지식관리시스템을 위한 의미형 한영 시소러스 구축에 관한 연구*

A Study on the Korean-English Semantic Thesaurus Construction for Knowledge Management System

남 영 준(Nam Young-Joon)**

<목 차>

초 록

본 연구에서는 지식관리시스템에 사용되는 의미형 한영 시소러스개발에 따른 개발방안과 원칙을 제시하였다. 기본적인 개발 방안은 1)용어 수집에 있어 디스크립터 추출의 정보원을 기존 문헌형태의 자료에서 내부문서와 웹에 등재된 외부문서 등도 포함하도록 하였다. 2)의미위주의 용어보다는 개념위주의 디스크립터가 필요하며, 이를 보완하기 위해 전거어 사전의 구성이 필요함을 제시하였다. 이를 위해 용어풀을 운영할 것도 제시하였다. 3)디스크립터의 적정한 규모를 15,000개 내외로 제시하였다. 4)관계설정은 수작업과 기계처리작업이 병행되는 하이브리드 방식을 제안하였다.

주제어 : 지식관리시스템, 한영 시소러스

Abstract

As the role of a library has changed to the integrated management system of knowledge, the library needs new information retrieval tools. The purpose of this study is to propose a method and principle of the Korean-English semantic thesaurus construction for a knowledge management system. The method and principle is as follows; 1) in collecting terminology, I included not only internal documents but external documents on the web as a source for the descriptors extraction. 2) conceptual descriptors are more needed than semantic ones. I also proposed the necessity of the authority files for complement. 3) I proposed the appropriate scale of the descriptors to be 15,000 in a thesaurus. And 4) I proposed a hybrid method that used both a manual and an automatic process in establishing the relationship.

Key Words : management system of knowledge, Korean-English semantic thesaurus

---

* 이 논문은 1999년도 한국학술진흥재단의 대학교수 해외파견 연구지원에 의하여 연구되었음.

** 전주대학교 사회과학대학 문헌정보학과 부교수(namyj@jeonju.ac.kr)

· 접수일 : 2001. 11. 12 · 최초심사일 : 2001. 11. 30 · 최종심사일 : 2001. 12. 7

I. 서 론

21세기 글로벌 문명시대에서는 지식의 활용에 따라 국가의 안위가 좌우된다고 판단되기 때문에 지식은 국가의 주요한 자산으로 평가받고 있다. 한편, 인터넷상의 정보는 매년 두배씩 증가하는 시점에서 기업들은 이와 같은 기업외 정보는 차치하고라도, 자신들이 생성한 기업내 정보도 효과적으로 활용하지 못하고 있는 실정이다. 이러한 문제가 있음에도 불구하고 기업들은 소속원이나 내부 부서에서 많은 양의 문서를 기계적으로 양산하고 있으며, 자신들이 양산한 정보에서 유효한 정보를 얻기 위해 시간적·경제적 투자를 수행하고 있는 모순에 직면하고 있다. 따라서 정보의 범람 속에 정보의 빈곤이라는 현대 정보화 사회의 딜레마를 해결하고 궁극적으로 기업의 경쟁력을 높이기 위해 기업들은 지식관리시스템을 개발하여 이러한 문제점을 해결하려고 한다.

이러한 지식관리시스템에는 반드시 지식구조체계와 이를 지원하는 검색도구가 개발되어야만 효율을 극대화할 수 있다. 그러나 기존 도서관에서 활용하고 있는 분류표 혹은 주제명 표목표와 같은 기존의 검색도구로는 적정한 검색효율을 얻을 수가 없는 실정이다. 왜냐하면 기존의 검색도구는 문헌자료와 같이 정형화된 자료를 검색하는 목적으로 개발되어, 기업이 필요한 비정형화된 정보나 지식을 검색하기에는 부적합하기 때문이다. 또한 언어적 의미 및 문화적 차이 때문에 영미계 자료들을 통합적으로 검색하기에도 어려움이 있다. 따라서 국내외의 주요 정보와 비정형화된 자료 및 지식까지도 통합적으로 검색할 수 있는 새로운 개념의 검색도구가 필요하게 되었다.

검색시스템에서 사용될 수 있는 검색도구가운데 시소러스는 개념간 체계적인 의미 구조와 연관관계를 갖고 있기 때문에 지식정보화 사회에 가장 효과적인 검색도구라 할 수 있다. 이러한 필요성에도 불구하고 국내외에 시소러스 개발은 기존 도서관 소장자료의 검색도구 수준으로만 개발되고 있다. 또한 시소러스개발에 대한 기본적인 국제 표준과 국내 표준지침만 제시되어 있어, 개발에 필요한 실질적인 구축 방법과 이론은 제시되지 못하고 있는 실정이다.

본 연구는 이 점에 착안하여 지식관리시스템에 필요한 의미형 한·영 시소러스를 구축할 수 있는 효과적인 방안과 기준을 연구하고자 한다. 이 연구결과는 향후 시소러스 구축에 소요되는 인력과 경비, 시간을 절약하고, 표준적인 의미형 시소러스를 개발할 수 있는 주요한 지식으로 활용될 수 있을 것이다. 궁극적으로는 이 연구가 지식기반 관리 시스템의 구축에 필요한 하나의 기준으로도 활용될 수 있을 것이다.

II. 지식관리시스템의 개관

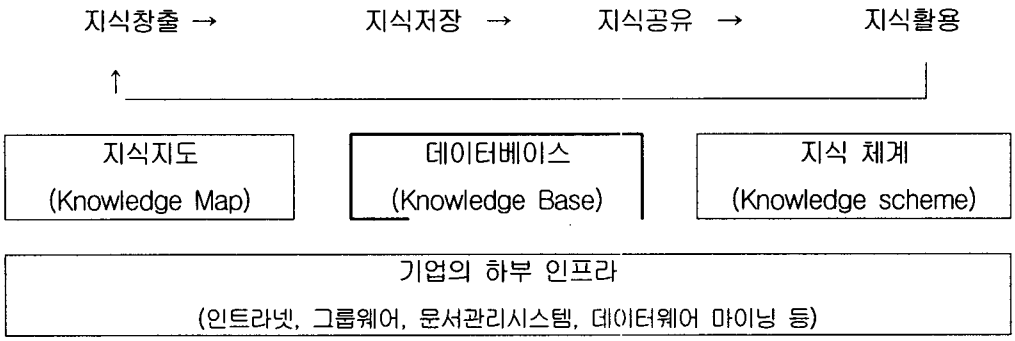
전통적으로 도서관은 국가 및 사회에서 생성된 주요 정보를 수집하여 이용자에게 정보를 제공하는 공공적 역할을 담당하고 있다. 따라서 도서관이 소장하고 있는 대부분의 정보나 데이터베이스들은 공공적 가치를 갖고 있는 정보가 주를 이루는 반면에, 일반 기업이 소장하고 있는 대부분의 정보나 데이터베이스들은 상업적 가치를 갖는 정보가 주를 이루고 있다. 특히 기업들은 소속 도서관들이 갖고 있는 정보의 가치를 극대화하기 위해 기존의 데이터베이스 외에 구성원들이 갖고 있는 유무형의 지식(know-how)까지도 주요한 정보로 간주하여 활용하고 있다. 이러한 변화는 지식이 국가 혹은 기관, 단체의 생산성 향상에 필요한 중요한 재화로서의 가치를 갖고 있음을 의미하는 것이다.

따라서 국내외 많은 기관과 단체에서는 필요하다고 판단되는 모든 정보를 수집하여 체계적으로 관리할 수 있는 지식관리시스템 구축에 많은 시간과 인력을 투자하고 있는 실정이다. 이와는 별도로 도서관들은 디지털 도서관 사업의 일환으로써 자관이 소장한 서지정보의 전산화를 오래 전부터 추진하여, 이미 대부분의 도서관들은 인터넷을 통해 서지정보를 제공하고 있다. 최근에는 자관들이 소장한 자료들까지도 전문형태(full-text)로 열람이 가능하도록 인터넷을 통해 열람서비스의 기회를 확대하는 봉사위주의 시스템으로 변화하고 있는 추세이다. 도서관도 단순히 소장자료만으로 서비스하는 수동적인 입장에서 웹에 기반한 통합 지식관리시스템으로의 변화를 시도하고 있는 것이다. 이러한 변화는 기업에서 의도하고 있는 지식관리시스템과 도서관이 추구하는 지식관리시스템이 부합될 수 있음을 의미하기 때문에 이 두 시스템의 차이를 분석할 필요가 있다.

