토픽맵을 이용한 시소러스의 구조화 연구*

A Study on the Thesaurus Construction Using the Topic Map

남 영 준(Young-Joon Nam)**

초 록

시소러스의 효율성을 유지하기 위해서는 지속적인 용어 관리가 절대적으로 필요하다. 실제적으로 특정 주제영역의 정보와 키워드들은 생성과 분화, 소멸 과정 등이 동적으로 이루어지기 때문에 시소러스의 효율적인 관리가 매우 어려운 실정이다. 따라서 시소러스의 구조와 관리를 유연하게 수행할 수 있는 방안이 필요하다.

이에 따라 본 연구에서는 토픽맵의 기본요소인 토픽과 대상물, 연관관계 등을 활용하여 시소러스 관리를 위한 구조화 방안을 제안하였다. 한편 구조체계의 맵핑 알고리즘과 구조체계의 병합 알고리즘을 이용한 시소러스 기본관계와 세부관계 표현 방법도 제안하였다. 또한 토픽 타입을 이용한 연결중심문서를 기준으로 디스크립터의 확장과 디스크립터의 대치 방안을 제시하였다. 특히, 고정된 개념을 통한 이중 용어관리라는 새로운 방안도 개발하였다. 이는 시간과 공간의 비종속적인 개념을 표현하는 용어를 고정시키고, 해당 개념의 범주에 속하면서 외부의 정보적 상황에 따라 디스크립터를 자유롭게 선정하는 방법이다.

ABSTRACT

The terminology management is absolutely necessary for maintaining the efficiency of thesaurus. This is because the creating, differentiating, disappearing, and other processes of the descriptor become accomplished dynamically, making effective management of thesaurus a very difficult task. Therefore, a device is required for accomplishing methods to construct and maintain the thesaurus.

This study proposes the methods to construct the thesaurus management using the basic elements of a topic map which are topic, occurrence, and association. Second, the study proposes the methods to represent the basic and specific instances using the systematic mapping algorithm and merging algorithm.

Also, using a hub document as a standard, this study gives the methods to expand and subsitute the descriptors using the topic type.

The new method applying fixed concept for double layer management on terms is developed, too. The purpose of this method is to fix the conceptual term which represents independent concept of time and space, and to select the descriptor freely by external information circumstance.

키워드: 토픽맵, 시소러스, 온톨로지, 시소러스 관리, 시소러스 병합
topic map, topic, thesaurus, thesaurus merging

---

* 본 연구는 한국과학재단 기초연구프로그램(R01-2003-000-11588-0)의 지원으로 연구되었음.
** 중앙대학교 문과대학 문헌정보학과 교수(namyj@cau.ac.kr)
■ 논문접수일자 : 2005년 7월 26일
■ 게재확정일자 : 2005년 9월 17일

1. 서 론

검색효율을 높이기 위한 노력은 과거로부터 지금까지 정보학 분야에서 추구하는 일관된 목표가운데 하나이다. 특히 정보가 무한에 가까운 규모로 증가하는 인터넷시대 검색효율은 21세기 정보화 시대가 지향하는 목표이다.

이에 따라 문헌정보학 분야에서는 이러한 문제점을 극복하고 대용량의 인터넷 정보시대의 안정적인 검색효율을 확보하기 위해 많은 연구를 경주하고 있다. 구체적인 방법으로 도서관은 전통적으로 주제명 표목표와 분류표와 같은 색인도구를 이용하여 효과적인 정보검색을 시도하였다. 전통적인 검색도구는 적정한 규모의 데이터베이스 검색에는 효과적이나 대용량 규모의 데이터베이스를 대상으로 하는 검색과정에서는 적정한 검색효율을 제공하지 못한 제한점을 갖고 있다.

이러한 제한점을 극복하기 위해 도서관은 기존의 검색도구를 의미적으로 구조화하였다. 시소러스는 주제명 표목표나 분류체계가 갖고 있는 평면적인 키워드 배열을 특정 용어에 대한 개념적 구조화를 표현한 진일보한 검색도구이다. 시소러스는 주제명 표목을 분류체계가 갖는 계층성과 연관성을 이용하여 전체 표목(디스크립터)을 개념적으로 구조화함으로써 기존 검색도구가 갖는 키워드 매칭과 포괄적 단순검색의 단점을 극복하였다.

이와 같은 시소러스는 기본적으로 인쇄자료의 효율적 검색을 의도한 검색도구이기 때문에 인터넷 자원 검색을 위해서는 인터넷자원의 특성을 고려할 필요성이 제기되었다. 예를 들면, 미래 시소러스는 XML로 표현되는 인터넷 자원의 구조적 정보와 전문검색에 필요한 언어적 및 의미적 표현을 수용할 수 있는 형태로 변화되어야 하는 것이다. 특히 대부분의 시소러스는 특정 주제 분야의 마이크로적 성격을 갖고 있기 때문에 디스크립터가 갖는 의미적 유한성과 다른 개념과의 연관성을 유지해야만 검색효율의 개선효과를 얻을 수 있다.

한편 XTM(XML Topic Map)은 토픽과 연관관계, 대상물을 이용하여 개념과 실물 정보, 타 정보와의 관계를 표현하는 토픽맵의 표준 온톨로지로써 특정 주제 영역의 용어간 개념을 구조화하여 의미적 관계성을 표현할 수 있는 온톨로지 표현언어이다. 한편 일반 시소러스의 디스크립터는 특정 영역의 자료를 검색할 수 있는 후조합 색인언어의 특성을 갖고 있으나, 토픽맵은 색인된 특정 용어에 배정된 정보자원도 검색할 수 있는 종합적인 검색도구이다. 따라서 토픽맵은 모든 분야를 수용하는 것보다 특정 분야의 개념 구조화에 상대적으로 유리하다.

본 연구에서는 이 점에 착안하여 시소러스의 디스크립터 구조화 과정을 토픽맵의 구축 알고리즘을 이용하여 특정 영역의 마이크로 시소러스를 구축할 수 있는 이론과 방법을 제안하고자 한다.

2. 토픽맵의 관련 연구

토픽맵은 2000년에 정식으로 세계 표준안이 만들어졌으며, 2002년 5월 그 개정판이 발표되었다(ISO/IEC 2002). 토픽맵은 온톨로지기반의 색인어 지도(index map)로써 컨텐츠 관리까지 가능하도록 하는 알고리즘이다. 초기 연구는 외국에서 주로 이루어졌으며, 토픽맵의 표현에 관한 것과 이를 이용한 검색 적용으로 구분할 수 있

다. 특히 토픽의 표현을 그래픽 형태의 트리 구조를 적용하여 토픽간 거리와 연관성을 이용하여 검색효율을 높일 수 있다는 연구(Benedicte Le Grand, Michel Soto 2000)가 있다. 이와 같은 이론에 대해 정준원(2003)은 그래프 기반탐색으로 정보의 연관 관계와 구조를 잘 표현할 수 있으나, 수많은 노드로 이루어진 토픽과 간선으로 인해 오히려 정보를 파악하기 어렵다는 부정적인 의견을 제시하고 있다. 이와 같은 지적은 시소러스의 표현형식가운데 국제 도로연구센터(International road research documentation)시소러스의 화살표 표시 방식과 유사한 표현 방식이 사용자의 가시성을 떨어트려 오히려 효율성을 저하시키는 것과 동일한 단점을 갖는다는 연구와 일치하고 있다(남영준 2002).

