시소로파셋

김태중
한국데이터베이스진흥센터 연구개발부장

시소로파셋은 비교적 오랜 1970년에 최초로 소개된 정보 가공 및 검색을 위한 도구이다. 시소러스와 분류 체계의 성질을 모두 갖고 있어 대단히 유용한 도구로 평가되고 있으며 시소러스의 구축에 많은 인력과 비용이 소요되는 데 반해 비교적 적은 자원으로 구축할 수 있다는 경제적 이점이 있다.

그러나, 국내에는 자료 속에서나 소개되어 있는 정도이며 이에 관한 연구는 거의 없는 실정이다. 시소로파셋은 분류와 시소러스의 중간 형태이면서 이 들의 기능을 모두 갖고 있어서 향후 GUI 및 멀티미디어 데이터베이스 구축과 관련하여서 대단히 유용하게 사용될 수 있을 것이다. 여기에서는 색인과 시소러스에 관한 일반적인 사항을 소개하면서 시소로파셋의 특성과 작성 방법에 대하여 개괄적으로 소개하였다.

1. 데이터베이스와 색인

(1) 색인의 정의

색인은 데이터베이스로 구축하려는 대상이 되는 정보에 포함되어 있는 사항 또는 그 정보로부터 유출해 낸 개념들을 체계적으로 안내하는 것으로 이들 사항이나 개념들은 검색이 가능한 형태로 표현된다.

(2) 색인의 목적

색인의 목적은 정보 이용자에게 정보의 知的 내용과 물리적인 소재를 알려주기 위한 것이다. 즉, 이용자가 알려고 하는 정보에 담겨진 내용과 해당 정보에 접근해서 입수하는 데 필요한 사항을 정보 이용자에게 알려 주기 위해서 색인을 한다.

(3) 색인의 종류

색인은 관점에 따라 여러 가지 형태로 나누어 볼 수 있으나 정보 검색을 위한 데이터베이스와 관련하여서는 형식 색인과 주제 색인으로 분류함이 가장 적절하다. 먼저, 형식 색인은 앞에서 언급한 색인의 목적에서 정보의 물리적 소재를 알려주는 것과 밀접한 관계가 있으며, 주제 색인은 정보의 내용을 이용자에게 알려주기 위한 것이다.

한편, 형식 색인은 대체적으로 외형적으로 나타나는 사항을 일정한 형식으로 정리하여 표현하는 것이어서 비교적 단순하고 이를 작성하는 데 많은 비용과 노력이 소요되지는 않는다. 반면에 주제 색인은 정보의 내용을 이용자에게 알맞은 형식으로 제공해 주어야 하기 때문에 상당히 고급 인력이 필요하며 이를 위한 준비 과정도 적지

않은 노력이 소요된다.

그래서 일반적으로 색인이라 함은 주제 색인을 의미한다.

(4) 주제 색인의 형태

하나의 정보에 대해 이용자가 정보의 내용을 쉽게 파악할 수 있도록 정보에 포함되어 있지 않는 적합한 용어나 특정 기호(주제 분류 기호 등)를 부여하는 할당 색인과 정보에 포함되어 있는 용어 등을 색인에 사용하는 추출 색인으로 색인의 형태를 구분할 수 있다.

현대의 색인에서는 대개 할당 색인은 시소러스와 같은 색인용 통제 어휘집을 사용하며 추출 색인은 할당 색인이 갖는 문제점을 보완하는 측면에서 해당 정보를 색별하는 데 유리한 용어나 기호를 사용한다. 그리고, 색인에 사용된 용어나 기호를 각각 디스크립터(Descriptor, 우선어 또는 주제어)와 아이덴티파이어(Identifier, 식별어)라 한다.

2. 시소러스

(1) 용어 통제

색용에 사용되는 언어 즉, 색인 언어는 정보 검색에서 재현율을 높이기 위해서 용어의 띄어쓰기, 외래어 표기법 등의 외형적인 통제는 물론 용어의 쓰임과 의미를 제한하는 통제가 필요하다. 동음이의어의 통제는 검색의 정확률에 그리고 동의어에 대한 통제는 재현율에 영향을 미친다. 색인에 쓰이는 용어를 통제함에 있어서 필요한 기준 내지는 규칙은 일관성과 예측성을 갖추고 있어야 한다는 점이다.