1. 기업의 지식관리시스템의 기본구조

기업에서 판단하는 지식관리시스템은 조직내의 인적 자원들이 갖고 있는 개별적인 지식을 체계화하여 공유함으로써 기업경쟁력을 향상시키기 위한 기업정보시스템을 의미한다. 즉, 조직 내·외부에서 생성되는 지식을 체계적·통합적으로 포착, 축적하여 이를 공유함으로써 업무의 생산성 및 의사결정 위험을 최소화하고 전문가적 지식활용을 통한 창의적 업무처리 능력을 향상시키기 위한 시스템을 의미한다. 이를 위해 각 조직들은 지식관리시스템을 1)지식창출 단계, 2) 지식저장 단계 3) 지식공유 단계 4) 지식활용단계와 같이 4단계로 구분하여 운영한다. 각 단계는 전문가 연관관계를 표현하고 있는 지식맵(knowledge map)과 주요 정보들

을 구조화한 데이터베이스(knowledge base), 분류체계와 같은 지식체계(knowledge scheme)에 기반한다. <그림 1>은 지식관리시스템을 구축하여 운영을 시도하고 있는 기관에서 판단하는 전형적인 지식관리시스템의 기본 모형이다.



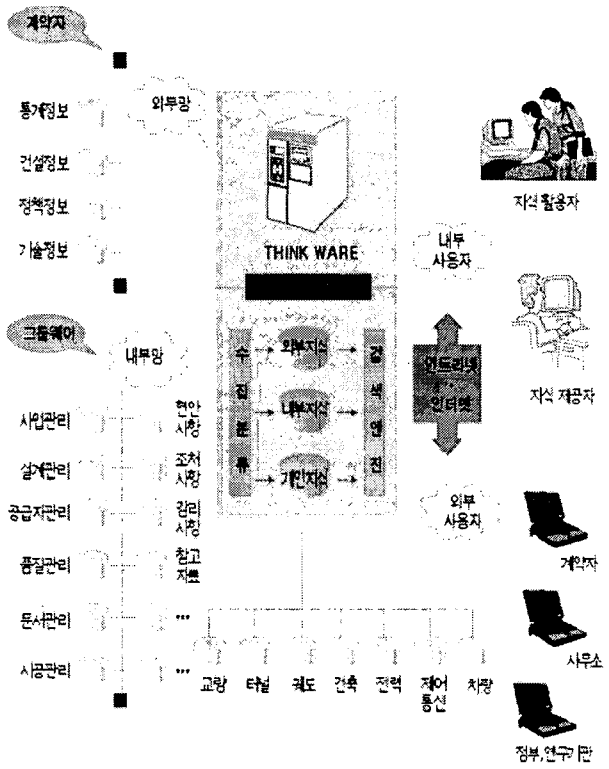
<그림 2> 지식관리 시스템의 기본 모형

2. 기업의 지식관리 시스템 사례

2.1 한국고속철도공단

한국고속철도공단은 고속철도 건설사업의 완공에 따라 건설관련 사업자료의 전자적 공유체계를 구축, 대 정부기관, 공단, 관련기업, 연구기관 등과의 지식정보공유를 통하여 업무 시너지 효과를 창출하고 향후 효율적인 철도건설과 시설유지보수를 위한 통합적인 지식관리시스템을 운영하고 있다. 한국고속철도공단은 지식관리의 토대가 되는 것으로 외부정보망, 내부정보망을 이용하여 수집된 자료들과 공단내부에 소장한 자료와 외부에서 생산되는 자료를 주요한 지식베이스로 간주하고 있다. 이와 함께 각 부서에서 생산되는 문서 및 자료를 비롯하여, 내부 구성원들이 업무에서 활용하는 개인적인 지식까지도 지식관리시스템의 주요 정보원으로 간주하고 있다.[1] 특히 초고속 철도사업의 결과에 따라 인수한 많은 보고서 및 문헌이 영어로 이루어진 토목설비 자료들이기 때문에 효과적인 관리와 검색을 위해 다국어가 지원되는 지식관리시스템의 형태로 구축되어 있다. <그림 2>는 한국고속철도공단에 구축되어 있는 지식관리시스템의 구축 모형이다.

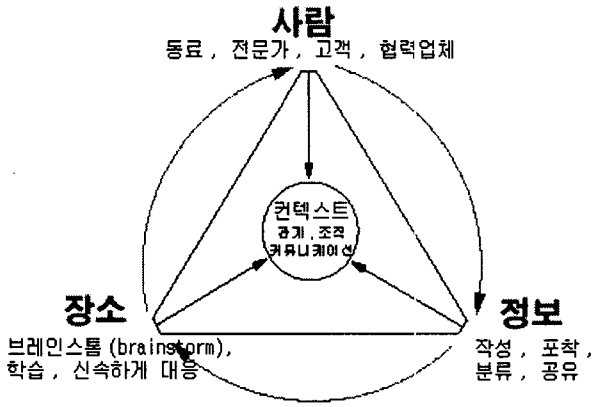
1) http://www.ktx.or.kr/kor/news/2001/7_8/26.html



<그림 2> 고속철도공단 지식관리시스템

2.2 IBM

IBM은 IBM의 인적 자원과 공간적 자원, 자원으로서의 정보를 효과적으로 연계하여 각 부서 혹은 팀간의 협력을 통해 경영의 시너지 효과를 창출하는 방향으로 지식관리시스템을 개발하였다. 즉, 기업에 소속된 모든 자원을 유기적으로 활용하여 기업의 이익을 창출하는 지식저장고의 역할을 수행하는 것이다. IBM은 지식관리의 토대가 되는 것으로 비즈니스 인텔리전스(Business Intelligence), 협업(Collaboration), 지식 이전(Knowledge Transfer), 지식 발견(Knowledge Discovery) 및 전문지식 찾기(Expertise Location) 등 5가지를 제시하고 있다. 즉, 1차적으로 구성원을 포함한 기관 소속 구성원들이 생산하고 있는 각종 정보뿐만 아니라 출판되지 않는 지식까지도 입수하여 해당 정보를 체계적으로 분류하여, 가능한 많은 내부 구성원들이 이러한 지식들을 공유할 수 있도록 지식관리시스템을 구조화하고 있다. 특히 지식 이전(knowledge transfer)은 e-Learning으로 신속하게 확장되고 있고, 지식 발견 및 전문지식 찾기는 이전에 접근하기 어려웠던 컨텐트 및 개인적 전문지식에 액세스하기 위한 새로운 방안으로 활용하고 있다.[2] <그림 3>은 이러한 지식공유를 의도한 지식관리시스템의 개념도이다.



<그림 3> IBM 지식관리시스템 개념도

3. 도서관의 지식관리 시스템 구조

기업의 지식관리 시스템은 기업의 외·내부에서 생성된 정보까지도 주요 정보원으로 처리하여 이를 체계적으로 수집하는 것에 비해, 도서관에서의 지식관리시스템은 전통적으로 물리적인 형태이던지 혹은 가상의 데이터라도 출판이 완료된 자료들을 주요 정보원으로 간주되는 점이 다르다. 특히, 도서관은 소장중인 자료이외에도 웹에 출판된 자료들도 주요 정보원으로 처리하여 자료관리의 범위를 가상공간의 자료까지 확대하고 있다. 이러한 시각은 최근 국내외에서 발표되는 많은 연구결과들이 이를 입증하고 있다. 브루스는 웹에 등재된 자료가운데 상당한 양의 데이터(웹문서)가 학술적 가치를 갖고 있으며, 기존 도서관에서 보관되어 일반 이용자에게 필요한 자료라고 판단되어 이의 학술적 활용가치와 방안에 대해 다양한 연구가 이루어지고 있음을 주장하였다. (Bruce. 1998)