국내에서의 토픽맵 관련 연구는 활용적인 측면과 기술적인 측면의 연구로 구분할 수 있다. 활용과 관련된 연구는 e-커머스분야의 활용이 상대적으로 많았다. 정원규(2003)는 동서양의 철학사상가운데 벤담의 사상을 토픽맵 개념으로 분류 및 연관관계를 토픽으로 처리하였다. 그는 토픽맵이 일련의 의미를 갖기 위해서는 컴퓨터가 직접 용어의 맥락과 의미를 이해할 수 없으므로 필요한 자료에 사람, 특히 전문가가 일일이 분류기호(tag)를 기입해 주어야 한다는 한계를 인정하였다. 고세영(2003)은 이 기종간의 상품분류체계를 통합하기 위한 도구로써 토픽맵을 이용하였다. 그는 이 연구에서 분류체계의 계층관계와 연관관계를 정의하고 이를 모델링하고 이를 통합하는 방법을 사용하였다. 고유미(2005)는 특허분야의 이종간 분류체계를 통합하는 방안을 제시하였다. 그는 특허문서의 서지정보를 토픽으로 설정하여 정보서비스 목적에 따라 온톨로지를 모델링하는 방법을 사용하였다.

한편 기술적인 연구로써 정호영 등(2003)은 토픽맵의 XTM을 이용하여 부가적인 메타 데이터-토픽, 어커런스, 토픽과 토픽간의 연관관계-를 기술함으로써 키워드 검색뿐 아니라 세미나 자료 지식에 대한 생성/유지/관리의 용이함을 지원하는 토픽들 간의 분류/연관관계에 의한 검색을 지원하는 방안을 제시하였다. 이은아(2003)는 시맨틱 네비게이션 시스템으로 토픽맵을 그래프로 브라우징할 수 있는 실험적인 네비게이션 시스템을 구축하여 이용자 중심의 토픽맵을 구성하고, 토픽을 효율적으로 보여줄 수 있는 토픽맵 어플리케이션을 개발하였다. 정준원(2003)은 지식맵의 효율적인서비스 방안을 제안하고자 토픽맵에 기반한 탐색 및 캐쉬를 이용한 정보전송 기법에 대해서 제안하였다. 그는 토픽맵을 연관성 중심으로 탐색을 지원하는 환경을 제안하고, 이 환경에서 연관된 정보의 캐쉬를 생성하여 전송하는 방법을 사용하였다. 유우종 등(2004)은 워드넷을 하나의 온톨로지로 적용하여 워드넷에 수록된 단어를 토픽으로 처리하고, 단어간 연결은 연결망으로 적용할 수 있는 방안을 제시하였다.

3. 토픽맵 모델

토픽맵은 ISO와 함께 IEC(the International Electrotechnical Commission, 국제전자기술협의회)에서 공동으로 개발한 지식구조화를 위한 국제 표준이다. 본 표준에서는 토픽맵을 다음과 같이 기본요소와 구조속성으로 구분하여 정의하고 있다.

3. 1 토픽맵의 기본요소

토픽맵은 특정한 개념을 나타내는 토픽의 개념과 연관성을 다른 개념구조간의 상호교환을 위해 정의한 일련의 표준화된 체계이다(ISO/IEC 2002). 즉 하나이상의 상호연관성을 갖는 문서의 집합을 토픽맵이라 한다. 일반적으로 토픽맵은 다음과 같은 가장 기본적인 구성요소를 갖는다(TopicMaps.Org 2001, ISO/IEC 2002).

- 토픽(topics) : 전자자원내에서 실세계의 주제를 대표하는 하나의 자원
- 대상물(occurrence) : 토픽에 해당되는 정보원의 주소
- 연관관계(association) : 토픽간의 관계 표시

즉 토픽맵은 기본적으로 토픽을 비롯하여 크게 연관관계, 대상자료 등 3개의 요소로 구성된다. 마지막으로 대상물은 하나의 토픽에 연관관계로 연결되는 자료를 의미한다.

3.1.1 토픽

토픽은 일상적으로 실생활에 사용하는 단어로 표현되며, 대상문헌의 핵심 개념들이 토픽이다. 예를 들면, 토픽은 유형의 모든 것과 인간이 생각할 수 있는 무형적인 것까지 단어로 표현될 수 있다. 이러한 단어들을 토픽이라고도 하며, 이를 토픽 개체명(names)으로도 표현한다. 이러한 토픽 개체명과의 상호관계가 성립할 경우, 그 각각의 관계는 연관관계로 표현될 수 있다(TopicMaps.Org 2001). 내용적으로 토픽 개체명은 토픽과 토픽의 연결을 나타내는 개념명이라 할 수 있으나, 형식적으로 개체명과 토픽은 분명하게 구별되지 않는 특성을 갖고 있다. 또한 토픽의 별명(alias)으로써 토픽개체명이 있다. 이를 세분하여 설명하면 다음과 같다.

- 토픽 : 토픽은 토픽맵에서 객체 혹은 하나의 노드를 나타낸다. 이 토픽은 토픽과 주제간에 일대일 관계를 가져야만 한다. 모든 토픽은 하나의 단일 주제를 나타내고 모든 주제는 유일한 토픽으로서만 표현되어야 한다.
- 토픽타입 : 토픽은 해당 개념에 갖는 주제종류에 따라 류구분을 할 수 있다. 하나의 토픽맵에서 토픽은 객체지향 모델에서의 토픽타입(topic type)에 대한 인스턴스이다. 이것은 책자형 자료의 복수색인에서 보여주는 분류속성과 동일한 개념이다. 즉 하나의 토픽과 그 형태간의 관계는 하나의 전통적인 류와 사례 관계로 표현할 수 있다. 토픽맵에 존재하는 토픽은 토픽 타입이 1개 이상의 여러 타입에 속할 수 있다. 따라서 토픽 타입은 그 자체가 하나의 토픽이다(정호영 2003). 따라서 토픽 타입간에도 상하관계와 연관관계가 존재할 수 있다. 토픽이 시소러스에서 용어와 의미, 주제분야로 설명할 수 있다. 소프트웨어 도큐멘테이션에서 토픽은 기능과 변수, 객체, 방법을 나타낸다. 토픽 타입은 토픽에 대한 일련의 선언 과정으로써 사용자 임의대로 규정할 수 있다.
- 토픽개체명 : 토픽은 일반적으로 대표하는 명칭을 갖고 있으며, 해당 주제영역에서 고착화되어 있는 명칭을 사용할 수 있다. 그 명칭에는 공식명과 간략명, 별명, 일반명 등이 포함되며 기준어가 표준화만이 채택될 필요는 없으나, ISO에서는 기본명(base name)과 표현명(display name), 정렬명(sort name)만으로 제한하며, 국제표준에 의해 등록된 것만 사용한다(ISO/IEC 2002).

따라서 토픽은 토픽타입과 토픽개체명과 엄격하게 구분할 수 있으나, 실제적으로 대부분 동일

한 의미로 사용하고 있다.

