콜론분류법에 바탕한 자동분류시스템의 개발에 관한 연구

- 농학 및 의학 전문도서관을 사례로 -

이 경 호*

I. 서 론

1.연구의 필요성 및 목적

도서관업무의 대부분이 자동화됨에 따라 이용자에 대한 정보봉사는 과거 어느 때보다도 신속, 정확하게 이루어지고 있다. 그러나 도서관업무 중에서도 분류업무만은 아직까지도 전통적인 수작업방법에 의존하고 있어 사서들은 분류도 다른 업무와 마찬가지로 자동분류시스템의 개발을 통하여 기계적인 처리가 이루어지기를 기대하고 있다.

이러한 자동분류에 대한 필요성은 도서관업무의 대부분이 자동화되어 있고 특정 주제분야의 자료로 명확히 정의된 이용자집단을 대상으로 전문적인 정보봉사를 수

* 대구대학교 문헌정보학과 부교수

행하는 전문도서관의 경우 다른 어떤 도서관보다 높게 나타나고 있다.

종래의 수작업 분류는 분류자가 도서의 주제나 형식을 분석하여 분류기호를 부여하는 방식이다. 그런데 분류자의 판단 잘못으로 인하여 분류의 일치율이 45퍼센트에서 76퍼센트[1]에 이르는 등 분류상에 오분류(誤分類)가 많아 신속, 정확성을 기하기 어려운 점이 있다. 그리고 분류도구인 분류표가 대부분 열거식이면서 인쇄매체로 되어있어 주제의 삽입이나 개정을 하고자 할때 상당한 시간과 경비를 요하는 등 문제점이 있다.

그러나 콜론분류법(Colon Classification; 이하 CC라 칭함)은 도서의 내용을 패시트(facet)라고 하는 구성요소로 분석하고, 분류할때 이들 개념을 합성하므로 다차원적인 주제를 하나의 선(線)상에 표현할 수 있어, 서가상의 배가는 물론 정보검색의 측면에서도 유용한 검색도구가 될 수 있는 장점을 지니고 있다.

CC는 랑가나단(S.R. Ranganathan)이 고안한 분석합성식 분류법으로 1933년 처음으로 발표되었다. CC는 복합주제를 분류기호로 표현할 수 있는 대표적인 분류표로 이미 널리 인식되어 있다. 따라서 이러한 장점을 지니고 있기 때문에 CC는 분류법 이론에 있어 새로운 연구대상으로 평가되고 있다.

이렇게 우수한 원리를 지니고 있음에도 불구하고 패시트 분류에 대한 지금까지의 연구는 도서관 자동화시스템에서 사용할 수 있을 정도의 분류시스템으로 발전하지 못하고 있다. 그러나 앞으로 패시트분류는 그 원리 자체가 우수하기 때문에 보다 더 실용적인 측면에서 연구가 이루어 질 것으로 생각되며, 나아가 이 원리를 이용한 도서분류 자동화도 향후의 연구과제 가운데 하나가 될 것으로 보인다.

종래의 수작업 분류법과 관련하여서도 도서의 서가상의 배가는 물론, 검색의 기능을 함께 지닐 수 있으면서 각 도서관마다 동일한 도서에 대해서는 같은 분류기호를 이끌어 낼 수 있고, 복합개념을 표현할 수 있으면서, 또 인간의 노력을 최소화 할 수 있는 자동분류에 대한 연구가 이루어져야 하리라고 본다.

특히 콜론분류법에 의해 도서분류 자동화를 실현하게 되면 수작업 분류와 비교할 때 다음과 같은 이점이 있다.

---

1) 丸山昭二郎, "分類作業の一致率," 情報の科學と技術 37 : 5(1987) : 198-199.

(1) 신속한 분류를 할 수 있다.

(2) 분류표의 개정이나 갱신이 간단하다.

(3) 복수의 개념을 하나의 분류기호 속에 표현할 수 있다.

(4) 분류자의 주관적 판단에 따른 분류가 아닌 객관적인 분류공식에 따라 분류를 하기 때문에 일치율이 높다.

(5) 이용자는 자기가 찾고자 하는 자료에 대해 관련 키워드를 통하여 분류기호를 합성하고 나서 검색을 행할 수도 있다.

(6) 대량의 자료를 단시간에 기계적인 처리로 분류할 수 있다.

(7) 주제배경이 없는 사서도 분류할 수 있다.

(8) 도서관업무의 토털시스템 구성에 기여할 수 있다.

(9) 도서관들은 동일한 프로그램 소프트웨어에 의해 기계적인 처리로 분류 할 수 있기 때문에 도서관간 정보유통에 기여할 수있다.

특히 도서의 표제를 통하여 분류를 자동적으로 생성하고자 한것은 다음과 같은 이유에서 였다.

엔도(遠藤英三)[2]가 그의 연구에서 조사한 바와 같이 『도서관협회선정도서총목록』에서 표제상에 주제를 나타내는 키워드의 포함비율이 1961년에는 89퍼센트로 나타난 반면 1967년에는 94퍼센트로 더욱 높게 나타나는 등 점차 도서의 표제가 도서의 내용을 대신하는 경향으로 나아가고 있는 점, KWIC(Keyword-in-Context) 색인법이 소개되면서부터 저자들은 과거 어느때보다도 기술적이고 의미있는 표제를 쓰려고 한다는 점,[3] 그리고 표제를 통하여 도서의 내용을 60퍼센트에서 70퍼센트는 파악할 수 있는 점[4]등 도서의 표제는 바로 내용과 직접적인 관련이 있다고 가정하였기 때문이었다.

따라서 본 연구의 목적은 CC의 패시트방식과 기호법을 약간 수정적용하여 전문

2) 遠藤英三, "主題の把握と その表現," 圖書館界 21 : 3(1969) : 82-87.

3) K. L. Kwok, "The Use of Title and Cited Titles as Document Representation for Automatic Classification," Information Processing & Management 11(1975) : 202.

4) 加藤宗厚, 件名作業, 東京 : 理想社, 昭和 32(1957) (이재철, 주제명목록의 연구, 서울 : 연세대학교 도서관학과, 1959. p. 80에서 재인용).

도서관에서 실용화할 수 있는 자동분류 시스템을 개발하기 위해 분류자동화가 가능한 분류데이타베이스를 설계, 구축하고 도서의 표제나 키워드를 컴퓨터에 입력함으로써 주제의 자동적인 인지는 물론 표제속에 있는 키워드를 CC의 기호조합방식으로 처리하여 분류기호를 자동적으로 만들어 내고자 하는데 있다.

2. 연구 방법

본 연구는 농학이나 의학도서관과 같은 전문도서관에서 분류자동화가 가능한지의 여부를 CC의 원리를 응용하여, 실험 및 검증하고자 한다. 이를 위해 분류자동화가 가능한 분류데이타베이스를 설계한 한 후, 실제 실험분야인 의학과 농학 분야를 대상으로 용어를 수집, 분석하여 분류데이타베이스를 구축한다. 그리고 도서의 표제입력(표제가 불명확 할때는 키워드)을 통하여 주제를 식별함과 아울러 분류기호도 함께 만들어낼 수 있도록 상기 두개 주제분야에 대해 분류자동화 원리를 유도하고 실제 프로그램을 작성한 후 실험을 행한다.

연구 방법은 다음과 같다.

(1) 농학과 의학 전문도서관에서 분류자동화가 가능한 분류데이타베이스를 설계한다.

(2) 분류데이타베이스의 구축을 위하여 CC[5]와 DDC 20판[6]및 MeSH(Medical Subject Headings)[7]를 주로 참고하여 용어를 수집하였으며, 수집된 용어 수는 농학 629개, 의학 1,033개로서 총 1,662개이나, 이 가운데 152개의 용어는 서로 중첩된 관계로 실제 용어 수는 총 1,510개 이다.

(3) 수집한 용어는 랑가나단이 제시하고 있는 주제분야별 카테고리로 분석한다.

(4) 분류자동화에 의한 자동화 가능성여부의 실험은 영문으로 된 농학과 의학도서로 한다.

---

5) S.R. Ranganathan, Colon Classification, 7th ed., Bangalore : Sarada Ranganathan Endowment fo Library Science, 1989.

6) Melvil Dewey, Dewey Decimal Classification and Relative Index, 20th ed., Albany : Forest Press 1989.

7) National Library of Medicine, Medical Subject Headings : Supplement to Index Medicus 30(1989) Maryland, 1989.

(5) 실험대상은 영남대학교 의과대학 도서관 소장 의학관계 도서를 목록함에서 일정한 간격으로 무작위 추출한 642권 가운데 생물학에 관련된 60권을 제외한 582권과, 영남대학과 대구대학교 중앙도서관이 소장하고 있는 전체 농학도서 가운데 임업(JX), 축산업(KX)과 같이 CC상에서 농학(J)으로 분류하지 않은 도서를 제외한 총 448권 (영남대 소장 336권, 대구대소장 143권, 중복 31권)을 대상으로 한다.

(6) 도서의 표제 (표제에 의해 주제인식이 안되어 분류기호가 생성되지 않을 경우는 인위적으로 키워드 입력)에 의해 주제식별과 분류가 가능하도록 분류자동화 원리를 유도하여 플로차트화 한다.

(7) 데이타베이스에 의해 분류자동화를 가능케 하는 분류자동화 원리는 CC의 기본 패시트 조합원리를 적용한다.

(8) 플로차트를 근거로 프로그램을 작성하여 농학분야 448권의 도서, 의학분야 582권의 도서를 대상으로 실제 데이타를 입력하여 실험을 한다.

(9) 실제 분류는 도서의 표제에 의한 분류를 원칙으로 한다. 만약 표제에 의해 주제인식이나 분류가 행하여 지지 않는 경우에는 내용목차나 본문속의 키워드를 입력하여 분류하도록 한다.

(10) 시스템의 운용을 위하여 분류데이타베이스를 언제든지 수정(edit), 삭제(delete), 추가(append) 및 색인(reindex)을 할 수 있도록 자동분류 시스템을 개발한다.

3. 연구의 제한점

연구의 제한점은 다음과 같다.

(1) 본 연구가 어디까지나 실험연구인 만큼, 용어의 망라적인 수집에 의한 완전한 분류데이타베이스를 구축하기보다는 자동분류가 가능한 실험용 분류표인 분류데이타베이스를 설계, 구축 하였으므로 실제로 모든 도서가 전부 분류되게끔 시도한 것은 아니다.

(2) 본 연구에 의한 시스템을 실제 적용함에 있어 농학이나 의학과 같은 전문도서관에 있어서는 실현 가능하나 일반 도서관에서 적용할 때는 다소의 문제점이 파

생될 수 있다.

4. 선행연구의 개요

문헌정보학에서 자동분류(automatic classification) 또는 분류자동화는 문헌분류(document classification) 또는 용어분류(term classification)의 자동화로 인식하는 경향이 있다. 이의 주된 이유 중의 하나는 지금까지 자동분류에 관한 연구가 주로 분류표에 의거하여 분류기호를 자동적으로 생성시키고자 하는 당초의 연구목적에서 벗어나, 컴퓨터에 의한 탐색의 효율성을 향상시키기 위하여 기계가독형 데이타베이스를 위해 적절히 그루핑하고, 배열하고자 하는 문헌분류나 용어분류에만 치우쳐 온 사실에 기인한 것이 아닌가 생각된다.[8]

따라서 자동분류(automatic classification)에 대한 지금까지의 연구는 주로 도서의 서가상의 배가를 위한 분류의 측면보다는 검색적인 측면에서 이루어져 왔다. 이러한 검색적인 측면에서 본 자동분류의 유형으로서는 크게 문헌분류(document classification), 용어분류(term classification), 잡지분류, 인용문헌분류 등이 있다.

그러나 이러한 측면에서의 자동분류는 본 연구에서 하고자 하는 자동분류, 즉 도서의 서가상의 배가와 검색의 기능을 동시에 갖는 코드화에 의한 분류와는 연구방법이 다르다.

본 연구의 자동분류와 관련성이 있는 국내외의 사례 및 선행연구를 개관하여 보면 다음과 같다.

본 연구와는 다소 상이한 점도 있으나 패시트원리에 입각한 분류시스템으로서는 영국의 양조회사에서 전산화 판매시스템에 적용하고있는 사례를 들 수 있다.[9] 이 회사에서는 판매하고자 하는 맥주의 구분에 따라, 용량 (capacity facet), 용기 (container facet) 및 술의 유형 (type facet)의 세가지 카테고리로 구분하고, 이

8) Rama N. Vashista, "Automatic Classification : Some Latest Development," Indian Librarian 32 : 2 (Sept. 1977) : 82.

9) Eric J. Hunter, Classification Made Simple, Adershot : Gower, 1979, pp. 7－9.

들을 다시 세구분하여 번호를 부여한 후 순서대로 조합하여 분류기호를 자동적으로 생성하고 있다.

분류기호의 조합은 반드시 용량---〉용기---〉 술의 유형 순으로 하여 분류하며,

	용 량	용 기	술의 유형
1	330 ml	1 can	1 mild
2	440 ml	2 bottles	2 bitter
3	550 ml		3 lager
4	1 litre		4 stout
5	3 1/2 litre		

440ml, cans, lager 를 213으로 분류하는 방법이다.