문경화와 남태우는 그들의 논문에서 "도서관에서 활용될 수 있는 정보를 이용자의 정보이용 목적에 적합한 지식정보를 제공하기 위한 것이다. 여기에는 도서를 중심으로하는 텍스트 외에도 이미지 파일이나 멀티미디어 자원을 포함하는 네트웍 상의 접근 가능한 모든 자원에 대한 효율적인 관리가 수행되어야 한다"고 주장하고 있다. 즉, 그들은 이미 인터넷에 존재하는 주요 정보를 도서관의 정보검색과정에서 주요 정보원으로 간주하고 연구를 진행하였다. (문경화, 남태우. 2001) 이란주는 공공도서관에서 지역정보봉사의 개선을 위해 의학정보사이트의 구축과 관리를 지원하는 통합 지식관리시스템의 모형을 제시하면서 인터넷상에서 제공하는 의료정보사이트의 활용을 역설하였다. 그는 도서관 지식관리시스템의 구축에 따른 주요

2) http://www.kr.ibm.com/e-innovation/km/

정보원으로 인터넷 자원을 활용할 수 있음을 보여주고 있다.(이란주 2001)

성기주는 국내 대학교에서 구축한 여성학 관련 웹사이트를 분석하여 콘텐츠로서 가치가 있음을 분석하였다. 이러한 연구는 각 대학에서 웹을 통해 제공하는 여성학 관련 콘텐츠가 주요한 정보원이라는 것을 전제로 한 것이다.(성기주 2001) 즉, 이와 같이 도서관 분야에서는 지식관리시스템의 범위를 자관이 소장하고 있는 장서이외에도 인터넷에서 입수할 수 있는 정보원들도 이용자들이 필요로 하는 주요정보원으로 판단하고 정보수집의 범위를 웹까지도 확대하고 있다. 이러한 변화는 필연적으로 소장자료 검색방법의 변화를 요구하게 된다. 그 변화에는 인적인 부분과 시스템적인 부분 등이 포함되며 특히 검색도구의 변화가 절실하게 요구되고 있다. 그 가운데 도서관에서 사용하고 있는 기존의 분류체계의 변화도 가속화되어 최근 수년사이에 국내외에서도 이에 대한 연구가 문헌정보학계에서도 다음과 같이 활발히 이루어지고 있다.

남영준은 주요 상용웹브라우저의 디렉토리 구조와 한국십진분류표의 구조를 비교하여 도서관에 소장중인 자료와 유효한 웹문서를 함께 분류할 수 있는 통합 분류표의 체계를 제시하였다. 그는 기존 도서관에서 소장하고 있는 장서와 웹에 존재하는 주요 정보를 도서관에서 통제하여야 할 자료로 간주하고 보다 적극적인 관점에서 웹정보를 활용할 것을 주장하였다.(남영준 1998) 한상길도 인터넷에 존재하는 정보가운데 주요 산업정보를 특정한 기준에 맞게 일관되게 분류하여 이용자들에게 제공하면 산업활동에 도움이 될 수 있다고 주장하였다. 그는 이를 위해 합리적이고 체계적인 분류체계를 제시하고, 이를 통해 웹자원을 보다 효율적으로 관리할 수 있는 가능성을 제시하였다(한상길 2001).

한편 기업입장에서 지식관리시스템은 소장자료와 각 개인들이 갖고 있는 무형의 지식 혹은 인지하지 못하는 노하우까지도 정보원으로 판단하여 이를 수집·관리하는 수준까지 개발하여 정보의 총체적인 수집과 관리가 이루어질 수 있는 시스템으로 정의하고 있다.

도서관 입장에서 지식관리시스템은 비소장자료로써 웹과 같은 외부 공간에서 얻을 수 있는 자료들을 체계적으로 수집하여, 이를 자관의 소장정보로 간주하여 지식으로 활용하는 소장장서를 외부에 있는 자원까지도 확대하여 운영할 수 있는 시스템으로 판단하고 있다. 즉, 내부 지식의 수집과 관리라는 것에 차이가 있으며, 공통점은 도서관과 기업 모두 자관이 소장하고 있지 않은 외부자원(인터넷 문서 등)도 수집할 주요 정보대상으로 간주하고 있는 것이다. 따라서 기업과 도서관은 소장정보와 함께 외부정보까지도 함께 구축된 하나의 거대한 지식베이스에서 효율적으로 검색할 수 있는 새로운 검색도구가 필요할 것이다.

4. 지식관리시스템의 검색도구

전통적으로 도서관에서는 소장자료의 관리와 검색을 위한 정보검색도구로써 분류표를 비롯한 주제명 표목집, 각종 통제어휘집, 시소러스 등이 사용되었다. 전통적인 도서관과 소장장서들을 대상으로 이러한 정보관리도구들을 이용한 검색과 색인은 이용자에게 적절한 검색효율을 제공하였다. 국내의 대표적인 지식범주화체계로는 듀이 십진분류체계를 원용한 한국십진분류표(KDC: Korea Decimal Classification)가 있으며, 여러 가지 문제점은 있으나 많은 도서관에서 이를 원용하여 사용하고 있다. 이는 아직 도서관에서는 인쇄형 자료위주의 지식관리와 검색이 이루어지기 때문이다.

한편 야후와 같은 상용 웹검색엔진들은 정보검색의 모델을 자연어로 이루어진 키워드나 쿼리(query)를 입력하여 정보를 검색하도록 하였다. 이러한 검색방법은 이용자에게 정련된 정보를 제공하지 못하고 질의어가 웹문서에 단순히 출현한 정보를 모두 제공하는 단점이 나타난다. 이러한 단점을 극복하기 위해 상용검색엔진은 수집한 주요 웹사이트나 정보들을 자체 구축한 분류체계에 따라 색인된 웹사이트를 링크하여 이용자에게 검색의 편의성을 제공하고 있다.

전통적인 도서관 분류체계가 지식의 분화발달과정에 따라 전개되어 학술적인 정보의 구조화를 의도한 것이라면, 웹검색엔진의 분류체계는 뉴스(hot news)와 감각적인 정보를 효율적으로 관리할 수 있도록 구조화되어 있다. 즉, 정보 검색욕구가 높은 것을 수용할 수 있도록 분류체계를 개발하고, 구조와 표제항의 변화는 매우 빈번하게 이루어지고 있다. 인터넷상에 웹정보를 구조화할 수 있는 대표적인 것으로는 Yahoo(Korea)의 분류체계가 있으나 이는 Yahoo(US)의 것을 그대로 번역한 수준에 머무르고 있어 국내를 포함하여 동양권 정보를 구조화하기에는 매우 부적절한 구조를 갖고 있다. 또한 국내에서 자체 개발된 웹검색엔진의 일부는 문헌분류테이블의 체계를 부분적으로 수용하였으며, 이는 문헌분류체계와 웹검색엔진의 분류체계가 최신의 정보를 효과적으로 관리하기에는 매우 부적절한 구조를 갖고 있음을 의미하는 것이다. 따라서 웹검색엔진의 분류체계는 이용자들에게 따라서 과거의 소급데이터와 학술데이터 등 유용한 국가지식이나 자원을 효율적으로 수용하지 못하는 특징을 갖고 있다. (남영준. 1998) 따라서 현재 기관이 소장하고 있는 책자형태의 자료와 내부문서, 웹으로 접근할 수 있는 외부자료 및 문서를 통합적으로 관리하고 검색할 수 있는 적절한 검색도구는 아직 개발되지 않은 실정이다. 또한 소장중인 영문자료와 인터넷을 통해 접근할 수 있는 외부 영문자료를 검색할 수 있는 다국어지원용 검색도구도 아직 개발되지 않은 실정이다.

Ⅲ. 의미형 시소러스 구축 방안

기존의 시소러스는 대부분 도서관과 같은 물리적인 공간에서 소장중인 물리적 형태의 자료를 대상으로 검색할 수 있도록 개발되었다. 왜냐하면 시소러스의 색인 요소는 용어이며, 일반적으로 시소러스의 용어는 사고를 전달하는 기본 의미 단위인 개념을 나타내는 데 사용되기 때문이다. 즉, 용어는 개별 단어나 단어 집단 혹은 명사구일 수 있지만, 대부분은 단일 단어이고, 단어가 가장 구체적으로 개념을 표현하는 품사이므로 용어는 기본적으로 명사를 사용하기 때문이다.(Baeza-Yates, Ribeiro-Neto. 1999) 따라서 시소러스에 있는 하나의 디스크립터만으로 기존 도서관 소장자료와 인터넷과 같은 가상공간에 존재하는 정보들을 효과적으로 통합검색하기에는 어려움이 있다. 또한 국내자료뿐만 아니라 영미계의 자료를 하나의 디스크립터와 그 대응 영어 디스크립터만으로 검색하는 것도 효율적인 검색이 이루어지지 않을 수 있다.