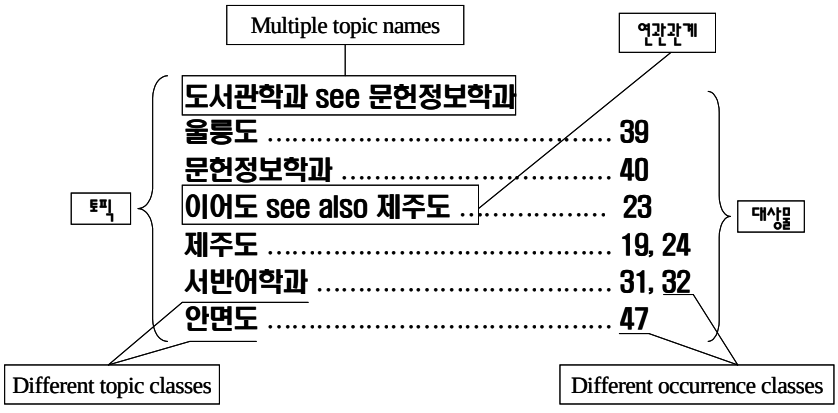
3.1.2 대상물(occurrence)

하나의 토픽은 하나이상의 관련있는 정보원과 연결되어질 수 있다. 이 때 관련있는 정보원이 토픽의 대상물이다. 대상물은 하나의 논문이나 백과사전내에 하나의 기사, 그림이나 비디오 등이 될 수 있다. 그 가운데 웹상에 존재하는 정보원은 전통적으로 웹주소(URI 혹은 URL 등)이다. 특히 대상물 관계는 토픽이 갖고 있는 두개의 연결구조를 분리할 수 있는 장점을 갖고 있다. 즉 특정 토픽에 관련되어 있는 대상물을 특성에 따라 구분할 수 있는 것이다. 예를 들면, 일정 토픽에 연관된 대상물을 단행본과 기사논문으로 구분하면, 대상물을 기존 토픽과 별도로 야구와 관련된 단행본과 축구와 관련된 단행본을 단행본으로써 야구관련 자료와 축구관련 자료로 분리하여 다른 관점(역할)으로 해당 정보원집단을 구분할 수 있다. 이러한 역할기호를 대상물 역할(occurrence role)이라 하며, ISO/IEC에서는 Hytime으로 표현하고, 페퍼(Pepper 2005)는 이를 대상물 역할과 별도로 대상물 유형(occurrence type)으로 확장하여 기술하고 있다.

3.1.3 연관관계(association)

연관관계는 두개이상의 토픽간에 각각의 관계를 정의한 것이다. 토픽은 사용자의 관점이나 토픽맵을 필요로 하는 사용자의 용도에 따라 개념을 군집화하거나 분리할 수 있다. 따라서 토픽은 해당 토픽들이 갖는 유형에 따라 연관관계가 설정될 수 있다. 연관관계의 유형은 "written__by"을 비롯하여 "takes__place__in" 등과 같이 의미연결망의 연결어 등이 존재한다. 이때 사용되는 연관관계의 유형은 해당 토픽맵을 구성하는 과정에서 표준으로 선언하여, 대상물 역할 등과 혼란이 발생하지 않도록 한다. 이와 같은 연결구조를 가짐으로써 토픽맵의 표현력을 극대화할 수 있다. 즉, 연관관계 지식은 추론 검색을 위한 지식베이스로 활용할 수 있는 것이다. 이와 같이 토픽간의 의미적 연관관계를 구조화함으로써 다양한 질의에 대해 효과적인 검색을 제공하는 지식베이스로 활용할 수 있는 것이다.

다음 〈그림 1〉은 토픽맵의 기본적인 관계를 나타낸 것이다. 토픽은 도서관학과를 비롯하여 울



〈그림 1 〉 토픽맵의 기본요소

릉도 등과 같이 특정 주제를 표현하는 단어이다. 이어도와 제주도는 다른 개념으로써 서로 독립된 개념이지만 연관성을 통해 의미적 관련성을 표시하고 있다. 이때 의미적 관계를 표현하는 'also'는 연관관계의 연결어가 된다. 이에 비해 위의 '도서관학과'는 도서관학과와 문헌정보학과가 의미적으로 동일한 개념을 표현하기 때문에 대상물정보를 갖지 않는다. 즉 이 관계는 연관관계가 아니라 이형동의어이며, 'see'는 동일 개념의 연결어로 정의되어 있다. 또한 각 토픽에 해당하는 소재정보(페이지)가 대상물에 해당한다.

3. 2 토픽맵의 구조

토픽맵의 구성은 해당 자료의 기본 프레임과 위치정보, HyTime에 정의된 하이퍼링크 모듈로 이루어진다. 따라서 토픽맵은 HyTime에서 정의하고 있는 'bounded object set(BOS)'이라 할 수 있다. BOS의 hub document(중심연결문서)에는 반드시 토픽맵의 구성을 지원하는 개념이 선언되어야 한다. 이를 위해 HyTime에서 정의된 하이퍼링크 구조가운데 하나인 variable link 개념을 필요로 한다.

3.2.1 토픽맵 구조적 형태

토픽맵 구조형식(Topic Map Architecture Form)에 정의된 요소는 *topicmap*과 *added themes*이다.

*topicmap*요소는 국제표준에 정의된 토픽맵 구조로 이루어진 모든 문서의 문서요소로 적용된다. 이에 비해 *added themes* 속성은 기 부여된 토픽특성에 참조할 수 있도록 부가적인 주제를 부여하는 역할을 수행한다. 이 경우에 *topicmap* 요소는 HyTime구조로 구성된 문서 요소에서 추출한다. 구조화를 위해 사용하는 모든 요소는 HyDoc에 적용되는 것을 원형대로 수용한다. 또한 이러한 구조에 따라 토픽 연결(topic link)을 위한 요소는 토픽연결 구조 형태(Topic Link Architectural Form)를 비롯하여 토픽명 구조 형태(Topic Name Architectural Form), 토픽 대상물 구조 형태(Topic Occurrence Architectural Form) 등 세가지 요소형태로 구성된다. 각각의 토픽구조 요소형태는 세부적인 요소를 갖고 있으며, 세부 요소들은 토픽맵 작성의 부수적이며, 설명적인 역할을 수행한다.

3.2.2 토픽맵의 구조 속성

토픽맵은 하나의 다차원적 토픽 공간으로 정의된다. 여기에서 공간은 토픽의 모여있는 하나의 개념사전이며, 특정 토픽간에 존재하는 토픽들이 일목요연하게 정렬되어 있어야만 하고, 하나의 토픽과 다른 토픽간의 경로가 정의되어 있는 관계가 설정되어 있는 것을 의미한다. 예를 들면 두 개의 토픽은 이 공간에서 하나의 연관관계로 연결되어질 수 있으며, 또한 하나의 대상물로써 연결되어질 수 있다.

또한 정보 객체는 성질(property)과 내부적으로 부여된 해당 특성의 값을 가질 수 있다. 이들 특성을 패싯 타입으로 표현한다. 예를 들면, 특정 단어의 패싯은 다면체의 한면 혹은 분석된 결과물, 혼합관점에서 한 측면이라 정의할 수 있다. 은유적인 표현으로써 특성은 특정 사상을 인식할 때 사용되며, 하나의 패싯은 새로운 관점을 생산하는데 사용되는 정보객체의 특성이 될 수 있다.