이와 유사한 시스템으로서는 부동산회사에서 매도하고자 하는 물건이나, 매수하고자 하는 사람들의 요구사항에 대하여 다음과 같이 (1) 침실의 수, (2) 주거형태, (3) 위치, (4) 가격으로 구분하여 분류하되 상기의 예와 같다.[10]

침실의 수	주거형태	위치	가격
1 4 bed.	1 semi-det. bungalow	1 Atwell	1 $ 30000- 40000
2 3 bed	2 flat	2 Blanford	2 $ 40000- 50000
3 1 bed	3 detached bungalow	3 Denby 4 Crosswood	3 up to $ 20000 4 over $ 50000
4 5 bed	4 semi-det.		

또 다른 분류자동화 시스템으로서는 기계부품인 나사(못)(machine bolts,screws)의 분류를 들 수 있다.[11] 이 시스템에서는 나사의 특징을 재료(material), 나선줄의 크기(thread size), 머리모양(head-shape), 마무리작업(finish) 등으로 다음과 같이 구분하여 분류하고 있다. 분류의 예를 보면, Chromium-plated brass

10) Ibid., pp. 11-22.

11) H.D.Clifton, Business Data Systems : a Practical Guide to Systems Analysis and Data Processing, London : Prentice Hall International, 1978, p. 233.

square-head bolts of 2 BA thread의 경우는 brass(first digit) / 2 BA(second digit) / square(third digit) / chromium plated(fourth digit)가 되어 분류기호는 2342가 된다.

.재료 (first digit)

1 = stainless stell

2 = brass

3 = steel

.머리모양(third digit)

1 = round

2 = flat

3 = hexagonal

4 = square

. 나선줄의 크기 (second digit)

1 = 0 BA

2 = 1 BA

3 = 2 BA, 등

. 마무리 (fourth digit)

1 = unfinished

2 = chromium plated

3 = zinc plated

4 = painted

그리고 본 연구에서와 같이 코드화에 의한 도서분류 자동화에 관한 연구로는 벤카트라만(S.Venkatraman)과 닐라메간(A.Neelameghan)에 의한 공동연구[12]가 있다. 이들의 연구는 컴퓨터의 자기테이프상에 CC의 주분류표(Basic Subject Schedule)와 주제분야별 분류표(Special Isolate Schedule) 및 공통구분 기호표(Common Isolate Schedule)를 CC분류표에서와 같이 분리하여 만들어 두고, 도서의 내용을 나타내는 핵심어 (Kernel Term)를 펀치카드로 입력하여 주분류표에서 해당 주제를 찾고, 다음으로 해당 주제의 주분류표와 공통구분기호표에서 개개 핵심어에 대한 분류기호를 탐색하여 분류기호를 합성하고자 한 연구이다.

12) S. Venkataraman : A. Neelameghan, "Formation of Isolate Number by Computer Using the Devices of Colon Classification," Library Science with a Slant to Documentation 6 : 1(1969) : 141－190.

- S. Venkataraman : A. Neelameghan, "Preparation of Schedule-on-Tape for Synthesis of Class Number by Computer," Library Science with a Slant to Documentation 6 : 1(1969) : 130－140.
- A. Neelameghan : S. Venkataraman, "Formulation of Kernel Terms for a Subject and Iso-late Terms for a Classification Schedule for Use in the Synthesis of class Number by Computer," Library Science whith a Slant to Documentation 6 : 1(1969):71－93.

셔리프(Sharif)[13]는 전문가시스템(expert system)적 접근방법으로 단행본도서에 대하여 분류 자동화를 시도하였다. 셔리프의 연구는 전문가시스템을 이용한 실험적 연구로서 분류시에 여러번 분류전문가의 의사결정이 요구되는 시스템으로서 기계적인 자동분류시스템과는 다소 거리가 있다.

한편 본 연구자는[14]는 CC의 원리를 응용하여 도서관학 및 서지학분야를 대상으로 이들 주제분야에서 사용되고 있는 키워드를 수집, 분석하여 데이타베이스화 하고, 표제상의 키워드를 입력하여 데이타베이스상에서 탐색한 후 CC의 분류원리에 의하여 이들 개념을 조합함으로써 자동분류를 실현하고자 하였다.

지금까지 살펴본 바와 같이 도서분류 자동화에 관한 연구는 단지 몇 편에 지나지 않는다. 그 중에서도 밴카트라만과 닐라메간의 연구가 CC의 원리를 이용하여 자동분류를 시도한 점은 본 연구와 유사한 점이 있다. 그러나 분류데이타베이스의 구성방법을 비롯하여 주제인지 방법, 데이타 입력방법 및 시스템 운영 등에서 본 연구와 상당한 차이가 있다.

그리고 본 연구자의 연구는 분류데이타베이스를 만든다는 점은 같으나, 단지 하나의 주제만을 대상으로 한 자동분류에 관한 연구이기 때문에 컴퓨터에 의한 주제의 자동인지가 불가능한 형태로 분류데이타베이스가 설계되어 있어 표제를 입력할 때 개개 도서마다 주제기호를 입력하여야 하는 등 본 연구와는 차이가 있다.

셔리프의 연구는 전문가시스템에 의한 연구인 만큼, 본 연구와는 연구방법이 다를 뿐 아니라 분류자의 의사결정에 따라 도서가 분류되기 때문에 지금의 수작업 분류방법을 컴퓨터에 의해 다소 변형시킨 것에 불과하다고 할 수 있다.

5. 용어의 정의

본 연구에서 사용하고 있는 용어를 정의 하면 다음과 같다.

(1) 패시트(facet) : 하나의 주제가 어떤 특징에 따라 구분될 때 생겨나는 구분

13) Carolyn A.Y. Sharif, Developing an Expert System for Classification of Books Using Micro-Based Expert System Shells, Boston Spa : British Library Research and Department, 1988.

14) 이경호, 도서분류의 자동화 : 도서관학 및 정보학분야 서지분류를 중심으로, 석사학위논문. 경북대학교 대학원, 1980.

의 전체를 의미한다.

(2) 도서분류(book classification) : 도서형태의 자료를 서가상에 배가하기 위해 일정한 기호를 부여하는 것을 의미한다.

(3) 문헌분류(document classification) : 유사한 문헌들의 집단을 형성함으로서 정보검색시에 한 문헌군만을 탐색하도록 제한시켜 검색의 신속성을 기하기 위한 분류이다.

(4) 용어분류(keyword classification) : 정보검색시 이용자의 질문과 문헌의 내용을 나타내는 용어가 서로 일치하는 범위를 넓히기 위하여, 개개 용어를 유사어군으로 분류하는 것이다.

(5) 분류데이타베이스 : 컴퓨터 프로그램에 의하여 주제의 자동인지에 이어 분류 기호가 자동적으로 생성될 수 있도록 분류에 필요한 각종 정보를 축적하여 둔 데이타의 집합을 의미한다. 본 연구에서의 분류데이타베이스는 AutoBC시스템을 위해 특별히 설계, 구축한 데이타베이스를 의미한다.

(6) AutoBC : 본연구에서의 자동분류시스템(Automatic Book Classification System)을 AutoBC 시스템이라 칭한다.

II. 분류데이타베이스의 설계 및 구축

1. 설계개요

정보시스템에서 사용하고 있는 분류표로는 열거식 분류표가 가장 많이 사용되고 있다. 이 중에서도 DDC가 가장 널리 사용되고 있다. 이의 주된 이유는 장기간 사용상에서 오는 탄력적인 점도 있겠으나 무엇보다도 주기적인 개정작업과 목록중심의 사용, 그리고 이를 대치할 만한 우수한 일반분류표가 없었기 때문으로 풀이된다.[15]

그러나 열거식 분류표는 자동화 시스템에 적용할 수 없을 뿐만 아니라 여러 개

15) Peter G.B. Enser, Automatic Classification of Book Material Represented by Back-of-the-Book Index, Ph.D. Thesis. University of Sheffield, 1983, p. 15.

념으로 되어있는 정보를 손실없이 일직선상에 기호로 표시할 수 없는 단점이 있다.

이러한 이유로 인하여 자동화 분류데이타베이스의 설계는 종래의 수작업용의 분류표와는 다른 차원에서 설계하여야 한다. 이 설계과정에 포함시켜야 할 요소로서는 컴퓨터가 표제나 키워드로 부터 주제를 인식하여 내고, 이를 근거로 분류기호도 만들 수 있어야 한다는 점이다.

예컨데, 오늘날 정보시스템에서 가장 많이 사용하고 있는 용어통제표는 시소러스와 주제명표목표이다. 이들 통제표에서는 특정 주제분야내에서의 용어간의 관계를 묘사하고 있을 뿐, 학문분야간의 관계는 묘사하지 못하고 있다.

그러나 분류데이타베이스상에서는 개개 용어마다 각 주제분야에서의 속성과 위치를 특정지워 주어야 한다. 한 예로서, 우리는 teachers, reading, boys, girls 등의 용어가 나타나는 문헌의 주제는 교육학으로 추측할 수 있다. 문제는 인간의 사고과정이 어떻게 이것을 추출하여 낼 수 있느냐 하는 점이다. 이점을 토대로 AutoBC시스템에서의 데이타베이스의 원리를 생각해 낼 수 있다. 이는 이러한 용어들의 속성이 모두 교육학분야에서 연구대상으로 하고 있는 용어라는 사실을 추론할 수 있기 때문에 교육학관계 문헌으로 인식이 가능한 것이다.

따라서 AutoBC시스템을 위한 데이타베이스의 설계는 종래의 주제 분야별 시소러스를 전부 통합한 성격을 지니면서, 각 주제분야별로 개개 용어마다 속성을 특징지워 줌으로써 주제식별이 가능한 형태로 설계하여야 할 것으로 생각된다.

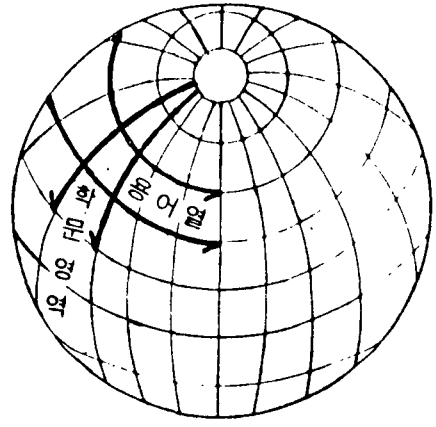
2. 설계원리

우리가 일상생활에서 사용하고 있는 언어를 모아 일정한 순서로 벌여 싣고 낱낱이 그 발음, 의의, 용법, 어원 등에 관하여 해설한 책이 사전(辭典)이다. 이 사전으로 우리는 단어의 의미를 알 수는 있으나, 학문분야간의 관계나 속성을 파악할 수는 없다. 즉 사전상의 단어의 의미는 표현상의 의미일 뿐이며, 주제구분의 의미는 없다. 그러나 하나의 단어는 주제분야에 따라 다른 속성, 다른 의미, 다른 위치를 가진다. 이러한 모든 요소들이 분류데이타베이스상에 묘사될 때 만이 자동분류가 가능할 것이다.

이러한 데이타베이스를 설계함에 있어서는 다음과 같은 두 가지 경우를 생각해 볼 수 있다.

(1) 지구의(globe)의 원리

우리가 일상생활에서 사용하고 있는 모든 키워드가 하나의 주머니나 등근공의 핵심부위, 즉 지구의의 중심부위에 그림2-1과 같이 모두 들어있다고 생각해 볼 수 있다. 그리고 지구의의 위도는 용어열이고, 경도는 각 학문영역으로 볼 때, 하나의 용어는 어떤 학문분야에서든지 그 속성을 지닐 수 있게 된다. 실제로 용어는 핵심부위에 있기 때문에 언제든지 방향만 전환하면, 다른 주제분야의 용어로 그 성격을 변환할 수 있다.

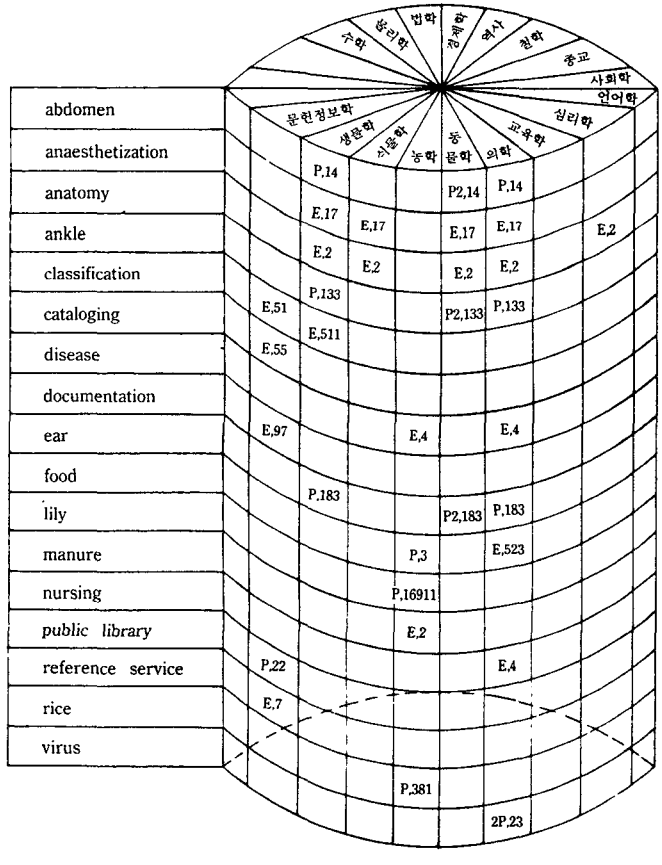


〈그림 2-1〉 지구의의 원리에 의한 설계

그리고 개개 주제분야는 그 주제분야에서 연구대상으로 하고 있는 용어만을 선택, 코드화시킬 수 있어 각 주제분야내에서의 특징을 묘사할 수 있게 된다.

