과학기술 전문용어의 다국어 의미망 생성과 분석

Building and Analysis of Semantic Network on S&T Multilingual Terminology

정 도 헌* · 최 희 윤**
Do-Heon Jeong · Hee-Yoon Choi

차 례

초 록

다국어로 구축된 학술정보 시스템의 통합검색 환경을 구현하기 위해서는 다국어 전문용어에 대한 해석을 제공하고 전문용어의 분야별 분류정보를 제공할 수 있는 시스템이 필요하다. 본 연구는 이러한 다국어 환경의 통합 정보검색 시스템을 운용할 수 있도록 기반시스템을 구축하는 것을 목적으로 한다. 다국어 의미망으로 상호 연결된 과학기술 전문용어 체계를 구축하는 방법과 다단계 연결노드에 대한 최단거리 탐색 기법을 소개하였다. 또한, 생성된 용어군집 결과를 해석하기 위한 기초분석을 수행하여 향후 심도있는 분석연구를 수행하기 위한 기반을 마련하고자 하였다.

키 워 드

과학기술 전문용어, 다국어 시스템, 용어 네트워크 분석, 의미망, 자동분류

* 한국과학기술정보연구원 지식기반팀 연구원
(Researcher, Knowledge Assets Team, Korea Institute of Science and Technology Information, heon@kisti.re.kr)
** 한국과학기술정보연구원 지식기반팀 책임연구원
(Principal Researcher, Knowledge Assets Team, Korea Institute of Science and Technology Information, hychoi@kisti.re.kr)
• 논문접수일자 : 2006년 11월 14일
• 게재확정일자 : 2006년 12월 14일

ABSTRACT

A terminology system capable of providing interpretations and classification information on a multilingual science and technology(S&T) terminology is essential to establish an integrated search environment for multilingual S&T information systems. This paper aims to build a base system to manage an integrated information system for multilingual S&T terminology search. It introduces a method to build a search system for S&T terminologies internally linked through the multilingual semantic network and a search technique on the multiple linked nodes. In order to provide a foundation for further analysis researches, it also attempts to suggest a basic approach to interpret terminology clusters generated with those two search methods.

KEYWORDS

S&T Terminology, Multilingual Systems, Terminology SNA, Semantic Networks, Auto Classification

1. 서 론

과학기술 전문용어 체계를 구축한다는 것은 이용자와 정보검색시스템 간의 정보검색 시 발생하는 격차를 해소할 수 있는 색인시스템을 제공한다는 의미이다. 즉, 용어의 해석이 가능하도록 지식베이스를 구축하는 것과 해석기법에 대한 연구가 용어시스템 구축의 중요한 사항이다. 문헌으로부터 추출한 색인어를 시소러스의 개념어와 매칭하여 자동 주제 및 분야할당을 시도하는 등의 유사연구(정한민 외 2006)도 이러한 관점에서 출발한 것이다.

그러나, 자동분류, 자동색인 실험을 비롯한 언어자원을 기반으로 한 검색성능 실험은 실제 대용량 데이터베이스 환경에서 수행하기 힘든 한계로 인해 주로 정제된 언어자원을 바탕으로 한 실험적 수준에서 이루어져 온 것이 대부분이다. 따라서, 본 연구에서는 다국어 의미망이라는 새로운 방법론을 제시함으로써 실제 대용량 데이터베이스로부터 상호 연계된 언어자원을 확보하는 기법을 제안하고자 한다. 용어간의 의미망 생성을 통해 다국어 용어의 네트워크를 구성하고, 이 의미망을 해석하기 위한 네트워크 분석방법을 소개할 것이다.

이러한 다국어 용어체계의 구축을 통해 얻을 수 있는 효과로 첫째, 다국어를 지원하는 자동 질의확장 시스템을 통해 이용자의 편의성을 증대시키고, 동북아 메타정보 공유시스템을 지원하는 CJK 상호검색기능을 제공하며, 대규모 학술 정보검색 시스템의 성능 고도화 및 지능

화 지원이 가능할 것이다. 둘째, 지속적으로 생산, 관리할 수 있는 용어 콘텐트 체계를 구축함으로써 이용자의 검색 행태를 분석하고 피드백하여 재적용(재학습)할 수 있는 용어체계를 구축할 수 있다. 향후 이러한 용어 콘텐트 체계 구축을 통해 언어자원에 대한 지속적이고 일괄성 있는 실험환경이 구축될 수 있을 것이다(김지영 외 2000). 마지막으로 과학기술 용어체계 및 과학기술 표준분류체계 등의 기초연구 수행을 통해 자동분류, 자동색인, 온톨로지 생성연구 등 차세대 응용연구 및 기술로의 발전을 도모할 수 있다.

2. 관련연구 및 한계

KISTI와 같이 수천만 건에 달하는 대용량 학술정보 데이터베이스를 서비스하는 기관은 이용자에게 필요한 정보를 적시에 선별적으로 제공하기 위한 맞춤형 정보제공 서비스를 수행하는 것이 매우 중요하다. 이를 위해, KISTI는 다양한 분야별 검색식을 미리 구축하고 이를 이용해 신규구축 데이터를 검색하여 맞춤형 정보를 제공하는 방법을 이용하고 있다. 이처럼, 선행 연구된 많은 자동분류 이론을 실제 시스템에 적용하기 어려운 점은 대규모 분류체계에 대한 정비가 선행되어야 하며, 시소러스 또는 용어체계와 같은 언어자원에 대한 기본적인 관리 시스템과 이를 응용할 수 있는 환경이 갖춰져야 하기 때문이다. 이러한 문제와 관련하여 사전에 필수적으로 검토되어야 하는 몇 가지 사항이 존재한다.

첫째, 실제 운용하는 로컬시스템에 필요한 언어자원의 구축 필요성에 관한 문제이다. KISTI의 경우에는 과학기술 전문용어 사전을 구축하는 업무와 과학기술 범용 시소러스 구축을 하는 업무 등이 이에 해당될 것이다. 언어자원 구축사업에 대한 동의가 중요한 이유는 그만큼 성공적인 수행을 위해서는 막대한 비용과 연구기간, 인력투자가 필요하기 때문이다.

둘째, 언어자원의 신뢰성을 어떻게 확보할 것인가 하는 점이다. 자원의 규모가 커짐에 따라 과거와 같이 전사적인 수준의 수작업 검증을 수행하는 것은 거의 불가능한 것으로 판단되고 있다. 이미 1970년대 초에 다국어 자원을 해석하고 분류하기 위한 연구가 있었다. 영어와 독일어의 교차언어 문헌검색의 실험적 연구(Salton 1970)가 이루어진 이후로, UMLS 메타시소러스를 이용한 스페인어와 프랑스어의 교차언어 검색연구(Eichmann and Ruiz 1998), 독일어와 영어로 구성된 다국어 시소러스(GIRT)를 이용한 교차언어 검색연구(Gey and Jiang 1999) 등 수많은 실험적 연구가 수행되어 왔으며, 국내에서도 역시 많은 연구가 수행되어 왔다. 그러나, 대부분의 용어처리와 관련된 실험들은 한정된 규모의 신뢰성 있는 자원을 바탕으로 수행한 경우가 많아 실제 시스템의 적용할 경우, 언어자원의 규모 및 신뢰성에 대한 문제는 여전히 남아있다.

셋째, 분야별로 수행된 연구결과(시소러스, 용어체계 등)를 어떻게 통합 또는 연계할 것인

가 하는 문제이다. 실제로 초기 구축 시에 통합 문제가 사전 검토되지 않는 경우에, 여러 분야별로 구축된 시소러스를 통합하여 매크로시소러스를 구축하는 것은 거의 불가능하다. 게다가 앞서 언급한 바와 같이 실제로 통합을 해야 할 만큼의 충분한 언어자원도 구축되어 있지 못한 실정이다. 이기종 시스템간의 통합문제는 데이터마이닝, MDR 표준화, 온톨로지 시스템 등에서도 공통된 논의거리이다.

