하이텔 메뉴검색용 시소러스의 개발에 관한 연구[1)]

Thesaurus Development for HiTEL Service

최석두(Suk-Doo Choi) * 김영환(Young-Hwan Kim) **
남영준(Young-Joon Nam) ***

목 차

초 록

정보검색시스템의 성능을 향상시키고 정보검색의 효율성을 높이는 데 활용할 지식베이스로서의 한글시소러스 개발알고리즘을 제시하고, 이 방법에 의한 실체개발사례를 제시한다. 본 연구를 위하여 개발된 시소러스 구축시스템의 기능과 함께 용어의 수집, 분류, 관계의 정의 등의 구축과정 에서 사용한 하이텔 메뉴, 용어사전의 이용방법 등에 대하여 논한다.

ABSTRACT

We present development results for a Hangul thesaurus which was provided to improve performance of the intelligent information retrieval system. The important stages and methods in the process of term acquisition, classification, creation of the consistency-effectiveness relationship using HiTEL menu and text of dictionary are described. To carry out our study we have built a thesaurus management system and also describe its utility functions.

1)본 연구는 한국전기통신공사 1995년도 연구지원비에 의해 수행되었음.
* 이화여자대학교 문헌정보학과
** 한국통신 멀티미디어연구소
*** 전주대학교 문헌정보학과
■ 논문 접수일 : 1996년 6월 24일

1. 서 론

현재 우리나라에서는 각 분야별로 생성되는 방대한 양의 정보자료를 수집, 가공, 조직하여 데이터베이스화하고, 첨단 과학기술정보뿐만 아니라 일상생활에 필요한 정보까지 검색하여 적시에 활용할 수 있는 정보서비스시스템들이 있다. 그 중 하이텔(HiTEL)은 대단위 종합통신망서비스로서 많은 가입자들을 보유하고 있으며, 생활정보를 비롯하여 전문정보, 금융관련정보, 기업정보, 취미정보, 오락 및 동호회 활동정보 등을 광범위하게 제공하고 있다.

하이텔과 같은 대규모 통신망서비스시스템에서 서비스에 개설되는 정보의 양과 종류는 시간이 흐를수록 더욱 다양하고 폭넓은 분야로 확대되고 있다. 하이텔은 적절한 정보제공을 위해 정보검색용 메뉴체계시스템을 채택하고 있다. 왜냐하면 서비스종류의 증가와 함께 정보검색메뉴의 종류와 계층의 깊이도 점점 증가하고 있기 때문에 이용자에게 많은 불편을 초래하고 있다. 따라서 필요한 정보의 탐색을 위하여 사용자들에게 탐색기법을 교육시키거나 혹은 정보검색시스템을 지능형으로 개발하여 쉽게 이용할 수 있는 방법을 강구하여야 한다. 이 가운데 사용자의 수는 점차 증가하고 사용연령도 넓어지는 경향이므로 모든 사용자들에 대한 탐색기법의 교육보다는 지능형 검색시스템이 훨씬 효율적인 방법이 될 것이다(이두영 등, 1995).

지능형 정보검색시스템이 개발되기 위해서는 색인시스템, 하부저장알고리즘, 인터페이스, 유사어사전, 시소러스가 반드시 함께 개발되어야 한다. 특히, 시소러스는 "지능형검색시스템"에서 지능에 해당하는 지식베이스이며, 정보검색에 있어서 시스템과 검색자간의 교량적 역할을 하는 정보검색도구이다. 또한 정보자료를 데이터베이스에 입력할 때의 색인어 선정과 정보자료를 검색할 때의 탐색어 선정을 지원해 줌으로써 정보검색시스템의 효율을 높이는 데 중요한 역할을 한다.

본 연구에서는 한국통신에서 개발하고 있는 지능형 정보검색시스템의 성능을 향상시키고 정보검색의 효율성을 높이는 데 활용할 지식베이스로서의 하이텔안내시스템용 시소러스 개발알고리즘과 개발에 이용한 시소러스구축 시스템에 대하여 논한다.

2. 시소러스의 위치

시소러스는 기능적인 측면과 구조적인 측면의 두 가지 측면에서 정의할 수 있다. 시소러스란 기능적인 측면에서는 문헌이나 색인자, 이용자가 사용하는 자연언어를 보다 통제된 시스템언어(도큐멘테이션언어, 정보언어)로 변환하기 위한 어휘통제장치이며, 구조적인 측면에서는 특정 분야의 의미론적, 속구조적 관련용어에 대한 통제되고 동적인 어휘집이다 (Kent et al., 1980 : v.30 ; 416). 전자는 시소러스 이용의 측면이며 후자는 시소러스 개발의 측면이다. 통제어휘개념을 포함하는 시스템에서는 동일한 시소러스를 사용하는 시스템이라 할지라도 이용방법과 이용수준에서 상당한 차이를 보이지만 기본적인 기능은 동일하다.