1. 의미형 시소러스의 정의

정보검색 분야에서 사용되는 전통적인 시소러스에 대해 이두영 등은 "후조합 검색을 위해 설계된 자연언어 용어들의 통제어휘집"로 정의하였으며(이두영 등. 1995), ISO에서는 "상위 및 하위 개념 사이의 전후관계를 명백하게 하기 위하여 공식적으로 조직·통제된 색인어의 어휘집", 정보학 사전에서 사공철 등은 "개념 사이의 관계를 명확하게 나타내기 위하여 일정한 형식으로 조직되고 통제된 색인어의 어휘집" 등과 같이 정의되고 있다 (사공철 2001). 즉, 시소러스는 개념보다는 통제 어휘를 구조화한 용어집으로 제한된 정의를 내리고 있다. 이러한 제한된 역할에 따라 정보검색분야에서 시소러스는 용어의 불일치성을 해결하려한 연구가 주를 이루었다. 그러나 최근에는 이용자가 입력한 탐색어와 시스템에 저장된 색인어와의 불일치성을 해결하고 단순한 용어확장 방법으로 성능을 향상시키는데서 더 나아가 개념 확장을 효과적으로 함으로써 이용자의 초기 탐색어와 가장 밀접한 용어들을 찾는 탐색문으로 포함시키려는 연구노력이 있어 왔다. (노영희 2000)

지식관리시스템의 완성은 도서관의 소장자료와 국내외 가상공간에 있는 디지털자료, 개인이 보유하고 있는 비정형화된 지식이 혼합된 하나의 거대한 데이터베이스가 구축되는 것을 의미한다. 이러한 대규모 혼합 데이터에서 이용자가 적절한 검색효율을 얻을 수 있도록 하기 위해서는 새로운 형태의 검색도구가 필요할 것이다. 이러한 검색도구는 문헌자료와 웹자료,

영문으로 기술된 자료를 모두 수용할 수 있는 형태의 지식베이스가 되어야 한다. 즉, 시소러스는 검색도구로서의 역할뿐만 아니라 색인도구의 역할과 지식구조화에 활용될 수 있는 지식창출시스템이 되기 위해서는 새로운 지식을 모니터링할 수 있는 지식분류체계와 색인도구, 시소러스로 구분되는 새로운 지식지도(Knowledge Map)가 구축되어야 한다.(이란주2001)

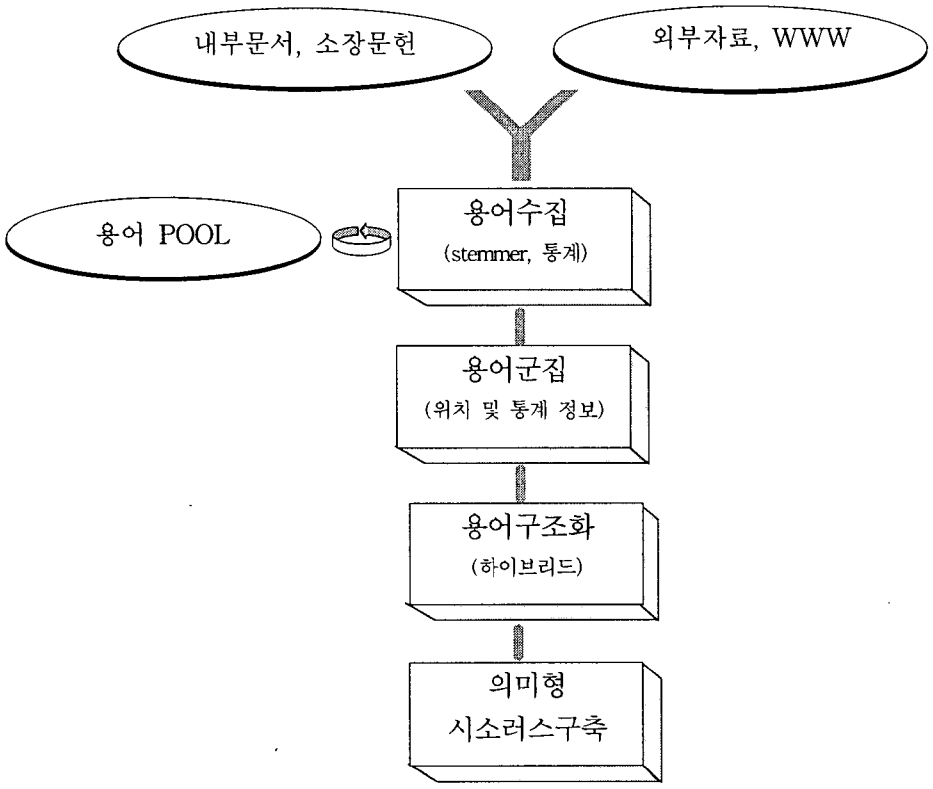
따라서 지식관리 시스템에서 사용될 의미형 시소러스는 의미에 기반하는 전통적 분류체계로서 계통분류표(Taxonomy)의 역할과 인적정보공유를 표현한 지식지도의 역할까지도 수행하는 새로운 형태의 정보검색 및 관리도구이다.

2. 의미형 시소러스의 개발

시소러스를 구성하고 있는 디스크립터는 크게 구체적인 것과 개념적인 것으로 구분할 수 있다. 구체적인 것인 개념보다는 대상물이 분명한 것을 나타낸 용어들이며, 개념적인 것은 구체적인 대상물보다는 추상적인 개념을 나타난 것으로 연속성을 갖고 있는 용어들이다.

구체적인 용어들은 시대상을 반영한 것으로 "MP3"와 "탄저균", "떼제베" 등과 같이 구체적인 대상물을 지칭하는 용어들이다. 개념적인 용어들은 인간이 갖고 있는 지식구조와 유사한 것으로 "음악파일"과 "생화학 무기", "고속철도" 등과 같이 어떠한 대상물을 총칭하는 용어들이다. 전자의 경우는 어떤 특정 기간동안 관련된 문헌과 정보들이 집중적으로 생산되어 일정 시기에는 매우 중요한 용어로 사용될 수 있다. 이러한 용어들은 대부분 시소러스나 분류체계에 표현되지 않은 것이 일반적이다. 후자의 경우는 특정 기간보다는 해당 시소러스를 개발한 시점에 그때까지 주요하다고 판단된 분야의 주요 용어들로써 새로운 개념이나 자료의 출현에 크게 좌우되지 않는 용어들이다. 의미형 시소러스는 기본 구조로서 개념위주의 용어들을 선택하고 구체적인 용어들은 디스크립터로 채택한다. 또한 개념위주의 용어들을 선택하고 시소러스의 유용성을 높이기 위해서는 문헌집합의 현재 내용을 정확하게 반영하는 동적 시소러스(dynamic thesaurus)의 구축이 필요하다. 왜냐하면, 시소러스의 구축 대상이 된 문헌집합은 문헌의 추가 및 삭제에 의하여 계속해서 변화한다. 이와 같이 동적으로 변화하는 문헌집합을 고정된 시소러스를 이용하여 검색할 경우 검색효율은 점차 떨어지게 되기 때문이다.(이재윤 1994)

<그림 4>는 의미형 시소러스의 개발과정에서 용어수집에서 시소러스 구축까지의 과정을 순차적으로 도식화한 것이다.