넷째, 용어통제에 관한 문제이다. 일단 시스템적으로든 수작업으로든 통제를 시작하면, "어떤 방법을 통해, 어느 수준까지, 언제까지 할 것인가"하는 문제와 직면하게 된다. 이와 비슷한 문제를 제기하여 언어통제를 취하지 않는 시스템을 통해 비용효과 측면에서 자연언어 통제어휘 혼합시스템을 제안한 연구도 있다(최석두 1993). 자언언어(비통제언어)와 통제어휘의 특성을 간략히 비교하면, 자연언어는 언어의 최신성이 반영되어 언어자원에 풍부한 어휘를 반영할 수 있어 표현력이 좋은 반면, 동의어와 유사 동의어 및 동음이의어 처리가 어려운 점이 있다.

다섯째, 시소러스 또는 용어체계와 정보검색 시스템(데이터베이스)과의 동조를 어떻게 맞출 것인가 하는 문제이다. 실제 운용되는 데이터베이스와 용어사업이 별개로 진행됨으로 인해 검색시스템과 언어자원 간의 의사소통이 원활치 못한 점에 기인하는 경우가 많아, 관련 시스템의 구축 후 상호 연계방안에 대한 지속적인 연구가 필요하게 된다(정한민 외 2006). 이기종 시스템의 통합과 상호연계에 관련된 분야 역시 많은 연구와 비용투자를 필요로 한다.

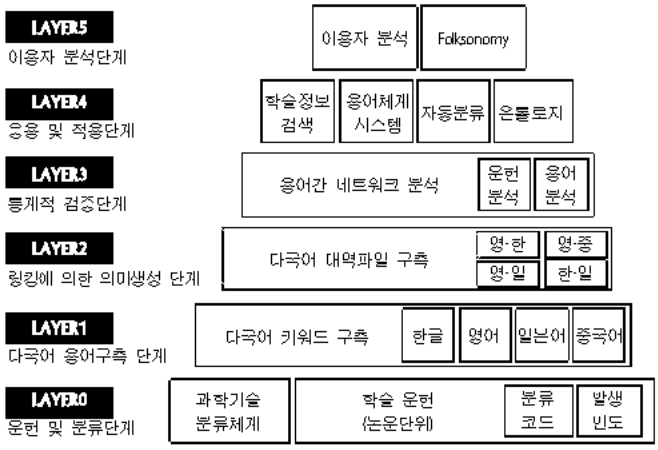
여섯째, 언어자원 시스템은 전문용어가 변화가 잦은 동적인 자원임을 감안하여 원천데이터의 변경 시 자동적으로 연계된 자원들이 갱신될 수 있도록 유연한 개방구조로 설계되어야 한다. 이것은 실제 이용자의 질의를 추적하고, 이를 피드백하여 재학습할 수 있도록 설계를 하는 것으로, 최근 이용자 협력형 태깅 시스템(Collaborative Tagging Systems)의 협동적인 데이터 구축방법을 참고할 수 있다(Cattuto et al. 2006). 또한, 초기 프레임워크 설계 단계에서 시스템의 향후 발전가능성 및 응용가능성에 대해 충분히 고려되어야 한다.

마지막으로, 용어시스템의 유니코드 환경에 대한 것으로, 다국어 시스템을 지원하는 경우에 필수적으로 채택해야 하는 것은 물론이지만, 기본적으로 자국어 환경을 지원하는 경우에도 향후 확장성을 위해서는 충분히 검토되어야 할 사항이다.

3. 과학기술 다국어 전문용어 체계 구축방안

3.1 과학기술 다국어 용어체계 개념 피라미드

<그림 1>은 전체적인 용어체계를 개념적인 레이어로 구분한 개념 피라미드이다. 실제 시스템에서는 각 레이어가 중첩적으로 구성되어



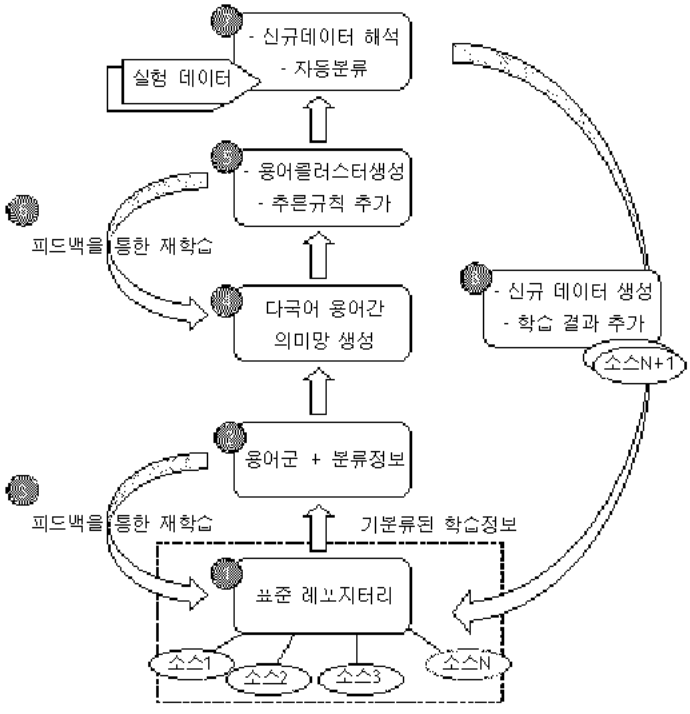
〈그림 1〉 과학기술 다국어 용어체계 개념 피라미드

있으며, 크게 L0을 백그라운드 단계, L1~L3를 의미생성과 분석검증 단계, L4~L5를 분석 및 응용단계로 구분할 수 있다. 본 연구에서는 전체 개념 피라미드상의 L0~L3 단계에 해당하는 기초 자원 구축에서 의미망 생성 및 구축결과를 검증하는 기초 분석단계까지를 수행하였다. 각 단계별로 주요 프로세스를 설명하면 다음과 같다.

1) L0 (문헌 및 분류단계): 정보시스템에 구축된 자원을 표준화된 포맷(유니코드 변환, 필드별 데이터 표준화)으로 재구축하고 분류, 빈도 등 기초통계를 산출한다. 또한, 과학기술 분류체계를 중심으로 도입자원의 분류에 대한 상호 매칭테이블을 생성한다.
2) L1 (다국어 용어구축 단계): L0으로부터 기계적으로 다국어 키워드를 추출한다.
3) L2 (연결에 의한 의미생성 단계): L0으로부터 다국어 키워드 대역정보를 추출한다. L0의 각 소스별로 대역파일 추출 알고리즘을 적용하여 데이터를 생성한다.
4) L3 (통계적 검증단계): L1~L2에 생성된 의미망 해석을 통해 구축된 자원간의 관계를 분석하고 통계적인 검증을 실시한다.
5) L4 (응용 및 적용단계): 다국어 검색시스템, 전문용어 분류시스템, 문헌 자동분류 및 온톨로지 시스템 등으로 응용 발전한다.
6) L5 (이용자 분석단계): 이용자의 피드백을 통한 자원 재해석 및 재구축 프로세스를 지원한다.

3.2 용어 재학습을 위한 시스템 프로세스

〈그림 1〉의 피라미드 개념도가 단계별 레이어를 중심으로 표현했다면 〈그림 2〉와 같이 프로세스 흐름을 중심으로 표현할 수도 있다. 이 그림은 L0~L3에 이르는 다국어 용어체계 구축을 기반으로 L4의 자동분류에 이르기까지의 프로세스를 나타낸 것이다. "④다국어 용어 간



〈그림 2〉 용어체계의 재학습 프로세스 흐름도

의미망 생성" 및 "⑤용어 클러스터 생성 및 기초분석" 단계까지를 본 연구에서 수행하며, 의미망 해석을 위한 추론규칙 생성과 피드백을 통한 재학습, 자동분류 등은 향후 심도있는 연구가 추가 수행되어야 한다.

〈그림 2〉의 단계별 주요내용을 설명하면 다음과 같다.

① 학술정보 데이터베이스로부터 추출한 원천 데이터를 유니코드 포맷의 표준화 스키마 구조에 저장한다.

② 표준 레포지터리로부터 각 데이터베이스별 키워드를 자동추출하며, 각 용어별 빈도 및 분류정보를 취합하여 후보분류를 추천한다.

③ 데이터베이스 품질오류를 수정하여 표준 레포지터리에 반영하며, 용어별 키워드 통계로 부터 추론된 분류정보를 이용해 임계치를 초과하는 오류추정 데이터에 대해 재학습 데이터를 추가하여 표준 레포지터리를 갱신한다.

