식물학문헌을 위한 자동분류시스템의 개발*

Developing an Automatic Classification System for Botanical Literatures

김정현(Kim Jeong-Hyen)** · 이경호(Lee Kyung-Ho)***

<목 차>

초 록

본 연구는 분류자동화를 위해 이미 연구된 바 있는 농학 및 의학분야의 AutoBC 시스템에 대한 계속적인 연구의 일환으로 식물학분야의 문헌에 대해 분류자동화가 가능한지의 여부를 CC의 원리를 응용하여 실험 및 검증한 것이다. 분류자동화를 위한 데이터베이스는 원통형과 행렬식의 원리에 의해 설계되었으며, 문헌의 표제나 키워드를 입력하여 자동적인 주제인지 및 분류기호가 생성될 수 있는 윈도우용 자동분류시스템을 새로이 개발하여 실험하였다.

주제어 : 분류데이터베이스, 자동분류, 자동분류시스템, 식물학분류

Abstract

This paper reports on the development of an automatic book classification system using the facet classification principles of CC(Colon Classification). To conduct this study, some 670 words in the botanical field were selected, analyzed in terms [P], [M], [E], [S], [T] employed in CC 7, and included in a database for classification. The principle of an automatic classification system is to create classification numbers automatically through automatic subject recognition and processing of key words in titles through the facet combination method of CC.

Particularly, a classification database was designed along with a matrix-principle specifying the subject field for each word, which can allow automatic subject recognition possible.

Key Words : Classification Database, Automatic Classification System, Database for Classification

---

* 본 연구는 2000년도 학술진흥재단의 협동과제인 "문헌분류 자동화시스템의 개발에 관한 연구"의 일부임. (KRF-2000-042-C00199)

** 전남대학교 문헌정보학과 조교수(jhgim@chonnam.chonnam.ac.kr)

*** 대구대학교 문헌정보학과 교수(khlee@biho.taegu.ac.kr)

· 접수일 : 2001. 11. 14 · 최초심사일 : 2001. 11. 30 · 최종심사일 : 2001. 12. 17

I. 서 론

1. 연구목적 및 필요성

도서관업무의 대부분이 자동화됨에 따라 이용자에 대한 정보봉사는 과거 어느 때보다도 신속, 정확하게 이루어지고 있다. 그러나 도서관업무중에서도 분류업무만은 아직까지도 전통적인 수작업 방법에 의존하고 있어 사서들은 분류도 다른 업무와 마찬가지로 자동분류시스템의 개발을 통하여 기계적인 처리가 이루어지기를 기대하고 있다. 즉, 종래의 수작업 분류법과 관련하여 문헌의 서가상 배가는 물론, 검색의 기능을 함께 지닐 수 있으면서 각 도서관마다 동일한 문헌에 대해서는 동일한 분류기호를 이끌어 낼 수 있으며, 또한 인간의 노력을 최소화 할 수 있는 자동분류에 대한 연구가 이루어져야 한다.

따라서 본 연구의 목적은 CC의 기본적인 분류원리를 이용하여 식물학분야 문헌을 위한 자동분류의 가능성 여부를 진단하는데 있다. 이를 위해 분류자동화가 가능한 분류데이터베이스를 설계, 구축하고 문헌의 표제나 키워드를 컴퓨터에 입력함으로써 자동적인 주제인지는 물론 표제 속에 있는 키워드를 CC의 기호조합 방식으로 처리하여 분류기호를 자동으로 만들어 내고자 하는데 있다.

2. 연구방법

본 연구는 이미 필자의 한사람이 농학과 의학[1], 문헌정보학[2] 분야를 대상으로 자동분류시스템을 설계, 실험한 바 있어, 이에 대한 계속적인 연구의 일환으로 식물학분야의 문헌에 대해 분류자동화가 가능하지의 여부를 CC의 원리를 응용하여 실험 및 검증을 하고자 한다. 이에 대한 연구는 향후 농학 및 의학문헌을 위한 자동분류시스템인 AutoBC(Automatic Book Classification) 시스템에 식물학을 추가하기 위한 사전 연구로서 이미 설계하여둔 AutoBC시스템의 골격을 그대로 유지하면서 식물학 문헌을 대상으로 분류데이터베이스를 설계한 후, 용어수집 및 분석을 하여 분류데이터베이스를 구축한다. 그리고 문헌의 표제입력을 통해 주제를 식별하여 분류기호를 생성할 수 있도록, 분류자동화 원리를 유도하고 실제 프로그램을

1) 이경호, 『콜론분류법에 바탕한 자동분류시스템의 개발에 관한 연구: 농학 및 의학전문도서관을 사례로』. 성균관대학교 박사학위논문, 1992.