<그림 4> 의미형 시소러스 개발과정

2.1 기본구조의 설정

시소러스의 구조는 평면시소러스구조를 포함하여 속(屬) 혹은 계층구조, 트리구조 혹은 패싯시스템, 최상위개념어구조, 양방향계층구조, 포괄적 계층구조 등으로 구분할 수 있다. 이 중 가장 일반적인 형태는 평면 시소러스를 포함한 계층구조이다. 이 가운데 포괄적 계층 구조는 자신의 모든 상위어(BT: Broader Terms)와 모든 하위어(NT: Narrower Terms)를 깊이에 상관없이 모두 자모순으로 배열하는 방식이다. 이 방식은 분야별시소러스(마이크로 시소러스)인 경우에는 좋으나 매크로시소러스인 경우에는 관계용어가 많아 적합하지 않다는 특징이 있다. 이 방법을 채택하는 경우, "과학"이라는 용어 아래에는 결국 수천개의 NT를 갖게 되어 참조에 어려움이 있다. 시소러스를 하나의 나무구조로 구축하는 경우에는 더 많은 NT를 갖게 된다. 또한 구축자의 입장에서 보면 전체를 참조할 수 있는 장점은 있으나 상하계층의 구별이 거의 되지 않아 정확한 계층관계를 정의할 수 없다는 단점이 있다. 장점으로는 기존의

분류표에 열거된 전개방식과 동일하기 때문에 이용자에 친숙하며, 인간의 지식구조의 형태와 유사하기 때문에 지식관리시스템의 구조화에 적합하다. 또한, 다른 구조방식보다 기계로 처리할 경우에 화면 표시가 단순하여 이용자의 가독성이 높아지는 장점이 있다. 전체가 제한된 디스크립터에 모두 NT의 형태로 링크되는 문제는 접기기능을 이용하여 처리할 수 있다. 따라서 의미형 시소러스는 분야별 시소러스(예: 기계 및 지정학 분야)로 개발되는 점과 기계로 처리하여 컴퓨터 화면을 통해 용어에 대한 관리와 열람이 이루어지기 때문에 포괄적 계층구조를 유지하는 것이 바람직하다.

한편, 다국어 시소러스의 경우 영어대응어를 비디스크립터로 처리하여 등가어(UF: Used For)로 처리하여 병행 기록하였으나, 의미형 시소러스에는 영어 대응어를 등가어로 처리하지 않으며 한국어판과 영미판 두 개의 시소러스로 구축한다.

다음은 본 연구에서 구축한 의미형 시소러스의 일부이다.

세계정치	geo-politics
.영국정치	.british-politics
..영국왕실	..british-royals
...영국여왕	...queen of england
..하원	..house of commons
...노동당	...labor party

2.2 용어의 수집

시소러스 개발과정에서의 용어 수집은 디스크립터와 비디스크립터, 후보디스크립터를 수집하는 과정이다. 이 과정은 해당 용어를 수집할 용어 추출 정보원(sources)을 참조로 하여 수작업 혹은 컴퓨터 처리과정을 통해 수행한다. 기계를 이용한 용어 수집은 크게 용어 추출 정보원의 수집과 자연언어처리방법을 이용한 후보 디스크립터의 추출로 과정이 크게 구분될 수 있다.

1) 용어추출 정보원의 수집

전통적으로 용어를 수집하는 정보원으로는, 1)기존의 유사주제분야의 시소러스 혹은 분류체계 등과 같은 구조화된 용어집과 2)백과사전을 비롯한 일반사전에 등재된 표제어 3) 전문분야의 전문 용어집, 4)색인지나 초록지, 기타 출판물의 권말색인, 5) 편람, 목록집, 교과서,

규격집 등 기타자료 (Aitchison, Gilchrist, 1972)를 활용한다. 이상에 제시된 각종 정보수집원은 책자형태의 자료이기 때문에 급변하는 정보화 사회에서 새로운 지식이나 개념을 실시간적으로 반영하지 못하고 있다. 지식관리 시스템에서 탑재될 의미형 시소러스에 사용될 용어 수집원은 애치슨(Aitchison)이 제시한 기본적인 정보원을 포함하여, 가상공간에 존재하는 디지털 문서와 해당 기관, 단체에서 소장하고 있는 내부문서 등까지도 포함되어야 한다. 왜냐하면, 지금까지 개발된 대부분의 시소러스는 문헌형태의 자료를 검색을 의도한 것이며, 인터넷을 포함한 가상공간에 표현된 문자정보와 전문인들이 갖고 있는 문자로 표현되지 않은 지식검색을 고려한 것은 아니기 때문이다. 따라서 지식관리시스템에 탑재될 검색대상 정보를 검색하는 의미형 시소러스는 용어 추출 정보원을 해당 기관에서 소장한 문헌자료 외에 내부문서, 소장하고 있는 기타자료, 구성원들이 생성한 유무형의 지식, 국내외의 유관 웹사이트 정보도 포함한다.

용어 추출 정보원의 확대는 수집원을 무한대로 확대하는 것이 아니기 때문에 적정한 규모의 정보원을 확보할 필요가 있다. 언어학에서는 실제 활용되고 있는 언어용례를 수집하여 사전의 표제어 선정과 사전의 용례를 확보하기 위해 주요 문장이나 자료들을 수집한다. 이를 언어학에서는 코퍼스(말뭉치)라 지칭한다. 시소러스에 실릴 디스크립터를 확보하고, 용어 추출과 용어간 관계요소를 파악하기 위해서는 코퍼스를 구축하여 활용하는 것이 바람직하다. 의미형 시소러스 구축용 코퍼스의 수집대상과 우선 순위는 다음과 같다.

① 해당 분야 국어 및 영어 시소러스 (전문용어사전 포함)
② 소장자료 (문헌, 보고서)
③ 내부문서
④ 외부 소장 문서

이렇게 용어 추출 수집원을 문서까지도 확대한 것은 지식관리 시스템에 사용되는 의미형 시소러스는 소급자료의 검색과 함께 최근에 생성된 인쇄자료뿐만 아니라 내부문서, 외국자료, 특정한 웹자료까지도 효과적으로 검색이 가능한 다이나믹한 디스크립터로 유지·관리되어야 하기 때문이다. 즉, 기존의 시소러스 혹은 주제명 표목집에 기술된 표제어와는 별도로 기존 인쇄형 시소러스가 출간된 이후에 나타난 새로운 개념(용어)을 수집하기 위해서는 해당 분야에서 새로이 출간된 혹은 생성된 자료들이 용어 추출 정보원으로 활용되어야 한다. 따라서, 용어 수집을 위해 참조해야할 문서의 양이 기하급수적으로 증가함에 따라 수작업으로 용어를 수집하여 처리하는 것보다는 자연언어처리기법을 도입하여 컴퓨터를 이용하여 디스크립터와 비디스크립터를 추출하는 것이 일반화되고 있다.

이와 같이 코퍼스를 이용한 용어 수집 방법은 필연적으로 용어 추출 정보원의 규모에 따라 적정 규모 이상의 디스크립터가 나타나기 때문에 디스크립터의 수를 적정하게 유지하기

위한 통제가 이루어져야 한다. 지금까지 선행연구에서 적정한 시소러스의 개수에 대한 연구는 거의 이루어지지 않았으나, 지금까지 개발된 주요 시소러스의 규모를 분석하여 유추할 수 있다. 대표적인 시소러스인 TEST(Office of Naval Research)의 디스크립터수는 23,364개(비디스크립터 5554개 포함)와 NASA 시소러스(NASA. 1985)의 경우는 40,738개, INSPEC 시소러스(INSPEC. 1991)의 경우는 약 13,000개(비디스크립터 6,000개 포함)의 디스크립터로 구성되어 있다. 국내에서 개발된 것으로는 국방과학기술 한글시소러스(국방과학연구소. 1994)의 경우에 23,126개(비디스크립터 5,211개 포함)의 디스크립터로 구성되어 있으며, 한국통신 시소러스(한국과학기술원. 1995)의 경우에는 9,031개(비디스크립터 687개 포함)의 디스크립터로 구성되어 있으며, 한국학 분야의 학술용 지시형 시소러스의 디스크립터수는 약 20,000개가 구축되었으며, 법률분야 관련어집(법원도서관. 1998)의 디스크립터수는 18,130개(비디스크립터 1,404개 포함)로 확인할 수 있었다. 일본의 닛케이시소러스(日本經濟新聞社. 1988)는 15,434개의 디스크립터로 구성되어 있었다. 시소러스가 검색에서 효과를 얻기 위한 적절한 규모는 구체적으로 제시된 것은 없으며, 다만 검색대상이 되는 분야의 수준에 따라 수록된 디스크립터의 수가 결정된다고 판단된다. 이를 분석하면 NASA 시소러스의 경우를 제외하고는, 대략 1만에서 2만 여개의 디스크립터 개수를 유지하고 있다. 또한 비디스크립터수를 제외할 경우에 대부분의 시소러스는 15,000개 내외의 디스크립터 항으로 이루어져 있다. 따라서 주제별 시소러스 개발은 비디스크립터수를 제외한 15,000개 이하의 수를 유지하는 것이 적합한 디스크립터의 수라고 판단한다. 왜냐하면 용어수가 많을 경우 용어의 특정성이 상대적으로 높아지기 때문에 적정한 재현율을 보장하지 못하기 때문이다. 한편 용어수가 적을 경우 용어의 일반성이 상대적으로 높아지기 때문에 적정한 정도율을 얻을 수 없다는 상반된 결과가 나타난다. 한편, 포괄적이며 이용자 위주의 검색이 가능하도록 이와는 별도로 대등관계를 갖는 전거어 파일이 개발되어야 다양한 이용자의 검색 패턴을 수용할 수 있을 것이다.