여러개의 토픽맵은 동일정보원에 대해 토픽구조정보를 제공할 수 있다. 토픽맵의 구조는 다른

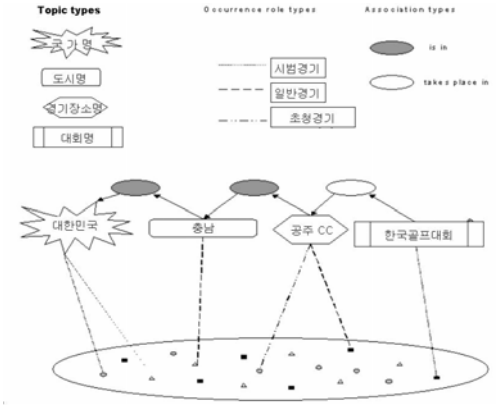
토픽맵을 복사하거나 수정할 필요없이 토픽맵간 병합(merging)할 수 있다. 왜냐하면, 토픽맵은 비종속적인 특성으로, 일련의 정보객체에 그대로 오버레이나 확장이 가능한 특성을 갖고 있기 때문이다. 토픽맵의 기본 기호(notation)는 SGML의 정의를 사용하고 있다. 따라서 통합 대상이 되는 토픽맵간에는 하나 이상의 토픽이 반드시 SGML형태로 이루어져야 한다. 이 때 통합이 필요한 별개의 토픽맵은 HyTime에서 정의하고 있는 BOS를 이용하여 구별되어질 수 있다. 즉 토픽맵의 통합과정은 hub document를 작성하여 이루어지고, 이 hub document는 통합 대상이 되는 BOS에 대해 기술한다. 또한 이 통합과정에는 토픽맵 문서들의 토픽명과 동일한 이름을 갖는 토픽링크가 포함된다. 통합된 토픽맵 문서 중 하나를 새로운 형태의 변경하는 과정은 새로운 별개의 BOS를 다시 처리하는 과정이다. 이 과정에서 두개의 다른 주제가 동일한 범주내에서 동일한 이름을 갖는다면, 두개의 토픽맵은 통합시 충돌이 발생한다. 이때에는 두 주제의 통합은 그들이 존재하는 토픽맵 문서들에 *addthms* 요소를 이용하여 서로 상이한 추가 테마를 적용함으로써 해결할 수 있다. 이러한 *addthms*에 의해 기술된 추가 테마는 동일 범주내에서 동일한 이름이 충돌하지 않도록 구별하는데 사용한다(ISO/IEC 2002).

3. 3 토픽맵의 연결 모델

토픽맵은 전통적인 색인화 도구인 분류표를 비롯하여 용어해설, 시소러스의 디스크립터간의 관계와 매우 유사한 구조를 갖는다. 예를 들면, 토픽 타입은 토픽맵 사이에서 유사 토픽간의 클러스터링 과정을 거친 후의 토픽군을 의미한다. 이때 하나의 토픽 타입을 작은 토픽맵으로 정의할 수 있다. 이러한 토픽타입은 그 내부적으로 토픽간의 관계가 설정될 수 있으며, 토픽 타입사이에도 연관관계나 계층관계가 설정될 수 있다(남영준 2005). 따라서 일련의 토픽맵을 하나의 BOS로 간주하고, 동시에 하나의 지식구조로 정의할 수 있다. 이와 같은 토픽과 주제간의 충돌문제를 해결할 경우에 두개의 토픽맵을 연결할 수 있는 것이다.

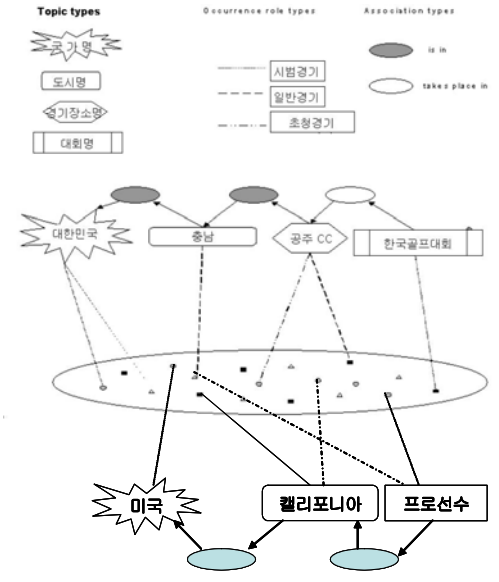
예를 들어, 박찬호와 박세리라는 두 개의 개념구조로 표현하면, 토픽은 '박세리'와 '박찬호'로 표현할 수 있으며, 대상물은 {골프, 유성, 골프공, 골프대회 … }와 {야구, 공주, 야구공, 야구대회… }로 나타낼 수 있다. 이 때의 연결관계 연결어로써 '프로운동선수'를 사용하여 두 개념을 연결할 수 있다. 또한 hub document의 개념으로써 '스포츠 연관관계'를 선언하여 각 토픽이 포함되어 있는 토픽맵을 연결할 수도 있다. 즉, 토픽맵은 특정 사물이나 집단을 분석하는 경우에 접근방법을 하나의 패싯(토픽)으로 설정하고, 이에 해당되는 패싯에 속해있는 요소(대상물)와 각각의 토픽을 연결하는 하나의 연결선을 연관관계로 설명할 수 있다. 한편 토픽맵을 기본요소의 유형들에 따른 내용을 3차원적인 연결구조로 표현할 수 있다.

다음 〈그림 2〉는 대한민국에서 개최된 골프대회가운데 충청남도 공주골프크럽에서 개최한 한국골프대회에서 참석한 선수를 하나의 토픽맵 모델로 도식화한 예이다. 토픽타입은 일련의 공간적 의미로써 공간명을 의미하고, 대상물 역할에 해당하는 것은 해당 공간에 존재하는 운동선수와 같이 대상물에 해당하는 유형을 지시한다.



〈그림 2〉 토픽맵 모델, 원용: empolis tutorial

〈그림 3〉은 앞의 〈그림 2〉를 기초로 충남의 운동선수인 박찬호를 하나의 hub document로 설정하고, 충남의 운동선수인 박세리가 포함된 BOS에 논리적으로 연결한 구조 모델이다. 즉, 미국에 거주하는 프로선수로써 캘리포니아에 거주하면서 고향이 충남인 박찬호를 기점으로 두개의 토픽맵 집합체를 연결한 모델이다.



〈그림 3〉 연결에 기반한 확장 토픽맵 모델

4. 토픽맵을 이용한 시소러스 구조화

전통적으로 문헌정보학에서의 정보검색은 주제명을 비롯하여 분류기호, 시소러스 등과 같은 단어기반의 통제방법과 용어간의 연관관계를 고려한 의미기반의 통제방법을 사용하였다. 이는 개념의 변화나 용어의 의미변이에 키워드의 가치의 변화에 따라 검색효율에 큰 어려움을 겪고 있다. 본 장에서는 용어의 소멸과 생성에 유연한 관리를 위한 토픽맵 요소를 이용한 시소러스 구조화방안을 제안한다.

4. 1 시소러스의 제한적 특성

시소러스의 효율과 활용방안에 대한 연구는 오래전부터 이루어졌다. 또한 국내외 여러 기관에서 검색과 색인의 효율성을 위해 시소러스의 연구와 개발이 지속적으로 이루어지고 있다. 현재 사용중이거나 혹은 개발중인 시소러스는 대부분 도메인에 종속적이며, 하나의 도메인 당 각각 하나의 시소러스로 구축되어 있다. 또한 현재 인쇄본 형태로 출간된 시소러스는 구성이 복잡하여 이용하기 어렵고, 한정된 분야에서만 이용이 가능하며, 보편화되어 있지 않아 일반인에게 쉽게 이용되지 않는다는 단점을 갖고 있다(최재황 1999). 즉 시소러스는 개발주체의 성격에 따라 용어의 의미부터 패싯 관점이 매우 상이하게 개발됨에 따라 이를 효과적으로 이용하기 위해서는 이에 대한 완전한 이해가 필요하다. 이는 일련의 시소러스가 특정주제 영역에서만 적용될 수 있는 매우 주제분야 의존적인 지식개념구조를 갖고 있기 때문이다. 따라서 하나의 도메인을 하나의 시소러스로 표현하는 경우, 시간이 지남에 따라 구

축된 시소러스가 방대해 짐으로써 관리 및 검색의 효율이 떨어질 수 있다(정규상 외 2004). 특히 현대와 같이 학문과 주제영역의 분화와 융합이 매우 빈번하게 이루어지는 외부적 환경에 서는 이와 같이 특정 영역에 고정화된 개념을 확장하거나 통합하는 것이 상대적으로 어렵다는 것을 의미한다. 이는 결국에 시소러스의 폭넓은 응용의 걸림돌이 되고 있다. 이와 같이 시소러스가 갖는 제한점을 요약하면 다음과 같은 공통점을 도출할 수 있다(남영준 2005).