하이텔안내시스템과 같은 메뉴체계에서도 시소러스의 기본적인 기능은 동일하다. 메뉴체계에서 색인의 대상은 각 단위메뉴가 되며, 모든 단위메뉴에 대하여 색인어를 부여하게 된다. 이것은 수작업색인일 수도 있고 자동색인일 수도 있으나 국제표준에 준거하여 만든 시소러스, 분류표 등의 색인언어에 있는 개념은 일반적인 색인지침에 따라 우선어로 번역하고, 새로운 개념의 용어는 해당 분야의 권위 있는 사전이나 백과사전을 이용하여 정확성과 수용가능성을 검토하게 된다(정영미, 최석두, 서혜란, 1996). 시소러스를 사용하는 경우에는 일반적으로 시소러스의 개념계층이 세분될수록 점점 특정 개념(좁은 개념)을 나타내게 되므로 메뉴의 수준이 깊어질수록 색인어는 시소러스의 세분된 개념을 사용하게 된다.

정보시스템 이용자가 정보검색을 위해 기계에 입력하는 검색문에 형태는 크게 다음 두 가지로 구분할 수 있으나, 검색어로는 은행과 이체라는 검색어가 사용되며 해당용어는 일련의 색인화일을 이용하여 필요한 정보를 검색한다.

i) 일정한 형식을 갖춘 검색식 (질의식)
예 : 은행 and 이체

ii) 일반적인 자연어 형태의 검색식 (질의문)
예 : 은행에서 은행으로 이체를 하고 싶다

이용자가 입력한 질의식(문)이 색인화일내에 매칭되는 용어(혹은 개념)가 있을 경우, 시스템은 이에 해당하는 관련메뉴를 알려주며 사용자는 이 중 특정 메뉴로 이동할 수 있다. 필요에 따라서는 관련 메뉴를 질의문에 적합한 정도에 따라 순위를 매겨 나열할 수도 있다. 그러나 질의문에 사용한 용어가 색인화일에 없을 때는 전거화일 혹은/및 시소러스를 점검하여야 한다. 전거데이터는 고유명사를 중심으로 구축되며, 시소러스는 일반용어를 중심으로 구축되는 것이 보통이나 전거데이터를 시소러스에 포함할 수도 있다.

예컨대, 상기 질의식 "무역회사 and 정보"를 이용자가 검색하고자 할 때 시소러스의 내용과 역할은 다음과 같다. "메뉴수준 n"에서 사용자가 "8. 기업/무역/농수산"을 선택한 "메뉴수준 n+1"의 결과는 〈그림 1〉과 같다.

1. HiTEL	2. HiTEL-POP / KT-MAiL	3. 전화번호안내
4. FAX 정보	5. 언론/인물	6. 금융/은행/카드
7. 증권/보험	***8. 기업/무역/농수산***	9. 교육/도서
10. 과학/기술	11. 법률/세무/부동산	12. 예약/주문/홈뱅킹
13. 생활/취업/문화	14. 관광/여행/교통	15. 취미/스포츠
16. PC통신서비스	17. 지역정보, INDITEL	18. 통신영업서비스
19. 해외 정보	20. 공공정보I (생활)	21. 공공정보Ⅱ (전문)
99. 공지사항		

(a) 메뉴수준 n

34301. KTMI 기업정보	30302. 중소기업정보은행
- 전체정보	- 전체정보
34303. 센서정보	35001. KOTRA-NET 서비스
- 센서/제어계측기기 상품정보	- 종합무역정보통신서비스
35002. 운송종합정보	37001. 농림수산정보센터
- 해운, 무역, 항만, 물류종합정보	- 농림수산종합정보서비스
64011. 농협일렉트로뱅크	75001. K - VAN
- 유통정보/농업기술/성공사례	- 컴퓨터 유통정보

(b) 메뉴수준 n+1

〈그림 1〉 하이텔 메뉴의 예

메뉴처리를 위하여 개발된 본 시소러스에서 "8. 기업/무역/농수산" 중의 용어인 "무역"에 설정된 용어관계를 보면 다음과 같다.

무역 0300
UF 대외무역
외국무역
NT 3국간무역
가공무역
간접교역
간접무역
간접수입
간접수출
역수입
역수출
연계무역
연불수출
우회수출
원자재수입
위장수입
위장수출
자유무역
지하자원수입
직접교역
특혜수입
플랜트수출
협동무역
호혜주의
RT 국제경제
국제무역위원회
국제수지
국제시장
무역상사 ——→ UF 무역회사
무역협상위원회 BT 기업
불공정무역국
자유무역지역
최혜국대우
한일통상회담

우선, 색인자가 "메뉴수준 n"의 "8. 기업/무역/농수산"을 "무역"이라는 디스크립터로 색인했다면 이용자가 "무역"의 UF 및 NT의 어느 것으로 검색하더라도 "8. 기업/무역/농수산"을 바로 제공할 수 있을 것이다.

또한 색인자가 "메뉴수준 n+1"의 "KTMI 기업정보"에 대하여 "기업정보, 기업, 정보"로 색인했다고 하자. 이용자가 "무역회사 and 정보"를 찾았을 때, 시소러스에서의 "무역회사 USE 무역상사"라는 관계와 "무역상사 BT 기업"에 의해 자동으로 기업정보가 있는 "메뉴수준 n+1"의 "34301. KTMI 기업정보"도 탐색결과에 포함되게 된다. 색인과 검색은 이와 같이 색인어와 시소러스구조를 이용하며, 고유명사의 異形表現이 있을 때에는 전거데이터를 이용하여 다양한 검색기능을 갖출 수 있게 된다.