④ 전문용어 데이터베이스의 상위 레이어에 다국어 대역파일을 생성하여 독립적인 키워드 데이터 간에 상호 연결된 다국어 의미망을 생성한다.

⑤ 용어군집화(클러스터링)를 통해 기계적인 데이터 검증 및 분석작업을 수행한 후 데이터 무결성 확보를 위한 새로운 의미망 생성

을 위한 추론규칙을 구축한다. 이것은 용어 군집의 재 클러스터링 작업 및 추론 규칙을 추가 생성(어의적 해석기능 학습)하는 과정이다.

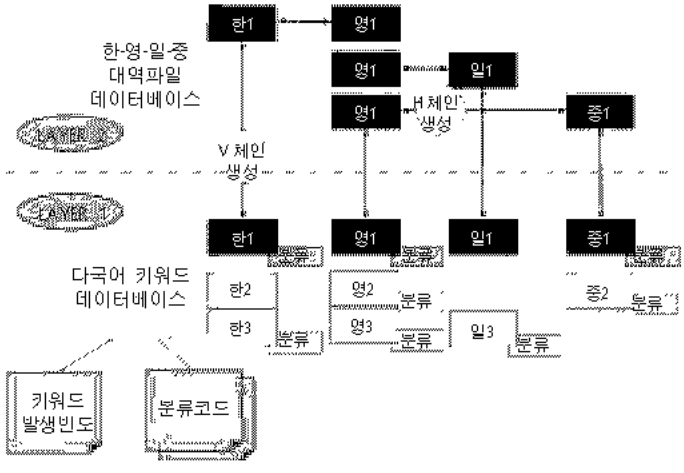
⑥ 용어의 다의적 해석 또는 동의어 집단 해석을 위한 추론규칙을 적용해 기 구축된 다국어 용어 간 의미망을 갱신한다.

⑦ 이 과정을 통해 얻은 다국어 용어체계에 신규 문헌 또는 키워드 데이터를 입력하면 시스템은 자동으로 후보 분류코드를 생성, 추천한다.

⑧ 시스템을 통해 얻은 실험데이터 결과를 표준 레포지터리에 신규 학습데이터(소스 N+1)로 추가하여 전체적인 시스템의 언어자원 규모를 증가시킴으로써 자동으로 시스템을 갱신하고 성장시킨다.

3.3 다국어 용어구축 방안(의미망 생성방안)

〈그림 3〉은 4장에서 기계적으로 처리할 용어 간 네트워크 생성방안을 도식화한 것이다. 데이터베이스로부터 추출한 키워드는 표준포맷으로 변환하여 구축(개념피라미드의 L1)하며, 다국어로 구성된 대역파일 레이어(L2)를 이용해 기 구축된 용어들을 상호 연계한다. L1, L2에는 용어 간의 수평체인(horizontal chain)이 구축되고, L1과 L2의 레이어 간에는 수직체인(vertical chain)이 생성된다. 이렇게 생성된 H체인과 V체인의 복잡한 네트워크를 "다국어 용어체계의 의미망"이라 한다. 이를 통해 데이터베이스 내에서 상호 연결정보가 없이 구축된 용어들이 추론이 가능한 상태로 바뀌게 되는데 〈그림 3〉에서 Layer 1의 일본어키워드 1에 비어있는 분류정보는 한국어키워드 1, 영어



〈그림 3〉 L1, L2간에 H체인, V체인으로 구성된 용어체계 의미망

키워드 1, 중국어키워드 1의 기 구축 정보로부터 후보 분류에 대한 추론이 가능해 진다.

3.4 데이터베이스 자원 및 구축현황

본 시스템의 가장 큰 특징은 실제 운용되는 대용량 정보시스템으로부터 직접 언어자원을 구축하였다는 점이다. 이를 통해 용어 시스템과 실제 데이터베이스간의 기계적인 언어해석기(일종의 미들웨어) 구현 문제와 같은 정보시스템의 의사소통 문제를 제거하였으며, 시소러스 또는 디스크립터 구축 시에 발생되는 정보

〈표 1〉 데이터베이스 자원 및 구축현황

구 분	주요 내용	구축량 (2006. 11 현재)
다국어 대역파일	-과학기술 전문용어 영-한 대역집 데이터	-영문-한글 153,255 쌍
	-원자력분야 한-영-일 용어사전 데이터	-영-한, 영-일 각 1,495 쌍
	-LCAS 중국메타정보, 영-중 대역파일(가공생성)	-영-중 102,952 쌍
	-과학기술학회마을, 영-한 대역파일 (가공생성)	-영-한 40,334 쌍
		총 298,036 쌍 (596,072 건)
한국어	-과학기술 학회마을 데이터베이스	-133,174 건
	-해외학술지(JAFO) 데이터베이스(BIST DB)	-994,651 건
	-원자력분야 한-영-일 용어사전 데이터	-1,729 건
	-과학기술 전문용어 영-한 대역집	-153,255 건
		총 1,282,809 건
영어	-과학기술 학회마을 데이터베이스	-275,055 건
	-해외학술지(JAFO) 데이터베이스	-628,968 건
	-LCAS 중국학술지 데이터베이스	-519,225 건
	-Inspec 해외도입 데이터베이스	-9,904 건
	-원자력분야 한-영-일 용어사전 데이터	-2,005 건
	-과학기술 전문용어 영-한 대역집	-153,255 건
		총 1,588,412 건
일본어	-JST 과학기술 문헌속보 CD-ROM 추출가공	-357,524 건
	-원자력분야 한-영-일 용어사전 데이터	-1,495 건
		총 359,019 건
중국어	-LCAS 중국학술지 데이터베이스	-465,060 건
		총 465,060 건

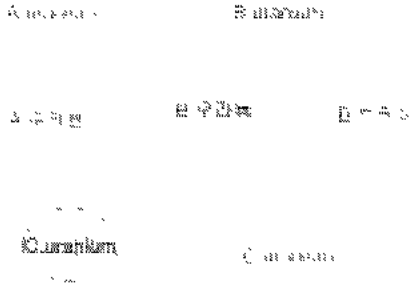
해석의 틈새문제를 해결하였다. 과학기술 전문 용어를 구축하기 위해 KISTI의 모든 가용자원을 이용하였으며, <표 1>에서 보는 바와 같이 대역파일은 약 59만6천여 건, 키워드는 약 337만2천여 건을 구축하고 있다. 향후 도메인과 데이터 량은 계속 증가할 것이다.

4. 다국어 전문용어 네트워크 생성과 탐색

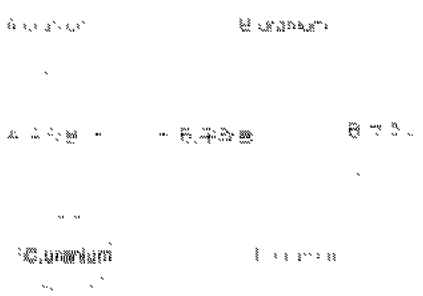
4.1 용어 간 네트워크(의미망) 생성

용어 간 네트워크를 생성하는 방법론은 본 연구의 핵심이 되는 부분이다. 비통제 자연언어 어휘를 키워드로 구축하고 있는 데이터베이스 시스템으로부터 원천데이터를 확보하고 기계적인 방법으로 의미관계를 자동 생성한다.