(2) 원통형의 원리

상기 지구의에 의한 원리는 가장 이상적이나, 좀더 이해하기 쉽게 도식화하여 보면 다음 그림2-2와 같이 원통형으로 설명할 수 있다.



〈그림 2-2〉 원통형의 원리에 의한 설계

이 원통형의 가장 중심부위 상하로 키워드가 집결(이해를 돕기위해 알파벳 순으로)되어 있다고 가정하고, 원통의 중심점을 기점으로 여러 조각으로 구분하고, 각 조각 부분을 그림에서와 같이 하나의 학문영역으로 생각한다.

또 좀 더 쉽게 이해하기 위해 중심부위에 상하로 배열되어 있는 용어를 그림2-2에서와 같이 좌측부위에 나열 시켜볼 수 있다. 이제 각 용어에 대하여 랑가나단이 주장하고 있는 방법으로 분석하여 해당용어에 그 용어의 분류코드를 배정한다. 즉 하나의 용어는 주제분야에 따라 그 용어가 지니고 있는 속성이 다르게 나타날 수 있어 주제분야별로 특정의 고유값을 가지게 되며, 때에 따라서는 전 주제분야에 동일한 하나의 분류코드를 부여할 수도 있다.

. 랑가나단은 Colon Classification의 색인부분에서[16] 다음과 같은 형식으로 용어간의 특징을 묘사하고 있다.

abdomen	G[P], K[P2], L[P], 14
anaesthetization	G, I, K, L[E], 17
anatomy	G, I, K, L, S[E], 2
ankle	G[P], K[P2], L[P], 133

상기에서 abdomen의 경우는 생물학(G)에 있어 [P], 동물학[K]에 있어 [P2], 의학(L)에 있어 [P]의 속성을 지니되, 그 값은 14라는 의미를 나타내며, anaesthetization은 생물학, 식물학, 동물학, 의학에서 다같이 [E]의 속성으로 17개의 값을 지닌다는 의미이다.

랑가나단의 이와같은 색인방식은 하나의 용어에 대하여 각 주제분야별 속성과 자리매김을 분명히 하여 줌으로써 권말 색인이 통합시소러스의 형태를 지니고 있는 예이다. 이들 용어에 대하여 기호화하여 보면 그림2-2와 같다.

이 그림2-2상에 표기된 용어를 설명하여 보면, anatomy는 생물학, 식물학, 동물학, 의학, 심리학에서 연구대상으로 하며, 용어의 속성은 [E]의 속성을 지니면서,

16) S.R. Ranganathan, Colon Classification, 6th ed. New York : Asia Publishing House, 1960, pp.2 126－129.

분류코드가 2로서 동일한 값을 지니고 있다. 또 abdomen은 생물학과 동물학, 의학에서 연구대상으로 하며, 생물학에서는 [P]의 속성, 동물학에서는 [P2]의 속성, 의학에서는 [P]의 속성을 지니면서 그 값은 다같이 14로 나타난다.

이상과 같은 분석 및 설계원리에 의거하여 우리가 일상생활에서 사용하고 있는 모든 용어를 수집, 분석하여, 데이타베이스를 구축 한다면, 키워드나 표제를 컴퓨터에 입력하여 자동적으로 주제를 인식케하고, 이에 따라 분류 기호도 자동적으로 만들어 낼 수 있을 것으로 본다. 따라서 이와 같은 원리에 입각하여 자동분류가 가능한 데이타베이스를 설계 하고자 한다.

3. 설계(안)

상기에서 언급한 원통형의 원리에 입각하되 농학과 의학만을 대상으로 컴퓨터 처리가 용이한 포멧으로 AutoBC시스템을 위한 분류데이타베이스를 설계하여 보면 다음 표2-1과 같다.

여기에서 두 주제분야 내의 제어코드1과 제어코드2는 가능한 한 CC원리에 의거하여 분류가 될 수 있도록 분류원리를 제어하는 기능을 수행케 하는데 그 목적이 있다. 그리고 이들 필드별 요소를 설명하여 보면 다음과 같고 그림으로 나타내면 그림2-3과 같다.

◈ 키워드 : 분류데이타베이스상의 용어

◈ 제어영역 : 키워드의 카테고리나 분류코드가 두개의 주제에 다같이 일치하는 경우 이들 값을 통제한다. 카테고리제어는 키워드의 카테고리를, 코드제어는 분류코드를 제어한다.

◈ 키워드속성 : 용어의 속성을 기술 한다.

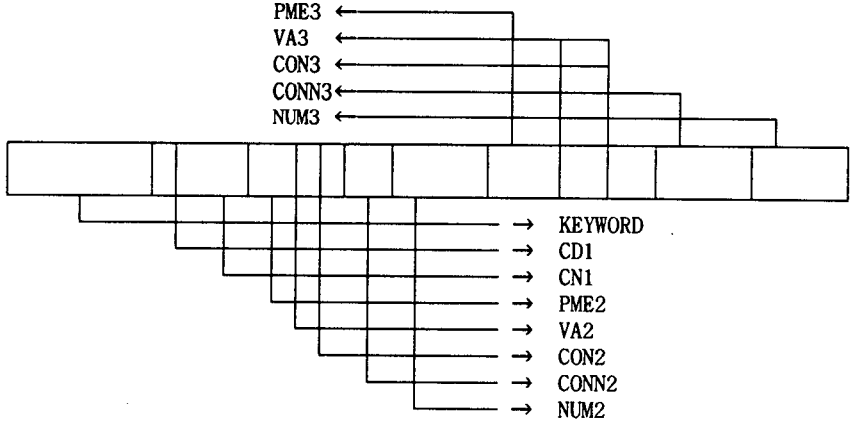
◈ 배열코드 : 한 주제내의 용어만을 출력하되 분류기호순으로 배열하고자 할 때 기호의 배열값을 나타낸다.

◈ 제어코드 1,2 : 분류원리에 의해 분류되도록 제어하는 키

◈ 분류코드 : 분류할 때 조합되는 분류기호

〈표 2-1〉 분류데이타베이스의 구성요소

필드내용	필드명	필드길이
◈ 키워드	Keyword	X(30)
◈ 제어영역		
카테고리제어	CN1	x(2)
코드제어	CD1	x(8)
◈ 제1주제 영역(농학)		
키워드속성	PME2	X(2)
배열코드	VA2	X(2)
제어코드1	CON2	X
제어코드2	CONN2	X(8)
분류코드	NUM2	X(8)
◈ 제2주제영역 (의학)		
키워드속성	PME3	X(2)
배열코드	VA3	X(2)
제어코드1	CON3	X
제어코드2	CONN3	X(8)
분류코드	NUM3	X(8)



〈그림 2-3〉 분류데이타베이스의 구조

4.용어수집 및 분석

가.용어수집

용어수집의 목적은 분류를 자동화 하는데 있어서 컴퓨터로 하여금 도서의 표제나 키워드에의한 탐색으로 자동적인 주제인식에 이어 분류원리에 입각하여 분류기호를 생성 시킬 수 있도록 분류용 파일을 만들기 위한 것이다.

데이타베이스 구축을 위한 용어의 수집은 그 분야에서 사용 되어질 수 있는 대부분의 용어를 수집하되 가능한 한 동의어나 유사어 등을 포함한 광범위한 수집이 바람직 하다. 그러나 본 연구가 어디까지나 분류데이타베이스의 설계를 통한 자동분류의 가능성 여부를 진단하고자 하는 실험연구인 만큼, 용어의 수집은 연구 대상분야인 농학과 의학분야의 용어를 망라적으로 수집하였다기 보다는 기본적인 정도의 용어만을 수집하였다. 그러나 무엇보다도 해당주제분야의 용어수집은 가능한 한 망라적인 수집이 바람직하기 때문에, 향후 실제적용을 위해서는 이들 주제분야의 전문가들과 계속적인 협력을 통하여, 보다 망라적인 차원에서 용어를 수집하여야 하리라고 본다.

본 연구에서 수집한 용어는 농학분야 580개, 의학분야 1,028개로서 총1,608개이다. 농학분야와 의학분야의 용어수집은 CC의 제6판과 7판의 농학분류표에 열거된 용어를 일차적으로 수집하고, 이 수집된 용어에서 누락된 부분은 DDC 20판에서 중요어라고 판단되는 용어를 추가하였다. 단 의학의 경우에는 MeSH의 계층구조표(tree structure)에서 중요하다고 판단되는 용어를 추가하여 수집하였다. 그리고 실제 분류과정에서 표제상의 주요용어가 매치되지 않을 때는 이들 용어를 분석, 추가하여 데이타베이스를 갱신하여 분류하는 방법으로 하였다.

나. 용어의 분석

수집한 용어의 분석은 용어자체가 지니고 있는 속성, 즉 용어의 성질을 분석하고, 기호화함으로써 그 용어에 대한 속성치를 기호로 나타내어 데이타베이스를 구축하기 위함이다.

용어의 분석은 동의어나 유사어가 아닌 경우에는 가능한 한 고유번호를 부여하였다. 그 이유는 고유번호를 부여하므로 고유의 분류번호가 만들어질 수 있어 분

류기호에 의한 배열 뿐만 아니라 검색할 경우 그만큼, 정도율을 향상시킬 수 있기 때문이다.

그러나 너무 고유번호에 치중하다보면, 용어의 수가 방대하여짐에 따라 분석된 용어에 대한 코드의 길이가 길어지게 되는 단점이 있고, 이것은 결국 분류번호가 길어지는 결과를 초래하게 된다. 따라서 분석된 용어의 코드화는 가능한 한 개개 용어마다 고유번호를 부여하도록 하되, 분류번호에 의한 배열 및 문헌검색적 측면을 고려하여 동의어 및 유사어 등을 한곳에 모아 코드화 하였다. 또 코드의 길이를 줄이기 위해 이미 CC표상에서 비십진 기호로서 사용하고 있을 때는 십진기호와 비십진기호를 혼용하여 사용하였다.

수집한 용어의 분석방법은 다음과 같다.

(1) 두개의 명사가 결합하여 하나의 개념을 형성하는 경우에는 분리하지 않는다.

예) air hygine M 558

(2) 동의어 및 유사어는 가능한 한 같은 기호를 부여한다.

예) human body P 1

body P 1

(3) 형용사+명사로 이루어져 하나의 개념을 형성하고 있는 경우에는 분리하지 않는다.

예) autonomic nerve P 78

(4) 형용사+복합명사의 경우에는 형용사를 제외한 복합명사만을 분석 한다.

(5) 복수의 형태는 단수로 취급한다. 단 반드시 복수형태로 사용되는 경우는 복수를 그대로 분석한다.

예) arteries -----〉 artery

citrus -----〉 citrus

(6) 데이타베이스의 개념배열상의 이해를 돕기 위하여 CC상에 표기된 용어를 그대로 코드화 한 경우도 있으며, 이러한 경우는 단지 표의 이해를 돕기위한 것이다.

예) gland of the skin

floor of mouth

(7) 분석한 용어의 개념카테고리는 의학분야는 [P], [MP], [E], [S], [T]와, 특수질병(specific disease)을 나타내는 별도의 카테고리 [D]를 생성시켜 분류하도록 하였으며, 농학은 [P1], [P2], [MP1], [MP2], [E], [2P], [2MM], [S], [T]의 9개의 카테고리로 분석하였다.

〈표 2-2〉 분류데이타베이스의 실례

KEYWORD	PME2	CON2	CONN2	NUM2	PME3	CON3	CONN3	NUM3
foetus					m			32
foliage	p			15				
food	p			3	m			573
food control					m			523
foot					p			132
foot bone					p			827
forage	p			2				
forage legume	p			261				
forehead					p			180
forearm					p			165
foreign matter	m1	4		48	m	4		48
forest	p			X				
forestry	p			X				
formosa	s			41v	s			41v
fracture					m	4		4775
france	s			53	s			53
frankincense	p			613				
front throax					p			153
fruit	p2	0	17	7				
fruit crop	p2			17				
full moon	t			13	t			13
functional disorder	m1	4		45	m	4		45
fungi					m	4		433
fungus					m	4		433
furniture					m			57197
gall bladder					p			292
gallstone					d		292	481
gangrene					m	4		476
garden	p			16				
garden legumes	p			175				
gardening	p			16				
garlic	p			1262				
gas	m1	4		488	m	4		488
gastric content					e			424
gastric fundus					p			242
general disease					m	4		41

이 데이타베이스상에 사용된 개개의 기호에 대하여 설명하여 보면 다음과 같다.

〈 농학분야 〉

◈ PME2 : 키워드가 농학분야의 용어인 경우, 농학분야의 카테고리 [P1], [P2], [M1], [M2], [E], [2P], [2M], [S], [T] 가운데 하나의 카테고리로 표시한다.

◈ VA2 : 농학분야내에서 분류표를 일정한 순서로 배열할 수 있게끔 부여한 아라비아 숫자.

◈ CON2 : [M1]과 [M2], [E]와 [2P], [E]와 [2M]의 속성이 결합하고자 할 때 결합의 가능성 여부를 결정하여 주는 제어번호. 주제분류기호와 동일한 분류기호를 갖는 용어에 대해 *(별표)로 표시하여 분류기호 합성시에 제외되도록 한다.

◈ CONN2 : 제어번호-2로서 농학분야에서는 미사용 상태임.

◈ NUM2 : 분류기호의 합성시에 사용되는 개개 용어의 기호.

〈 의학 분야 〉

◈ PME3 : 키워드가 의학분야 용어인 경우, 의학분야의 카테고리 [P], [D], [M], [E], [S], [T] 가운데 하나의 카테고리로 표시한다.