3. 시소러스 구축시스템

3.1 설계원칙

시소러스의 구조와 내용은 시간의 흐름에 따라 변화되는 것이 보통이다. 이것은 개념간의 관계가 새롭게 정의되거나 새로운 정보의 필요성이 생길 수 있기 때문이다. 관계지시기호를 확장, 다국어시소러스로의 변환, 용어사전의 참조, 언어정보 및 디스크립터정보(분류기호, 등록 및 갱신정보, 데이터베이스정보 등)를 포함하는 새로운 정보의 추가, 표시방법의 변경 등을 들 수 있다. 따라서 시소러스 구축시스템은 용어의 추가, 삭제, 수정기능뿐만 아니라 이와 같은 신규정보의 추가에도 융통성있게 대처할 수 있는 기능이 필수적이며, 충분한 기능을 갖는 구축시스템 없이는 좋은 시소러스의 개발은 불가능하다. 이 시소러스 구축시스템에 색인 및 검색 관련기능을 추가하면 시소러스 관리시스템이 될 수 있다.

본 시스템에서는 기본적으로 ISO 2788-1986(E), ISO 5964-1985(E), ANSI / NISO Z39.19-1991(최석두, 정동열, 1994)을 기준으로 시소러스를 작성할 수 있는 시소러스 구축시스템을 개발한다. 한글의 특성상 한글기입어 부분(한정어 포함)으로 모든 디스크립터를 식별할 수 있도록 하는 것을 원칙으로 하며, 다른 요소(예를 들면, 분류기호, 한자, 영문 등)는 참조데이터로 처리한다. 다만 필요시에는 다른 요소를 식별요소로 사용할 수 있도록 한다.

또한 시소러스와 고유명사의 전거데이터는 동일한 구조를 가지고 있으며 서로가 관련이 깊어 양자를 구별할 수 없는 경우가 많다. 예를 들면, "과학자"의 사례관계에는 인명이, "금융기관"의 사례관계에는 은행명이 올 수 있다. 따라서 시소러스와 통합이 가능하므로 고유명사에 대한 전거데이터의 작성용으로도 사용할 수 있도록 한다.

3.2 기능

전술한 설계원칙에 따라 추가, 삭제, 수정, 브라우징과 같은 기본적인 기능 이외에 다음과 같은 기능을 가지고 있다. 이 시소러스의 구축시스템은 특정 데이터베이스에 독립적인 시스템이며 종속적인 시스템으로의 변환시에

는 색인 검색기능과 통합되어야 할 것이다. 차후에는 색인 검색시스템과의 연계를 고려하여 멀티시소러스의 참조기능, GUI개념을 도입한 디스플레이형식, 타 용어의 동시참조, 계층이 방대한 용어의 효율적인 디스플레이, 마이크로시소러스의 생성기능 등 보다 고차적인 기능의 보완이 필요하게 될 것이다.

우선, 한글기입어 및 대응되는 영어기입어에 의한 조정기능을 갖는다. 각각 별도로 정렬하여 조정할 수 있다. 또한 다음 예와 같이 관계의 충돌이 일어나는 경우에는 입력 즉시 경고메시지를 내어 주의를 환기시킨다.

A	A
USE A'	BT A'
BT B	NT B
NT C	RT A'

둘째, 관련 용어가 없는 "고립어"도 입력이 가능하다. 순수한 고립어인 경우도 있으나 추후에 관계를 설정하기 위하여 입력해 두어야 하는 경우도 있다. 또한 용어의 절단검색이 가능하다. 용어의 입력시 "필름스트립, 스트립필름", "통제경제, 경제통제"와 같이 어느 용어로 사용했는지가 모호한 경우가 많다. 이때, 좌측절단검색이 유용하다. 또한 참조하고 있는 용어의 의미를 알고 싶을 때는 링크되어 있는 용어사전의 설명문을 참조할 수 있으며, 설명문내에는 이미지데이터가 포함될 수 있다.

셋째, 자신의 모든 BT와 모든 NT를 깊이에 상관없이 모두 자모순으로 배열하는 포괄적 계층구조방식(EJC, 1993 ; 최석두, 1994)과 자신의 상하 1수준만을 보여주는 방식을 모두 혼용한다. 전자의 방식은 분야별시소러스인 경우에는 좋으나 매크로시소러스인 경우에는 관계용어가 많아 적합하지 않기 때문이다. 전자의 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. 반대로 후자의 방식은 자신의 상하계층 전체를 한꺼번에 참조할 수 없다는 단점이 있다. 관련 용어가 많아지면 그래픽표시도 무리가 있으므로 차후 CD-Plus사의 OVID(MEDLINE)가 사용하고 있는 접기기법(fold-in method)을 이용하여 두 기능을 통합할 계획이다.

넷째, 다국어시소러스의 개발을 지원하기 위하여 별도의 언어별 관계지시기호를 설정할 수 있다. 색인자는 1개국어로 색인하나 탐색자는 1개국어의 용어를 입력하여 같은 주제의 다국어자료를 검색할 수 있도록 하기 위함이다. 화면에서의 선택란은 하나를 두고 언어기호를 입력하게 한다. 입력결과의 예를 보이면 다음과 같다.