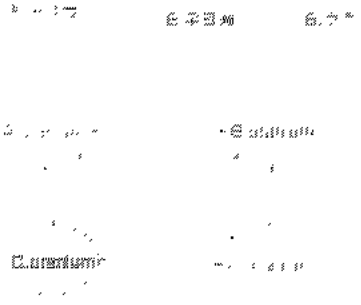
<그림 4>에서 A, B 도메인은 대역파일 형태로 존재하고 C, D 는 분류정보와 문헌 내 발생빈도 등의 정보를 담고 있다. "우라늄"에 해당하는 다국어 키워드를 추출한 결과, 4개의 도메인에서 7개의 용어가 검색되었는데, 이 중 두개의 그룹은 이미 대역파일 정보가 일부 생



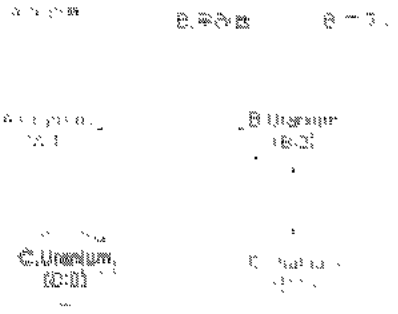
<그림 4> 자연구축 상태의 용어 '우라늄' 관련어



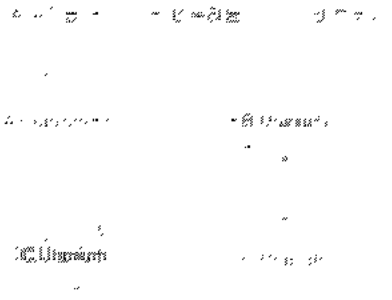
<그림 5> '(도메인)우라늄'으로 의미망 생성



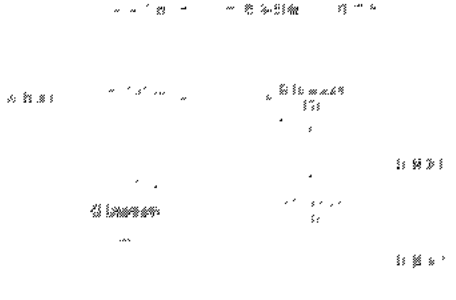
<그림 6> 복잡한 상호참조



<그림 7> 우선어 생성규칙을 적용



〈그림 8〉 객체 간 상호 연결체인(의미망) 생성



〈그림 9〉 신규 용어 생성 및 추가로 확장된 다국어 용어체계 네트워크

성되어 있었다. "A.우라늄"과 "B.우라늄"은 분류정보와 빈도정보가 서로 다른 도메인에 존재하므로 고유한 값(URI)으로 취급한다. 도메인 이하의 객체 값(Object Value)을 연결하는 방법을 통해, "A.우라늄 = B.우라늄" 의 관계를 생성하면 〈그림 5〉와 같이 "A.uranium"을 통해 "B.ウラン"을 찾을 수 있게 된다. 현재 노드 간 거리(의미망의 거리)는 4차(단계)이다.

〈그림 6〉과 같이 "uranium"을 통해 모든 상호참조를 생성하면, 용어 N개에 대해 N(N 1)/2의 체인이 생성되어 비효율적인 관리구조를 갖게 된다. 〈그림 6〉에서는 모두 4(4 1)/2=6 회의 연결체인 생성이 일어나며 노드가 늘어날수록 기하급수적으로 체인이 증가한다. 따라서 〈그림 7〉과 같이 참조노드의 수를 이용해 대표용어에 (N 1)회의 링크 생성과정을 거치면 심플한 연계체인을 생성할 수 있다. 이로써 B.Uranium의 중간매개 역할(정보력)이 증가하게 된다. 이 경우 4 1=3회의 링크 생성으로 해결된다. 만약 모든 용어의 참조노드 수가 같다면 임의로 지정하거나 자동으로 선택하도록 한다.

〈그림 8〉은 〈그림 5〉와 〈그림 7〉을 모두 적용한 결과이다. 이것은 본 용어체계에서 발생하는 일반적인 연결구조로 자동 링크생성 프로세스를 수행한 결과로 나타나는 공통모델이다. A.uranium ⇒ B.ウラン 을 찾아가는 거리는 가장 먼 거리가 4, 가장 짧은 거리는 2이다. 연결체인을 해석한 결과, 가장 짧은 거리를 찾아내는 탐색 알고리즘을 통해 2를 계산하게 된다. 〈그림 9〉처럼 용어 집단에 A.용어1, D.용어1, D.용어2 등이 계속 추가될 수 있으며, 네트워크상의 어느 노드에 생성되어도 모든 용어는 동일하게 검색될 수 있다.

4.2 네트워크 탐색: 최단거리 탐색기법

기계적으로 생성된 용어 간 의미망 모형은 각 객체 간 최단거리 탐색기법을 통해 서로 상호 참조하게 된다. 최단거리 탐색기법은 실시간 정보검색 시 전체 의미망을 검색하는 탐색시간을 최소화하여, 객체 간에 최적의 상호 의미해석 환경을 제공한다. 본 연구에서는 추론

규칙의 기본인 "if A → B and B → C, then A → C"를 처리하기 위해 연결체인 해석(의미망 해석) 엔진을 개발하였다. 이후 모든 프로세스는 이와 같은 추론규칙을 이용하여 진행한다. 네트워크 탐색을 위한 공식은 아래와 같다. 네트워크 탐색과정에서 용어탐색이 중복적으로 일어나므로 반복 발생하는 노드 값을 제어하기 위해 집합간의 연산처리를 실행하였다. N단계의 노드 값인 Rn은 네트워크를 무한 탐색하여 결과 값이 공집합이 될 때 연결차수를 반환하며 종료된다.

$$R_n = X_n - S_{n-1} \quad (S_n = \sum_{i=0}^{n} X_i)$$

또는,

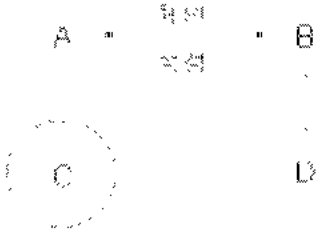
$$R_n = X_n - \sum_{i=0}^{n-1} X_i$$

X_n: n 단계에서의 이웃 노드 값의 집합 (단, X_0는 자신을 취함)

R_n: n차 링크노드 값(집합)

가. 네트워크 탐색의 예 1 (Single)

이하의 과정은 용어 A C, B D 두 개의 군집 간에 A B 체인이 생성된 경우, 용어 C에서 D까지 최단거리 검색 수행을 하는 예이다.



1) C로부터 1차링크 해석

self 집합 X_0 = {C}

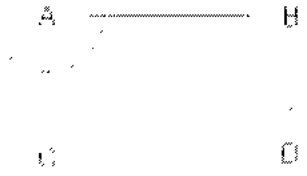
1차 참조 집합 X_1 = {A}

1차 연결어 X_1 X_0 = {A} ⊓ ⌐{C} = {A}

1차 링크체인을 따라 A로 이동

R_1 = {A}

∴ A로 이동



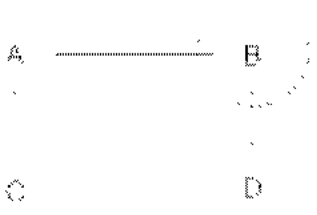
2) C로부터 2차 링크해석(1차 연결체인을 따라 A로 이동 후 연결체인 해석)

X_0 = {C}, X_1 = {A}

X_2 = {B, C}

2차 연결어 R_2 = X_2 (X_0+X_1) = {B, C} ⊓ ⌐{A, C} = {B}

∴ B로 이동



3) C로부터 3차 링크해석(2차 연결체인을 따라 B로 이동 후 연결체인 해석)

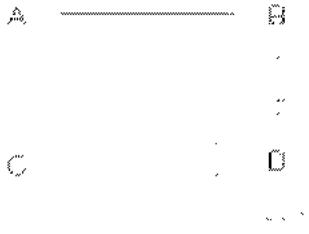
X_0 = {C}, X_1 = {A}, X_2 = {B, C}

X_3 = {A, D}

3차 연결어 $R_3 = X_3 \quad (X_0+X_1+X_2) = \{A, D\}$

$\sqcap \ \lceil \{A, B, C\} = \{D\}$

∴ D로 이동



4) C로부터 4차 링크해석(3차 연결체인을 따라 D로 이동 후 연결체인 해석)

$X_0 = \{C\}$, $X_1 = \{A\}$, $X_2 = \{B, C\}$, $X_3 = \{A, D\}$

$X_4 = \{B\}$

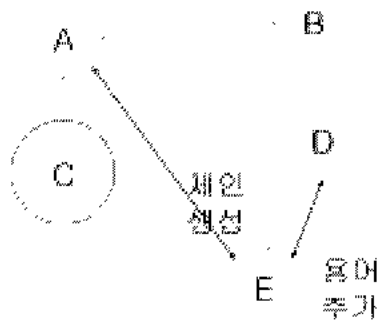
4차 연결어 $R_4 = X_4 \quad (X_0+X_1+X_2+X_3) = \{B\}$

$\sqcap \ \lceil \{A, B, C, D\} = \{\ \}$

따라서, 탐색중인 객체리스트 집합이 공집합으로 반환되는 시점인 4차에서 탐색 프로세스가 중지되며 최종해석 결과는 "최단거리 3차 연결어"이다.