2) 이경호, "문헌정보학 문헌을 위한 자동분류시스템의 개발", 『慶北大學校 文獻情報學科 創立二十周年紀念論文集』. 1994. pp. 365-422.

작성한 후 실험을 행한다. 구체적인 연구방법은 아래와 같다.

(1) 식물학 문헌에 대해 분류자동화가 가능한 분류데이터베이스를 설계한다.

(2) 분류데이터베이스의 구축을 위해 주로 CC[3]와 DDC 21[4]의 분류명사 또는 분류주기에 나타나 있는 용어를 대상으로 수집하였으며, 수집된 용어의 수는 총 670개이다.

(3) 수집한 용어는 CC 7판에서 다루고 있는 주제분야별 카테고리인 [P], [M], [E], [S], [T]로 분석한다.

(4) 분류자동화에 대한 실험은 K대학 도서관이 소장하고 있는 380권의 영문으로 된 식물학문헌을 대상으로 한다.

(5) 문헌의 표제입력(표제가 불분명할 때는 내용목차나 본문의 키워드를 인위적으로 추출하여 입력)을 통해 주제인지를 한 후, CC의 패싯 조합원리에 따라 분류가 가능하도록 한다.

(6) 시스템의 운용을 위해 분류데이터베이스를 언제든지 수정, 삭제, 추가 및 색인을 할 수 있도록 자동분류시스템을 개발한다.

3. 연구의 제한점

연구의 제한점은 다음과 같다.

(1) 본 연구는 어디까지나 실험연구인 만큼, 용어의 망라적인 수집에 의한 완전한 분류데이터베이스를 구축하기보다는 자동분류가 가능한 실험용 분류표인 식물학분야 분류데이터베이스를 설계, 구축하였으므로 모든 주제분야의 문헌이 전부 분류되게끔 시도한 것은 아니다.

(2) 본 연구는 CC의 5개 기본카테고리에 의한 패싯 공식만을 대상으로 하였으며, 공통보조기호 등은 적용하지 않았다.

II. 분류데이터베이스의 설계

1. 설계원리

3) S.R. Ranganathan, *Colon Classification, Vol. 1(Schedules for Classification).* 7th ed. revised and edited by M.A. Gopinath. Bangalore : Sarada Ranganathan Endowment for Library Science, 1987.

4) Melvil Dewey, *Dewey Decimal Classification and Relative Index.* 21th ed.. Albany : Forest Press, 1996.

우리가 일상생활에서 사용하고 있는 언어를 모아 일정한 순서로 배열한 후, 하나하나 그 발음, 의의, 용법, 어원 등에 관하여 해설한 책이 사전이다. 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 분류데이터베이스에 묘사될 때만이 자동분류가 가능할 것이다.

이러한 분류데이터베이스를 설계함에 있어서 입체모양의 지구의(globe)와 원통형(cylinder), 그리고 평면상의 행렬식(matrix) 원리를 생각해 볼 수 있다.

첫째, 우리가 일상생활에서 사용하고 있는 모든 키워드가 하나의 주머니나 둥근 공의 핵심부위 즉, 지구의의 중심부에 모두 들어 있다고 생각해 볼 수 있다. 여기에서 지구의의 위도는 용어열이고, 경도는 각 학문영역으로 볼 때 하나의 용어는 어떤 학문분야에서든지 그 속성을 지니게 된다. 실제로 용어는 핵심부위에 있기 때문에 언제든지 방향만 전환하면, 다른 주제분야의 용어로 그 성격을 변환할 수 잇다. 그리고 개개 주제분야는 그 주제분야에서 연구대상으로 하고 있는 용어만을 선택, 코드화 시킬 수 있어 각 주제분야 내에서의 특징을 묘사할 수 있게 된다.