1) 정보확장을 수용할 수 있는 신규용어의 한계
2) 검색대상의 대용량화에 따른 재현율의 불필요한 향상
3) 적정한 정도율 확보의 어려움
4) 유사시소러스(외국어시소러스 포함)와의 통합의 제한점
5) 개발된 시소러스의 디스크립터에 대한 유지보수의 어려움

이러한 제한점은 시소러스가 갖는 구조적인 단점이기 보다는 정보의 폭발시대에 급증하는 정보행태라는 외부적 요인에 크게 좌우되고 있다. 이러한 점을 극복하기 위해서는 시소러스의 유지관리를 유연하게 수행할 필요성이 있다. 예를 들면, 신규개념이나 신규용어의 수용을 자유롭게 유지하며, 불필요한 용어나 유용성이 극히 저하되는 디스크립터의 삭제나 축소를 용이하게 할 수 있도록 한다. 이러한 것은 매우 이론적인 것으로써 기존에 부여된 색인어로써 디스크립터에 대한 관리와 적정한 규모의 디스크립터를 유지해야하는 현실과는 괴리감이 발생할 수 있다. 즉 시소러스 구축에 따른 기본원칙은 디스크립터의 안정성과 적절한 규모유지로 대별할 수 있기 때문에, 대용량 데이터베이스 시대에 이 두가지 원칙을 수용하기 어려운 실정이다. 이러한 상반된 것을 극복하기 위해서는 개발 방향이 적정한 규모의 시소러스 개발이 필요하다.

매크로 시소러스의 경우는 사전에 수록된 표제항을 망라적으로 나열·구조화하여 전주제 영역에 대한 적정한 검색효율을 제공할 수 없는 단점을 갖고 있다(남영준 2002). 따라서 특정 학문의 개념체계를 정교하게 수용하기 위해서는 해당 주제 분야의 마이크로 시소러스의 필요성이 높아지고 있다. 마이크로 시소러스는 단일 주제를 개발범위로 하여 해당 분야의 개념만을 디스크립터로 선정하여 구성된 것을 의미한다. 그러나 실제적으로 마이크로 시소러스는 정확히 하나의 주제만을 설정하여 개념을 수집하고 디스크립터를 구조화할 수 없다. 왜냐하면 대부분의 주제분야는 여러 주제가 합쳐져서 이루어져 하나의 주제분야를 이루는 것이 일반적이기 때문이다. 예를 들면, 경영학 시소러스라고 하여도 여기에는 순수 경영학 이외에 경제학이나 무역학, 외교 통상분야의 용어가 반드시 일정부분 포함되어 있기 때문이다. 한편 정보의 양이 급증함에 따라 시소러스의 형태도 마이크로 형태로 세분화되며 특정 영역의 정보수집을 위한 이용도구로 사용되는 것이 보편화되고 있다(남영준 2002). 즉 시소러스에 수록된 디스크립터의 수가 제한적인 것도 대용량 시소러스의 무용성을 입증하는 것이다.

4. 2 시소러스의 구조화 표현

앞서 조사한 바와 같이 시소러스는 검색을 위한 구조화된 지식으로 그 유용성 때문에 이에 대한 연구와 개발이 지속적으로 이루어지고 있다. 또한 최근에는 그 활용성을 극대화하기 위해 기

존 시소러스의 장점을 수용하며, 전자 형태의 원문기반 정보검색을 위해 시소러스의 기능과 표현을 확장 또는 병합하는 연구도 급증하고 있다. 이러한 연구는 시소러스의 주제적 고정성과 의미표현의 한계성을 극복하기 위한 방법으로 제안되고 있다.

가장 일반적인 연구로써 디지털 자원의 효과적인 검색을 위해 구조화된 마크업 언어의 형태로 시소러스의 구조와 디스크립터를 변환하여 검색효율을 개선하려는 연구가 지속적으로 이루어지고 있다. 남영준 등은 디스크립터에 패싯의 개념을 도입하여 특정 의미를 표현한 개념어와 이를 지칭하는 용어를 구분하여, 시소러스에 수록된 적절한 규모의 디스크립터를 유지하는 방법을 제안하였다(남영준, 이두영 2004). 그들의 연구에서는 디스크립터의 선정과정에서 패싯분류의 개념체계를 디스크립터 선정과정에 적용하는 방안을 제안하였다. 한편, 일반적으로 모든 토픽은 시소러스의 디스크립터와 토픽간의 관계를 시소러스의 기본 관계개념으로 표현이 가능하다. 따라서 토픽맵의 언어로써 온톨로지에 쓰인 어휘들은 여러 가지 방법으로 체계화할 수 있는데, 가장 손쉬운 방법이 'is-a' 관계에 의한 시소러스 형태의 체계화를 구성할 수 있다. 또한 'part-of'나 여러 다른 방법으로 어휘간 관계를 정의할 수 있기 때문에 토픽맵은 시소러스를 이용한 재구성 가능성을 갖는다(권혁철 2004). 또한 시소러스의 일반관계도 온톨로지 OWL Lite를 이용하여 계층관계의 경우는 subClassOf와 subPropertyOf 등이, 대등관계를 equivalentClass를 비롯하여 equivalentProperty, sameAS 등으로, 연관관계는 ObjectProperty을 비롯하여 DatatypeProperty, inverseOf 등을 이용하여 시소러스의 일반구조를 표현할 수 있다(조현양, 남영준 2004). 이는 시소러스의 구조를 온톨로지 언어로 표현하는 방법을 제시함으로써 기존에 구축된 시소러스의 통합 보다는 시소러스의 개념을 일련의 메타언어를 이용하여 새롭게 구조화하는데 적합한 연구였다. 이러한 구조화가 가능한 것은 시소러스의 표층적인 구조화로 계층관계와 연관관계, 대등관계로 간략하게 표현하고 있기 때문이다. 온톨로지와 같이 의미표현에 있어, 다양한 어휘로써 개념망을 표현할 수 있는 지식도구로는 시소러스의 표층구조화를 용이하게 표현할 수 있다. 특히 시소러스의 사례관계와 같은 심층적인 표현도 온톨로지를 이용하여 표현이 가능하기 때문에 토픽맵 알고리즘을 이용한 시소러스 구조화 표현도 가능하다. 따라서 시소러스와 토픽맵, 시멘틱 웹 등과 같은 지식체계는 단어들간의 개념적 구조화를 추구하고 있기 때문에 구성요소와 활용처가 매우 유사하다. 한편 토픽맵과 시소러스의 근본적인 차이는 색인과 검색결과에 대한 통제 시간이다. 전자의 경우, 해당 토픽에 해당하는 대상물(occurrence)의 소재정보와 함께 링크가 설정되어 있는 전조합색이라 할 수 있다. 후자는 후조합색인의 통제도구로써 디스크립터에 해당하는 검색결과를 설정하지 않고, 검색된 결과를 대상으로 검색의 넓이와 깊이를 조정하는 도구라는 색인시점에 차이가 있다.