나. 네트워크 탐색의 예 2 (Multiple)

이하의 과정은 용어 E가 추가되고 다중체인이 생성된 경우, 최단거리 검색 수행을 하는 예이다.



1) C로부터 1차 링크해석

self 집합 $X_0 = \{C\}$

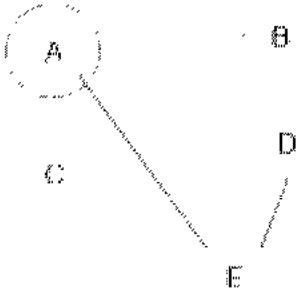
1차 참조 집합 $X_1 = \{A\}$

1차 연결어 $X_1 \quad X_0 = \{A\} \sqcap \ \lceil \{C\} = \{A\}$

1차 연결체인을 따라 A로 이동

$R_1 = \{A\}$

∴ A로 이동



2) C로부터 2차 링크해석(1차 연결체인을 따라 A로 이동 후 연결체인 해석)

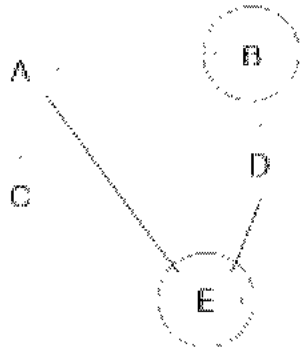
$X_0 = \{C\}$, $X_1 = \{A\}$

$X_2 = \{B, C, E\}$

2차 연결어 $R_2 = X_2 \quad (X_0+X_1) = \{B, C, E\}$

$\sqcap \ \lceil \{A, C\} = \{B, E\}$

∴ B와 E로 분기하여 이동



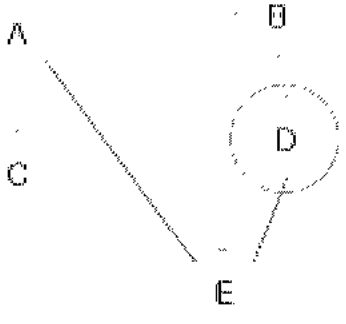
3) C로부터 3차 링크해석(2차 연결체인을 따라 B와 E로 분기된 후 각각 연결체인 해석)

X_0 = {C}, X_1 = {A}, X_2 = {B, C, E}

X_3 = {A, D}

3차 연결어 $R_3 = X_3 \quad (X_0+X_1+X_2)$ = {A, D} ∩ ⌐{A, B, C, E} = {D}

∴ D로 이동



4) C로부터 4차 링크해석(3차 연결체인을 따라 D로 이동 후 연결체인 해석)

X_0 = {C}, X_1 = {A}, X_2 = {B, C, E}, X_3 = {A, D}

X_4 = {B, E}

4차 연결어 $R_4 = X_4 \quad (X_0+X_1+X_2+X_3)$ = {B, E} ∩ ⌐{A, B, C, D, E} = { }

따라서, 탐색중인 객체리스트 집합이 공집합으로 반환되는 시점인 4차에서 탐색 프로세스가 중지되며 최종해석 결과는 "최단거리 3차 연결어"이다.

4.3 동의어 군집 생성

용어체계를 구축함에 있어 동의어 처리는 매우 어려운 문제이다. 한글의 경우는 띄어쓰기에 따라서, 영어의 경우는 약어표기나 두문자 표기 등으로 인해 많은 이형(異形)이 존재할 수 있으며, 모든 패턴에 대한 기계적인 인식이 실제로는 거의 불가능하게 된다. 용어시스템은 이러한 용어에 대한 패턴처리 뿐만 아니라 의미적으로도 동의어군을 해석해 낼 수 있어야 한다. 본 연구에서는 자연언어 상태의 키워드군으로부터 기계적인 방법으로 대역파일을 생성하고 대역파일과 키워드 간에 의미망 생성을 통해 전체 동의어 집합을 생성하였다. 이러한 용어군집을 처리하기 위해서 의미망을 생성하고 해석하는 의미망 처리용 엔진개발이 필요하였다. 의미망을 생성하기 위해 앞서 설명한 의미망 생성규칙을 적용하였으며 의미망을 해석하기 위해 의미망 탐색기법을 적용하였다. 이것은 유니코드를 완벽하게 지원하는 대용량 용어 데이터베이스의 처리 및 관리를 위한 두 가지 핵심엔진이라 할 수 있다. 이를 통해 전체 데이터베이스 내 용어 간 군집도를 계산하며,

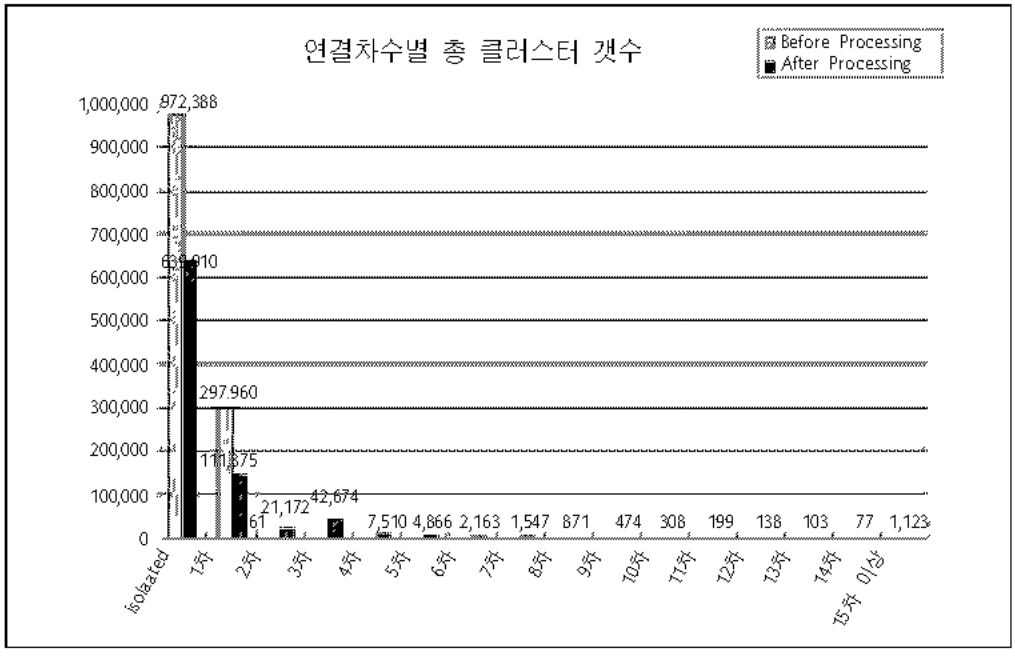
용어 간 평균 체인 수, 군집도, 용어 군집 수 등을 산출할 수 있다.

용어의 군집을 생성하고 이를 분석하는 일련의 프로세싱 결과는 〈표 2〉와 같다. 프로세싱을 통해 링크생성은 약 60만개에서 142만개로 증가하였고, 이에 따라 평균 연결체인의 차수와 클러스터 내 평균용어의 개수가 증가하였다. 또한, 1차 연결 상태로 흩어진 집단을 군집화한 결과, 전체 클러스터의 수는 7만개 이상 감소하였다.