둘째, 이러한 내용을 원통형으로 설명할 수도 있다. 원통형의 가장 중심부위 상하로 키워드가 집결(이해를 돕기 위해 알파벳순으로)되어 있다고 가정하고, 원통의 중심점을 기점으로 하여 여러 조각으로 구분하고, 각 조각부분을 하나의 학문영역으로 생각한다. 이와 같은 방법으로 모든 영역의 용어들을 각 학문의 속성에 따라 하나의 원통 속에 체계적으로 표현할 수가 있다.

셋째, 입체모양의 지구의와 원통형을 보다 알기 쉽게 하기 위해 평면상의 행렬식으로 나타낼 수도 있다. 즉, 원통형에서 중심부위에 상하로 나열되어 있는 용어를 좌측부위(행)에 나열하고, 원통의 중심을 기점으로 원형으로 배열되어 있는 모든 학문영역을 우측부위(열)에 배열하면 <그림 1>과 같다.

여기서 각 용어에 대해 랑가나단이 주장하고 있는 방법으로 속성을 분석하여 해당용어에 그 용어의 분류코드를 배정한다. 이때 하나 하나의 용어는 주제분야에 따라 그 속성이 다르게 나타날 수 있어 주제분야별로 특정의 고유값을 가지게 되며, 경우에 따라서는(예를 들면, 시대 또는 지역구분) 전 주제분야에 동일한 하나의 분류코드를 부여할 수도 있다.

랑가나단은 CC 6판의 색인부문에서 용어간의 속성을 개념카테고리와 함께 분류기호로 나타내고 있다. <그림 1>의 내용은 CC 7판에 의해 분석한 것이며, 이것을 색인 형태(CC 7판은 색인부분이 아직 출판되지 않음)로 나타내면 <그림 2>와 같다.

학문영역 / 용 어	문헌정보 (2)	생물학 (G)	식물학 (I)	농 학 (J)	동물학 (K)	의 학 (L)	화학 (E)	...	
⋮									
blood		P1, 35			P2, 35	P, 35			
cell		P1, 11	P2, 11		P2, 11	P, 11			
circulation	M1, 8				P1, 3	P, 3			
classification	P2, 5	M1, 115	M1, 115		M1, 115				
collecting	M1, 3	E, 17	E, 17	E, 2	E, 17				
flower			P2, 16	P1, 36					
food				P1, 3		M1, 573			
fungi					P1, 33				
hair		P1, 88			P2, 88	P, 88			
hormone		M2, 86	M2, 86		M2, 86		P, 86		
Korea	S, 41U1	S, 41U1	S, 41U1	S, 41U1	S, 41U1	S, 41U1	S, 41U1		
microscope		E, 18	E, 18		E, 18				
seed			P2, 18	P1, 58					
selection	M1, 31			M1, 65					
virus		M1, 423	M1, 423	M1, 423	M1, 423	2P, 23			
2001	T, P01	T, P01	T, P01	T, P01	T, P01	T, P01	T, P01		
⋮									

<그림 1> 행렬식의 원리에 의한 분류데이터베이스

색인어	색인번호
blood	I[P1], K[P2], L[P], 35
cell	G[P1], I[P2], K[P2], L[P], 11
circulation	2[M1], 8. K[P1], L[P], 3
classification	2[P2], 5. G, I, K[M1], 115
collecting	2[M1], 3. G, I, K[E], 17. J[E], 2
flower	I[P2], 16. J[P1], 36
food	J[P1], 3. L[M1], 573
fungi	K[P1], 33
hair	G[P1], K[P2], I[P1], 88
hormone	G[M2], I[M2], K[M2], E[P], 86
Korea	2, G, I, J, K, L[S], 41U1
microscope	G, I, K[E], 18. L[E], 325
seed	I[P2], 18. J[P1], 58
selection	2[M1], 31. J[M1], 65
virus	G, I, J, K[M1], 423. L[2P], 23
2001	2, G, I, J, K, L[T], P01
⋮	