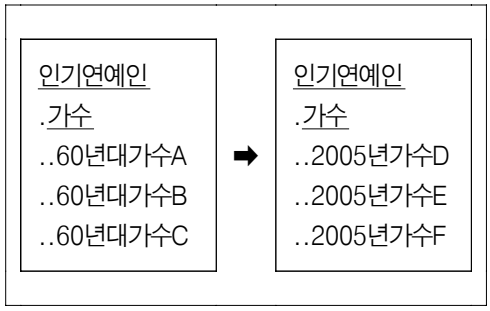
4. 3 토픽맵 기반의 시소러스 구조화

시소러스 효용성 유지의 필수적인 요소는 용어의 관리이다. 용어의 관리는 신규용어 및 사용되지 않은 용어의 채택과 삭제라는 과정과 함께, 유사 주제 영역의 지식구조를 통합하는 과정이다. 이와 같은 용어관리 방안은 기존의 인쇄기반 시

소러스로의 적용은 현실적으로 불가능하다. 왜냐하면 열거식 분류체계가 갖는 단점과 같이 용어의 잦은 변경은 시소러스의 구조를 혼란스럽게 할 수 있기 때문이다. 본 장에서는 이와 같은 제한점을 극복할 수 있는 방안으로써 개념의 고정과 토픽맵 구조의 수용방안을 제안한다.

4.3.1 개념의 고정

미래 지향적인 시소러스는 개발 주제를 특정 영역만을 수용하는 마이크로 시소러스의 형태로 유지해야 하며, 특정 주제영역과 관련이 있는 인접영역의 일부분을 수용하는 최소한의 매크로적인 성격을 유지해야 한다. 또한 디스크립터의 선정에 있어 패싯의 개념을 도입하여, 주요 개념은 디스크립터로써 개념어를 선택하고, 해당 개념과 관련있는 디스크립터는 일반 용어를 선택하여 디스크립터의 유연성을 확보해야 한다. 한편 기존 시소러스 구축을 위한 국내외 기준은 후보 디스크립터의 구조화와 그 사례 중심이었다. 이에 따라 디스크립터 선정의 기준방법은 용어의 출현빈도나 혹은 입력한 검색어와 검색결과와의 로그파일을 분석한 결과에 기반하고 있다. 이러한 디스크립터 선정절차에서 후보 디스크립터는 해당 주제 영역에서 사용되는 주요 단어들이 채택되었다. 또한 선정의 일반적인 절차로는 디스크립터의 품사적 성질을 고려하는 것이다. 즉, 디스크립터의 형태는 명사와 대표 명사어구로 제한하여 해당 주제영역에서 자주 사용되는 명사어구에 제한되고 있다. 이러한 일반적 선정기준은 한번 구축된 시소러스로 하여금 정보의 급증과 함께 신규용어의 출현, 기존용어의 소멸 등과 같은 외부적 상황에 적극적으로 수용하지 못하는 약점을 갖게 하였다. 따라서 이러한 제한점을 개선하기 위해 시소러스 선정에 있어 토픽타입을 해당 마이크로 시소러스에 선언하는 것이다. 예를 들면, 인기연예인과 가수와 같은 디스크립터는 검색어로써 일련의 특정성을 함유하고, 이를 의미하는 개념은 크게 변화하지 않는다. 예를 들면, 〈그림 4〉과 같이 60년대의 인기연예인의 종류가운데 가수라는 개념은 21세기초 가수라는 것과 개념적으로 큰 변화가 없다. 다만 가수라는 개념내에 이를 대표하는 용어들(가수 A, B, C, D, E, F 등)이 시간과 공간(한국과 미국 등)에 따라 앞의 개념어보다 변화가능성이 상대적으로 높은 특성을 갖고 있다. 따라서 현재 시소러스에는 정보의 생산성을 근거로 60년대 가수들은 2005년에 관련 문서나 정보들이 거의 없기 때문에, 시소러스 운용과정에서 용어 A, B, C는 비디스크립터로 처리하거나 혹은 삭제할 수 있다. 즉 현재 정보생산성이 높은 용어 D, E, F를 디스크립터로 처리함으로써 시소러스의 규모를 항상 적정규모로 유지하여, 시소러스의 최신성을 지속적으로 유지할 수 있다.



〈그림 4〉 디스크립터 선정의 예

이를 위해서는 디스크립터의 선정수준을 두가지 수준으로 운영한다. 첫 번째는 인기연예인과 같이 시간이나 공간의 통제를 받더라도 의미와 용어가 변하지 않는 용어를 개념어로 고정시키는

것이다. 개념어는 일련의 토픽형태로써 하위 디스크립터를 대표할 수 있는 용어이다. 두 번째는 가수A나 가수F와 같이 시간이나 공간의 변화에 따라 상대적으로 가변적인 형태의 용어를 유연하게 사용하는 것이다. 이 용어는 일반적인 토픽으로써 차상위 개념어의 범위에 속하는 용어수준이다. 이와 같은 이중 용어관리는 시간과 공간의 비종속적인 개념을 표현하는 용어를 고정시키고, 해당 개념의 범주에 속하면서 외부의 정보적 상황에 따라 디스크립터를 자유롭게 선정하는 방법이다. 또한 이 개념어를 hub document의 역할을 수행하는 이종간 시소러스의 병합을 수행할 수 있는 병합고리로 활용한다. 이러한 일련의 방법을 통해 시소러스의 유지관리를 유연하게 수행할 수 있다.

4.3.2 토픽맵 구조의 수용

토픽맵에서 선언하는 토픽은 인간이 용어로 표현될 수 있는 모든 개념을 의미한다. 일반적으로 토픽은 사람을 비롯하여, 사물, 개념, 의미 등 실제 존재하는 것, 또는 특정 속성이나 어떤 의미 등이 될 수 있다. 보통 명사형으로 표현된다. 즉 특정 문서의 토픽은 그 문서의 작성자가 나타내고자 하는 주제를 표현할 수 있는 단어들로 구성된다(정호영 외 2003). 이는 토픽맵의 토픽이 시소러스의 디스크립터와 동일한 용도로 활용되고 있음을 알 수 있다.

한편, 표준 토픽맵 온톨로지를 시소러스에 적용하는 연구에서(Ahmed 2003), 두가지 방법을 사용하였다. 하나는 주제에 하나의 모델명을 제공하는 것이고, 하나는 사물에 하나의 이름을 제공하는 것이다. 전자의 경우에는 하나의 주제에 대해 일련의 용어를 부여하는 것이고, 후자는 각 주제에 존재하는 개념에 용어를 부여하는 것이다. 특히 그는 시소러스를 토픽맵으로 변환하는 과정에서 다음과 같은 일련의 원칙을 설정하였다.

- 각 디스크립터는 토픽과 같으며, 비디스크립터는 하나 이상의 USE 관계를 갖지 않는다.
- USE 관계로 연결된 용어(우선어와 비우선어)는 동일한 개념에 대해 서로 다른 용어를 사용하여야 한다.
- 설명주기(SN: scope note)는 occurrence __type의 "scope note"를 의미한다.
- RT관계는 association__type에서 "related term"으로 표현된다.
- 계층관계표시는 association__type에서 "broader/narrower"로 표현한다. 이때 각각의 용어에 상위 개념과 하위 개념에 대한 역할어를 부기한다.

이런 과정은 앞서 설명한 바와 같이 시소러스의 구조를 온톨로지 개념을 이용하여 토픽맵의 표층적인 변환방법에 적용되는 것이다. 이 변환방법은 매우 단순하지만, 효율적인 알고리즘이라 할 수 있다. 이를 기반으로 토픽맵의 관계요소를 확대하여 시소러스를 심층적으로 표현하는 방법을 제안할 필요가 있다.