◈ VA3 : 의학분야내에서 분류표를 일정한 순서로 배열할 수 있게끔 부여하는 숫자. 현재는 미사용 상태임.

◈ CON3 : 분류기호 합성시에 합성의 가능성 여부를 제어하는 제어키로서, [M] 카테고리내에서 산부인과학에 관련된 개념은 분류기호 합성시에 [P] 카테고리의 CON3가 f인 개념과 합성하여 분류하도록 제어하며, 주제분류기호와 분류코드의 값이 동일한 경우에는 *(별표)를 표시하여 분류기호 합성시에 제외되도록 제어한다.

◈ CONN3 : 의학에서의 CONN3는 질병의 [D]카테고리와 같이 [P]카테고리와 [M] 카테고리가 결합된 형태로 나타날 때 [P]카테고리의 분류기호가 표시된다. 예) Cancer d 11 1에서 앞의 숫자 11은 CON3의 값이며, 이것은 데이타베이스상의 [P] 카테고리의 분류기호가 11 cell,tissue 의 분류기호와 일치한다. 즉 Cancer는 세포나 조직에 관한 질병이란 의미이다.

◈ NUM3 : 분류기호 합성시에 사용되는 개개 용어의 기호.

1) 丸山昭二郎,"分類作業の一致率,"情報の科學と技術 37:5(1987): 198-199.

2) 遠藤英三, "主題の把握と その表現," 圖書館界 21:3(1969):82-87.

3) K.L.Kwok,"The Use of Title and Cited Titles as Document Representation for Automatic Classification," Information Processing & Management11(1975) : 202.

4) 加藤宗厚, 件名作業, 東京: 理想社, 昭和32(1957) (이재철,주제명목록의 연구, 서울:연세대학교 도서관학과, 1959. p.80에서 재인용).

5) S.R. Ranganathan, Colon Classification, 7th ed., Bangalore: Sarada Ranganathan Endowment for Library Science, 1989.

6) Melvil Dewey, Dewey Decimal Classification and Relative Index, 20th ed., Albany: Forest Press, 1989.

7) National Library of Medicine, Medical Subject Headings : Supplement to Index Medicus 30(1989), Maryland, 1989.

8) Rama N. Vashista, "Automatic Classification : Some Latest Development, "Indian Librarian 32:2(Sept.1977):82.

9) Eric J. Hunter, Classification Made Simple, Adershot: Gower,1979, pp.7-9.

10) Ibid., pp.11-22.

11) H.D.Clifton, Business Data Systems:a Practical Guide to Systems Analysis and Data Processing, London: Prentice Hall International, 1978, p.233.

12) S.Venkataraman; A.Neelameghan, "Formation of Isolate Number by Computer Using the Devices of Colon Classification," Library Science with a Slant to Documentation 6:1(1969): 141-190.

S.Venkataraman ; A.Neelameghan, "Preparation of Schedule-on-Tape for Synthesis of Class Number by Computer," Library Science with a Slant to Documentation 6:1(1969): 130-140.

A.Neelameghan; S.Venkataraman, "Formulation of Kernel Terms for a for a Subject and Isolate Terms for a Classification Schedule for Use in the

Systhesis of Class Number by Computer," Library Science with a Slant to Documentation 6:1(1969): 71-93.

13) Carolyn A.Y. Sharif, Developing an Expert System for Classification of Books Using Micro-Based Expert System Shells, Boston Spa: British Library Research and Department, 1988.

14) 이경호, 도서분류의 자동화: 도서관학 및 정보학분야 서지분류를 중심으로, 석사학위논문. 경북대학교 대학원, 1980.

15) Peter G.B. Enser, Automatic Classification of Book Material Represented by Back-of-the-Book Index, Ph.D. Thesis. University of Sheffield, 1983, p.15.

16) S.R. Ranganathan, Colon Classification, 6th ed. New York:Asia Publishing House, 1960, pp.2.126-129.

III. 자동분류 원리의 유도

1. 자동분류 원리유도의 개요

도서관이나 정보센터등에서 서가상의 배열이나 검색이 가능한 코드화를 위한 분류의 자동화란 컴퓨터의 도움으로, 체계적으로 편성된 분류데이타베이스에서 한 도서의 내용, 주제 또는 형식에 일치하거나 유사한 분류번호를 탐색하여 그 도서에 자동적으로 분류번호를 배정하는 행위를 뜻한다.

그러나 자동분류를 위한 가장 중요한 부분은 분류데이타베이스를 만들기 전에 자동분류가 가능한 원리, 즉 분류가 행하여지는 과정을 하나의 표인 플로차트로 도식화 하여 내는 일이다. 이 플로차트로 분류과정을 도식화하기 위해서는 먼저 다음과 같은 전제조건이 충족되어야 한다.

(1) 분류의 방법은 열거식이 아닌 조합식이 바람직 하다.

(2) 각 용어마다 개개 주제분야내에서 성격과 위치가 분명하여야 한다.

(3) 조합하고자 하는 각 개념은 일정한 성격을 표현할 수 있는 기호가 있어야 한다.(CC의 P.M.E.S.T등)

(4) 개념의 조합에는 일정한 원리인 분류공식이 있어야한다.(주제 분야마다 동일할 필요는 없다.)

(5) 여러개의 개념은 조합원리에 따라 일직선상에 표현할 수 있어야 한다.

(6) 개념의 속성에 따라 주제의 인식과 분류가 함께 이루어질 수 있어야 한다.

(7) 각 주제분야의 분류공식은 하나의 일관성 있는 도표로서 나타낼 수 있어야 한다.

이상과 같은 조건이 충족 되면 각 학문분야마다 필요한 모든 용어를 수집하고 분석하여 분류용 파일을 만들어 자동분류가 가능할 수 있다.

그러나 지금까지의 많은 분류표 가운데에서도 CC는 그 원리 자체가 분석합성식에 근거하고 있기 때문에, 학문분야마다 분류공식이 다르지만 사용하는 개념들의 성격이 비교적 명확하고 또 조합의 원리가 일정하게 명시되어 있어 향후 분류 자동화의 가능성이 어떤 분류방식 보다도 높다고 하겠다.

2. 콜론분류법의 기본원리

콜론분류법(Colon Classification)은 랑가나단이 고안한 최초의 분석합성식 분류표이다. 1933년에 초판을, 1960년에 6판을 간행하였으며, 1987년에 제7판이 개정, 간행되었다.

CC는 개정될 때 마다 새로운 기호나 패시트를 추가시킴과 아울러 패시트의 조합방식도 개정하고 있다. 제6판까지만 하더라도 학문의 전문영역을 42개로 구분함과 아울러 각 학문분야에서 연구되어질 수 있는 현상, 즉 기본 카테고리를 P(Personality), M(Matter), E(Energy), S(Space), T(Time)로 구분하면서 이들 개념의 논리적인 조합순서를 PMEST로 정의 하였다. 그러나 7판에서는 M(Matter)패시트를 다시 MP(Matter-Property Isolate)와 MMt(Matter-Material Isolate)로 세분하여 표현하고 이들의 논리적인 결합방법도 6판의 규정과는 다르게 나타나고 있다.

이들 카테고리내의 개개 패시트 조합방법은 6판 이전의 5개의 카테고리 하에서

는 [P], [M], [E], [S] 및 [T] 순으로 카테고리를 조합하되 [P]앞에는 콤마(,), [M]앞에는 세미콜론(;), [E] 앞에는 콜론(:), [S]앞에는 마침점(.), [T]앞에는 아포스트로피(')를 를 부여하되 주제구분기호에 이어서 나오는 [P]패시트 앞에는 콤마를 부여하지 않는 분류방식이었다.

그러나 7판에서는 주제구분기호에 이어서 나오는 [P]패시트 앞에도 콤마를 사용하고 있다. 그리고 [MP]와 [MMt]카테고리가 있을 때는 이들 카테고리 앞에도 각각 세미콜론을 부여한다. 농학과 의학분야 분류의 특징은 다음과 같다.

가. CC의 농학분야 분류의 특징

CC 7판의 농학분야 분류표는 p.216에서 p.219에 걸쳐 있다. 분류공식은 J,[1P1], [1P2];[1MP1];[1MP2], [E][2P1];[2MM1]로서 비교적 분류기호의 조합이 타 주제분야에 비하여 복잡하게 나타나고 있다.

이들의 특성을 살펴보면 다음과 같다.

◈ J (Agriculture) : 농학의 주제분류 기호

◈ [1P1] : 개개의 식물(Plant)을 쓰임새에 따라 1. Decoration, 2. Feed, 3. Food, 4. Stimulant, 5. Oil, 6. Drug, 7. Fabric 8. Dye, Tan, 91. Adhesive, 92. Manure, 93. Vegetable, 94. Sugar producing 등으로 구분하고, 이들 개념을 다시 세구분하고 있다.

◈ [1P2] : 상기 [1P1]의 개개 식물에 대하여 각 부분(Organ)별로 구분하고 있는 카테고리로서 1. Sap, 2. Bulb, 3. Root, 4. Stem, 5. Leaf, 6. Flower, 7. Fruit, 8. Seed, 97. Whole plant 등으로 구분하고 있다.

◈ [1MP1] : 농학에 있어서 Property의 속성을 지닌 카테고리로서, 대표적인 개념을 보면, 1. Soil, 2. Manure, 3. Propagation, 4. Disease, 5. Development, 6. Breeding, 7. Harvest 등이 있다.

◈ [1MP2]: 품질(Quality)에 관련된 속성의 카테고리로서, 표상에는 2.Etiology, 3. Symptom, 4. Pathology, 7. Yield, 8. Use의 4개의 개념만 열거하고 있다.

◈ [E] : 행위(Action)에 속하는 카테고리로서, 1. Preparation, 2. Collection, 3. Application, 5. Prevention, 6. Treatment, 7. Improvement, 8. Storing ...

등이 있다.

◈ [2P1] : [E] 패시트 다음에 이어서 올 수 있는 성질의 [P]로서 Second Round Personality라고 하며, 이들 두개의 개념이 조합될 때는 연결 기호의 삽입없이 이루어진다. 표상에는 별도의 개념을 열거하지 않고 L(Medicine)의 [2P1]을 적용하도록 지시하고 있다.

◈ [2MM1] : [E]패시트 가운데 5. Prevention 과 6. Treatment의 경우에 있어서만 결합이 가능한 패시트로서, 3. Spray, 5. Dusting, 6. Injection의 3개의 개념이 있다.

나. CC의 의학분야 분류의 특징

CC 7판의 의학분야 분류표는 p.229에서 p.234에 걸쳐 나타나며, 분류시 패시트의 조합공식이 L,[1P1];[MP]:[E]로 표기하고 있다. 여기에서 이들의 내용을 보면 다음과 같다.

◈ L (Medicine) : 의학의 주제분류기호

◈ [P] (Organ) : 랑가나단이 제시하고 있는 카테고리 가운데 Personality의 속성을 가진 개념을 의미하며, 의학내에서는 주로 인간의 신체적 부위나 각종의 기관을 나타낸다.

◈ [MP] (Property) : [MP]는 Matter-Property 의 의미로서 의학 내에서는 조직학, 생리학, 산과학, 질병, 공중위생 등이 여기에 해당한다.

◈ [E] (Action) : Energy의 속성으로서 간호, 진단, 병리학, 치료학, 및 외과적 수술이 여기에 해당한다.

특히 의학분야 분류에서 특이한 점은 [P]와 [MP], [E]의 개념들은 다른 주제분야와 마찬가지로 열거하고 있으나, 특정질병에 대하여는 이미 조합된 분류번호를 배정하고 있는 점이다. 이의 조합은 [P]카테고리와 [MP]카테고리가 사전 조합된 형태로 나타나는 경우와, [MP] 4의 Disease 와 이의 세구분이 조합된 형태로 구분 (CC 6판에서는 [E] 4 Disease 하에 [2P]로서 분류하도록 하고 있음) 된다. 이에 특정질병의 분류번호의 구성을 보면, L;423 pox와 같이 분류하여 둔 경우와 L,12;26 Obesity와 같이 분류하여둔 2가지 경우로 대별된다. 위의 Pox의 경우 분류번호

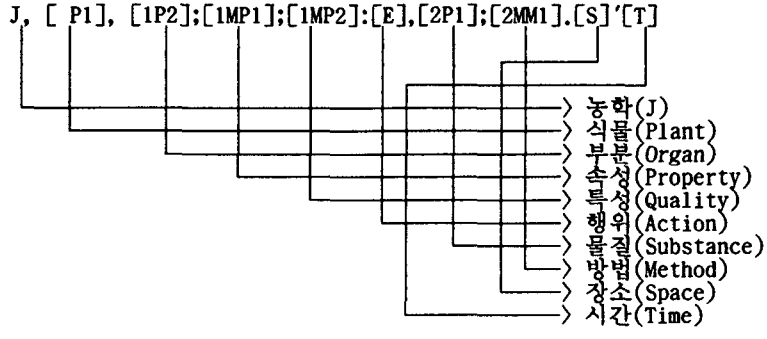
L;423은 개념의 조합이 L/MP로서 Medicine(L)/ Disease(MP)의 상세구분인 경우와 같게 나타나며, L12;26 Obesity는 조합이 L/P/MP로서 Medicine(L)/Tissue(P)/ Morphology(MP)와 같은 형태이다. 그러나 단지 특정 질병의 경우는 이들 개념을 조합하여 분류하는 형식을 취하지 않고 바로 표상에 열거하고 있는 점이 특이하다.