<그림 2> CC 7판의 색인 예

위의 색인에서 'classification'의 경우, 문헌정보학(2)에 있어 [P2]의 속성을 지니되 그 값은 5이며, 생물학(G)과 식물학(I), 동물학(K)에 있어서는 [M1]의 속성을 지니되 그 값은 115라는 의미를 지닌다. 그리고 'Korea'의 경우 모든 주제에서 다같이 [S]의 속성으로 41U1의 값을 지니며, '2001'의 경우는 마찬가지로 모든 주제에서 다같이 [T]의 속성으로 P01의 값을 지닌다는 의미이다. CC에서 이와 같은 색인방식은 하나의 용어에 대하여 각 주제분야별 속성과 자리 매김을 분명히 하여 줌으로써 권말색인이 통합시소러스의 형태를 지니고 있는 예이다.

이와 같은 분석 및 설계원리에 입각하여 우리가 일상생활에서 사용하고 있는 모든 용어를 수집, 분석하여 데이터베이스를 구축한다면, 키워드나 표제를 컴퓨터에 입력하여 자동으로 주제를 인식하게 하고, 이에 따라 분류기호도 자동으로 조합해 낼 수 있다.

2. 분류데이터베이스의 구조

위에서 언급한 원통형과 행렬식의 원리에 따라 컴퓨터 처리가 용이한 포맷으로 자동분류 시스템을 위한 분류데이터베이스를 설계하여 보면 <그림 3>과 같다. 여기서는 편의상 식물학과 동물학으로 한정하여 데이터베이스를 구성하였지만, 필요한 경우 주제를 추가하여 원하는 분야의 분류를 할 수 있도록 데이터베이스를 확장하여 설계할 수 있다.

	Subject 1 (botany)							Subject 2 (zoology)				
X(30)	XX	X(8)	XX	XX	X	XX	X(8)	XX	XX	X	XX	X(8)
KEYWORD	CN1	CD1	PME2	VA2	CON2	CONN2	NUM2	PME3	VA3	CON3	CONN3	NUM3

<그림 3> 분류데이터베이스의 구조

· 키워드 : 분류데이터베이스상의 용어(KEYWORD)

· 제어영역 : 키워드의 카테고리나 분류코드가 두 개의 주제에 다같이 일치하는 경우 이들 값을 통제한다. 카테고리제어(CN1)는 키워드의 카테고리를, 코드제어(CD1)는 분류코드를 제어한다.

· 키워드속성 : 주제분야 용어의 속성을 기술한다(PM2, PM3)

· 배열코드 : 한 주제내의 용어만을 출력하되 분류기호순으로 배열하고자 할 때 기호의 배열값을 나타낸다(VA2, VA3)

· 제어코드 : 분류원리에 의해 분류되도록 제어하는 키(CON2, CONN2, CON3, CONN3)

· 분류코드 : 분류할 때 조합되는 분류기호(NUM2, NUM3)

3. 용어수집 및 분석

용어수집은 분류자동화를 하는데 있어 문헌의 표제나 키워드에 의한 분류기호 생성의 바탕이 되는 분류데이터베이스를 만들기 위한 것이다. 즉, 종래의 분류체계에 의한 분류는 분류표를 근거로 이루어지지만, 자동화된 시스템의 경우에는 이 분류데이터베이스를 근거로 분류가 행하여진다. 본 연구에서는 CC 7판과 DDC 21의 식물학분야를 대상으로 약 670개(이 가운데는 분류실험을 위해 동물학 용어 120개도 함께 포함되어 있음) 용어를 수집하여 분석대상으로 하였다.

수집된 용어의 분석은 용어자체가 지니고 있는 속성을 [P], [M], [E], [S], [T]의 카테고리로 분석하고 기호화함으로써, 그 용어에 대한 속성치를 CC의 분류기호로 나타내어 데이터베이스를 구축하기 위함이다. 수집한 용어 분석방법을 요약하면 다음과 같다.