(2) 용어 통제 수단

용어를 통제하기 위한 수단으로 典據表(Authority List, Preferred Keyword List), 주제명 표목표(Subject Heading List), 시소러스(Thesaurus) 등이 있다. 이 가운데 가장 널리 사용되고 있는 형태가 시소러스이다. 이 가운데 전거표(典據表)는 1980년대 초에 가끔 사용된 적이 있으나 근래에는 거의 볼 수 없다. 극히 소수의 인력으로 데이터베이스를 구축하려 한다면 사용해 볼만한 수단이기는 하나 시소러스와 주제명 표목표의 작성 방법이 보편화되어 있고 이들의 좋은 점이 충분히 있는 만큼 가급적이면 시소러스 또는 주제명 표목표를 만들어 사용함이 바람직하다.

(3) 시소러스 정의

ISO2788이 시소러스의 기능과 구조를 가장 적절하게 표현하고 있는 정의하고 있다. 이에 따르면 구조적인 면에서는 상위 개념의 용어와 하위 개념의 용어를 의미론적으로 밝힌 어휘집이며, 기능적인 면에서는 색인자 또는 정보 이용자가 자연어를 통제어로 변환하는 데 사용되는 어휘집이다 라고 하고 있다.

다시 풀어 본다면, 색인 및 검색에 사용되는 어휘를 표준화하고 정보의 축적 및 검색 시스템에서 동의어, 상위 개념어, 하위 개념어, 관련어 등의 선정이 용이하도록 각 용어의 개념적 관계를 밝힌 통제 어휘집이다.

(4) 용어간 개념 관계

시소러스에서는 색인에 사용될 용어(디스크립터)간의 의미 관계를 체계적으로 밝혀 놓아 색인자 및 이용자로 하여금 용이하고 정확하게 필요한 용어를 선정하게 한다. 용어간의 관계를 나타내기 위해서는 보통 4-5가지의 기호를 사용하며

대표적인 표시와 의미는 다음과 같다.

1) USE, UF(Use For)

- 동의 또는 유사 관계
- USE: 비디스크립터를 해당 디스크립터로 안내
- UF: 비디스크립터를 알려 준다.
- 동의 개념에 속하는 용어들은 다음과 같은 경우이다.
  ①같은 의미이나 어원이 다른 용어
  ②일반명과 학명
  ③일반명과 상표명
  ④새로 생긴 개념을 설명하기 위해 쓰이는 명사
  ⑤현대어와 고대어
  ⑥철자법 또는 표기법 차이
  ⑦문화적 배경의 차이에 따는 다른 용어
  ⑧표준어와 저속어
  ⑨약어와 원어
  ⑩분할된 용어와 분할되지 않은 용어

2) BT(Broad Term), NT(Narrow Term)

- 디스크립터간의 계층 관계
- BT: 표시어가 도입어보다 광의 또는 상위 개념어
- NT: 표시어가 도입어보다 협의 또는 하위 개념어
- 계층관계는 다음과 같은 경우에 해당된다.
  ①분류학상 종속 개념의 용어
  ②전체-부분 관계: 지역, 신체 기관, 학문이나 지식 체계, 사회 구조

3)RT(Related Term)

- 관련 관계: 상호 설명 또는 보완 관계를 나타낸다.
  ①학문 분야와 연구 대상
  ②행위와 결과
  ③동작 또는 과정과 기구, 시약
  ④행위와 행위의 수동체
  ⑤성질과 관련된 개념: 독물-독성
  ⑥출처와 관련된 개념
  ⑦종속적으로 연결되는 개념: 질병-병균
  ⑧개념과 상대되는 시약 등: 곤충-살충제
  ⑨함께 사용되어야 의미가 있는 명사구와 명사구의 중심어: 모형배-배

4)TT(Top Term), HE(Hierarchy Entry)

- 용어간 계층 구조에서 최상위 개념어를 안내하기 위해 사용
- 일부 시소러스에서 보조적으로 사용

(5) 시소러스 작성 방법

시소러스를 작성할 때 반드시 고려해야 할 사항으로 시소러스의 규모(디스크립터의 수), 용어 통제의 수준, 주제 분야의 성질 등이 있다. 시소러스를 작성하는 방법에는 근본적으로 2가지로 구분된다.