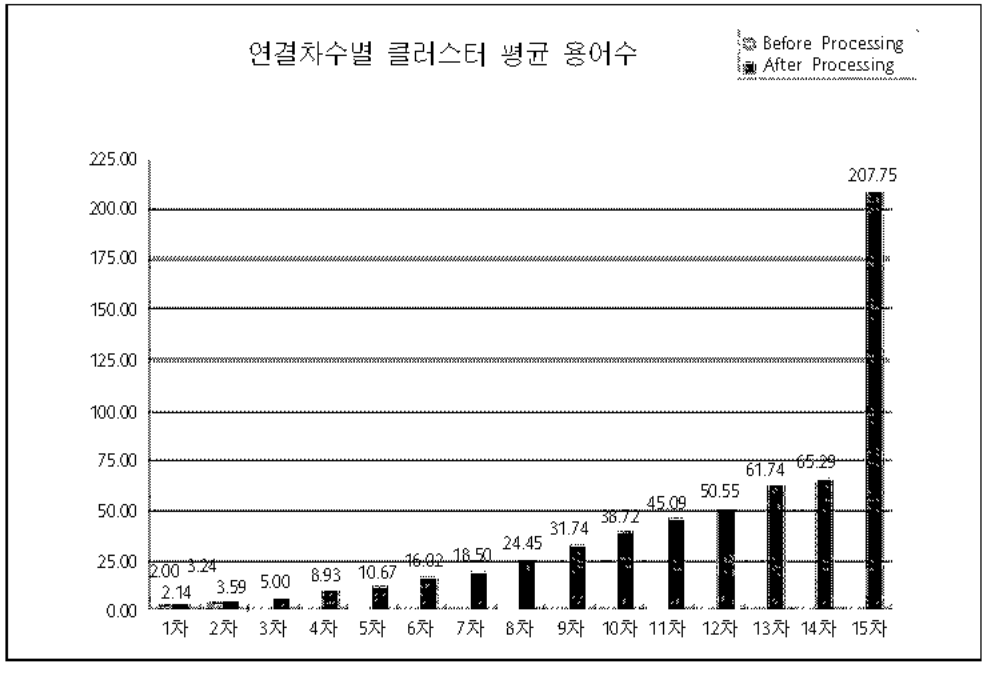
〈그림 10〉과 같이 상호체인이 없이 독립된 용어와 1차 링크어가 감소하고 2차이상의 연결체인을 갖는 군집수가 증가하였다. 15차 이상(16차, 17차 등 포함)의 용어 군집수가 상당히 많은 것을 알 수 있다. 〈그림 11〉은 1차, 2차 연결 군집수의 평균용어 수를 나타낸 것이다. 군집의 규모(또는 링크 연결차수)에 따라 평균용어의 수가 완만하게 증가함을 알 수 있다. 그러나 연결 차수별 용어의 증가 추이에서 15차(15차만 측정) 링크어의 평균용어 수가 급증하였는데, 이러한 현상은 대역파일의 생성 시 데이터 품질에 문제가 있었거나 범용적인 용어를 군집화한 것으로 그 원인을 추정할 수 있을 것이다. 따라서 현재의 용어체계에서는 14차 이하의 연결수준이 유효하며 그 이상에 대해서는 용어체계에 대한 검증이 필요한 것으로 확인되었다. 이 과정은 〈그림 2〉에서 "⑥ 피드백을 통한 재학습"에 해당하는 프로세스로 향후 지속적인 연구수행을 통해 해결해야 하는 부분이다.

〈표 2〉 다국어 의미망 생성 전과 후의 결과 비교표

	BEFORE	AFTER	비 고
대상 데이터 건수	1,569,919건	1,569,919건	3,981,872 건 중 빈도1인 키워드 삭제
대상 용어 수	597,531건	930,009건	링크가 존재하는 용어의 수
전체 클러스터 수	298,021개	224,600개	※ 용어군집을 생성하고, ※ 용어군집 스캐닝 알고리즘을 적용한 결과
평균 연결체인 차수	1.00차	1.90차	
클러스터 내 평균용어 개수	2.00개	4.82개	
상호체인 수	599,086개 체인 (초기대역파일구축 결과)	1,426,420개 상호체인	노드간 연결라인 수는 "체인수÷2" (A⇒B, B⇒A)



〈그림 10〉 연결차수별 총 클러스터 개수



〈그림 11〉 연결차수별 클러스터 평균 용어 수

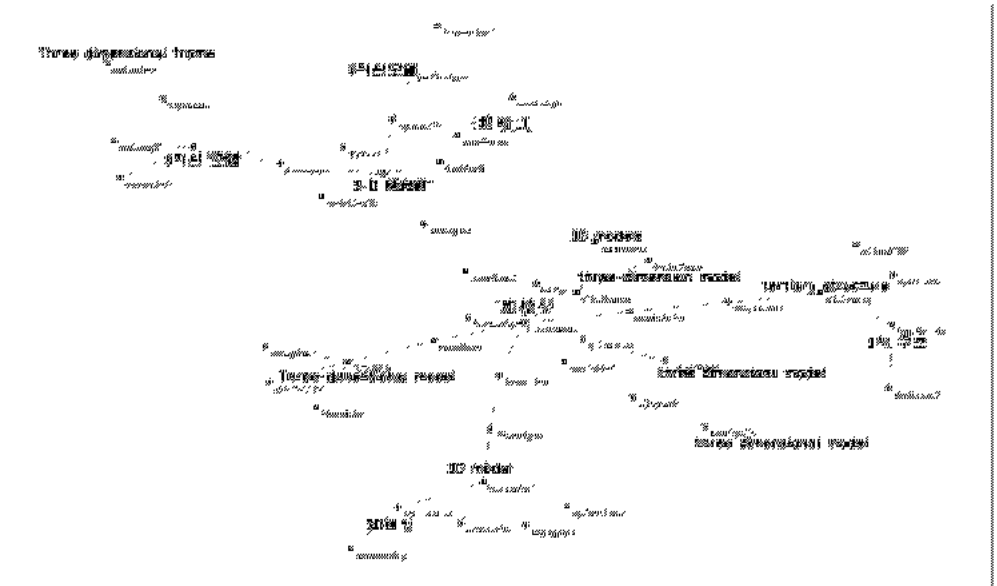
5. 다국어 전문용어 네트워크 분석

용어 네트워크를 도식화하고 이를 검증하기 위해, 구축된 용어 데이터베이스로부터 48개 용어, 10차 링크어로 구성된 적당한 규모의 샘플군집을 임의로 추출하였다. 48개 용어의 고유한 ID(URI)를 기준으로 네트워크 그래프를 그린 후, 용어문자열 기준 분석, 연결된 노드 수 기준분석, 중심성 분석 등을 차례로 수행하였다. 네트워크 드로잉을 위해 pajek을 사용하였고, pajek 표준입력 포맷에 맞도록 네트워크 해석도구 및 포맷변환 도구를 직접 개발하였다. 특히 다국어 문자열 처리를 위해 유니코드를 지원하는 중심성 분석도구 개발이 필요하였다.

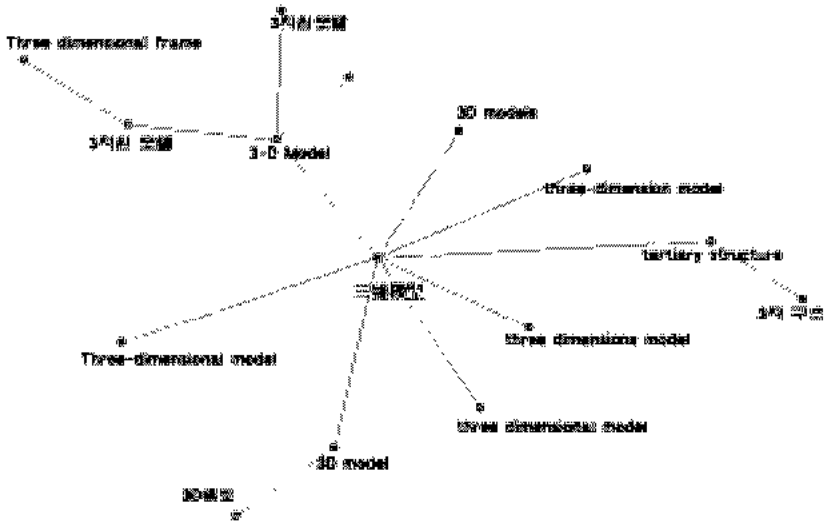
5.1 용어 ID(URI) 그래프와 용어 문자열 그래프

〈그림 12〉는 48개 용어를 모두 표시한 그래프로 각 ID별로 분류코드와 발생빈도 등의 고유한 정보를 담고 있다. 네트워크상의 각 노드마다 후보 분류코드를 추천할 수 있으며, 각 용어별 분류, 전체 네트워크의 분류 측정 등 다양한 수준에서 정보를 추출할 수 있다. 〈그림 13〉은 복잡한 네트워크를 용어의 문자열 값을 기준으로 15개 용어로 단순화 한 것으로, 검색 및 관리 시에 유용한 그래프 형태이다.