예 : 무조음악 〔無調音樂〕 atonal music ——영어
GER Atonale Musik ——독어
FRA musique atonale ——불어
노오 〔能〕
JAP のう ——일본어

다섯째, 네트워크를 통하여 시소러스를 공동으로 갱신할 수 있다. 시소러스를 한 사람이 만든다는 것은 거의 불가능하며 여러 분야의 전문가가 하나의 데이터베이스를 이용하여 공동으로 갱신할 수 있어야 한다. 한편, 다수의 개발자가 개별적으로 시소러스를 구축하여 이를 통합할 경우에 발생할 수 있는 문제도 해결할 수 있다. 여섯째, 패싯인디케이터(facet indicator)를 사용할 수 있으며, 각 용어가 자신의 갱신일자(갱신되지 않은 것은 등록일자)와 갱신자명을 기록할 수 있다. 이는 개발이 공동의 기억공간에서 이루어짐으로써 발생할 수 있는 용어의 조정문제를 해결할 수 있다. 예를 들면, 개발자 A가 "카패시터"(capacitor)를 "축전기"로 설정하고 개발자 B는 이를 다시 "카패시터"로 조정할 수 있다. 각 디스크립터에서 이러한 변경내용을 시간정보와 함께 보여주므로 개발자 B는 다시 "카패시터"로 바꾸기 전에 개발자 A와 상의할 수 있다.

일곱째, 참조하던 용어의 관련용어를 복사하여 다른 용어 아래에서 사용할 수 있다. 예를 들면, 용어 "011"의 관계용어와 동일한 새로운 용어 "017"이 추가되었을 때, "011"의 관계용어를 그대로 복사하여 "017"의 관계를 설정할 수 있다. 에러를 줄일 수 있으며 처리가 빨라진다.

예 : 011

RT 디지털통신
셀룰러통신
한국이동통신
휴대용무선전화기

⟶

017

RT 디지털통신
셀룰러통신
한국이동통신
휴대용무선전화기

여덟째, 분류번호를 입력할 수 있도록 하여, 자모순 시소러스, 계층시소러스 및 분류번호순 시소러스를 생성하는 기능을 갖는다. 계층시소러스의 예를 들면 다음과 같다. 계층관계가 있는 모든 용어를 최상위개념어 아래 배열하였으며, 마침표(.)는 하나의 계층을 의미한다.

예 : 유럽 900
· 동유럽
·· 루마니아
·· 슬로바키아
·· 체코
·· 폴란드
·· 헝가리
· 북유럽
· 서유럽

마지막으로, 해당 시소러스의 내용을 분석할 수 있다. 예를 들면, 디스크립터의 수, 비디스크립터의 수, NT의 최대계층수, 한정어 부기용어의 수 등에 대한 통계정보를 알 수 있다.

4. 시소러스의 구축

4.1 대상분야의 선정 및 분류

시소러스의 개발시 대상 주제분야를 선정하는 일은 매우 중요하며 어려운 일이다. 왜냐하면 일반적으로 특정분야나 특정시스템의 처리범주가 정해지지만 언어상의 개념이란 수치

적으로 확연히 구별되는 것이 아니기 때문이다. 본 연구에서는 하이텔 초기화면에 나타나는 22개의 메뉴(그림 1 참조)를 참조하여 본 시소러스의 대상분야를 선정하였다. 〈그림 1〉의 초기메뉴에 나타나는 용어들을 그룹화하여 대표개념을 결정하고 그 대표개념을 최상위개념어(top term)로 간주하여 기존 메뉴에 속해 있는 용어들의 분포도를 조사하였다. 분포조사에 대한 결과, 특정 메뉴에서는 이용도와 용어의 수가 다른 메뉴에 비해 월등히 많았음을 알 수 있었다. 이 결과로서 최상위메뉴의 평준화를 위해 본 시소러스는 다음과 같이 10개 메뉴로 새로운 하이텔 메뉴를 설정하였으며, 이것을 類分類로 사용하였다.

0000 정보산업⇒홈쇼핑, 홈뱅킹, 예약, PC통신, 정보통신, 뉴미디어, 신문(인물 동정)

0100 교육⇒교육기관, 단체, 교육정책, 교육문제, 교재, 교육자, 교육지도, 초등교육, 학습, 진로, 어학, 훈련, 교양, 논문, 유학, 국가시험

0200 정치⇒행정부, 사법부, 입법부, 지방정부, 지역정치, 지역정보

0300 경제⇒재정, 금융(증권), 기업, 무역, 노동, 노무(노사문제), 투자, 경영, 세무, 취업

0400 산업⇒농수산(원예), 임축산, 에너지, 운수, 교통(자동차, 수송), 토목, 건설, 서비스업, 상업(쇼핑), 특허, 광고(홍보)

0500 사회⇒사회복지, 환경(공해), 의료, 건강, 생활, 가정, 세대, 어린이, 청소년, 취미, 오락, 방송(연예), 여성, 요리, 패션, 출산

0600 문화⇒문화재, 종교, 미술, 음악, 연극(영화), 무용, 문헌, 도서관, 박물관, 학술도서, 출판, 잡지, 동화, 만화

0700 과학⇒인문과학, 사회과학, 자연과학, 기술과학(원자력공학, 첨단공학, 식품, 생명공학 등)

0800 레저⇒스포츠(아마추어/프로), 관광, 여행, 숙박(콘도, 호텔), 바캉스, 레크레이션

0900 해외⇒아시아, 일본, 미주, 미국, 유럽, 해외동포, 북한, 영문정보

이상과 같이 설정한 10가지의 주제범주를 기준으로 하이텔메뉴에서 분석한 용어들의 출현빈도를 조사한 결과는 〈표 1〉과 같다. 〈표 1〉에서 A는 하이텔메뉴의 류클래스와 강클래스를, B는 InfoSHOP의 류클래스와 강클래스를 분석한 결과를 나타낸다.

