉과 같다.



〈그림 1〉 개념 기반 프레임워크를 통한 시소러스 연결의 예
(비고: 괄호 안의 숫자는 계층구조 내에서 용어가 위치한 수준을 의미함)

〈그림 1〉에서 보는 바와 같이, PACS 시소러스 내의 'Transduction'이라는 용어와 PIRA 시소러스의 'sound reproduction'은 '음성 혹은 소리의 변형'이라는 공통된 의미를 지니고 있다. 하지만, 이 동일한 의미는 각 시소러스에서 상이한 용어를 사용하여 표현하고 있으므로, 용어의 레이블에 기반한 시소러스 통합에서는 이 두 가지 용어를 동일한 의미를 지닌 것으로 통합하기가 어려워진다. 반면, 개념에 기반한 패싯 프레임워크에서는 이들 상이한 용어가 지닌 근본적인 의미를 추출하여 이를 패싯 프레임워크의 계층구조에 적용시킴으로써 'Wave'라는 기본패싯 하위의 패싯 'Acoustics'를 통해 이 두 가지 용어를 의미적으로 상호 연결시킬 수 있게 된다.

이와 같이 패싯으로 변환된 개념에 기반해서 구축된 프레임워크는 상이한 시소러스를 서로 연결시킴으로써 물리학 분야에서 사용되는 개념들을 검색하고 조직하는 지식기반으로서 사용될 수 있다. 이 개념에 기반한 패싯 프레임워크에 수록된 패싯들은 용어의 레이블로서의 기능을 하는 것이 아니라 용어로부터 추출된 개념으로서의 기능을 하며, 관련된 개념이 어떤 용어를 사용해서 표현되던지 간에 동일한 개념이면 서로 연결시킬 수 있는 중개자로서의 역할을 하게 된다. 따라서, 각각의 개념은 각 시소러스에 수록된 용어와 의미적으로 연결된다. 또한, 새로운 주제분야가 나타날 경우, 각각의 새로운 주제가 지닌 개념을 확인하고 그 의미적 범위를 설정함으로써, 상위의 기본패싯이나

하위의 잠재적 패싯으로 추가할 수 있으며, 패싯 프레임워크의 계층구조가 세 단계 이상의 보다 깊은 수준으로 확장될 수 있는 확장성을 지니게 된다. 이 확장성은 연결되는 시소러스가 지닌 상황에 따라 혹은 상이한 시소러스를 사용하는 이용자의 목적에 따라 변용될 수 있기 때문에, 이용자들의 다양한 정보요구에 즉각적으로 부응할 수 있는 융통성을 지니게 되며, 보다 효과적인 정보의 조직 및 검색을 지원할 수 있게 된다.

기존의 매핑을 통해 시소러스를 통합하는 접근방법은 또 하나의 새로운 시소러스를 형성하게 되지만, 이 개념적인 패싯 프레임워크는 새로운 시소러스를 형성하는 것이 아닌, 상이한 시소러스를 서로 연결시킴으로써 기존의 시소러스의 특성은 그대로 유지하면서 두 시소러스를 의미적으로 연결시켜 주게 된다. 이 프레임워크를 이용하면 상이한 두 시소러스에 수록된 동일한 개념을 서로 연결시키고 서로 관련되어 있는 개념들을 다른 시소러스의 계층구조 안에 개념적으로 포함시킬 수 있게 된다. 따라서, 상이한 시소러스 사이의 중개자로서의 기능을 하여 시소러스 사이의 상호운용성을 확보함으로써 물리학 분야의 정보자원을 보다 효과적으로 검색하고 조직할 수 있도록 해준다. 또한, 시소러스의 구축과정에서 비롯된 고유한 특징 및 용어들 사이의 특정적인 관계들을 잃지 않고 그대로 유지하면서 이들 상이한 시소러스를 통합적으로 사용할 수 있는 활용성을 제공해 준다. 그러므로 개념에 기반한 패싯 프레임워크를 구축함으로써 해당 도메인에서의 정보를 보다 효율적으로 조직하고 검색할 수 있는 보다 효과적인 접근방법이 될 수 있다. 이외에도, 용어의 레이블을 통한 것이 아닌 각각의 용어가 지닌 근본적인 개념에 기반을 두고 있기 때문에, 상이한 언어로 구축된 다양한 시소러스를 통합적으로 운용하는데에도 효율성을 발휘할 수 있을 것으로 기대된다.

V. 결 론

현재의 정보조직이나 정보검색이 당면한 여러 가지 문제를 해결할 수 있는 방법 가운데 하나로 시소러스와 같은 통제된 용어의 집합에 대한 필요성이 점차 증대되어 왔다. 하지만, 시소러스가 지닌 정적이고 고정적인 구조와 동일 도메인 내에서 각각의 시소러스가 지닌 의미적인 범위의 차이, 각 시소러스가 지닌 이질적인 특성 등으로 인해 여러 시소러스의 공존이 정보의 조직 및 검색에 있어서의 어려움을 초래하고 있다. 이와 같은 시소러스가 지닌 문제들을 해결하기 위해, 최근 문헌정보학 분야에서는 시소러스를 통합하여 기존의 시소러스가 지닌 한계를 극복하려는 움직임이 많이 나타나고 있다. 하지만, 시소러스를 통합하기 위한 기존의 접근방법은 정적이고 고정된 분류체계에 기반을 두고 있으며, 시소러스에 수록된 용어의 레이블 간의 매핑에 초점을 맞추고 있기 때문에 디지털 정보와 같은 동적이고 변화가 심한 정보자원을 조직하는데 있어서 많은 어려움을 겪고 있다. 이에 본 연구는 기존의 정적이고 고정된 분류체계를 통한 정보조직이 아닌 동적이고 융통성

있는 특성을 지닌 개념기반 패싯 프레임워크를 구축하여 상이한 시소러스를 효과적으로 연결할 수 있는 방안을 제시하였다.