또 [MP]의 3은 생리학(Physiology)과 산과학 (Obstetrics)이 동시에 나타나고 있으나, 산과학의 경우에는 L9F의 여성의학 (Female Medicine)하에서 결합 되도록 하고 있어 분류시 주의가 요망되고 있는 점 등을 들 수 있다.

3. AutoBC시스템의 자동분류 원리

가. 농학분야 도서의 자동분류 원리

농학에 있어 분류는 J, [P1], [P2];[M1];[M2]:[E], [2P];[2M]으로 조합이 되게 끔 한다. ([P1]은 [1P1], [P2]는 [1P2], [M1]은 [1MP1], [M2]는 [1MP2], [2P]는 [2P1], [2M]은 [2MM1]을 간략하게 묘사한 것임) 이러한 분류공식을 그림으로 나타내면 그림3-1과 같다.



〈그림 3-1〉 농학도서의 분류시 조합방법

이렇게 분류의 과정을 하나의 도표로서 묘사할 수 있다는 점은 바로 기계적인 처리가 가능하다는 점이며, 이로 인하여 향후의 분류시스템은 수작업 방식의 분류시스템을 개선할 수 있는 가능성이 있다고 하겠다.

농학분야의 도서는 다음과 같은 과정을 거쳐서 개념의 통제를 행한다.

1) 탐색된 개념 가운데 2개 이상의 서로 다른 개념에서 키워드속성과 분류코드 값이 동일한 개념이 있는지 확인하고, 있을시는 이들중 하나의 개념을 삭제한다.

2) CON2 필드에 *(별표)가 있을시는 이 개념을 분류에서 제외한다.

3) CON2필드에 0과 4의 용어가 동시에 사용될때는 0의 개념을 분류상에서 제외한다.

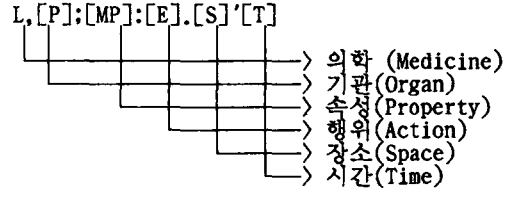
4) 개념의 전반적인 조합방법은 개념 [P2]는 [P1]이, [M2]는 [M1], [2P]는 [E], [2M]은 [E]의 개념이 나타날 때만 조합이 가능하며, 이들 선행 카테고리가 나타나지 않을 때는 뒤에 조합되는 개념들은 조합이 불가능하도록 하였다. 다만 [P2]는 [P1]이 출현하지 않으면, CONN2값으로 조합한다.

5) 위의 과정을 거치고 나서 잔류하고 있는 개념에 대한 통제는 [P1], [P2], [M1], [M2], [E], [2P] [2M]순으로 하되 최초의 개념은 2개까지 허용하도록 하고, 나머지 개념은 개념별로 하나의 개념만 허용되도록 하였다.

나. 의학분야 도서의 자동분류 원리

의학분야에서 자동분류의 원리는 [P], [M] ([MP]를 간략하게 표현한 것임), [E]의 3개의 기본 패시트와 특정질병을 나타내는 [D] 및 [S] 와 [T]의 6개의 패시트 조합으로 이루어지도록 하였다.

이러한 패시트의 배열순서 및 조합할 경우 기호사용을 그림으로 나타내면 그림3-2와 같다. ([D]패시트는 [P]와 동등한 관계이므로 표상에서 생략 하였음)



〈그림 3-2〉 의학도서의 분류시 조합방법

분류공식에 따른 패시트의 조합방법은 반드시 이러한 순서로 분류기호화 되지는 않는다. 이의 대표적인 예가 [D]카테고리가 조합된 경우이다.

한 예로서 개념의 조합이 L/P/D/E로 나타나면, 실제 분류시에는 L/P/E와 L/D/E

와 같이 2개의 분류기호가 생성되도록 하였다. 이의 주된 이유는 [D]패시트의 경우는 이미 [P]패시트까지 사전 조합이 되어 있어 [P]를 다시 조합하게 되면 분류기호가 길어질 뿐만 아니라, 이중으로 조합되는 문제가 발생하기 때문이다.

그리고 자동화원리에 의한 개개 개념의 조합방법은 도서의 표제상에서 입력된 정보를 근거로 하여, 분류데이타베이스상에서 탐색하여 해당 도서의 주제를 인식하고 난 후, 의학도서로 결정되면, 다음과 같은 과정을 통해서 탐색된 개념의 수를 통제하고 나서 분류기호를 합성한다.

1) 탐색된 개념가운데 서로 다른 2개의 개념에서 키워드속성과 분류코드가 같은 개념이 있는지 확인하고, 있을 때는 이들중 하나의 개념을 삭제한다.

2) CON3 필드에 *(별표)가 있을 때는 이 개념을 분류에서 제외한다.

3) [M]카테고리의 CON3가 0인 용어와 4인 용어가 동시에 나타날 때는 CON3가 0인 용어를 분류상에서 제외한다.

4) [P]와 [D]의 개념은 분리하여 합성하되 이하의 개념은 다같이 조합한다.

5) [M]카테고리로서 CON3가 f인 개념은 [P]카테고리의 CON3가 f인 개념과 조합이 되게끔 한다.

6) 위의 1)과 2)의 과정을 거치고 나서 잔류하고 있는 개념에 대한 통제는 개념의 우선순위를 [P], [D], [M], [E], [S], [T]순으로 하되 최초의 개념은 분류시에 절대적인 영향을 미치기 때문에 두개까지 허용되도록 하였으며, 나머지 개념은 개념별로 하나씩만 허용되도록 하였다.

IV . AutoBC 시스템의 운용

1. AutoBC시스템의 개요

AutoBC시스템은 크게 다음과 같이 도서의 표제나 키워드를 입력하는 시스템, 입력된 데이타를 근거로 AutoBC 분류데이타베이스에서 키워드를 탐색하는 시스템, 탐색된 키워드를 주제분야별 출현빈도에 따라 주제를 인지하는 시스템, 주제인지

후 주제분야별 분류기호 조합원리에 따라 분류기호를 생성하는 시스템 및 분류데이타베이스의 갱신시스템의 5개의 하부시스템으로 구성된다. 이들을 간단히 설명하면 다음과 같다.

(1) 입력 시스템: 분류하고자 하는 도서의 표제나 표제에 의한 분류가 불가능할 때 키워드를 입력하는 시스템이다. 입력의 화면은 그림4-1과 같다.

(2) 탐색 시스템: 입력된 데이타를 근거로 AutoBC 분류데이타베이스에서 해당 키워드를 탐색하는 시스템이다. 입력이 끝난 후 리턴키(Enter)를 치면 바로 탐색한다.

(3) 주제인지 시스템: 탐색된 키워드를 근거로 주제분야별 키워드의 출현빈도를 계산하여 출현빈도가 높은 주제를 해당주제로 인지하는 시스템이다. 탐색과 동시에 주제가 결정되면 화면상에 주제가 바로 나타나고, 주제인지가 되지 않으면 주제를 결정하도록 지시한다.

(4) 분류기호 생성시스템: 주제인지를 행한 후, 각 주제분야별 자동분류 원리에 따라 분류기호를 조합해 내는 시스템이다.

(5) 분류데이타베이스 갱신시스템: AutoBC 분류데이타베이스를 수정하거나, 새로운 용어가 출현하여 추가하고자 하는 경우 및 기존데이타를 삭제하거나 검색하고자 할 때 사용하는 시스템이다.

2. AutoBC시스템의 분류규정 및 분류데이타 입력방법

AutoBC시스템에서 분류는 표제에 의한 분류를 원칙으로 하되 다음과 같은 과정으로 입력하여 분류한다.

첫째, 표제만으로 완전한 주제인지와 분류기호가 생성되는 경우에는 표제만으로 분류한다.

예) world soils J;1.1
soil conservation. J;1:96
microsurgery of the glaucomas. L,185;44:71

둘째, 키워드의 출현빈도가 두 주제분야에 같은 비율로 나타나면 분류자가 주제

를 결정하여 분류한다. 이 때 분류자가 주제분야를 결정하고 나서 생성된 분류기호가 완전하다고 판단하면 표제만으로 분류한다.

예) disease, 1974. L;4'N74
nutrition and the climatic environment. J;46
mechanisms of virus infection. L;423;42

셋째, 표제만으로 분류기호가 생성되지 않을 때와 분류자의 판단에 따라 부제를 포함시켜 분류하는 것이 더 바람직하다고 판단되는 경우에는 표제에 이어 부제를 포함하여 분류한다.

예) made in washington : food policy and ... J,3.7361
the yearbook of agriculture 1969. food. J,3'N69
exploring agriculture : an introduction to food. J,3

넷째, 표제와 부제까지 포함시켜 분류하여도 분류기호가 생성되지 않을 때는 먼저 내용목차에서 키워드를 추출하여 이를 추가시켜 분류하고, 이것으로도 불가능할 때는 본문 중에서 키워드를 추출하여 분류한다.

이상과 같은 규정에 따라 그림4-1의 입력화면상에 분류데이타를 입력하되 입력방법은 다음과 같다.

1) 도서상의 표제를 그대로 입력하는 것을 원칙으로 한다.

2) 모든 단어는 전부 소문자로 입력하는 것을 원칙으로 한다. 다만 첫자를 대문자로 입력하면 대문자로 탐색하고 탐색이 되지 않으면 소문자로 탐색한다. 이것은 첫자를 반드시 대문자로 표현하여야 하는 경우를 고려한 것이며, 인명이나 지명이 여기에 해당한다.

3) 단어와 단어사이에는 반드시 한 칸의 여백을 두고 입력한다.

예) disease ˘ of ˘ greenhouse (˘는 한칸의 여백을 의미함)

4) 입력문내에 cancer, diet and nutrition... 등과 같이 단어 뒤에 이어 오는 콤마(,)는 그대로 입력하되 콤마 후에는 한칸의 여백을 두고 입력한다.

예) cancer, ˘ diet ˘ and ˘ nutrition

5) 표제에 이어 부제를 입력하고자 할 때는 부제앞의 콜론(:) 전후에 한칸의 여백을 두고 입력한다.

예) farm management ˘ : ˘ planning, ˘ control, ˘ and ...

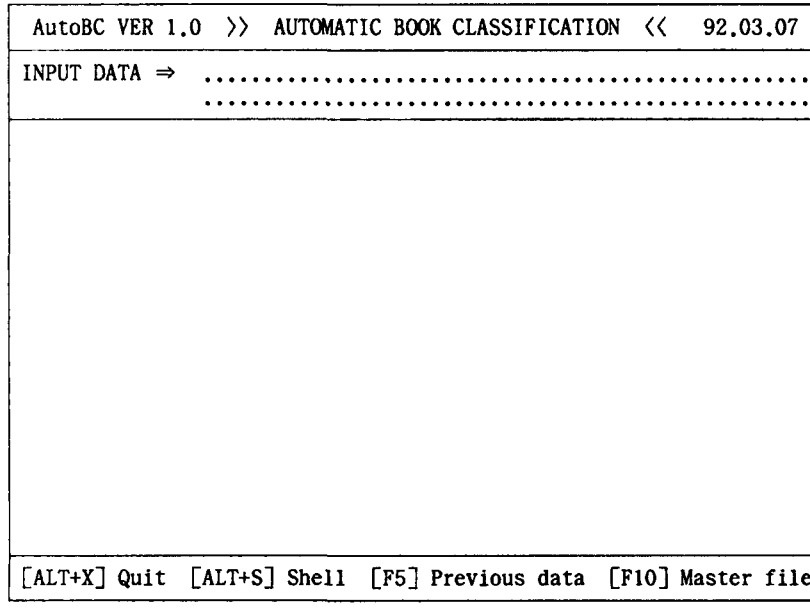
6) 복수는 단수로 입력함을 원칙으로 한다. 단순히 ..s로 끝난 단어는 복수형태로 입력하여도 무방하나, ...us가 i로된 형태나, ...um이 ..a로 된형태 및 ..y 가 ies로 된 형태와 같이 단순히 ..s로 끝난 형태가 아닌 단어는 반드시 단수로 입력한다. 예) potatoes -----〉 potato로 입력한다.

7) 지역명이나 국명은 명사형으로 입력하되 첫자를 대문자로 입력할 수 있다.

예) Korean -----〉 korea로 입력한다.

8) Shearer's 와 같이 소유격의 단어는 단어와 소유격 사이에 한칸의 여백을 두고 입력한다.

예) shearer ˘ ' ˘ s ˘ manual ˘ of ˘ human ˘ dissection



〈 그림 4-1 〉 AutoBC시스템의 입력화면구성

9) 입력문의 마지막의 마침점은 입력하여도 무방하다.

10) 표제만으로 분류가 되지 않을 때는 목차나 본문상에서 주요 키워드를 적절히 선택하여 분류자의 주관에 의하여 분류한다.

3. 데이타베이스에서의 탐색

키보드에서 입력한 표제는 프로그램에 의해 AutoBC데이타베이스상에서 용어를 탐색한다. 도서의 표제가 다음과 같이 a study of rice harvesting and storage in korea 인 경우를 예로 들어보면, 도서의 표제상에서 최대

(1) a study of

(2) a study

(3) a

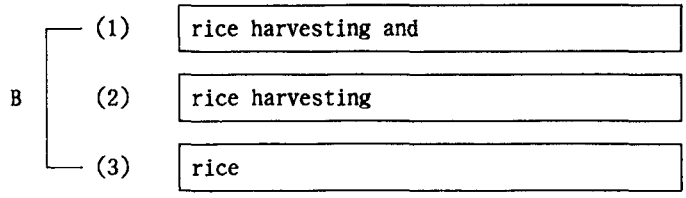
30글자 수의 범위내에서 3단어까지 읽어 (1)에, 2단어까지 읽어 (2)에, 1단어를 읽어 (3)에 옮겨 놓는다. 탐색은 (1)에서 (3)순으로 탐색하여, 어느 것이든지 먼저 탐색이 되면 메모리상에 기억시키고, 탐색된 용어의 길이만큼 도서의 표제를 앞으로 이동시킨다.