〈표 1〉 용어출현빈도

분류코드	분야	A(%)	B(%)
0000	정보산업	26.8	31.2
0100	교육	5.3	4.9
0200	정치	5.7	6.8
0300	경제	13.8	13.4
0400	산업	8.8	15.5
0500	사회	18.1	12.8
0600	문화	10.8	10.0
0700	과학	1.6	1.0
0800	레저	6.8	3.3
0900	해외	1.9	1.0

〈표 1〉에서 알 수 있는 바와 같이 0200 정치분야와 0700 과학분야의 용어 출현빈도가 상대적으로 다른 분야에 비해 떨어지는 것을 확인할 수 있었다. 이는 하이텔의 특성상 전문성이 강한 분야의 정보보다는 일상정보가 압도적으로 많기 때문이라고 판단된다. 이 두 분야는 앞으로 접속할 정보가 많아질 것으로 판단되어 하이텔내에 출현한 용어와는 별도로 과학일반과 정치 사회에 속하는 중 고등학교 교과서를 분석하여 용어를 추가하였다.

0100 교육분야와 0900 해외분야도 분포도가 떨어지나, 0100 교육분야는 현재와 같은 입시형태와 각 대학에서의 정보를 계속하여 업로드시키는 추세라면 메뉴의 수는 급속도로 증가할 것이다. 0900 해외분야는 본 시소러스가 한글 위주의 시소러스이기 때문에 영문으로 된 메뉴와 데이터들이 제외되었다. 따라서 용어분포가 상대적으로 떨어지게 되었다. 또한, 북한 관련 정보를 0900 해외분야에 배당한 것은 향후 증가하게 될 북한과의 교류에 따른 정보증가를 예측한 것이다.

4.2 시소러스의 구축과정

4.2.1 구축 방법

시소러스 구축방법에는 연역적 방법과 귀납적 방법이 있다(ISO 2788-1986(E)). 연역적 방법은 먼저 시소러스 작성 대상 주제분야의 주요 카테고리를 설정한 후에 대상 용어 전체를 수집하는 방법으로, 이 방식은 필요할 것으로 예측되는 모든 카테고리와 계층을 모두 유추해야 하는 어려움이 있다. 한편 귀납적 방법은 대상 주제분야의 용어를 다양한 정보원으로부터 수집하는 시점에서 용어간의 개념 및 관계를 정의하고 시소러스에 추가해 가는 방법으로, 이 방식은 연역적 방법보다 신뢰성이 높은 것으로 평가되고 있다. 그러나 본 연구에서는 이 두 가지 방법을 혼합적으로 사용하였다. 즉, 연역적 방법으로 수집하여 클러스터링된 용어를 귀납적 방법으로 관계를 설정하였다.

4.2.2 용어의 수집

우선, 하이텔의 메뉴의 모든 용어를 단일개념으로 분리하여 처리대상 용어로 삼는다. 다만, 전술한 바와 같이 0200 정치분야와 0700 과학분야의 용어가 다른 분야에 비해 상대적으로 적으나 향후 하이텔에 접속할 정보가 많다고 판단하여, 이 두 분야에 대한 용어를 하이텔내에 출현한 용어와는 별도로 수집하였다. 이를 위하여 과학일반과 정치 사회분야의 중고등학교 교과서의 용어를 수집하여 추가하였다.

메뉴용어의 예로, 〈그림 1〉의 "메뉴수준 n"의 용어를 단일개념으로 분리한 결과는 다음과 같다. 이 분리된 용어를 전술한 10가지 주제범주의 하부범주로 분류하며, 두 가지 이상의 범주에 속하는 경우도 인정한다.

FAX
HiTEL
HiTEL-POP
KT-MAiL
과학

교육
금융
기술
기업
농수산
도서
무역
법률
보험
부동산
세무
안내
언론
예약
은행
인물
전화번호
정보
주문
증권
카드
홈뱅킹

4.2.3 관련용어의 추출 및 관계설정

제1단계에서 추출된 용어를 해당 주제전문분야의 용어사전을 참조하여 관련개념어를 추출한다. 용어사전은 해당 분야의 전문가들이 만든 용어해설이므로 이 해설에서 관련용어를 추출한다는 것은 연역적 시소러스작성법과 유사하다. 또한 일반적으로 용어사전은 정보이용자가 접근할 수 있는 정도의 전조합수준이므로 용어의 분할에 대하여 별도로 고려하지 않아도 무리가 없다는 장점이 있다(金泰中, 1989). 만약 해당 용어가 사전의 표제어로 등록되어 있지 않을 경우, 광의의 개념으로 용어를 통제하여 표제어를 찾는다.