① 두 개의 명사가 결합하여 하나의 개념을 형성하는 경우, 분리하지 않는다.
예) root hair → [P2] 138

② 형용사+명사로 이루어져 하나의 개념을 형성하는 경우, 분리하지 않는다.
예) structural disorder → [M1] 47

③ 모든 명사는 단수 형태로 한다. 단, 반드시 복수형태로 사용되는 경우는 그대로 분석한다.

④ 동의어 및 유사어는 가능한 같은 기호를 부여한다.

⑤ 개념 하나의 최대 문자수는 30자로 제한한다.

4. 분류데이터베이스의 구축

분류데이터베이스는 앞서 언급한 원통형과 행렬식의 설계원리에 의거하여, 입력된 용어에서 컴퓨터가 특정주제의 인식과 더불어 분류기호를 자동으로 조합할 수 있도록 설계되어 있다. <그림 4>는 분류데이터베이스의 예를 나타낸 것이다.

이 데이터베이스에 사용된 개개의 기호에 대해 설명하면 다음과 같다.

① KEYWORD : 키워드인데, 대부분 소문자 단수형으로 이루어져 있다.

② PM2 : 키워드가 식물학분야의 용어인 경우, [P1], [P2], [M1], [M2], [E], [S], [T] 가운데 하나의 카테고리로 표시한다. PM3은 키워드가 동물학분야인 경우이다.

KEYWORD	PME2	CON2	CONN2	NUM2	PME3	CON3	CONN3	NUM3
taxodonta					p1			7121
taxonomy	m1			115	m1			115
teeth					p2			214
thallophyta	p1			2				
tissue	p2			12				
transpiration	m1			36	m1			36
united ststes	s			73	s			73
variation	m1			62	m1			62
vegetation	p1	*		1				
vermes					p1			6
virus	m1			423	m1			423
winter	t			964	t			964
zoology					p1	*		1
2001	t			P01	t			P01
2000	t			P00	t			P00
1999	t			N99	t			N99

<그림 4> 분류데이터베이스의 예

예) thallophyta → [P1]

transpiration → [M1]

③ CON2 : 분류기호 합성시에 결합의 가능성 여부를 결정하여 주는 제어키이다. 학문분야에 따라 다양한 기능을 수행한다. 주제분류기호와 동일한 분류기호를 갖는 용어에 대해 *(별표)로 표시하여 분류기호 합성시에 제외하도록 한다. CON3은 동물학분야의 제어키이다.

예) vegetation → * [P1] 1

zoology → * [P1] 1

④ NUM2 : 분류기호의 합성시에 사용되는 식물학분야 용어의 기호이며, NUM3은 동물학분야를 나타낸 것이다.

예) thallophyta → 2

transpiration → 36

Ⅲ. 자동분류원리의 유도

1. 자동분류원리 유도의 개요

분류자동화란 도서관이나 정보센터에서 서가상의 배열 및 검색이 가능하도록 컴퓨터에 의해 체계적으로 편성된 분류데이터베이스에서 한 문헌의 내용, 주제 또는 형식이 일치하거나 유사한 분류번호를 탐색하여 그 문헌에 자동으로 분류기호를 조합하여 배정하는 것을 말한다.

그러나 자동분류를 위한 가장 중요한 부분은 분류데이터베이스를 만들기 전에 자동분류가 가능한 원리 즉, 분류가 이루어지는 일련의 과정을 하나의 플로차트로 도식화하는 일이다. 이 플로차트로 분류과정을 도식화하기 위해서는 먼저 다음과 같은 전제조건이 충족되어야 한다.

① 분류의 방법은 열거식이 아닌 조합식이 바람직하다.

② 각 용어마다 개개 주제분야 내에서 성격과 위치가 분명해야 한다.

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

④ 개개의 조합에는 일정한 원리인 분류공식이 있어야 한다(주제마다 동일할 필요는 없다).

⑤ 여러 개의 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 충족되면 각 학문분야마다 필요한 모든 용어를 수집하고 분석하여 분류용 파일을 구축함으로써 자동분류가 가능할 수 있다. 그런데 지금까지의 많은 분류표중에서도 CC는 그 원리 자체가 분석합성식에 근거하고 있기 때문에 학문분야마다 분류공식이 다르지만 사용하는 개념들의 성격이 비교적 명확하고, 또한 조합의 원리가 일정하게 명시되어 있어 장차 분류자동화의 가능성이 어떤 분류방식보다도 높다고 하겠다.