1) 구조 체계의 맵핑

시소러스의 기본구조는 표층관계로써 3대 관계인 대등관계, 계층관계 및 연관관계를 구분할 수 있다. 심층적인 관계로는 ANSI에서 계층관계의 유형으로 속관계(Generic Relationship)를 비롯하여, 전체-부분 관계, 사례 관계, 다계층 관계 (Polyhierarchical Relationship), 계층내 노드 레이블 (Instance Relationship) 등 다섯가지의 형태를 열거하고 있다. 연관관계의 유형으로

는 동일범주내 중복 자매어 (Overlapping sibling terms)를 비롯하여, 동일범주내 상호 독점적 자매어 (Mutually Exclusive sibling terms), 동일범주내 파생 관계 (Derivational sibling terms), 관련어용 노드 레이블 등으로 구분하고 있다. 대등관계의 유형으로는 동의어 관계, 변형어휘 관계, 유사동의어 관계, 복합명사에서 상호참조 관계 등으로 세부적으로 구분하고 있다(ANSI 1999). 이를 토픽맵의 요소로 설명하면 다음과 같이 맵핑처리를 수행할 수 있다.

① 계층관계에 속하는 요소로는 계층적인 관계를 갖는 디스크립터와 패싯의 개념을 갖는 속관계로 구분할 수 있다. 토픽타입은 일련의 범주화를 선언하는 요소이다. 이를 사용하여 일련의 토픽들의 개념을 대표할 수 있으므로 시소러스의 계층관계에서 사례 관계, 계층내 노드 레이블을 표현할 수 있다.

② 계층관계에 있어 속관계를 비롯하여 전체-부분 관계를 표현하는 것으로는 Association types의 is__in(속관계), born__is(전체__부분관계)를 사용하여 시소러스를 나타낼 수 있으며, association(of broader/narrower)를 사용할 수도 있다. 이 이외에 Association role로써 연관관계(written__by 등)의 일반적인 유형을 모두 표현할 수 있다.

③ 대등관계에 속하는 요소로써 이형동의어로써 USE, UF의 관계를 표시할 수 있는 요소로는 Topic names를 사용한다. 이를 사용하여 디스크립터에 대한 비디스크립터의 이형동의어를 표현할 수 있다. 또한 이형동의어의 내용별 유형(별명, 변형어 등)로도 세분화할 수 있다.

④ 관련어용 노드레이블은 일련의 패싯개념을 적용하여, Association roles을 이용하여 표현할 수 있다. 예를 들면, 특정 디스크립터에 대해 역할을 한정함으로써 연관관계 내에서 하나의 노드 레이블이 될 수 있다. 예를 들면, 푸치니와 베르디에 대해 Association role을 사용하여 사람으로서 영향을 준 사람(작곡가로서)과 영향을 받은 사람을 구분할 수 있기 때문에 서로간의 역할 연관성을 표현할 수 있다. 이는 콜론분류법의 패싯 요소인 P·M·E·S·T의 역할한정어와 매우 유사한 역할을 수행한다. 즉, Association role이용하여 시소러스의 연관관계를 심층적으로 표현할 수 있다.

⑤ 시소러스의 SN은 Scope note를 이용하여 처리할 수 있으며, 토픽맵의 Scope note가 하나의 단어로나 혹은 대응 외국어로 기술되었을 때에는 이 관계과 설정된 용어간에는 대등관계로도 적용될 수 있다(남영준, 2005).

2) 구조 체계의 병합

시소러스는 그 효용성을 유지하기 위해서는 지속적인 디스크립터의 관리를 필요로 한다. 이 경우에 디스크립터의 관리는 크게 두가지 관점으로 접근할 수 있다. 하나는 특정 개념을 표현하는 용어의 관리이며, 다른 하나는 새로운 개념을 표현하는 새로운 개념과 그에 따른 용어의 수용이다. 토픽맵을 이용한 구조 체계의 병합은 새로운 시소러스를 개발하는 측면보다 기존의 시소러스를 관리하는 방법을 의미한다. 즉, 기존의 시소러스와 별개의 구조화된 지식의 병합을 통해 새로운 형태의 시소러스를 구축하는 것이다.

한편 토픽맵의 병합은 이름을 기반으로 하는 병합(name-based merge)과 주제를 기반으로 하는 병합(subjectbased merge)으로 대별된다. 이름을 기반으로 하는 통합은 서로 동일한

〈baseName〉을 가지는 경우, 둘 이상의 토픽을 하나로 통합하는 방식이고, 주제를 기반으로 하는 통합은 서로 동일한 〈subjectIndicator〉를 가지는 경우 둘 이상의 토픽을 하나로 통합하는 방식이다. 그러나 대부분의 주제를 표현할 때 다양한 이름이 존재하므로 이름을 기반으로 통합하는 방식은 자칫 오류를 범하기 쉽다(고세영 2003). 따라서 정확한 통합을 위해서는 주제를 기반으로 하는 통합 방식이 이종간 병합에 유리하다. 즉, 결론적으로 시소러스의 구조에 관현 표현이 일련의 토픽 요소를 이용하여 구축되어 있을 경우에, 시소러스의 병합이 가능하다. 특히 구조의 병합은 개념계층의 차이와 이형동의어에 대한 혼란을 최소화하기 위해서는 주제로써 토픽타입의 역할을 하는 토픽(개념어:"제주도특산물")을 중심으로 이루어져야 한다. 왜냐하면 무형물의 개념을 hub document와 같은 허브단어(hub term)로 설정함으로써 용어의 병합이 용이하기 때문이다.

- 디스크립터의 확장: 두개의 서로 다른 시소러스를 제주도 특산물이란 무형의 개념어를 topic type으로 설정하여 두개의 시소러스를 합쳐 디스크립터를 확장할 수 있다.

기존의 구조		새로운 구조		통합된 구조
제주도 특산물 nt 바나나 파인애플	+	제주도 특산물 nt 망고 파인애플	=	제주도 특산물 nt 바나나 망고 파인애플

- 디스크립터의 대치: 특정 용어(topic type)의 의미적 범위는 동일하나 이를 표현하는 용어의 활용빈도를 기준으로 새로운 용어를 디스크립터로 승격시키고, 기존의 용어를 비디스크립터로 변환한다. 이때 위치 변환의 근거기준은 대상물(occurrence)의 수와 대상물의 역할을 기준으로 활용한다.

기존의 구조		새로운 구조	수		변경된 구조
1) 파인애플 is a member_of 제주도특산물 2) 파인애플, 망고 is a member_of 열대과일	:	혁재 bought a fruit, 망고 이숙 bought a fruit, 파인애플	100 9	➡	망고 is a member_of 제주도 특산물

일례로 '제주도 특산물'이란 무형의 개념어는 '파인애플'이나 '망고'와 같은 특정 개체명보다 포괄적이다. 따라서 '제주도 특산물'을 하나의 기준으로 설정하고 이와 관련된 용어들이 상황에 따라 바뀌어도 개념체계의 변이는 거의 나타나지 않는다. 즉, 개념수준의 용어 변동은 컨텐츠 수준의 용어(파인애플과 망고 등)보다 안정적으로 유지된다. 또한 안정적인 개념수준과 관련된 컨텐츠 수준의 용어는 상황에 따라 바뀌어도 개념구조에 큰 변화는 없다. 한편 컨텐츠 수준의 용어의 변화는 해당 토픽(디스크립터)에 연결된 대상물의 수의 과다와 역할에 따라 조정한다.

단, 위의 모든 경우는 XTM에서 선언하는 TNC[1] 규칙을 준수해야 한다(Rath 2003). 즉 색인자료와 계층구조가 완전하게 일치하지 않는

1) topic naming constraint(TNC) 규칙: 복수개의 토픽맵을 병합할 경우에는 중복되는 토픽과 연관관계, 대상물은 모두 삭제하여 하나로 통합한다. 별명을 설정하는 것은 두개의 토픽에서 지정하는 대상물은 동일한 topic type(혹은 class)과 동일한 자원이어야 한다. 또한 연관관계는 반드시 동일한 topic type(혹은 class)와 동일한 역할관계와 역할관계어를 가져야 한다.