2) 후보 디스크립터의 추출

기계를 이용한 후보 디스크립터의 추출은 용어추출기(stemmer)를 사용한다. 이는 주로 문장내에서 명사(구)를 인식하여 추출하는 일종의 프로그램이다. 용어추출을 질을 결정하는 형태소 해석과정과 문장의 품사를 결정하기 위한 태깅능력이다. 또한 복원된 단어의 형태를 파악하기 위한 용어사전과 불용어사전의 질이 추출된 용어의 최종 질을 결정한다. 본 연구에서 추출의 대상이 되는 한글 후보 디스크립터의 형태는 다음과 같다.(남영준. 1995)

- 명사(구)
- 복합명사(구)
- (복합)명사+의,적,성+(복합)명사

이와 같이 추출될 디스크립터의 형태를 고정하는 것은 기존 정보검색도구에 수록된 디스크립터의 형태를 고려한 것이며, 상용웹검색엔진의 이용자의 로그파일[3]을 분석한 결과이다. 추출될 디스크립터의 형태를 고정함으로써 형태소해석과정과 태깅과정에 정확도를 높일 수 있다.

3) 용어 풀

용어 풀은 용어 추출 정보원에서 추출된 후보 디스크립터 형태가운데 디스크립터나 비디스크립터로 채택되지 않았으나 의미가 있다고 판단된 용어에 대한 임시저장소이다. 용어풀은 개념적인 정보보다는 구체적인 정보로서 용어로서의 생명주기가 상대적으로 짧은 용어들을 보관한다. 여기에 보관되어 있는 용어들은 계속해서 생성되는 내부 자료와 새로이 입수되는 외부자료들을 분석하여 적정 규모의 빈도값을 갖게 될 경우에 의미형 시소러스에 기재하여 의미형 시소러스의 새로운 개념의 추가 용이하도록 한다. 또한, 의미형 시소러스에서 색인이나 검색에서 거의 사용되지 않은 디스크립터를 용어 풀로 이동하여 보관한다. 따라서 용어풀에 있는 모든 용어들은 관련 디스크립터와의 링크정보를 유지하도록 하여, 용어의 추가와 삭제가 용이하도록 하고 이에 대한 히스토리 파일을 유지하여 적정한 디스크립터의 수를 유지할 수 있도록 한다.

이와 같이 코퍼스와 용어추출기를 이용한 용어 수집 방법은 국내에서 법원시소러스(법원도서관. 1998) 개발시에 처음으로 도입되었으며, 그 후 국내에서 개발된 많은 시소러스들이 이 방법을 사용하여 개발이 이루어지고 있다.(대한민국 국회.2001)(고려대학교 민족문화 연구소(2001 개발중).

2.3 용어의 통제

용어 수집과정을 거쳐 수집된 정보는 지식베이스로 변환되기 위해서는 일련의 통제과정이 필요하다. 첫 번째 통제과정은 유사 용어들을 서로 군집화하는(clustering) 과정이다.

두 번째 과정은 군집화된 용어가운데 디스크립터로 채택될 수 있는 용어를 선정하는 작업이다. 세 번째로는 선정된 디스크립터간의 관계를 설정하는 과정이다.

---

3) 라이코스 코리아의 이용자 로그파일 (2001년 1월에서 7월까지)
야후 코리아의 이용자 로그 파일 (2001년 1월에서 7월까지)
한미르의 이용자 로그파일 (2001년 1월에서 7월까지)

1) 군집화

군집화는 유사한 개념 혹은 서로 관련이 있는 용어들을 하나이상의 주제에 모아두어 용어의 선정과 관계설정을 할 수 있도록 하는 사전 작업이다. 군집화를 위한 방법으로는 전통적으로 통계적인 연관도 계산에 유사계수 공식을 사용하며, 중요한 공식으로는 다이스계수, 자카드계수, 코사인계수, 중복도계수, 타니모토계수 등이 있다. 이 공식들은 모두 비슷한 결과를 산출하므로 어느 공식을 선택할 것인가는 문제가 되지 않지만, 특히 코사인계수와 다이스계수가 비교적 널리 사용되고 있다(정영미 1993).

그러나 본 연구에서는 유사도 측정 방식보다는 추출 정보원의 주제구분에 따라 용어를 군집화하는 방법을 택한다. 왜냐하면 유사도 알고리듬은 후보 디스크립터의 수가 적정한 규모를 유지할 경우에만 효과적으로 활용할 수 있기 때문이다. 즉, 1)길이가 긴 문서는 일반적으로 단어(term)들이 반복적으로 나타나기 때문에 길이가 짧은 문서에 비하여 비교적 질의와 높은 유사도를 나타낸다. 또한 2) 길이가 긴 문서는 일반적으로 여러 개의 서로 다른 단어들이 나타나기 때문에 길이가 짧은 문서에 비하여 질의와 높은 유사도를 나타낼 뿐만 아니라 검색될 가능성도 그만큼 높아지게 되기 때문이다. 즉, 문헌의 길이가 불규칙한 경우에는 유사도의 수치가 안정적이지 못하기 때문이다. 지식관리 시스템의 소장 대상인 자료의 경우에는 책자 형태의 자료와 내부문서, 웹문서 등과 같이 문헌의 길이가 서로 상이한 경우가 대부분이다. 이렇게 용어 추출원의 길이가 차이가 있기 때문에 의미형 시소러스의 디스크립터 군집화는 용어 추출에 유사도 측정방법을 사용하지 않는다. 왜냐하면, 마이크로 시소러스 구축을 위한 디스크립터 추출 정보원은 이미 주제별로 대체적인 구분이 이루어져있기 때문이다. 즉, 개발하여야 할 주제가 결정되어 있고 해당 주제에 관련된 용어 추출을 위해서는 이미 추출을 위한 코퍼스가 형태별 혹은 주제별로 구성이 되어 있기 때문에 추출된 용어의 군집화는 이미 일차적으로 이루어졌다고 간주한다. 예를 들면, 기계학분야의 진동(vibration)관련 학술논문에서 추출한 용어들은 진동과 이미 관련이 있다고 간주하고, 추출된 용어들은 진동이라는 중심어에 연결되도록 한다.

2) 디스크립터 선정

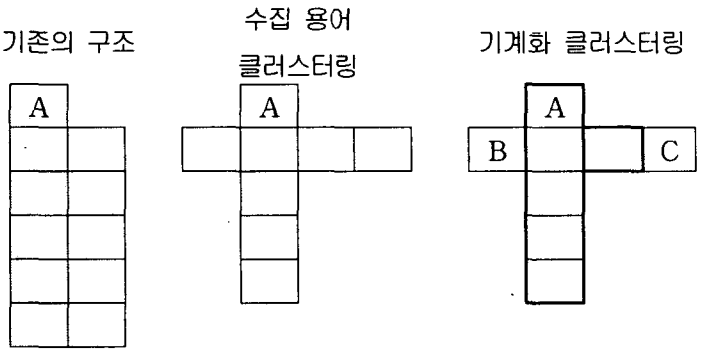
군집화된 용어들가운데 디스크립터의 선정은 상대빈도를 이용한다. 상대빈도는 단순빈도 값을 각각 문헌빈도, 장서빈도, 문헌길이 등으로 나누어줌으로써 빈도의 값을 표준화한 것이다. 특히 단순빈도 값을 문헌의 전체합으로 나누어 용어출현빈도의 값을 표준화하여 그 값의 대소로 선정의 기준 순위를 결정하는 방법인 역문헌빈도가 일반적으로 상대빈도에 사용된다. 이러한 역문헌빈도는 문헌빈도가 낮은 단어 즉, 적은 수의 문헌에 나타난 단어에 높은 중요도를 부여하는 방법을 사용한다. 내부문서의 경우는 한 건을 하나의 문서단위로 보지 않고

해당 부서의 업무 분장에 따른 업무에 관련된 문서파일을 하나의 문서단위로 간주한다. 왜냐하면 일반적으로 하나의 문서가 갖고 있는 정보의 양은 극히 제한적이기 때문이다.