2. CC의 식물학분야 분류의 특징

CC는 1933년 랑가나단이 고안한 분석합성식 분류표이며, 1987년에 간행된 제7판이 최신판이다. 제6판까지만 하더라도 학문의 전문영역을 42개로 구분함과 아울러 각 학문분야에서 연구되어질 수 있는 현상 즉, 기본카테고리를 P(personality), M(matter), E(energy), S(space), T(time)로 구분하면서 이들 개념의 논리적인 조합순서를 [P], [M], [E], [S], [T]로 정의하였다. 그러나 7판에서는 M패싯을 다시 MP(matter-property isolate)와 MMt(matter-material isolate)로 세분하여 표현하고, 이들의 논리적인 결합방법도 6판의 규정과는 다르게 나타내고 있다.

CC 7판의 기본적인 분류공식은 주제,[P];[M]:[E].[S]'[T]와 같으며, 식물학분야(I)의 분류공식은 I,[1P1],[1P2];[1MP1];[1MM1]:[E].[S]'[T]와 같이 제시되어 있다. 즉, 이것은 I,[Natural

group],[Organ];[Property];[Varies with 1MP1 isolate]:[Action]과 같은 의미이며, 이들의 특성을 보다 구체적으로 살펴보면 다음과 같다.

I : 식물학(botany)의 주제분류기호

[1P1] : 식물을 자연그룹(natural group)에 따라 1. Cryptogamia, 2. Thallophyta, 3. Bryophyta, 4. Pteriodophyta, 5. Phanerogamia, 6. Gymnosperm, 7. Monocotyledon, 8. Dicotyledon 등의 8개 항목으로 구분한 다음, 총314개 세목으로 있다.

[1P2] : [1P1]의 개개식물에 대하여 각 기관(organ)별로 구분하고 있는 세목인데 11. Cell, 12. Tissue, 13. Root, 14. Stem, 15. Leaf, 16. Flower, 17. Fruit, 18. Seed 등 총15개 세목으로 구분하고 있다.

[1MP1] : 식물학에 있어서 Matter-Property의 속성을 지닌 세목인데 생물학의 [1MP1]내용을 대부분 그대로 적용하고 있다. 즉, 1. Preliminaries, 2. Morphology, 3. Physiology, 4. Disease, 5. Ecology, 6. Genetics, 7. Development 등 7개 항목으로 구분한 다음, 총83개 세목으로 구분하고 있다.

[1MM1] : 식물학에 있어서 Matter-Material의 속성을 지닌 세목인데 물질(substance)에 따라 구분하고 있으며, 생물학에서와 마찬가지로 화학분야(E)의 'E1 General chemistry'내용을 그대로 적용한다. 이는 [1MP1]의 세목에 따라 구분되므로 [1MP1]이 함께 나타나지 않으면 적용하지 않는다.

[E] : 행위(action)에 속하는 세목인데 생물학의 [E]내용을 그대로 적용하여 17. Collecting, 18. Microscopy, 182. Sectioning, 185. Fixing, 186. Mounting, 188. Staining, 195. Microphotography 등 7개 항목으로 구분하고 있다.

[S] : 지리(공간)에 속하는 세목인데 모든 주제에 적용할 수 있다. 분류표의 패싯공식 아래에 열거되어 있지 않으며, 별도의 공통세목으로 마련된 지리구분표(space isolate)를 참조하면 된다.

[T] : 시대(시간)에 속하는 세목인데 모든 주제에 적용할 수 있다. [S]와 마찬가지로 분류표의 패싯공식 아래에 열거되어 있지 않으며, 별도의 공통세목으로 마련된 시대구분표(time isolate)를 참조하면 된다.