그리고 (1), (2), (3)에서 어떠한 탐색도 이루어지지 않으면 최종적인 (3)의 한 단어만큼 앞으로 이동시켜 다시 반복한다. 따라서 위의 경우는 탐색이 되지 않으므로 (1), (2), (3)에 다음 A와 같이 데이타를 옮겨 놓는다.

A
(1) study of rice
(2) study of
(3) study

이때도 역시 탐색되지 않기 때문에 위의 과정을 반복한다. 탐색이 되는 시점은

B에서와 같은 상태에서 (3)에서 최초로 탐색이 이루어 지게 된다.



이와 같은 과정을 반복하여 탐색을 행하며, 탐색방법은 다음과 같다

1) 입력된 데이타의 첫자가 소문자이면 소문자만으로 탐색하고, 대문자이면 대문자와 소문자 탐색을 병행한다.

2) 개개의 단어를 스페이스(space)에 따라 분리하여 처리한다.

3) 입력된 단어 가운데 복수형의 단어는 단수형의 단어로 먼저 검색하고, 검색이 되지 않을 때는 복수형으로 검색한다.

4) 입력문장내에서 콤마(,)가 나타나는 문장은 콤마 이전과, 콤마 이후로 분리하여 검색한다.

5) 만약 입력데이타에서 어떠한 탐색도 이루어지지 않으면, 이때는 "Not found any keyword"라는 메시지를 출력하도록 함으로써 어떠한 용어도 탐색되지 않았음을 분류자에게 지시하여 준다. 이 때는 이 용어를 분석하여, 코드를 배정한 후 분류데이타베이스에 추가시킨 다음 분류를 계속한다.

4. 주제인지 및 분류

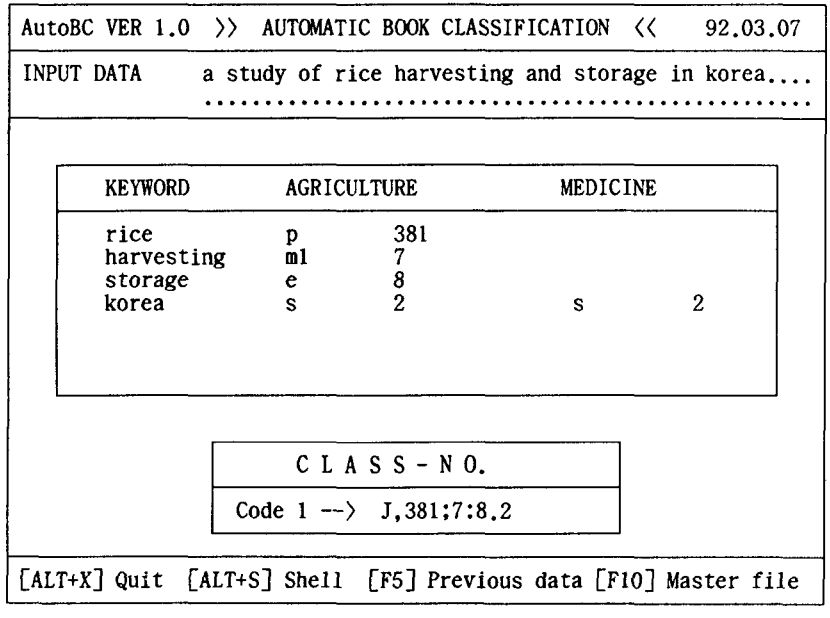
이제 이 탐색결과를 근거로 어느 주제에 어느정도의 용어가 출현하였는지를 계산하여 주제를 결정한다. 상기의 예에서는 농학에 치우치고 있어 쉽게 주제를 결정할 수 있으나, 주제의 결정이 되지 않는 경우에는 분류자로 하여금 주제를 입력하도록 컴퓨터가 요구하는 때도 있다. 이 때는 반드시 주제를 입력하여야만 분류가 가능하게된다.

이렇게 주제인식이 되고나면 분류는 매치된 이들 용어를 근거로 해당 주제분야

의 분류조합공식인 자동화원리에 의해 조합이 되게 된다. 상기 도서는 주제가 J (농학)로 결정됨으로써 J/P/M1/E/S와 같은 형태로 조합되어 분류기호는 J,381;7:8.2가 된다.

5. AutoBC 데이타베이스의 운용 및 경신

분류용 데이타베이스는 언제든지 필요한 용어에 대해 수정(edit), 추가(append), 삭제 (delete), 검색(retrieval) 및 색인(reindex)이 가능하다. 이에 대한 방법은 다음과 같다.



〈그림 4-2〉 AutoBC 시스템에 의한 분류기호의 출력 예

1) 수정(Edit): 분류용 데이타베이스는 필요한 경우에 언제든지 어떠한 부분이라도 수정, 보완할 수 있다. 여기에는 데이타베이스를 구성하고 있는 용어의 철자정정을 비롯하여, 개개 용어의 카테고리나, 분류코드 등을 수정할 수 있다. 특정 용어에 대하여 수정을 할 경우에는 분류시의 화면상태에서 [F10]키를 눌러 마스터

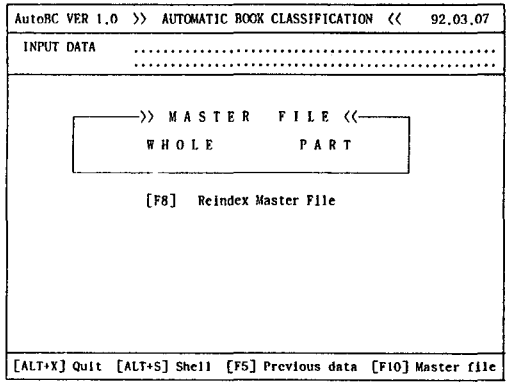
파일로 간다.

다음으로 데이타베이스내의 탐색부분을 결정한다. 전체(Whole)를 탐색하고자 WHOLE을 선택하고 부분적인 탐색을 하고자 하면 PART를 선택한다. 이때 특정부분을 탐색하기 위해 PART를 선택하면, 탐색하고자 하는 용어의 시작부분을 4자로 입력하도록 요구한다.

예를들어 food control이란 용어를 수정하고자 한다면 food를 입력하고 enter키를 누르면 화면상에는 데이타베이스상의 용어가 알파베트순으로 배열된 상태에서 food로 시작하는 용어가 그림4-4와 같이 알파벳상의 다른 용어와 함께 나타난다.

이 때 food control로 접근할 수 있으며, 필요하다면 모든 사항을 일목요연하게 볼 수 있을 뿐만 아니라, 바로 이 시점에서 수정할 수 있다. 수정의 방법은 수정하고자 하는 용어의 필드위치에 커서를 두고 enter키를 누르면 화면상에는 수정할 수 있는 화면이 나타나며, 수정 후에는 다시 enter키를 치면 수정이 끝나게 된다.

2) 추가 (Append):새로운 용어가 출현한 경우이든지, 아니면 기존의 용어가 수집 및 분석에서 제외되어 데이타베이스에 수록되어 있지 않은 용어에 대해서는 이들 용어를 데이타베이스에 수록할 필요가 있다.



〈그림 4-3 〉 [F10]키에 의한 데이타베이스 검색 초기 화면

이러한 경우에는 데이타베이스의 화면상태에서 [F6]키를 눌러 새로운 용어를 추가할 수 있다. F6키를 누르면, 그림4-4와 같이 용어를 추가할 수 있는 상태가 나타난다. 이때 키워드를 입력하고 나서 나머지 입력부분은 수정(edit)의 방법과 같다.

>> M A S T E R F I L E <<

KEYWORD	CN1	CD1	PME2	VA2	CON2	CONN2
food			p	1		
food control						
foot						
foot bone						
forage			p	1		
forage grass			p	1		
forage legumes			p	1		
fore head						
forearm	INPUT :					
foreign matter						
forest				1		
forestry			p	1		
formosa	s	41v	s			
fracture						
france	s	53	s			
frankincense			p	1		
front throax						
fruit			p2	0	0	
fruit crop			p	1		

[Paup/dn] Page skip [F4] Delete [F6] Append [ESC] Quit

〈그림 4-4〉 AutoBC 마스터파일 용어 추가 예

3) 삭제(Delete): 데이타베이스상에서 잘못 입력된 용어나 코드를 삭제하고자 할 때는 [F4]키로 삭제할 수 있다.

4) 검색(Retrieval): 용어의 검색은 크게 데이타베이스의 특정부분의 검색과, 특정용어의 검색으로 나누어진다. 특정부분의 검색은 a로 시작하는 용어, 아니면 abc로 시작하는 용어를 화면상으로 보고자 하는 경우이며, 특정 용어의 검색은 apple과 같이 오직 하나의 용어만을 완전하게 검색하고자 하는 경우이다.

AutoBC시스템에서의 검색은 검색하고자 하는 용어의 처음 4자만을 입력하여 탐색하도록 하였다. 따라서 banana를 검색하고자 하면 처음 4자인 bana를 입력하고 enter키를 치면, 화면상에는 bana로 시작하는 용어들이 사전체배열 형태로 나타나므로 화면상에서 해당 용어를 탐색할 수 있다.

>> M A S T E R F I L E <<

KEYWORD	CN1	CD1	PME2	VA2	CON2	CONN2
food			p	1		
food control						
foot						
foot bone						
forage			p	1		
forage grass			p	1		
forage legumes			p	1		
forehead						
forearm						
foreign matter			m1	3	4	
forest			p	1		
forestry			p	1		
formosa	s	41v	s			
fracture						
france	s	53	s			
frankincense			p	1		
front throax						
fruit			p	0	0	
fruit crop			p	1		

[Paup/dn] Page skip [F4] Delete [F6] Append [ESC] Quit

>> M A S T E R F I L E <<

VA2	CON2	CONN2	NUM2	PME3	VA3	CON3	CONN3	NUM3
1			3	m				573
				m				523
				p				132
				p				827
1			2					
1			25					
1			261					
				p				180
				p				165
3	4		48	m		4		48
1			X					
1			X					
			41v	s				41v
				m		4		4775
			53	s				53
1			613					
				p				153
0	0	0	7					
1			17					

[Paup/dn] Page skip [F4] Delete [F6] Append [ESC] Quit

〈그림 4-5 〉 AutoBC 마스터파일의 출력 예

V. AutoBC시스템에 의한 분류 결과 및 분석

1. 자동분류의 결과

이미 언급한 바 있는 분류데이타베이스 설계원리에 따라 데이타베이스를 구축하고, 자동분류 원리에 의거하여 프로그램을 작성한 후 농학과 의학도서를 대상으로 실제 도서의 표제를 입력하여 분류하였다. 농학관계 도서로서 조경학, 임업, 축산업 등 CC 7판에서 농학으로 취급하지 않고 별도의 학문으로 취급하고 있는 도서를 제외한 전체 479권 가운데, 중복되는 31권을 제외한 전체 448권의 도서를 분류한 결과의 일부분은 표5-1과 같다.

의학관계 도서는 총 642권으로서 이 가운데 생물학에 관련된 60권을 제외하였으므로, 실제 분류대상 도서는 582권이다. 의학관계 도서를 분류한 일부분은 표5-2와 같다.

분류기호의 배열은 분류기호속에 포함된 기호의 코드값과 개념의 조합에 사용된 기호에 따라 배열된다. 따라서 분류기호순으로 배열하기 위해서는 개념조합에 사용된 기호의 코드값을 변경할 필요가 있다. 분류기호의 배열은 분류기호를 조합할 때 사용되는 별표(*), 역방향 화살표(←), 이중아포스트로피(''), &기호, 하이픈(-), 동등기호(=), 플러스기호(+), 화살표(→), 콤마(,), 세미콜론(;), 콜론(:), 마침점(.) 및 아포스트로피(')의 값을 * < ← < ''< & < ' < . < : < ; < ,< - < = < + < →의 순으로 하되 숫자의 최저치(0)보다 작은값을 부여 한 후 배열한 것이다.

그러나 실제 분류를 행한 후 코드값의 변경없이 분류기호를 배열하면, 표상의 배열과는 다른 형태가 된다. 그 이유는 컴퓨터내의 ASCII(American Standard Code for Information Interchange)코드값에 의하면, 아포스트로피(')와 콤마(,) 및 마침점(.)은 숫자보다 작은값으로 나타나는 반면, 콜론(:)과 세미콜론(;)은 숫자보다 큰값으로 나타나기 때문이다. 이로인해 [P]패시트 앞에 부여하는 콤마(,), [M]패시트 앞에 부여하는 세미콜론(;), [E]패시트 앞에 부여하는 콜론(:), [S]패시트 앞에 부여하는 마침점(.) 및 [T]패시트 앞에 부여하는 아포스트로피(')의 값이 다

같이 숫자보다 크든지 혹은 적으면 분류기호의 배열시 관련주제가 인접될 수 있겠으나, 이와같은 상태에서는 관련주제의 인접배열이 어렵게 된다. 따라서 관련주제의 인접을 위해서는 코드값의 변환이 필요하다.