디스크립터의 경우에는 그 디스크립터를 별개의 것으로 간주하며, 해당 디스크립터는 특정개념을 표현하는 차상위 디스크립터의 하위어 관계로 전개하여야 한다.

다음은 이상의 경우에 대해 XTM으로 표현한 토픽 병합을 위한 문법적 구조를 기술한 것이다.

```
〈topicMap id="열대과일"〉
...
〈mergeMap xlink:href="http://www.jeju.go.kr/aaa.xtm"〉
〈resourceRef xlink:href="http://www.jeju.go.kr/aaa.xtm/"〉
〈subjectIndicatorRef xlink:href="http://www.jeju.go.kr/aaa.xtm#fruit"/〉
〈topicRef xlink:href="#english"〉〉
〈/mergeMap〉
...
〈/topicMap〉
```

5. 결 론

시소러스가 검색도구로써 지속적인 효용성을 유지하기 위해서는 용어의 관리가 필수적이다. 그럼에도 불구하고 시소러스의 지속적인 관리가 현실적으로 어려운 것은 시소러스가 갖는 고정적인 구조의 특성 때문이다. 본 연구에서는 이러한 제한점을 극복하기 위한 방법으로써 토픽맵의 표현 요소를 이용하여 시소러스를 유지관리하는 원칙과 방법을 개발하였다. 이를 요약하면 다음과 같다.

1) 개념의 고정: 디스크립터의 선정을 두가지 수준으로 운영한다. 하나는 시간이나 공간의 통제를 받더라도 의미와 용어가 변하지 않는 개념어를 디스크립터로 채택한다. 개념어는 일련의 토픽타입으로써 하위 디스크립터를 대표할 수 있는 용어이다. 다른 하나는 시간이나 공간의 변화에 따라 외부상황을 정확하게 표현하는 가변적인 형태의 용어를 의미한다. 이 용어는 일반적인 토픽으로써 차상위 개념어의 범위에 속하는 용어수준이다. 이와 같이 디스크립터 선정시 토픽형태에 해당하는 개념어를 디스크립터로 선정하고, 해당 디스크립터의 개념적 범주를 다양하게 표현하는 용어를 하위 디스크립터로 사용한다.

2) 토픽맵 구조의 수용: 시소러스의 구조를 토픽맵의 요소로 표현하여 시소러스에 일반적인 관계를 표현하기 위해서는 구조체계의 맵핑 알고리즘과 구조체계 병합의 알고리즘을 이용한다.

- 구조체계의 맵핑

· 계층관계에 속하는 요소로는 계층적인 관계를 갖는 디스크립터와 패싯의 개념을 갖는 속관계로 표현한다.

· 토픽타입을 시소러스의 계층관계에서 사례관계, 계층내 노드 레이블을 심층적으로 표현한다.

· 속관계와 전체-부분 관계를 Association types의 is__in(속관계), born__is(전체__부분관계)등과 같이 요소로 표현한다.

· Association role로써 연관관계의 일반적인 유형을 표현한다.

· USE, UF의 관계를 표시할 수 있는 요소로는 토픽개체명을 사용한다.

- 구조 체계의 병합: 두개의 이종간 시소러스를 토픽타입을 이용하여 병합함으로써 디스크립

터를 확장할 수 있었다. 또한 토픽맵의 요소(topic type)를 고정시키고, 특정 용어(topic)의 활용빈도를 기준으로 시소러스의 질을 유지할 수 있다. 토픽맵으로 연결되어 있는 대상물의 수와 대상물의 역할을 기준으로 한다.

- 기타 표현

・설명주기(SN: scope note)는 occurrence_type의 "scope note"를 표현한다.

・RT관계는 association_type에서 "related term"으로 표현된다.

・계층관계표시는 association_type에서 "broader/narrower"로 표현한다.

이상과 같이 토픽맵 구조요소와 표현방법을 통해 시소러스의 유지관리에 적용하는 원리를 제안하였다. 이는 전통적인 시소러스가 갖는 한계점을 극복하고 디스크립터 유지관리의 편의성과 개념의 안정성을 모두 확보하여 시소러스의 효율성을 유지할 수 있는 중요한 기준자료가 될 것이다. 단, 본 연구는 이론적 측면에서의 시소러스의 유지 및 개발에 대한 방안으로써 실제 토픽맵 표현을 활용하고 토픽맵 구축기를 이용한 실험이 함께 이루어지지 않았다는 한계성을 갖고 있다. 따라서 토픽맵 형태의 시소러스 변환과 병합을 위한 실제적인 연구와 실험이 지속적으로 이루어져야 할 것이다.

참 고 문 헌

Ahmed, Kal. 2003. Topic Map Design Patterns For Information Architecture. XML. (December 2003). [cited 2005.7.1]. 〈http://www.techquila.com/tmsinia.html〉.

ISO/IEC. 2002. ISO/IEC 13250:Topic Map. ISO/IEC.

Le Grand, Soto, M.. 2000. "Information Management Topic Maps Visualization", *XML Europe 2000*, Paris, France.

NISO/ASI/ALCTS. 1999. Workshop on Electronic Thesauri: Planning for a Standard, Washington, DC. [online]. [cited 2005.7.10]. 〈http://www.niso.org/news/events_workshops/thesau99.html#issues〉

Pepper, Steve. The TAO of Topic Maps : Finding the Way in the Age of Infoglut. [cited 2005. 7. 14]. 〈http://www.ontopia.net/topicmaps/materials/tao.html〉.

Rath, H. Holger. 2003. 『The topic map handbook』. empolis.

TopicMaps.Org. 2001. XML Topic Maps(XTM) 1.0, TopicMaps.Org [cited 2005.7.1]. 〈http://www.topicmap.org/xtm/index.html〉.

고세영. 2003. 『토픽맵을 이용한 이 기종 상품분류체계 온톨로지 통합에 관한 연구』. 석사학위논문, 숙명여자대학교 대학원, 문헌정보학과.

고유미. 2005. 『토픽맵 기반의 특허정보 서비스를 위한 시스템 구축에 관한 연구 : 항체이용

기술 분야를 중심으로』. 석사학위논문, 숙명여자대학교 대학원, 문헌정보학과.

권혁철. 2004. 시맨틱 웹의 가능성과 한계.『지식정보인프라』. 15호: 15-19.

남영준, 이두영. 2004. 로그데이터를 이용한 디스크립터의 외형적 특성 분석.『정보관리학회학술대회논문집』. 11: 61-6.

남영준. 2002.『고속철도 시소러스 개발』. 쓰리소프트.

남영준. 2005. 시소러스와 토픽맵의 연관성 연구.『정보관리학회학술대회논문집』. 12: 261-8.

유우종, 김진우, 권주흠. 2004. 워드넷 온톨로지를 이용한 토픽맵 메핑.『한국정보과학회 학술발표논문집』: 175-177.

이은아. 2003.『XML 토픽맵(XTM)을 이용한 시맨틱 네비게이션 시스템 구현』. 석사학위논문, 숙명여자대학교 대학원, 문헌정보학과.

정규상, 김원중, 양재동. 2004. 객체기반 다중 시소러스 시스템의 설계 및 구현.『한국정보과학회 학술발표논문집』 2004년도(추계 Ⅰ) : 181-4.

정원규. 2003. 벤담-도덕 및 입법의 원리 서설,『철학사상』 제2권 제8호 별책.

정준원. 2003.『XML 기반의 지식맵 캐쉬 기법』. 석사학위논문, 서울대학교 대학원, 컴퓨터공학과.

정호영 외. 2003. XTM 기반의 지식맵.『데이터베이스연구』. 19(1) : 38-94.