3. 식물학분야 문헌의 자동분류원리

식물학에 있어서 분류기호는 앞서 분석한 분류공식에 따라 I,[P1],[P2];[M1];[M2]:[E]와 같이 조합이 되도록 한다. 여기서 [P1]은 [1P1], [P2]는 [1P2], [M1]은 [1MP1], [M2]는 [1MM1]을 간략하게 나타낸 것이다.

이러한 기본적인 분류공식에 근거하여 개념조합의 경우의 수는 각각의 개념이 중복없이 출현할 경우에 32개의 유형으로 나타나며, 실제 분류상에서 조합되고 있는 [S]와 [T]개념을 포함하여 조합하게 되면 경우의 수는 <표 1>과 같이 32×2×2=128로 나타나 훨씬 늘어나게 된다. 또한 실제분류에서는 각 카테고리마다 많은 세목들이 있기 때문에 조합의 경우 수는 엄청나게 증가하게 된다.

한편 식물학분야 분류특성에서 살펴본 바와 같이 [M2]의 개념은 [M1]의 세목에 따라 구분되므로 [M1]이 함께 나타나지 않으면 적용하지 않는 것으로 처리한다.

<표 1> 식물학문헌의 개념조합 수

(○: 해당개념이 있음, X: 해당개념이 없음)

번호	P1	P2	M1	M2	E	S	T	분류기호의 조합
1	X	X	X	X	X	X	X	I
2	X	X	X	X	X	X	○	I/T
3	X	X	X	X	X	○	X	I/S
4	X	X	X	X	X	○	○	I/S/T
5	X	X	X	X	○	X	X	I/E
6	X	X	X	X	○	X	○	I/E/T
7	X	X	X	X	○	○	X	I/E/S
8	X	X	X	X	○	○	○	I/E/S/T
9	X	X	X	○	X	X	X	I
10	X	X	X	○	X	X	○	I/T
11	X	X	X	○	X	○	X	I/S
12	X	X	X	○	X	○	○	I/S/T
13	X	X	X	○	○	X	X	I/E
14	X	X	X	○	○	X	○	I/E/T
15	X	X	X	○	○	○	X	I/E/S
16	X	X	X	○	○	○	○	I/E/S/T
17	X	X	○	X	X	X	X	I/M1
18	X	X	○	X	X	X	○	I/M1/T
19	X	X	○	X	X	○	X	I/M1/S
20	X	X	○	X	X	○	○	I/M1/S/T
·	·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·	·

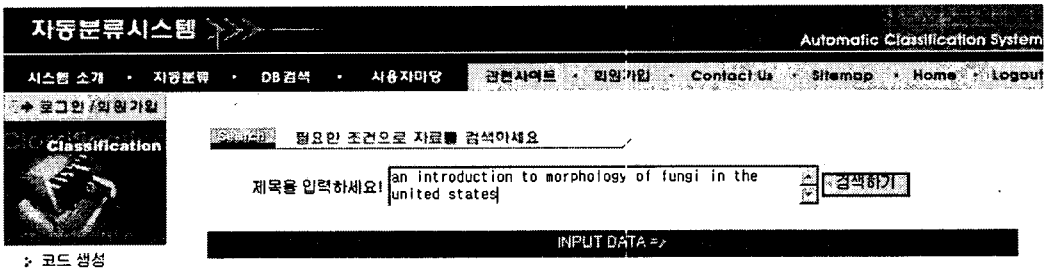
번호	P1	P2	M1	M2	E	S	T	분류기호의 조합
·	·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·	·
·	·	·	·	·	·	·	·	·
109	○	○	X	○	○	X	X	I/P1/P2/E
110	○	○	X	○	○	X	○	I/P1/P2/E/T
111	○	○	X	○	○	○	X	I/P1/P2/E/S
112	○	○	X	○	○	○	○	I/P1/P2/E/S/T
113	○	○	○	X	X	X	X	I/P1/P2/M1
114	○	○	○	X	X	X	○	I/P1/P2/M1/T
115	○	○	○	X	X	○	X	I/P1/P2/M1/S
116	○	○	○	X	X	○	○	I/P1/P2/M1/S/T
117	○	○	○	X	○	X	X	I/P1/P2/M1/E
118	○	○	○	X	○	X	○	I/P1/P2/M1/E/T
119	○	○	○	X	○	○	X	I/P1/P2/M1/E/S
120	○	○	○	X	○	○	○	I/P1/P2/M1/E/S/T
121	○	○	○	○	X	X	X	I/P1/P2/M1/M2
122	○	○	○	○	X	X	○	I/P1/P2/M1/M2/T
123	○	○	○	○	X	○	X	I/P1/P2/M1/M2/S
124	○	○	○	○	X	○	○	I/P1/P2/M1/M2/S/T
125	○	○	○	○	○	X	X	I/P1/P2/M1/M2/E
126	○	○	○	○	○	X	○	I/P1/P2/M1/M2/E/T
127	○	○	○	○	○	○	X	I/P1/P2/M1/M2/E/S
128	○	○	○	○	○	○	○	I/P1/P2/M1/M2/E/S/T