즉, 용어를 수집한 후에 개념을 정립하여 정리하는 방법과 개념을 체계적으로 정립한 후에 체계에 맞추어 용어를 수집하는 방법이다. 전자는 현실 세계에서 실제로 사용되고 있는 용어를 수집한 후에 용어 또는 용어간의 개념 및 관계를 정의하는 귀납적 방법이며, 후자는 학문 주제 분야별 전문가들로 위원회를 구성하여 여기서 부문별 개념을 체계화하고 용어를 정해진 개념의 틀에 맞추는 연역적 방법이 있다.

그러나, 일반적으로 이들 두 가지 방법이 혼용되는 것이 바람직하다. 왜냐하면 각각의 다음과 같은 장단점이 있기 때문이다. 귀납적 방법은 현실성은 있으나 주제의 포괄성이 결여되기 쉽고, 연역적 방법에서는 포괄성에는 문제가 되지 않으나 준비된 틀에 맞추기 위해 현실성이 떨어진 용

어가 수록될 우려가 있다.

(6) 시소러스의 규모

시소러스에는 일정한 수준의 디스크립터가 수록되어야 효과적으로 사용될 수 있다. 너무 많은 디스크립터는 적절한 색인어나 검색어를 골라내기 어렵고, 지나치게 적은 수의 디스크립터가 있는 시소러스는 검색시 적합률을 떨어뜨려 시소러스의 효용성을 낮추는 결과를 가져온다. 적절한 규모의 시소러스에 관하여는 오래 전부터 연구되어 오고 있다. 축적 또는 축적하려는 정보량과 정보 1건당 부여하려는 디스크립터의 수가 시소러스의 규모를 결정하는 데 영향을 미치는 주요 변수라고 보는 Houston 등이 연구한 다음과 같은 수식을 참고하면 비교적 알맞은 규모의 시소러스를 작성할 수 있다.

[시소러스의 규모(디스크립터의 수)] = 3,300 log {(정보량 × 색인어의 수) + 10,000) - 12,600

예를 들어 정보량이 100,000건이고, 정보당 색인어를 평균 10개씩 부여한다고 한다면, 시소러스의 규모는 7,214가 된다. 즉, 3,300log (100,000×10 - 12,600 = 7,214이다.

(7) 용어의 분할

시소러스를 작성할 때 고려해야 될 중요한 일 가운데 하나가 용어의 분할 문제이다. 용어 분할 문제는 예들 들면, "기업 정보 검색"이라는 용어가 있을 때 어떠한 형태로 디스크립터로 삼는 것이 가장 효율적이냐 하는 문제이다. "기업 정보 검색"은 우선 쉽게 1개의 용어 "기업 정보 검색"으로 다룰 수 있으며 또한 "기업", "정보", "검색"의 3개 용어로 나누어 볼 수 있다.

물론 "기업 정보"와 "검색" 그리고 "기업"과 "정보 검색"으로도 나누어 볼 수 있다. 어떠한 용어는 나누는 것이 효율적이고 또 다른 용어는 그렇지 않다.

이러한 용어의 분할(Factoring)에 관하여 영국 표준(BS: British Standard 5723)에서 제시하고 있는 방안이 가장 합리적이라고 생각한다. 시소러스 작성에 관한 국제 표준(ISO 2788)에서도 이 방법을 권장하고 있다. 그 내용을 간단히 소개하면 다음과 같다.

1) 의미론적 분할은 바람직하지 못하며 구문론적 분할이 타당하다.

- 온 도 계 = 온도 + 측정기 (×)
- 정보 관리 = 정보 + 관 (O)

2) 색인어는 단일 개념을 나타내야 한다.