이 패싯 프레임워크를 구축하기 위해서 물리학 분야에서 널리 사용되고 있는 시소러스인 PACS와 PIRA를 선정하여 이들 시소러스의 구조와 특성을 분석하고, 각 시소러스에 수록된 용어로부터 근본적인 개념을 추출하였다. 이 추출된 개념을 기반으로 기본패싯과 보다 세부적인 개념의 잠재적 패싯을 구성하여 패싯화 된 개념적인 계층구조를 구축하였다. 이를 통해 각 개념들 사이에 설정할 수 있는 여러 가지 관계를 확인하여 물리학 분야에서 사용되는 주제를 표현할 수 있는 하나의 개념적인 지식기반을 제공하였다. 또한, 주제를 표현하는 용어의 레이블이 아닌 용어가 지닌 근본적인 개념에 기반을 두게 됨으로써, 표현상의 차이로 인해 매핑과정에서 발생할 수 있는 용어의 누락을 방지할 수 있으며 보다 효율적이고 신뢰성 있는 시소러스의 통합을 이끌어낼 수 있다. 이외에도, 기존의 통합을 위한 접근방법은 용어 사이의 일대일 매핑을 통해 상이한 시소러스를 직접적으로 통합함으로써 정적이고 고정적인 또 하나의 시소러스를 생성한다는 문제가 있지만, 개념기반 패싯 프레임워크는 시소러스가 지닌 개념에 기반하여 구축한 체계로, 상이한 시소러스를 직접 통합하는 것이 아닌 시소러스 간에 존재하는 관련된 개념들을 개념적으로 상호 연결시켜 주는 역할을 한다. 따라서, 각각의 시소러스가 지닌 고유한 특성을 변경하지 않고, 하나의 시소러스를 통해서 여러 가지 다른 시소러스를 동시에 활용할 수 있는 매개체로서의 기능을 하게 된다. 반면, 패싯 프레임워크에 사용되는 개념의 추출과정은 각 시소러스가 지닌 계층구조에 기반을 두고 있기 때문에, 계층구조가 아닌 평면적 구조를 지닌 시소러스에 패싯 프레임워크를 적용할 경우, 추출하는 개념들 사이의 관계가 명확하게 나타나지 않을 수 있다는 한계를 지니고 있다. 하지만, 패싯 프레임워크의 구축은 개념들 사이의 매핑이 아닌 개념적인 연결에 기반한 것으로, 동일한 의미를 지닌 용어들은 개념적으로 연결되어 각각의 시소러스가 통합적으로 활용될 수 있는 기반을 제공해 준다.

결론적으로, 본 연구에서 제안한 개념기반 패싯 프레임워크는 기존의 시소러스 통합이 지닌 단점을 보완하고 여러 가지 상이한 시소러스를 보다 효율적으로 통합할 수 있는 대안적인 접근방법을 제시한다는 점에서 중요한 의미를 지니게 될 것으로 기대된다.

〈참고문헌은 각주로 대신함〉

Appendix A. 메인 카테고리와 하위 카테고리로 사용되는 패싯

메인 카테고리	하위 카테고리(2nd level)
mechanics(11)	measurement, motion in dimensions, relative motion, newton's law, statistics of rigid bodies, torque, gravity, work and energy, linear momentum, rotational dynamics, properties of matter,
fluid(5)	surface tension, statics of fluids, dynamics of fluids, weather dynamics, radiative transfer
acoustics(5)	oscillations, wave motion, waves, instruments, sound reproduction
thermodynamics(9)	thermal properties of matter, change of state, kinetic theory, distribution functions, entropy, distribution functions, chemical reactions, free energy, engines and refrigerators
astronomy(11)	planetary astronomy, stellar astronomy, cosmology, astrometry, galactic astronomy, radio astronomy, solar system, ionosphere, spacecraft, auroral physics, magnetosphere
heat(6)	heat transfer, convection, conduction, radiation, adiabatic process, temperature
state(8)	change of state, phase transition, cooling by evaporation, dew point, humidity, vapor pressure, sublimation, critical point
gas(8)	gas law, constant pressure, constant temperature, constant volume, ideal gas, fermi gas, bose gas, van der waals gas
magnetic(9)	magnetic properties, magnetic fields, inductance, electricity, electrostatics, capacitance, resistance, semiconductors, electromagnetic radiation
optics(11)	geometrical optics, photometry, diffraction, interference, color, spectroscopy, polarization, eye, laser physics, matrix methods, modern optics
nuclear(9)	radioactivity, nuclear reactions, elementary particles, relativity, quantum physics, quantum electrodynamics, solid state, nuclear astrophysics, nuclear structure
equipment(3)	support systems, electronic instruments, mechanical instruments

* 비고: 괄호 안의 숫자는 하위 패싯의 수를 의미함.