예컨대, 해당용어가 "당좌예금"이라 하자. 용어사전의 설명문과 그 설명문에서 추출된 관련용어는 다음과 같다.

당좌예금 : 〈current deposit〉 〔경〕 은행 예금의 하나. 은행이 예금자의 요구에 따라서 어느 때나 예금액을 지불한다는 약속하에 지급하는 예금. 지급요구를 할 때는 예금자가 수표를 발행함. 곧, 수표지급의 자금이 되는 예금. (준)당좌

당좌예금 - 은행예금, 은행, 예금자, 예금액, 지급, 약속, 지급요구, 수표, 수표지급, 자금, 예금, UF 당좌, BT 경제

이 중 "UF 당좌, BT 경제"는 설명문에서 "〔경〕"과 "(준)당좌"로 명백하게 상위개념어와 동의어라고 명시하고 있으므로 이 단계에서 관계를 설정할 수 있다. 명백하게 동의어나 상 하위개념어로 정의되는 용어를 제외하고 추출된 모든 용어(은행예금 등)는 표제어(당좌예금)와의 초기관계를 연관관계(RT)로 설정한다. 이 단계에서 디스크립터로 적합치 않다고 생각되는 용어(예금자 등)는 삭제한다.

당좌예금
UF 당좌
BT 경제
RT 금융

수표
예금
은행
은행예금
자금

4.2.4 타 용어와의 관계조정

1) 타 용어관계에 의한 관계조정

계층관계는 기존에 구축된 시소러스, 용어사전 및 용어그룹의 구성 등을 참조하여 조정한다. 상기 예를 보면, "금융"에서 "금리"를, "예금"에서 "저축"을, "자금"에서 "자본"을 관련관계로 만들어 완성된 관계는 다음과 같다. 명확하게 계층관계가 설정되지 않는 것은 관련관계로 설정하며, 관련어만 있는 경우도 인정한다. 이 관계들은 새로운 용어가 추가되면서 다시 바뀌게 될 것이다.

당좌예금
UF 당좌
BT 경제
금융
예금
은행예금
RT 금리
수표
은행
자금
자본
저축

2) 각 관계에 대한 대응관계의 조정

각 용어관계는 〈표 2〉와 같은 대응관계를 갖게 된다. 〈표 2〉에서 보는 바와 같이 NT, BT에 대하여 보다 세분된 관계지시기호를 사용할 수 있다. ISO 2788 - 1986(E)에서 권장하고 있는 NT, BT의 세분된 관계지시기호는 BT(상위개념어), BTI(상위 사례관계), BTG(상위 屬관계), BTP(상위 전체/부분관계), NT(하위개념어), NTI(하위 사례관계), NTG(하위 屬관계), NTP(하위 전체/부분관계)이다. 시스템에 따라서 더 세분할 수도 있을 것이다.

〈표 2〉 관계지시기호의 대응관계

A BT B	→ B NT A	및 B NT A	→ A BT B
A BTI B	→ B NTI A	및 B NTI A	→ A BTI B
A BTG B	→ B NTG A	및 B NTG A	→ A BTG B
A BTP B	→ B NTP A	및 B NTP A	→ A BTP B
A RT B	→ B RT A	및 B RT A	→ A RT B
A USE B	→ B UF A	및 B UF A	→ A USE B

또한 패싯인디케이터를 사용하는 경우에도 동일한 대응관계를 갖게 된다. 패싯인디케이터가 없는 그룹이 있어도 좋으며 아예 사용하지 않아도 좋다. 사용례를 보면 다음과 같다.

예 : 자동차
(사용연료)
NT 가솔린자동차 → 가솔린자동차
디젤자동차 BT자동차 (사용연료)
알콜자동차
전기자동차
……

(용도)

NT 경주용자동차 → 경주용자동차

승용차 BT 자동차 (용도)

트럭

.....

패싯인디케이터를 사용하여 그룹화하고, 다시 사례관계, 속관계, 전체/부분관계의 각 그룹내에서 자모순으로 배열하게 되면 배열이 복잡해져 혼란을 가중시키게 된다. 그러므로 표시의 방법은 다음과 같이 동일한 계층 내에서는 패싯인디케이터별로 전체를 자모순으로 배열하고, 사례, 속, 전체/부분관계는 해당 기호(i, g, p)만을 해당하는 용어의 앞부분에 부여한다.

은행

NT 간사은행

i 국민은행

국책은행

보통은행

이와 같은 상호관계는 시소러스구축시스템에서 자동처리되므로 완벽하게 대응관계가 정립되었다고 간주할 수 있다. 심지어는 A UF B인 관계가 정의된 후, 다른 용어 아래에서 B를 입력하면 시스템은 A로 대체한다.

다만 현재의 시스템에서 처리의 효율화를 위하여 각 용어가 서로 하나의 포인터를 갖고 있어서 상 하위개념어 전체를 보여주는 기능에서 전파갱신(propagation updating)을 하지 않고 있으므로 삭제시에는 주의해야 한다. 예를 들면 다음과 같다(진한 글자체 부분은 키보드입력을 나타낸다).