3) 관계 설정

시소러스의 용어 관계는 크게 계층관계와 연관관계, 대등관계로 구분될 수 있다. 서휘는 시소러스 용어들간의 계층을 미리 설정하지 않는 방법으로 시소러스의 자동 구축방법을 제시하였다. 그는 여기서 용어간의 관계표현은 자동으로 이루어질 수 있는 방법을 제시하였으며, 용어의 선정과 연관관계설정, 계층관계설정에서 적정한 결과치를 얻어 시소러스의 자동생성에 대한 가능성을 제시하였다.(서휘 1999), 한편, 남영준 등은 문장구조적인 특성을 이용하여 용어간의 관계설정을 시도하였으나, 용어의 추출과정을 거치지 않고 대상 문장에 있는 주요어간의 관계를 실험하였다. 그는 시소러스 관계 자동설정은 반드시 제한적 전제조건이 있어야하며, 그 범위는 극히 제한적이고, 계층관계의 정립은 지금까지의 자연어처리기법으로는 기타관계설정보다 상대적으로 어렵다는 것을 제시하였다(남영준 등 1995). 또한 서휘의 연구 결과에서도 자동 구축된 계층 결과에 대해 86.1%가 만족을 한다고 응답하였지만, 13.9%의 잘못 구조화된 결과에 대해서는 수작업으로 처리해야 함을 의미한다.

이러한 제약점 때문에 국내에서 개발된 시소러스의 계층관계를 비롯한 관계설정과정에서는 기계작업과 수작업이 병행되는 하이브리드 방법을 택하고 있다. 일차 계층 및 연관관계 설정은 기계를 이용하여 주요 용어가 출현한 빈도와 용어가 동시에 출현한 위치정보로써 용어간 관계의 유사도를 측정하는 용어간 공기정보(word co-occurrence) 알고리듬을 활용한다. 이 과정은 특정한 용어들이 하나의 문장내에서 출현한 것과 하나의 문단에서, 하나의 장서에서 동시에 출현한 각각의 공기정보를 이용하여 용어간 관계의 순위를 결정한다. 또한 용어가 동시에 출현한 위치에 따라 가중치를 부여한다. 즉, 한 문장내에서 용어가 동시에 출현하였을 경우에 가장 높은 연관계수를 부여하고, 문헌내에 동시에 출현하였을 경우에는 상대적으로 유사계수를 낮게 책정한다. 한편, 공기정보의 추출범위를 문헌까지 확대하였을 경우에만 단어쌍을 갖는 용어들은 문장내 관계설정을 하지 않는다. 두 번째 과정은 첫 번째 과정에서 기계를 이용하여 일차 구조화된 시소러스를 해당 분야 전문가들이 수작업으로 관계의 적절성 여부를 평가한다. 계층에 관한 정보는 기존에 구축된 유사분야의 한글시소러스구조를 이용하였다. 이를 그림으로 설명하면 <그림 6>과 같다.(남영준 2000)



<그림 6> 구조 자동화의 순서도

위 그림에서 하나의 눈금은 하나의 용어를 의미하며, A는 기준이 되는 용어이다. 수집된 용어 클러스터링을 기존의 시소러스 구조(하나의 표준 클러스터링)에 적용하여 기계적으로 클러스터링을 이루도록 한다. 세 번째 눈금의 좌우 용어(B와 C)는 기존의 계층구조와는 매칭이 되지 않으므로 용어 A와 서로 관련은 있되 계층관계에 속하지 않는 용어이기 때문에 용어 A와 연관관계(RT)로 처리한다. 이렇게 일차 구조화된 정보는 수작업으로 최종 검증과정을 거치며 이차 수작업 과정에서 디스크립터 선택의 검증과 평가가 다시 이루어진다.

4) 영어디스크립터 선정 및 링크

지식관리시스템에서 활용되어야할 의미형 시소러스는 반드시 영어 대응 디스크립터가 있어야 한다. 왜냐하면 지식관리시스템은 기본적으로 자관이 소장하고 있지 않은 정보도 검색하는 것을 목적으로 하기 때문이다. 영어 디스크립터는 영문으로 이루어진 자료를 대상으로 앞서 언급한 의미형 시소러스구축 방안에서 제시한 수집방안을 사용하지 않는다. 왜냐하면, 용어 추출과 선정에 있어 영문디스크립터의 구조는 한글 디스크립터의 구조와 차이가 있기 때문이다. 예를 들면, 한글 디스크립터의 경우는 대부분 명사 혹은 명사구로 이루어진 반면에 영어 디스크립터는 명사와 명사구 이외에 어절 형태의 디스크립터도 존재하기 때문이다.

영어 디스크립터의 선정은 한글 의미형 시소러스가 완성된 이후에 대응 디스크립터를 영미계에서 출판한 시소러스의 디스크립터를 활용한다. 한글 시소러스에 대응하는 영어 디스크립터가 없을 경우에는 상용웹검색엔진에서 제공하는 디렉토리에 열거된 표제항을 사용한다. 인쇄형 영어 시소러스에 등재되지 않은 용어는 신조어 혹은 사용빈도가 최근에 나타난 용어가 대부분이다. 따라서 해당 용어들은 상용웹검색엔진의 분류체계의 항에서 나타날 확률이 높기 때문이다.

이렇게 수집된 영어 디스크립터는 구축이 완료된 한글 시소러스의 디스크립터에 링크하여

구축한다. 왜냐하면 하나의 한글개념어에 복수의 영어개념이 존재할 경우에는 이를 대등어 혹은 전거어로 처리할 수 있으나, 세분화된 영어개념에 대한 대응 한글시소러스가 없을 경우에는 시소러스 구조의 혼란이 발생할 수 있기 때문이다. 다음은 미국의 "예술 및 유산 시소러스 (AAT: Arts and Architecture Thesaurus)"에 열거된 디스크립터가운데 한글의 단일개념에 세분화된 영어개념이 존재한 예이다. 예에서 볼 수 있듯이 영어 디스크립터를 기준하여 한글개념(디스크립터)을 부가하는 방법을 사용하였을 경우에는 대응 용어의 선정과 구조가 상대적으로 어렵게 된다. 즉, 적절한 대응어가 없기 때문에 용어추출정보원에 없는 용어를 생성해야 하는 문제가 발생할 수 있다. 그러므로 A의 과정을 B로 처리하기보다는 C에 관련된 영어대응어 D를 부여하는 것이 훨씬 효율적이다. 따라서 의미형 영어 시소러스의 개발순서는 의미형 한글 시소러스를 완성한 후에 각 개념에 해당하는 영어디스크립터를 한글디스크립터에 부여하는 방식을 채택한다.

A	B	C	D
영어디스크립터	한글대응어	한글디스크립터	대등처리
ceramic porcelain	자기 없음	자기(瓷器)	ceramic porcelain

IV. 결 론

새로운 지식기반 시대에서 지식관리는 생산성 향상 및 장기적인 경제발전의 주 원동력이며, 이를 어떻게 활용하는 지가 국가 성장의 주요한 결정요인으로 활용되고 있다. 따라서 국내외적으로 생산되는 지식을 체계적이고 구체적으로 생성, 획득, 확산, 활용하는 것이 국가경쟁력의 주요한 관건이 될 것이다. 전통적으로 정보서비스를 제공하던 도서관은 소장자료외에 자관내외에서 생산되는 모든 정보를 지식화하여 지식서비스를 제공하는 지식관리의 주체로 역할이 변화하게 되었다. 이는 필연적으로 보다 유효한 정보검색도구가 필요하게 되었으며, 정보검색의 범위가 국내자료외에 전세계 정보의 많은 부분을 차지하고 있는 영미계 자료까지도 검색할 수 있는 검색도구가 필요하게 되었다.

본 연구에서는 지식관리시스템의 효율적인 관리 및 검색을 위해 의미형 한영 시소러스가 개발되어야 하는 이유를 조사하였으며, 또한 의미형 시소러스 개발에 고려해야할 기준과 요소를 제시하였다. 그 내용을 요약하면 다음과 같다.