〈표 5-1〉 농학도서의 분류결과

분류기호	표 제
J;24	fertilizer manusl.
J;24	fertilizers fertilization : introduction and practical guide to crop fertilization.
J;24;1	fertilizers and soil amendments.
J;24;1	fertilizers and soil fertility.
J;24;7	fertilizing for maximum yield.
J;243;1	nitrogen in agricultural soils.
J;424:907	bacteriology for dairy students.
J;424;438	practical insect pests of temperature.
J;424;56	handbook of pests, disease, and weed of quarantine significance.
J;438	fundamentals of applied entomology.
J;438;424	aricultural insect pests of temperature.
J;438;424	biologicals insect pest suppression.
J;438;424	introduction to insect pest management.
J;46	nutrition and the climatic environment.
J;5	earth manual : a water resources technical publication.
J;5	operations research in agriculture and water resources.
J;52	irrigation principles and practices.
J;52.1	a world geography of irrigation.
J;551	solar application in agriculture.
J;56	microbial control of weeds.
J;56	weed biology and control.
J;56	weed control : a sicience.
J;56	weed control handbook.
J;56	weed science : principles and practice.
J;56	weed science : principles.
J;7	no easy harvest : the dilemma of agriculture in underdeveloped countries.
J;7	the harvest of the years.
J;7.1	fourth international symposium on preharvest sprouting.
J,09.74	agricultural development in mexcican tropics.
J,09.74	uxpanapa : agricultural development in the mexcican tropics.
J,0916	rural develpoment and human fertility.
J,092	agricultural policy in an affluent society.

〈표 5-2〉 의학도서의 분류결과

분류기호	표 제
L;4	microbial diseases : notes, reports, summaries, trends.
L;4	the hazard from dangerous exotic diseases.
L;4	the history and geography of diseases.
L;414	fever : the hunt for a new killer virus.
L;417	the control of chronic pain.
L;42	anaerobic infections in childhood.
L;42	biologic and clinical basis of infectious.
L;42	clinical concepts of infectious diseases.
L;42	principles and practice of infectious diseases.
L;42:33	biological basis of chemotherapy of infections and infestations.
L;42:56	the immunology of parasitic infections : a handbook for physicians.
L;42;4242	infectious diseases and medical microbiology.
L;421;54	tuberculosis and its prevention.
L;4221;422	syphilis and other venereal diseases.
L;423;42	mechanisms of virus infection.
L;423;42	viral and rickettsial infections of man.
L;423;42:3	diagnosis of viral infections : the clinical laboratory.
L;423;42:3	diagnosis of viral infections : the role of the clinical laboratory.
L;2430	clinical virology : the evaluation and management of human viral infection.
L;4230	essentials of medical virology.
L;4230	handbook of medical virology.
L;4230	molecular virology.
L;4230	recent advances in clinical virology.
L;4230:34	diagnostic methods in clinical virology.
L;424	bergey's manual of determinative bacteriology.
L;424;4241	bacteriology virology and immunity : for students of medicine.
L;424;4241	fundamentals of medical bacteriology and mycology.
L;4241	manual of clinical mycology.
L;4241	practical laboratory mycology.
L;4242	review of medical microbiology.
L;4242:56	current topics in microbiology and immunology.
L;4242;42	microbiology and infectious disease.

2. 자동분류 결과에 대한 분석

1) 표제에 의한 분류결과 분석

본 연구의 실험결과 표제상에 나타난 키워드의 수를 조사한 바에 의하면 표5-3과 같다. 표5-3에 나타난 주제별 도서의 키워드 수를 보면, 농학도서는 전체 448권 가운데 55.8퍼센트인 250권의 도서가 1개의 키워드를 사용하고 있었으며, 34.38 퍼센트인 154권의 도서가 2개의 키워드를, 그리고 8.48퍼센트인 38권의 도서는 3개의 키워드를 사용하고 있는 등 도서 한 권당 평균 1.56개의 키워드를 사용하고 있는 것으로 나타났다.

의학도서에 있어서는 전체 582권 가운데 1개의 키워드가 사용된 도서가 전체의 57.90퍼센트인 337권, 2개의 키워드가 사용된 것이 31.10퍼센트인 181권, 3개의 키워드가 사용된 것이 8.93퍼센트로서 52권, 4개의 키워드가 사용된 것이 0.67퍼센트로서 3권 등으로 나타나, 도서 한 권당 평균 1.55개의 키워드를 사용하고 있는 것으로 나타났다.

〈표 5-3〉 도서 한권당 키워드 출현 수

	1	2	3	4	5 -	계	평균
농 학	250	154	38	3	3	448	1.56
(비 율)	55.80	34.38	8.48	0.67	0.67	100.00	
의 학	337	181	52	10	2	582	1.55
(비 율)	57.90	31.10	8.93	1.72	0.34	100.00	
계	587	335	90	13	5	1,030	1.56
(비 율)	56.99	32.52	8.74	1.26	0.49	100.00	

본 연구에서 표제만으로 완전히 분류한 도서가 1,030권 가운데 84퍼센트인 865권이며, 주제인지가 불가능(탐색된 용어가 농학과 의학에 같은 빈도로 출현한 경우)하여 분류자가 주제결정을 행한 후 분류한 도서가 4퍼센트로서 41권, 분류데이타베이스의 키워드 부족으로 인하여 키워드 삽입후 분류한 것이 12퍼센트인 124권

이었다.

이와 같이 자동분류는 비록 표제상에 나타난 키워드에 근거한 분류이기 때문에 수작업 분류처럼 관련 주제가 밀접되도록 분류하기는 어렵다 하더라도 수작업 분류에서의 오분류를 방지할 수 있고, 신속한 분류를 할 수 있는 점, 그리고 분류데이타베이스의 개정이 쉬운점 등 많은 장점이 있으므로 수작업 분류보다 더 나은 효과를 가져올 수 있다 하겠다.

2) 주제식별의 관점에서 본 분석

컴퓨터가 도서의 표제를 근거로 주제를 자동인지 하기란 쉬운일이 아니다. 본 연구에서의 분류데이타베이스 설계에서는 단지 몇개의 용어만으로도 주제인식이 가능하다. 그러나 도서의 표제가 주제성이 희박하거나 여러 학문 분야의 용어를 사용하고 있을 때는 주제인지가 어렵게 된다. 이 때는 컴퓨터에서 주제를 입력하라는 메시지를 제공한다. 분류자가 해당주제(main class)의 기호를 입력하면 분류를 계속한다.

분류자가 컴퓨터로 하여금 자동적으로 주제를 인식케 하는 방법으로서 가장 기본적인 방법은 도서의 표제 입력이다. 이 방법은 가장 기초적인 방법이며, 이것에 의하여 주제인지 및 분류가 되지 않을 때는 목차나 본문상에서 주요 키워드를 분류자가 인위적으로 추출하여 입력하는 방법이 있을 수 있다.

주제결정 후의 분류기호는 가능한 한 표제상의 키워드에 의하여 분류하는 것이 좋다. 목차나 내용상의 키워드에 의하여 분류를 하면, 동일한 도서라 하더라도 분류자의 오분류로 인하여, 도서관에 따라 서로 다른 분류기호가 생겨나기 때문이다.

그리고 주제식별의 정확성을 향상하기 위해서는 무엇보다도 주제분야별 망라적인 용어수집과 철저한 분석에 기인한 데이타베이스의 구축이 선결과제라고 하겠다.

3) 분류기호에 대한 분석

도서의 표제(표제만으로 주제인지가 되지 않을 때는 목차나 본문상에서 입력한 키워드)에 의해 주제가 인지 되면, 다음은 분류단계로 접어든다. 분류는 데이타베이스상에서 탐색된 용어를 자동분류의 원리에 의거하여 일정한 순서로 조합함으로써 이루어 진다. 이러한 분류기호의 조합상에 고려해야 할 사항은 다음과 같다.

가. 동일 패시트가 연속 출현하는 경우

용어의 조합시에 나타날 수 있는 사항은 같은 패시트를 가진 용어가 2회 이상 출현하는 경우이다. 농학의 경우 표제가 apple and orange라면 개념의 조합이 J/P/P로 나타나 실제 분류시는 J,371,372이나, 이때는 J,372,371로 부출지시가 반드시 필요하며, 이와같은 경우는 이하의 패시트에서도 다같이 적용될 수 있다. 그러나 동일한 패시트가 자주 반복하여 나타나면, 분류기호가 길어질 뿐만 아니라 혼란이 야기되기때문에 두번이상 출현하는 개념에 대해서는 통제방법이 수반되어야 하리라고 본다. AutoBC 시스템에서는 개념의 순서를 PMEST로 볼때 최초 출현하는 개념만은 2개까지 조합이 허용되도록 하였다.

나. 분류기호에 영향을 미치는 용어 수의 제한

분석합성식 분류는 개념조합의 수에 따라 분류기호가 길어진다. 이때 하나의 도서를 기계적인 방법으로 자동처리 하기 위해서는 분류기호에 영향을 미치는 전체 개념 수의 제한 및 동일개념을 어느정도까지 제한하여 줄 필요가 있다. 한권의 도서를 분류하는데 있어, 전체 개념의 수를 몇개 이내로 할 것인지 통제가 필요하다.

여기에는 동일 개념 수의 통제와 전체 개념 수의 통제의 두가지 유형으로 구분된다. 예를 들어, 한 도서에 출현한 개념 수에서 최종적으로 분류되는 개념의 수를 5개 이내로 한다면, 이러한 경우에는 동일 개념이 4개나, 5개까지도 사용 가능하게 되며, 단지 5개를 넘는 경우에만 통제를 하는 방법이다. 그리고 동일 개념의 경우에는 위에서 언급한 바와 같이 전체 개념 수에 관계없이 동일 개념의 출현회수를 통제하는 방법이다.

본 연구의 AutoBC시스템에서는 동일개념이 두개 이상인 경우는 최초 출현하는 개념은 2회까지 출현을 허용함으로써 동일 개념의 수와 전체 개념의 수를 동시에 통제하도록 하였다. 그리고 동일 개념이 2회 사용된 경우에는 이들 개념을 순환시켜, 또 다른 분류기호를 생성시킴으로써 검색상의 검색요소를 늘리는 방향으로 설계하였다. 그러나 이와같은 개념통제의 문제는 시스템의 운영에 관한 내용인 만큼, 도서관의 분류정책 방향에 따라 수정, 보완이 가능하기 때문에 개개 도서관의 실정에 맞게 선택할 수 있는 문제라 하겠다.

3. 자동분류상의 문제점

1) 표제의 표기

도서가 무엇에 관하여 기록하고 있는지 가장 간단 명료하게 표현하고 있는 것이 도서의 표제이다. 그러나 모든 도서가 표제에 의해 자동적으로 주제인지가 되고, 분류될 수 있도록 표제를 정확히 표현하고 있지는 않기 때문에 표제에 있는 용어만으로 주제식별 및 분류가 불가능한 경우도 주제 분야에 따라서는 상당 수에 이를 것으로 생각된다.

따라서 향후 분류 자동화를 여러 주제분야에 확대 적용하기 위해서는 도서의 표제표기에 있어 상당히 신중을 기하여야 할 것으로 보인다. 표제의 표기를 가능한 한 특정성이 있는 학술적인 것으로 기술하도록 권장하되, 저자나 출판관계자 모두가 관심을 가져야할 사항이다. 단지 하나의 대안으로서는 도서를 출판할 때 ISBN (International Standard Book Number)을 부여하는 방식과 같이 서명에 이어 도서의 내용을 나타내는 몇개의 키워드를 의무적으로 표기하도록 하는 방안이 있을 수 있다. 이렇게 되면 도서관 및 정보센터 등에 있어서 자료조직 특히 분류문제에 있어서는 큰 변혁을 가져올 것으로 보이며, 나아가 BSO(Broad System of Ordering)와 같은 중개언어(switching language)를 대신할 수 있어 양자의 기능을 겸비한 진일보한 시스템이 될 수 있을 것으로 보인다.

2) 각 학문분야별 자동분류원리 정립

자동분류는 도서의 표제나 키워드 등을 컴퓨터에 입력하고, 컴퓨터내에서 이들 데이타를 근거로 주제를 인식함으로써 시작되며, 이때 분류되는 조합방식도 상이하지만, 각 개념의 속성이 분야마다 일치하지 않는다. 한 예로서 대부분의 주제분야내에서 지리적인 개념은 [S]패시트를 사용하고 있으나, 역사학(V) 에서는 분류의 조합방법이 V,[1P1];[MP1]:[E]:[2MP1]으로 [P]의 속성을 지닌 [1P1]가 지리구분에 해당하는 국가에 해당한다.

개개 학문분야별 자동분류 원리의 정립은 그 학문의 특성에 따라 별개의 분류원리를 정립하여야 한다. 원리의 정립에는 랑가나단이 제시하고 있는 개개 학문분야별 분류공식을 참고할 수 있겠으나 CC의 분류공식도 향후의 자동화분류를 위해서

는 상당한 부분을 수정, 보완하여야 하리라고 본다.

3) 종합 분류데이타베이스 구축

분류데이타베이스의 구축은 해당 학문의 특성에 따라 자동분류가 가능한 자동분류원리를 정립하고 난 후 가능하다. 본 연구에서와 같이 농학과 의학에 한정된 두 개의 주제분야가 아닌 다수의 학문분야를 대상으로 자동분류를 실현하기 위해서는 자동화 원리의 정립에 이어 광범위한 범위에 걸쳐 용어를 전문적으로 수집, 분석하고 난후, 데이타베이스를 구축하는 작업이 필요하다 하겠다.