가가	나가	가가	나가	가가
NT **나가**	BT 가가	NT 나가	BT 가가	NT 나가
다가	NT **나나**	나나	NT 나다	나나
나다	나다			나다
		다가		다가
(a)	(b)	(c)	(d)	(e)

〈그림 2〉 용어의 입력 및 수정결과

우선 "가가"의 용어관계를 입력한 것이 (a)이다. 하위개념어인 "나가"의 하위개념어를 입력한 후(b), "가가"를 디스플레이하면 (c)와 같은 상태가 된다. 레벨에 관계없이 모든 하위개념어를 NT로 갖는다. (b)에서 "나가"의 하위개념어 "나나"를 삭제하고(d), "가가"를 디스플레이하면(e) 하위개념어로 "나나"가 그대로 남아 있다. "가가"에서도 "나나"를 삭제하거나 "나나"를 디스플레이하여 관계용어를 일괄적으로 삭제해야 한다.

4.2.5 다의어 및 동의어의 발견과 처리

1) 다의어와 동의어의 발견

한글 기입어 부분을 정렬함으로써 다의어를 발견할 수 있다. 이때 영문을 한정어로 사용하는 경우와 한글을 한정어로 사용하는 경우가 있다. 다의어를 발견했을 경우에도 어느 방식을 사용할지는 시소러스개발의 정책에 따라 달라지지만 본 시스템에서는 영문과 한글을 다 사용하나 한글이 우선한다. 즉, 한글로 구분되지 않는 것을 영문으로 구분한다.

영어로는 한 가지 형태의 용어이나 한글기입어로는 여러 가지 용어로 사용되는 경우가 많다. 다음과 같이 영문대응어를 부기하고, 영문한정어를 중심으로 배열함으로써 동의어를 발견할 수 있다(남영준, 1996).

capacitors	카패시터
capacitors	축전기
window 95	윈도우 95
window 95	윈도 95

2) 다의어와 동의어의 처리

기본적인 동의어류는 본 연구에서 별도로 개발한 동의어파일의 내용을 시소러스에 포함시키고 있으며, 구축과정 중 시소러스 내에서 동의어를 발견했을 때는 일반적으로 알려진 규칙에 따라 어느 한쪽을 따르고 다른 한쪽을 UF로 처리함으로써 끝난다.

다의어는 시소러스구축, 색인, 검색과정에서 많은 문제를 야기시킨다. 분야가 넓어지거나 개념의 깊이가 깊어지면 명확한 다의어 뿐만 아니라 개념간의 뉘앙스가 미묘하게 되므로 처리상 문제가 된다. 예를 들어, 다음의 "밀도"는 원래 다의어는 아니나 관점에 따라 다르게 사용된 예이다.

밀도(비율/면적) 〔密度〕 density(rate / area)
밀도(수/크기) 〔密度〕 density(number / volume)
밀도(질량/체적) 〔密度〕 density(mass / volume)

우리말은 영어와 달리 대부분의 글자가 독립적인 개념을 가지고 있다. 예컨대, "우리말은 영어와 달리 대부분의 글자가 독립적인 개념을 가지고 있다"라는 앞 문장에서 1문자가 독립적인 개념을 갖지 않는 것은 "와, 립, 있" 정도일 것이며, 한자어나 방언까지 포함시킨다면 더 적어질 것이다. 특히, 한글용어는 일반적으로 용어열이 짧을수록 다의성이 늘어가는 특징을 가지고 있다. 예를 들면, "말"은 다음과 같은 다의성을 가지고 있다.

말(도량형)	〔斗〕	
말(동물)	〔馬〕	horse
말(식물)	〔藻〕	pondweed
말(언어)	〔言語〕	language
말(장기)		chessman

"말"을 그대로 색인했을 때, 이와 같은 여러 가지 개념이 혼합되어 검색되므로 데이터의 규모가 커질수록 잡음률은 높아지게 될 것이다. 그러나 자연언어의 자동색인에서 이와 같은 개념을 문맥만으로 식별한다는 것은 매우 어려운 일이다. 어렵지만 색인시에는 이들을 구별하여 색인하여야 하며, 시소러스 없이는 이들의 구별이 거의 불가능하게 된다. 따라서 시소러스에 등록된 용어가 동물의 "말" 하나밖에 없다할지라도 이용자가 혼동할 수 있으므로 "말(동물)"과 같이 한정어를 붙여야 한다. 한정어를 포함하는 한글부분이 각 기입어를 유일하게 식별하므로 다른 용어 아래에서 그 용어를 참조할 때 한글 이외의 부분은 입력하지 않아도 자동으로 갖고 올 수 있다.

동음이의어를 구별하는 방법은 한글, 한자음(예 : 학교-學校), 한자의미(예 : 귀-耳), 영문 등을 한정어로 사용할 수 있다. 그러나 한

글시소러스에서는 한글부분이 모든 용어의 유일한 식별어가 되어야 한다. 따라서 본 연구에서는 기입어의 내용 중 한글부분을 "unique string"으로 삼아 이 부분만으로 모든 기입어를 식별하는 것을 원칙으로 한다. 다만, 상기 예 중에서 "말(장기) - chessman"의 한정어 "장기"는 "長技, 將棋, 臟器, …" 등과 같이 다의성이 있으므로 "놀이, 바둑, 체스, …" 중에서 어느 것을 사용할지는 별개의 문제이나 가능한 한 학술용어보다는 그 의미를 직관적으로 이해할 수 있는 일반용어(예를 들면, "화훼"보다는 "꽃"을, "재봉"보다는 "바느질" 등)를 사용한다.