1) 지식관리에 대한 관점은 도서관계와 기업간에 차이가 있다. 도서관계에서 지식관리시스템은 소장자료와 웹과 같은 접근가능한 자료를 하나의 공간에서 구조화하여 정보서비스를 제공할 수 있는 포털 정보제공 사이트로 간주하고 있다. 기업에서는 도서관의 역할외에 소속원들이 생성한 유·무형의 비정형데이터까지도 수집·관리할 통합형 데이터관리 시스템으로 간주하고 있다.

2) 의미형 시소러스는 주제별로 개발되며, 개발 주제 영역이 분명한 마이크로 시소러스이기 때문에 포괄적 계층구조 체계로 개발한다. 디스크립터의 수준을 가능한 개념적인 표현이 있는 용어(예: 생화학무기)를 선택하고, 구체적인 내용을 표현하고 있는 용어(예: 탄저균)를 제외한다. 구체적 표현이 이루어지고 있는 용어들은 전거어 사전에 수록하여 처리한다.

3) 후보 디스크립터수집을 위해서는 유사 시소러스의 디스크립터와 같은 전통적인 정보원외에 해당 분야 한·영 시소러스와 기관소장자료, 내부문서, 외부 소장 문서를 활용한다. 이러한 정보원의 확대는 디스크립터의 증가로 이어지나, 주제분야별 시소러스는 최대규모인 NASA의 40,738개에서 한국통신의 9,031개이며 이를 제외할 경우에, 대부분의 시소러스는 디스크립터만을 계수하였을 경우에 약 15,000개 내외의 표제항수를 유지하고 있다. 따라서 마이크로 시소러스의 디스크립터수는 비디스크립터를 제외하고 15,000개의 디스크립터를 유지하는 것이 바람직하다. 다만 다양한 정보접근점이 제공과 이용자 위주의 검색시스템이 되기 위해서는 반드시 대등관계를 갖는 전거어사전이 개발되어야 한다.

4) 용어풀을 이용하여 추출된 후보디스크립터 가운데 시소러스 등재에 누락된 단어와 시소러스에서 사용빈도가 낮은 용어들을 관리한다. 이는 시소러스 개발시점 이후에 해당 용어의 관련된 외부 지식과 정보가 증가할 경우에 우선적으로 시소러스의 디스크립터나 비디스크립터로 활용할 수 있는 용어 저장소의 역할을 수행한다.

5) 용어간 관계설정은 수작업과 기계처리작업을 병행하는 하이브리드 방법을 사용한다. 특히, 일차 계층 및 연관관계 설정은 기계를 이용하여 주요 용어가 출현한 빈도와 용어가 동시에 출현한 위치정보로써 용어간 관계의 유사도를 측정하는 용어간 공기정보(word co-occurrence) 알고리듬을 활용한다. 1차구조화된 시소러스는 해당 분야 전문가에 의해 수작업으로 조정하는 2차 과정을 거쳐 완성한다.

6) 영어 디스크립터의 선정은 한글 의미형 시소러스가 완성된 이후에 대응 디스크립터를 영미계에서 출판한 시소러스의 디스크립터를 활용한다. 한글 시소러스에 대응하는 영어 디스크립터가 없을 경우에는 상용웹검색엔진에서 제공하는 디렉토리에 열거된 표제항을 사용한다. 즉, 의미형 한영시소러스의 개발의 순서는 의미형 한글 시소러스를 완성한 후에 각 개념에 해당하는 영어디스크립터를 부여하는 방식을 채택한다.

본 연구에서는 국내외 시소러스의 구축에 대한 여러 가지 이론과 실험에 대한 것을 정리

하였다. 이를 바탕으로 국내외 모든 기관에서 구축하고 있는 지식관리시스템의 지식베이스로 활용되고 있는 의미형 시소러스에 대한 실제적인 개발방안을 제시하였다. 이는 의미형 시소러스 개발의 주요한 지침이 될 것이며, 시소러스 개발 소요되는 시간과 경비를 줄일 수 있는 하나의 지식으로 활용될 수 있다. 궁극적으로 이 연구 결과는 국가 지식기반 관리 시스템의 주요한 지식베이스로 활용될 수 있을 것이다. 단, 본 연구에서는 이러한 개발 기준에 따라 개발된 시소러스의 효율성에 대한 객관적인 수치데이터를 제시하지 못했기 때문에 이론적인 연구에 머무를 수 있다. 이 연구가 지식기반사회의 시소러스개발 지침으로 자리잡기 위해서는 이에 대한 후속적인 연구가 시급히 이루어져야 할 것이다.

참 고 문 헌

고려대학교 민족문화 연구소(2001 개발중). 『한국학 시소러스』. (http://kcrc.korea.ac.kr/tul/nul7_start.htm)

국방과학연구소. 1994.『국방과학기술 한글시소러스』. 동연구소.

남영준, 최석두, 이두영. 1997. "시소러스 자동생성에 관한 실험적 연구",《한국정보관리학회 학술대회 논문집》제4권. pp. 25-30.

남영준. 1995. 색인어 형태분석에 의한 한국어 자동색인기법연구. 중앙대학교 대학원 문헌정보학과 박사학위논문.

남영준. 1998. "웹문서 분류체계의 새로운 설계",《한국문헌정보학회지》32권, 3호. pp. 207-230.

남영준, 안동언. 2000. "다국어 시소러스의 개념 동일화 작업에 관한 실험적 연구",《지식처리 연구》1권, 2호. pp. 57-78.

노영희. 2000. "개념기반 검색을 위한 시소러스 관계의 효과적 활용방안에 관한 연구",《정보관리학회지》17권, 4호. pp. 47-65.

대한민국 국회. 2001.『국회도서관 시소러스(정치, 경제 편)』. 국회도서관.

문경화, 남태우. 2001. "디지털 도서관의 효율적인 지식컨텐츠관리에 관한 연구",《정보관리학회지》18권, 3호. pp. 41-61,

법원도서관. 1998.『법률분야 관련어집』. 법원도서관.

성기주. 2001. "국내 대학의 여성학 웹 사이트 콘텐츠에 관한 연구",《한국문헌정보학회지》35권, 3호. pp. 191-210.

신동민. 2001. "인터넷 검색엔진의 디렉토리 구성에 관한 연구", 《정보관리학회지》 18권, 2호. pp. 143-163

오재익. 1997. 영한대역 시소러스의 문제점에 관한 연구. 중앙대학교 대학원 문헌정보학과 석사학위논문.

이두영, 최석두, 남영준. 1995.『지능형 정보검색에 관한 연구』. 한국통신 연구개발단 95 장기기초연구과제 최종보고서.

이란주. 2001. "공공도서관 의학정보서비스를 위한 지식관리시스템 설계에 관한 연구", 《정보관리학회지》 18권, 3호. pp. 63-86.

이재윤. 1994. 동적 시소러스의 구축에 관한 실험적 연구. 연세대학교 대학원 문헌정보학과.

이창수. 2000. "정보통신기술분야 인터넷 자원의 분류체계에 관한 연구", 《한국도서관정보학회지》 31권, 4호. pp. 111-138.

日本經濟新聞社. 1988.『日經シソラス』. 同新聞社

정영미. 1993.『정보검색론』. 개정판. 서울: 구미무역 출판부.

한국과학기술원. 1995.『한국통신 시소러스』. 한국통신 연구개발원.

한상길. 2001. "산업 분야 인터넷 자원의 분류체계에 관한 연구", 《정보관리학회지》 18권, 3호. pp. 285-309.

American National Standards Institute. 1980. *Guidelines for Thesaurus Structure, Construction, and Use.* New York : ANSI,

Bruce, H. 1998. "User Satisfaction with Information Seeking on the Internet", *Journal of the American Society for Information Science,* Vol. 49, No.6. pp. 541-556.

Galinski, Christian, Goebel Jurgen W. 1996. *Guide to Terminology Agreements. Inforterm. Termnet.*

H. P. Luhn, 1957. "A Statistical Approach to Mechanized Encoding and Searching of Library Information", *IBM Journal of Research & Development,* Vol. 1, No. 4. pp. 309-317.

INSPEC. 1991. *THESAURUS 1991.* The Institution of Electrical Engineers.

NASA. 1985. *NASA THESAURUS.* NASA.

Office of Naval Research. 1967. *Thesaurus of Engineering and Scientific Terms.* United States Department of Defense.