4) 분류번호의 길이

분석합성식에 의한 분류는 종래의 열거식 분류방법에 비하여 분류번호가 길어지는 경우가 많다. 그 이유는 여러개의 개념조합으로 분류번호가 형성되기 때문이다. 그러나 대부분의 분류시스템에서 분류기호의 길이는 장서의 규모에 비례하여 길어지는 것이 보편적이지만 자동분류를 행하게 되면 출현하는 키워드의 수에 따라 길어지는 단점이 있다.

따라서 서가분류시는 분류기호의 길이를 줄이기 위하여 기계적인 방법으로 절단을 행하고 있으나 기계적인 절단보다는 의미성을 우선적으로 고려하여야 할 것으로 생각된다. 이러한 이유로 인하여 분석합성식의 분류방식에 익숙하지 못한 사람들은 분류번호가 길기 때문에 서가상에 배가하기에는 부적합 하다고 생각하고 있다.

그러나 대부분의 도서는 표제상에 두 세개 이상의 개념이 출현하는 경우가 드물기 때문에 서가상에 배가하기에 부적합한 정도의 분류기호는 그렇게 많지 않을 것으로 생각된다. 또 긴 분류번호라 하더라도 종래의 열거식분류표상의 분류기호와 같이 분류번호가 연속된 상태가 아니고, 여러 부분으로 조합되어 있기 때문에 서배(書背)상에는 계층적으로 표기하고, 배열시에는 개념단위로 배열할 수 있는 이점이 있어 해결 가능성은 있다. 이미 랑가나단은 마드라스대학 도서관 장서 30,000권을 분류한 예[17]에서도 이를 증명해주고 있다 하겠다.

17) Susan Artandi, ed., Colon Classifcation. 6th ed., New Jersey : The Rutgers University Press, 1965, p. 15.

5) 보다 복잡한 분류공식의 적용

이 연구에서는 개념의 기본 카테고리에 의거하여 분류기호를 생성시키는 방법을 취하였으나, CC상에서 언급하고 있는 공통구분기호나 열거식 분류상에 나오는 형식구분기호 등과 같은 규정을 적용할 수도 있다. 이러한 규칙을 적용하기 위해서는 적용규칙이 증가함에 따라 컴퓨터 자체의 프로그램적인 처리에만 의존하기에는 어려운 점이 수반된다. 이러한 규칙을 적용함에 있어서는 상세한 데이타베이스의 설계나 다소의 인위적인 입력조작이 필요할 것으로 생각된다. 한 예로서 입력키워드에 백과사전이란 용어로 encyclopedia란 용어가 입력될 경우 이것이 사전인지 아니면 사전에 관하여 기술한 도서인지 컴퓨터 측에서 판단하기에는 어려운 점이 많다.

따라서 이렇게 복잡한 규칙을 현 시점에서 적용하고자 하면 도서의 표제나 키워드를 입력할 때 태그(tag)를 입력하는 방법을 이용할 수도 있다. 단지 이러한 방법을 적용하면 이러한 분류시스템은 완전 자동화라기 보다는 반 자동화적인 성격을 띠게 되고, 분류는 그만큼 상세하게 분류할 수 있다. 향후의 연구는 이러한 측면에서도 필요 하리라고 본다.

VI. 결 론

본 연구에서는 전문도서관에서 도서분류자동화를 실현하기 위해 농학과 의학분야를 대상으로 관련분야의 키워드를 수집하여, 분류데이타베이스를 구축하고, 컴퓨터를 통해 도서의 표제나 키워드를 입력함으로써, 컴퓨터에 의한 자동적인 주제인지 및 분류기호의 생성이 이루어질 수 있는지에 대하여 실험을 통해 입증하고자 하였다.

연구의 결론을 요약하면 다음과 같다.

(1) 콜론분류법의 패시트원리를 이용하여 도서의 표제만으로 분류기호를 84퍼센트 자동 생성시킬 수 있다.

(2) 분류하고자 하는 도서에 대한 자동적인 주제인지는 지구의의 원리를 응용한 데이타베이스 설계와 개개 용어에 대해 연구대상 주제분야를 명시하여 둠으로써 가능하다.

(3) 입력데이타에 의한 주제의 자동적인 인지는 탐색된 용어의 주제분야별 출현회수 측정으로서 인지할 수 있다.

(4) 분류기호의 조합은 주제분야별 분류공식을 플로차트화 함으로써 이루어 질 수 있다.

(5) 분류데이타베이스상에 제어코드를 설계하여 둠으로써 분류기호를 효율적으로 통제할 수 있다.

(6) AutoBC시스템에 의한 자동분류는 단일 분야만을 다루는 전문도서관에 있어서는 실현 가능하나, 일반도서관에서 적용하기에는 다소의 문제가 있을 수 있다.

향후 정보관리기관에서 자동분류가 실제로 적용되기 위해서는 보다 많은 연구가 요망되며, 특히 다음과 같은 내용에 대한 연구가 이루어지기를 기대한다.

(1) 본 연구에서 설계한 데이타베이스 설계원리에 입각하여, 다수의 주제분야를 대상으로 데이타베이스를 구축하게 되면 이러한 주제에 대해서도 분류가 가능하기 때문에 이에 대한 연구가 계속되어야 할 것이다.

(2) 각 주제분야별로 분류자동화원리를 유도, 정립하는 연구가 필요할 것이다.

(3) 본 연구에서는 CC의 5개의 기본 카테고리만을 대상으로 분류하였으나, 향후의 보다 나은 분류를 위해서는 CC에서 사용하고 있는 공통보조기호에 의한 분류도 연구되어져야 할 것이다.

(4) 개념조합에 의한 자동분류에 있어서는 분류기호가 길어짐에 따른 개념통제 방안에 대한 연구가 이루어져야 할 것이다.

(5) 영문도서만이 아닌 한국어로 된 도서에 대해서도 자동분류의 실현 가능성에 관한 연구가 이루어져야 할 것이다.

참 고 문 헌

이 경 호, 심 의 순. "도서분류자동화 원리 유도." 圖書館學論集第 11集(1984): 32-66.

이 재 철. 주제명목록의 연구. 서울:연세대학교 도서관학과, 1959.

遠藤英三, "主題の把握と その表現," 圖書館界 21:3(1969): 82-87.

木原通夫 ; 志保田務 ; 高鷲忠美, 資料組織法, 第II版. 東京 : 第一法規, 1988. pp.14 -15.

櫻井宣隆."自動分類法に 關する 研究ノート(1)."圖書館短期大學紀要9(1975):49-58.

--------."自動分類法に關する研究ノート(2)." 圖書館短期大學紀要10(1975):59-77.

永 田 清 一. "圖書分類作業について 小考." 圖書館界 33:4(1981):192-196.

丸 山 昭二郎. "分類作業の一致率." 情報の科學と技術 37:5(1987):198-199.

Artandi, Susan, ed. Colon Classification. 6th ed. New Jersey : The Rutgers University Press, 1965.

Atherton, Pauline." Ranganathan's Classification Idea:An Analytico-Synthetic Discussion." Library Resources & Technical Services 9:4 (Fall 1965): 463-464.

Batty, C.D. Introduction to Colon Classification. Hamden :Archon books, 1966.

Coates, E.J. "Classification in Information Retrieval : the Twenty Years Following Dorking." Journal of Documentation 34:4(Dec.1978): 288-299.

Cochrane, Pauline Atherton. Redesign of Catalogs and Indexes for Improved Online Subject Access:Selected Papers of Pauline A. Cochrane. Phoenix, Arizona: The Oryx Press,1985.

Enser, Peter G.B. Automatic Classification of Book Material Represented by Back-of-the-Book Index. Ph.D.Thesis. University of Sheffield, 1983.

Farradane, J. "Concept Organization for Information Retrieval." Information Storage & Retrieval 3(1967): 297-314.

Fernandez, Cheryl Wise. "Semantic Relationships Between Title Phrases and LCSH." Cataloging & Classification Quarterly 13:1(1991): 51-77.

Harris, Kevin. " A Faceted Classification for Special Literature." International Library Review 19:4(Oct. 1987):335-344.

Hunter, Eric J. Classification Made Simple. Adershot : Gower,1979.

Kaula, P.N. " Library Science : Schedule of Class Number." Herald of Library Science 6:1(Oct.1967): 224-230.

Kumer, Krishan. The Thoery of Classification. New Delhi : Vikas Publishing House, 1979.

Kumer, P.S.G. "Internationl Study Conference on Classification Research (3)." Herald of Library Science 14:1 (Jan.1975):8-26.

Kwok, K.L. "The Use of Titles and Cited Titles as Document Representation for Automatic Classification." Information Processing & Management 11 (1975): 201-206.

National Agricultural Library. AGRICOLA Subject Catagory Codes with Scope Notes. Beltsville : National Agricultural Library, 1990.

National Library of Medicine. Medical Subject Headings; Supplement to Index Medicus V.30(1989). Maryland : National Library of Medicine, 1989.

National Library of Medicine. National Library of Medicine Classification : a Scheme for the Shelf Arrangement of Books in the Field of Medicine and its Related Sciences. 4th ed., Rev. Bethesda,Md.: U.S.Dept. of Health and Human Services, Public Health Service, National Institutes of Health National Library of Medicine, 1989.

Navalani, K.; N.N.Gidwani. A Practical Guide to Colon Classification. New Delhi : Oxford & IBH Publishing Co.,1981.

Ranganathan, S.R. Colon Classification. 6th ed. New York : Asia Publishing House, 1960.

----------. Colon Classification. 7th ed. Bangalore : Sarada Ranganathan Endowment for Library Science, 1989.

----------."Colon Classification Edition 7 (1971) : A Preview." Library Science with a Slant to Documentation 6:3(1969): 193-242.

Ray-Jones, Alan and Daviv Clegg. CI/SfB Construction Indexing Manual. 3rd rev.ed. London : Riva, 1976.

Richmond, P.A. "The Future of Classification." Drexel Library Quarterly 10:4 (1974): 105-117.

Seetharama, S. " Human Digestive System: Depth Classification Version of CC." Library Science with a Slant to Documentation 10:1(March 1975): 17-26.

----------. "Human Disease: Depth Classification Version of CC." Library Science with a Slant to Documentation 8:4(Dec.1971) : 335-368.

----------."Human Nervous System : Depth Classification Version of CC." Library Science with a Slant to Documentation 10:1(March 1973): 30-89.

----------." Human Skeletal System: Depth Classification Version of CC." Library Scienc with a Slant to Documentation 10:3(sept.1973): 344-392.

Sharif, Carolyn A.Y. Developing an Expert System for Classification of Books Using Micro-Based Expert System Shells.(British Library Research Paper 32) Boston Spa, Wetherby,Yorkshire : British Library Reaearch and Development Department, 1988.

Vashista, Rama N. " Automatic Classification : Some Latest Development." Indian Librarian 32:2(Sept.1977):73-83.

Venkataraman, S.; A. Neelameghan. " Formation of Isolate Number by Computer Using the Devices of Colon Classification." Library Science with a Slant to Documentation 6:1(1969):141-190.

----------. "Preparation of Schedule-on-Tape for Synthesis of Class Number by Computer."Library Science with a Slant to Documentation 6:1(1969): 130-140.

Yahmai, N. Shahla; Jacqueline A. Maxin. "Expert Systems: A Tutorial." Journal of American Society for Information Science 35:5(Sept.1984): 297-305.

Developing an Automatic Classification System Based on Colon Classification : with Special Reference to the Books housed in Medical and Agricultural Libraries

Abstract

Lee, Kyung-Ho

The purpose of this study is (1) to design and test a database which can be automatically classified, and(2) to generate automatic classification number by processing the keywords in titles using the code combination method of Colon Classification(CC) as well as an automatic recognition of subjects in order to develop an automatic classification system (AutoBC System) based on CC which can be applied to any research library.

To conduct this study, 1,510 words in the fields of agriculture and medicine were selected, analized in terms of [P],[M], [E], [S], [T] employed in CC, and included in a database for classification.

For the above-mentioned subject fields, the principle of an automatic classification was specified in order to generate automatic classification codes as well as to perform an automatic subject recognition of the titles included.

Whenever necessary, editing, deleting, appending and reindexing of a database can be made in this automatic classification system. Appendix I shows the result of the automatic classification of books in the fields of agriculture and medicine.

The results of the study are summarized below.

1. The classification number for the title of a book can be automatically generated by using the facet principles of Colon Classification.

2. The automatic subject recognition of a book is achieved by designing a database making use of a globe-principle, and by specifying the subject field for each word.
3. The automatic subject-recognition of input data is achieved by measuring the number of searched words by each subject field.
4. The combination of classification numbers is achieved by flowcharting of classification formular of each subject field.
5. The effieient control of classification numbers is achieved by designing control codes on the database for classification.
6. The automatic classification by means of AutoBC has been proved to be successful in the research library concentrating on a single field. The general library may have some problem in employing this system.

The automatic classification through AutoBC has the following advantages:

1. Speed of the classification process can be improve.
2. The revision or updating of classification schemes can be facilitated.
3. Multiple concepts can be expressed in a single classification code.
4. The consistency of classification can be achieved with the classification formular rather than the classifier's subjective judgement.
5. A user's retrieving process can be made after combining the classification numbers through keywords relating to the material to be searched.
6. The materials can be classified by a librarian without subject backgrounds.
7. The large body of materials can be quickly classified by means of a

machine processing.

8. This automatic classification is expected to make a good contribution to design of the total system for library operations.
9. The information flow among libraries can be promoted owing to the use of the same program for the automatic classification.