5.2 용어 문자열 그래프와 용어별 노드 수 그래프



〈그림 12〉 용어 ID(URI) 기준: complex network 48개 용어

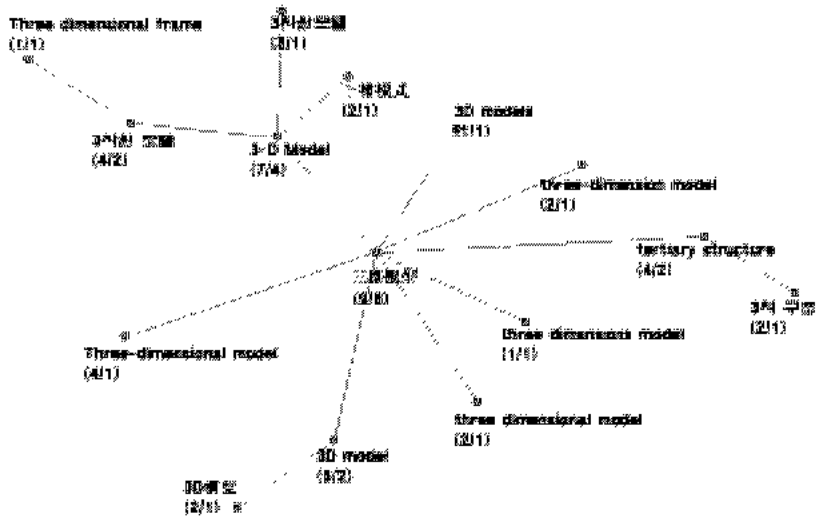


〈그림 13〉 용어 문자열 기준: simple network 15개 용어

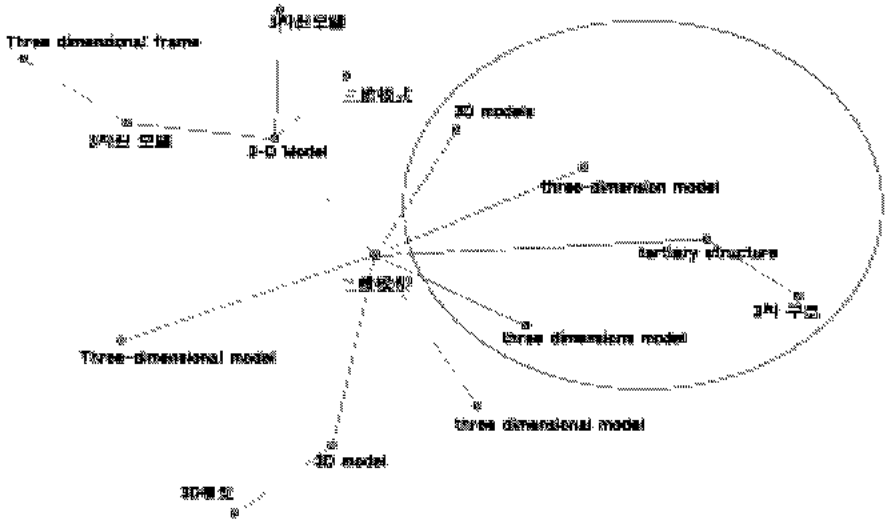
용어 문자열 기준 그래프와 용어에 각 용어에 직접 연결된 노드 수를 나타낸 그래프를 〈그림 14〉와 같이 "용어 문자열 (동일 문자열 개수 / 이웃한 용어 노드 수)" 형태로 표시하였다. 이를 통해 전문용어 정보량 측정과 대역용어 효용성 측정이 가능하다. 즉, 전문용어로서의 정보량이라 함은 데이터베이스 내에서 발생한 빈도를 기준으로 보다 많이 발생한 용어는 네트워크 내에서 전문용어로서 의미가 있다는 것이다. 또한 대역어로서의 효용성 측정이라 함은 실제 대역파일을 많이 가지고 있는 용어를 측정하는 것이다.

따라서 전체 네트워크에서 평균 정보량인 "ID기준 용어 수/문자열 기준 용어 수(48/15) = 3.2"를 기준으로 그 이상인 용어(그림에서 6개 용어가 해당)가 전문용어로서 유용하다고 볼 수 있으며, 평균 대역양인 "전체 노드 수/문자열 기준 용어 수(28/15) = 1.867"을 기준으로 그 이상인 용어(그림에서 5개 용어가 해당)가 대역어로서의 효용성이 높다고 할 수 있다. 그러나 임계치의 측정은 응용시스템을 위한 질의확장 등에 적용될 경우 검색성능 결과에 따라 변할 수 있는 값이므로 절대적인 측정 기준은 아니다. 응용분야에 따라 다양한 지표를 개발하는 것이 용어네트워크 분석을 위해 향후 수행되어야 할 과제이다.

또한, 〈그림 13〉에서 나타나듯이, '3차원 모델'과 '3차 구조'는 동의어군으로 보기에는 무리가 있다. 이는 중국어 대역파일 생성 시 나타난 문제로 결국 대역파일의 품질문제로 인한



〈그림 14〉 용어 문자열 기준, 용어별 노드 수 측정

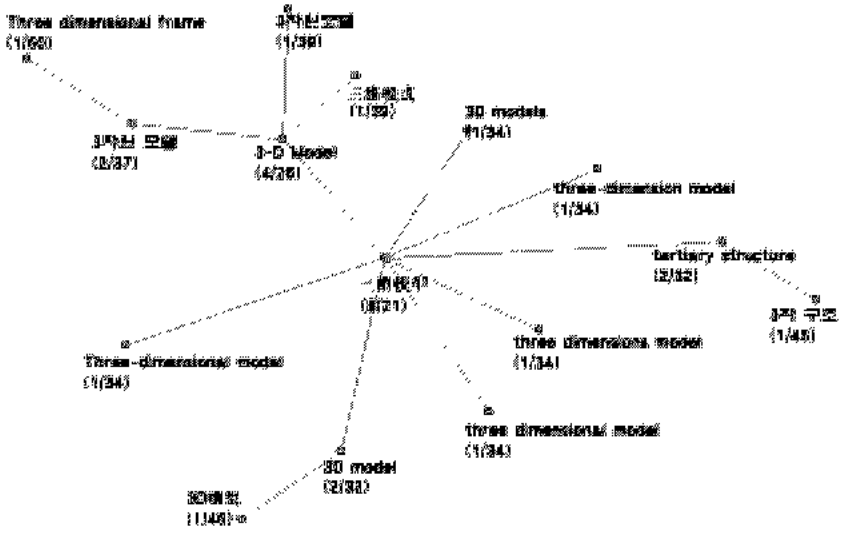


〈그림 15〉 중국어 대역파일 중 발생빈도 2인 용어군

것이다. 〈그림 15〉와 같이 중국어 대역파일의 생성 임계치를 3 이상으로 올릴 경우, 붉은 원 안의 데이터는 전체 용어군에 포함되지 않는다. 대역파일 생성을 위해서는 원천데이터 별로 적합한 임계치의 기준 값이 필요하다. 실제 전문용어 환경에서는 시스템 관리자가 이와 같이 인지적으로 용어 간 관계의 타당유무를 판단하기 힘들기 때문에 일정한 기준에 따른 기계적인 처리가 불가피하다. 그러나 시스템에 대한 분야별 전문가의 검수는 시스템의 성능과 신뢰성을 높이는데 필수적인 사항으로, 의미망의 유효수준을 평가하기 위한 다각적인 실험이 향후 반드시 수행되어야 한다.