3) 중심점(Fcus)과 차이점(Difference) 논리에 의한 분할 규칙을 적용

- 중심점(focus)
  - "兒童 病院"에서 "病院"
  - "合成 樹脂 容器"에서 "容器"
- 차이점(difference)
  - "兒童 病院"에서 "兒童"
  - "合成 樹脂 容器"에서 "合成 樹脂"

4) 분할하지 않는 용어

- 고유명사 및 고유명사와 함께 쓰이는 용어
  - 국제 연합, 정보통신부, 파킨스씨 병
- 차이점이 본래의 뜻을 잃어버린 경우
  - 무역풍
- 차이점이 비유 역할을 할 때
  - 나비 밸브
- 다른 명사와 같이 쓰여야만 의미가 있는 용

어: 차이점이 중심점과 같은 부류에 속하지 않음
- 장난감 병정

5)분할해야 하는 용어

● 중심점이 부분 또는 성질을 나타내고, 차이점이 소유자가 되는 경우

차이점(Difference)	중심점(Focus)
문	경첩
학교	교사
토양	산성도

● 중심점이 타동적 행위를 나타내고, 차이점이 목적이 되는 경우

차이점(Difference)	중심점(Focus)
도서관	관리
하천	오염
도서	제본

● 중심점이 자동적 행위를 나타내고, 차이점이 수행자가 되는 용어

차이점(Difference)	중심점(Focus)
조류	이동
파충류	동면

3. 시소로파셋

(1) 개요

영국의 여러 도서관에서 시소러스를 작성하였으나 English Electric Company의 경우를 제외하고는 미국 EJC의 "Thesaurus for Engineering Terms"(後에 TEST-Thesaurus of Engineering and Scientific Terms가 됨)처럼 큰 규모는 아니었다. 영국의 여러 CRG(Classification Research Group)이 최초의 대규모 패싯 분류 체계를 만들었으며 1961년에 제3판을 내게 되었다.

그후 몇 년이 지나면서 세밀한 개정 작업이 필요하게 되었다. 한편으로 기업체의 도서관 등에서는 컴퓨터를 이용하는 방안과 후조합 색인에 관해서 면밀히 검토하기 시작하였으며 이러한 결과 시소러스와 결합된 분류 체계가 개발되었다.

1970년에 발표된 이 분류 체계를 다소 이상한 이름인 시소로파셋(Thesaurofacet)이라 하였으며 주제 색인 이론과 실제에 많은 기여를 하였다. 이 것의 정식 명칭은 "Thesaurofacet: a thesaurus and faceted classification for engineering and related subjects"이다.

(2) 특징과 장점

시소로파셋은 분류와 시소러스의 2가지 도구의 기능을 모두 갖추고 있어서 이 들이 보완적으로 사용되게 되어 아주 좋은 결과를 가져올 수도 있다. 과거와는 달리 정보가 다루는 주제가 복합적이 경우가 많아서 순수한 분석적인 분류 체계보다는 분석적이면서 종합적인 분류가 이용하기 좋다. 시소로파셋은 이러한 면에서 대단히 경제적이고 유용하게 사용될 수 있다.

(3) 작성 방법

시소로파셋을 작성하는 순서는 패싯을 만든 후에 필요한 경우 각 패싯에 범위(Scope Note)를 달아주고, 사전에 수집된 용어군으로 부터 동의 관계어를 정리한다. 이어서 광의어, 협의어, 관련어 등의 순으로 각 용어를 패싯에 맞게 분류, 정리한다.

패싯분류는 각 주제를 몇 개의 패싯으로 분석하고 일정한 방식으로 다시 조합하여 하나의 주제를 표현하는 것으로 분석합성식(Analytico-

Synthetic) 분류라고도 하며 1933년 인도의 S.R. Ranganathan이 제창한 콜론 분류가 대표적인 예이다. 아무리 복잡한 주제라 할지라도 여러 패싯을 복합적으로 사용하여 표현할 수 있다는 장점이 있다.