시소러스에서 다의어를 명확하게 식별하고 이를 색인과정에서 사용한다면 탐색자가 "말"을 탐색하는 경우에 이와 같은 다의어군을 디스플레이하여 정확한 주제를 선택하게 할 수 있을 것이며, 결과에 대하여 보다 정확하게 순위를 매길 수 있을 것이다. 이 방법은 디스크립터의 배열시 "한국사"가 "한국(韓國)" 보다 먼저 배열되는 것도 막을 수 있다. 다만, 처리과정 중에서는 한자나 영문으로 구별되는 경우도 있다. 본 연구의 범위 외이지만 시소러스 디스크립터의 표기문제로 남아 있는 것은 우리말의 옛말과 북한어로 인해 생기는 다의어의 구별이다. 이 문제도 상기 기본원칙을 적용하면 무리 없이 해결할 수 있다.

4.3 구축결과

본 연구결과, 전술한 바와 같이 10개 분야를 대상으로 하이텔안내시스템용 시소러스를 구축하였다. (전주대학교, 1995) 주된 구축분야로는 일반 시사용어와 생활정보분야를 설정하였으며 디스크립터의 수는 8,278개이며, 비디스크립터의 수는 377개다. 비디스크립터의 수가 기타 정보검색용 시소러스보다 상대적으로 적은 것(4.5%)은 본 연구와는 별도로 유사어사전과 동의어사전(한영균, 1995)이 개별적으로 연구되었기 때문이다. 본 연구에서 구축된 결과는 유사선행연구의 결과(이두영 등, 1995)와 통합하여 하이텔내의 지식베이스로 활용될 것이다.

5. 결 론

하이텔을 대상으로 하는 지능형 정보검색시스템용 시소러스의 개발에 대한 방법 및 과정과 아울러 본 연구를 위하여 개발된 시소러스 구축시스템의 기능에 대하여 논하였다. 본 연구에서 개발한 시소러스는 해당 시스템에서 정보접근점으로서의 역할을 충분히 다하리라 생각된다.

다만 본 한글시소러스는 시소러스 구조에 대한 연구, 용어관계에 대한 한글의 언어학적 연구, 전문용어의 최신성 유지에 대한 연구, 나아가서는 용어의 전거통제를 위한 후속 연구에 의해 계속적으로 보완되어야 할 것이다. 또한 다음과 같은 두 가지 측면이 특히 강조되어야 할 것이다. 우선, 용어를 망라해야 한다. 해당 분야에서 사용되고 있는 모든 일반용어와 그에 대응되는 다국어 뿐만 아니라 방언까지도 가져야 하며, 가능하다면 고유명사까지도 가져야 한다. 둘째, 각종 언어정보를 가져야 한다. 자연언어처리에 필요한 구문,

의미, 문맥, 共起, 사례, 통계정보 등을 가져야 할 것이다.

시소러스의 개발은 시간, 비용, 내용 측면에서 매우 어렵지만 어려워도 개발해야 한다. 색인 및 검색효율을 높이기 위한 시소러스의 활용방법은 무한하기 때문이다. 단순화시킨 관계만을 갖는 전통적인 시소러스가 아니라 해당 도메인의 거의 모든 용어에 대하여, 대응되는 다국어용어, 세분된 개념관계, 각종 언어정보, 해설 등을 갖고 있으며, 고유명사에 대한 전거데이터까지 망라하고 있어야 한다. 이때 비로소 시소러스는 정보처리분야에 없어서는 안될 중요한 기본 지식베이스로서의 역할을 다할 수 있을 것이다.

참 고 문 헌

金泰中(1989). 우리말 시소러스 作成에 관한 硏究. 성균관대학교 경영대학원 석사학위논문, 미간행.

남영준(1996). 시소러스 구축 알고리즘. 한국정보관리학회 1996년도 정보관리강좌 - 시소러스구축 알고리즘 -.

이두영 등(1995). 시소러스. 지능형 정보검색에 관한 연구 -별책부록- . 한국통신.

전주대학교(1995). 지능형 하이텔안내시스팀용 시소러스개발에 관한 연구. 한국통신 연구개발원.

정영미, 최석두, 서혜란 편역(1994). 〔색인지침〕. 서울 : 문헌정보처리연구회 (동연구회 시리즈 4).

최석두(1994). 시소러스의 표시형식에 관한 연구. 〔94年度 韓國情報管理學會 全國論文大會(第1回)論文集〕, 105-108.

최석두, 정동열 공역(1994). 〔시소러스 개발지침〕. 서울 : 문헌정보처리연구회 (동연구회 시리즈 3).

한영균(1995). 정보검색용 전자사전. 지능형 정보검색에 관한 연구. 한국통신. 115-169.

ANZI / NISO Z39.19-199X. Guidelines for the Construction, Format and Management of Monolingual Thesauri.

EJC(1993). Thesaurus of Engineering and Scientific Terms. Engineers joint Council.

ISO 2788 - 1986(E). Guidelines for the Establishment and development of Monolingual Thesauri.

ISO 5964 - 1985(E). Guidelines for the Establishment and development of Multilingual Thesauri.