5.3 연결정도(degree) 중심성과 근접성(closeness) 중심성 측정

중심성 분석이란, 행위자가 중심에 근접한 정도를 평가하는 작업을 의미한다. 용어 네트워크를 분석함에 있어 각 용어의 중심성을 측정하는 것은 매우 중요하다. 중심성을 측정하는 방법은 매우 다양한데 한 점이 전체 네트워크의 중심에 위치하는 정도(한 점과 네트워크 전체 점들과의 거리)를 측정하는 포인트 중심성 측정이 일반적인 방법이다. 본 연구에서는 한 노드에 직접 연결된 이웃 노드의 수로 중심성을 측정한 연결정도(degree) 중심성 분석과 한 노드와 네트워크의 모든 노드 사이의 거리를 측정하여, 최소 값(최단거리)을 갖는 아이템



〈그림 16〉 용어별 연결정도 중심성, 근접성 중심성 측정

〈표 3〉 연결정도 중심성분석과 근접성 중심성분석에 따른 순위그룹 비교

중심성 순위그룹	연결정도 중심성 (용어 : 값)	근접성 중심성 (용어 : 값)
1	二維模型:8	二維模型:21
2	3-d model:4	3-d model:26
3	tertiary structure:2 3d model:2 3차원 모델:2	tertiary structure:32 3d model:32
4	3d models:1 three dimensions model:1 three-dimensional model:1 three dimensional model:1 three-dimension model:1 3차원모델:1 二維模式:1 3d模型:1 3차 구조:1 three dimensional frame:1	3d models:34 three dimensions model:34 three-dimensional model:34 three dimensional model:34 three-dimension model:34
5		3차원 모델:37
6		3차원모델:39 二維模式:39
7		3d模型:45 3차 구조:45
8		three dimensional frame:50

을 네트워크의 중심 아이템으로 평가하는 근접성(closeness) 중심성 분석의 두 가지 방법으로 네트워크상의 용어 중심성을 측정하고 결과를 비교하였다. 분석결과를 〈그림 16〉의 간략한 네트워크 구조상에 "(연결정도 중심성/근접성 중심성)"의 형태로 표시하였으며, 이를 다시 〈표 3〉과 같이 순위화 하였다.

〈표 3〉에서 보는 바와 같이, 연결정도 중심성의 3그룹에 위치한 '3차원 모델'은 근접성 중심성을 측정한 결과, 5그룹에 나타나게 되어 4그룹과의 순위가 바뀌었다. 연결정도 중심성을 기준으로는 15개 용어를 4그룹으로 구분할 수 있으나, 근접성 중심성을 기준으로는 8개 그룹으로 구분할 수 있어 더욱 세밀한 거리 측정이 가능하다. 또한 근접성 중심성을 이용하면 전체 네트워크상의 아이템 위치를 보다 정확하게 산출할 수 있다.

본 샘플 용어군에서는 중국어가 대표어(또는 우선어)로 뚜렷하게 나타나는데 그 이유는 중국 학술데이터로부터 추출된 '중국어 영어' 대역파일의 이형(異形)이 많이 나타났기 때문이다. 영문데이터를 중심으로 하여 전거데이터

를 확보하기 위해서는 중심성을 '영어'로 바꿀 필요가 있을 수 있다. 단, 본 연구에서는 자동 구축된 데이터를 기반으로 임계치 조절을 통한 용어 클러스터 통제의 방법을 사용하고 있다. 이후 용어군집에 대한 다양한 분석 및 통제기법이 사용되어야 시스템의 신뢰성과 기능성을 모두 높일 수 있을 것이다.

6. 결 론

본 연구에서는 다국어 전문용어 체계를 구축하기 위한 새로운 방법론을 제시하였으며, 향후 다양한 응용분야를 지원하기 위한 시스템 발전방안에 대한 부분까지 전체적인 프레임워크 설계를 하였다. 의미망을 구축하고 해석하는 과정을 통해 전체 의미망으로 연결된 언어자원에 대한 기초분석과 검증이 부분적으로 수행되었다. 이 과정에서 중요한 점은 용어집단에서 채택 가능한 대역용어와 전문용어의 유효수준을 측정하기 위한 임계치를 산출하는 것이었다. 그래프로 용어군을 도식화하는 과정에서 제시한 몇 가지 측정방법 뿐만 아니라 다각적인 연구가 향후 지속적으로 수행되어야 한다.

과학기술분야의 다국어 교차검색을 수행하기 위해서는 대규모 범용 언어자원이 완벽하게 구축되어야 함은 물론 이를 바탕으로 한 다양한 검색기법에 대한 연구 등이 수반되어야 한다. 또한 용어체계의 다국어 대역파일을 계속 주기적으로 보완하여 구축하는 작업과 함께 질의어 번역시의 의미적인 모호성 문제를 효과적으로 해결하는 기법들이 더욱 연구되어야 한다. 향후 의미망 분석, 해석과정 및 추론규칙의 생성에 대한 연구와 방법론 개발이 진행되어야 한다. 본 연구는 이러한 정보검색의 근간이 되는 언어적인 기반자원을 구축하는 단계로 수행된 것으로, 언어자원을 생성하고 이를 관리할 수 있는 체계를 구축하는 데 초점을 두었다. 이러한 언어자원 시스템은 다국어 검색 시스템의 구현과 검색엔진의 기능고도화를 통한 학술정보 서비스의 고도화에 기여하며, 기타 정보시스템과 관련된 업무에 상보적인 기능을 제공하여 업무효율화에 기여할 수 있을 것이다. 또한 동북아 CJK 다국어 환경지원, KISTI 내 모든 정보 분석 및 응용연구 지원, 자동분류를 비롯한 온톨로지 생성연구 등 차세대 정보기술 발전을 지원하는 등의 다양한 효과를 제공할 수 있을 것으로 기대한다.

참고문헌

김지영, 장동현, 맹성현, 이석훈, 서정현, 김현. 2000. 한국어 테스트 컬렉션 HANTEC의 확장 및 보완. 『2000 한국인지과학회 춘계학술대회지』, 210 215.

손동원. 2002. 『사회네트워크 분석』. 서울: 경문사.

정도헌, 김태수. 2003. 시소러스를 기반으로 한 온톨로지 시스템 구현에 관한 연구. 『정보관리학회지』, 20(3): 155 176.

정영미. 1993. 『정보검색론』. 개정판. 서울: 구

미무역(주).

정한민, 강인수, 성원경. 2006. 시소러스와 분야분류체계를 이용한 과학기술문헌에의 주제 및 분야할당. 『제7회 한국언어정보학회 하계학술대회』, 2006년 6월 16 17일. [춘천: 강원대학교].

최석두. 1993. 무전거시스템에 관한 연구. 『한국문헌정보학회지』, 25: 233 264.

최석두, 조혜민. 2001. 다국어 시소러스의 설계. 『한국정보관리학회 제8회 학술대회 논문집』, 5 10.

Antoniou, G. and F. V. Harmelen. 2004. A *Semantic Web Primer*. Cambridge: The MIT Press.

Cattuto, C., V. Loreto and L. Pietronero. 2006. Collaborative Tagging and Semiotic Dynamics. [cited 2006. 9. 10]. 〈http://arxiv.org/abs/cs.CY/0605015〉.

Dejean, H., E. Gaussier and F. Sadat. 2002. "An approach based on multilingual thesauri and model combination for bilingual lexicon extraction". Proceedings of the 19th international conference on Computational linguistics, Volume 1.

Eichmann, D and M.E. Ruiz. 1998. "Cross Language Information Retrieval with the UMLS Metathesaurus". In Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. [cited 2006. 10. 5]. 〈http://citeseer.ist.psu.edu/eichmann98crosslanguage.html〉.

Gey, F.C. and H. Jiang. 1999. "English German Cross Language Retrieval for the GIRT Collection Exploiting a Multilingual Thesaurus". The Eighth Text REtrieval Conference (TREC 8). 219 234. [cited 2006. 10. 20]. 〈http://citeseer.ist.psu.edu/347595.html〉.

Networks/Pajek Program for Large Network Analysis. [cited 2006. 10. 15]. 〈http://vlado.fmf.uni lj.si/pub/networks/pajek/〉.

Salton, G. 1970. "Automatic processing of foreign language document". *Journal of the American Society for Information Science*, 21: 187 194.

Soergel, D. 1996. "SemWeb: Proposal for an Open, Multifunctional, Multilingual System for Intergrated

Access to Knowledge about Concepts and Terminology". *Knowledge Organization and Change*. Proceedings of the Fourth International ISKO Conference 15 18 July 1996, Washington D.C. Edited by Rebecca Green. (Advanced in Knowledge Organization, vol. 5). Frankfurt/Main: Indeks Verlag. 165 173.

Staab, S. and R. Studer. 2004. *Handbook on Ontologies*. New York: Springer.