랑가나단은 기본 범주를 [P]: personality, [M]: matter, [E]: energy, [S]: space, [T]: time 등의 5가지로 분류하여 이들의 조합으로 정보의 내용을 분류하였다. 예를 들어 "Welding of steel drums in Madras in 1987"라는 정보는 [P]: drums, [M]: steel, [E]: welding, [S]: Madras, [T]: 1987의 형태로 분류·정리된다.

패싯은 ① 문헌 조사를 통해 핵심 주제와 주변 주제를 설정, ② 유사한 개념을 그룹 지어 배열, ③ 일정한 규칙으로 작성, ④ 분류 체계의 작성 등의 차례로 만들어 간다.

(4) 활용과 전망

정보를 다루는 데 있어서 가장 중요한 요소는 3D이다. 즉 정보 처리(Data Processing), 정보 통신(Data Communication), 데이터베이스(Database)이며, 이 들은 각각 고성능 정보 처리(High Performance Computing - Supercomputer), 초고속 정보통신망(Information Super Highway), 멀티미디어(Multimedia)로 발전해 나아가고 있다. 멀티미디어 데이터베이스에서는 종래의 문자 중심의 데이터베이스와는 달리 이미지와 화상을 중심으로 정보가 축적되고 검색되도록 구성될 것이다. 이러한 멀티미디어 데이터베이스의 가공과 이용에 가장 적합하게 활용될 수 있는 도구가 시소로파셋이라고 생각한다.

왜냐하면, 정보를 체계적으로 분류하는 분류체계와 정보를 용어로 구체적으로 표현하는 시소러스가 합쳐진 형태로 이 들의 장점을 모두 지니고 있기 때문이다.

다시 말하면, 체계적이나 구체적이지 못한 분류와 구체적이나 체계적이지 못한 단점을 상호 보완하는 형태로서 멀티미디어 또는 GUI 환경에서 대단히 유용한 도구일 것이다. DC

〈참고 자료〉

1. A.C. Foskett, "The Subject Approach to Information", Clive Bingley, 1988
2. Jean Aitchison 외, "Thesaurus Construction: A Practical Mannual", Aslib, 1987
3. ANSI Z39.4 - 1984, American National Standard for library and information science and related publishing practices - basic criteria for indexes
4. BS 6529 - 1984, British Standard recommedations for examining documens, determining Their Subjects and selecting indexing terns
5. Aslib 교육 자료, "Building your own classificaiton scheme", Aslib, 1989
6. BS 5723 -1987, British guide to establishment and development of monolingual thesauri
7. 김명옥, "자료분류법", 구미무역(주) 출판부, 1986
8. 김태중, "우리말 시소러스 작성에 관한 연구", 성균관대학교 석사학위 논문, 1989

원고를 모집합니다

「DATABASE」의 세계로-
「데이터이스월드」는 독자 여러분께 그 문을 활짝 열어 독자들이 공감하고 같이 동참하는 우리 모두의 「광장」이기를 원하고 있습니다.

1. 원고내용

- 데이터베이스 관련 연구논문
- DBMS신기술
- 데이터베이스서비스
- 데이터베이스산업정책 및 정보표준화
- 데이터베이스기술 동향
- 데이터베이스법령 해설
- CD-ROM 및 멀티미디어
- 독자투고
- DB진흥센터에 바란다
- 기타 데이터베이스 및 정보화 관련 등 정보화사회 인식제고에 기여할 수 있는 글

2. 분 량

30매내외 (200자 원고지)

3. 마 감

매달 10일

4. 보낼곳

서울시 종로구 수송동 146-1 이마빌딩 8층 (재)한국데이터베이스진흥센터
홍보출판과 데이터베이스월드 담당자 앞
(전화) (02)725-3751/3, (팩스) (02)725-3750
E-MAIL: 이용자번호: DPCK(천리안, 하이텔, 나우콤)

5. 기 타

- 도착된 원고는 반환치 않으며 게재된 원고에 한해 원고료 지급
- 원고 제출시 주소, 주민등록번호, 온라인번호, 약력, 사진1매 등을 작성 제출요망