Ⅳ. 자동분류시스템의 운용 및 분류결과

자동분류시스템은 크게 문헌의 표제나 키워드를 입력하는 시스템, 입력된 데이터를 근거로 분류데이터베이스에서 키워드를 탐색하는 시스템, 탐색된 키워드를 주제분야별 출현빈도에 따라 주제를 인지하는 시스템, 주제인지이후 주제분야별 분류기호의 조합원리에 따라 분류기호를 생성하는 시스템 및 분류데이터베이스의 갱신시스템 등 다섯 개의 하부시스템으로 구성된다.

1. 분류데이터 입력방법

자동분류시스템에서 분류는 문헌의 표제에 의한 분류를 원칙으로 하여 설계되어 있으므로 표제를 그대로 입력한다. 다만, 표제만으로 분류기호가 생성되지 않을 때는 분류자의 판단에 따라 목차나 본문에서 키워드를 추출하여 이를 추가 입력한 후에 분류하며, 키워드의 출현빈도가 두 주제분야에 같은 비율로 나타나면 분류자가 주제를 결정하여 분류한다. <그림 5>는 자동분류시스템의 입력화면을 나타낸 것이며, '제목을 입력하세요'라는 메시지 다음에 분류하고자 하는 책의 서명이나 키워드를 입력하면 된다.



<그림 5> 자동분류시스템의 입력화면

2. 분류데이터베이스에서의 탐색

키보드에서 입력한 표제는 프로그램에 의해 자동분류의 분류데이터베이스에서 용어를 탐색한다. 문헌의 표제가 "An introduction to morphology of fungi in the United States"인 경우를 예로 들어 설명하면 다음과 같다.

① an introduction to, ② an introduction, ③ an

문헌의 표제에서 최대 30글자 수의 범위로 3단어까지 읽어 ①에, 2단어까지 읽어 ②에, 마지막 1단어를 읽어 ③에 옮겨 놓는다. 탐색은 ①에서 ③의 순서로 하며, 어느 것이든지 먼저 탐색이 되면 메모리에 기억시키고, 탐색된 용어의 길이만큼 문헌의 표제를 앞으로 이동시킨다. 그리고 ①, ②, ③에서 어떠한 탐색도 이루어지지 않으면 최종적인 ③의 한 단어만큼 앞으로 이동시켜 다시 반복한다. 따라서 위의 경우는 탐색이 되지 않으므로 ①, ②, ③에 다음 A와 같이 데이터를 옮겨 놓는다.

A: ① introduction to morphology, ② introduction to, ③ introduction

이때도 역시 탐색이 되지 않기 때문에 위의 과정을 반복한다. 탐색이 되는 시점은 B의 ③에서 최초로 탐색이 이루어지게 된다.

B: ① morphology of fungi, ② morphology of, ③ morphology

이와 같은 과정을 반복하여 탐색을 하며, 만약 입력 데이터에서 어떠한 탐색도 이루어지지 않으면, 이때는 'Not found any keyword'라는 메시지를 출력하도록 함으로써 어떠한 용어도 탐색되지 않았음을 분류자에게 지시하여 준다. 이때는 용어를 새로이 분석하여 코드를 배정한 후 분류데이터베이스에 추가시킨 다음, 분류를 계속한다.

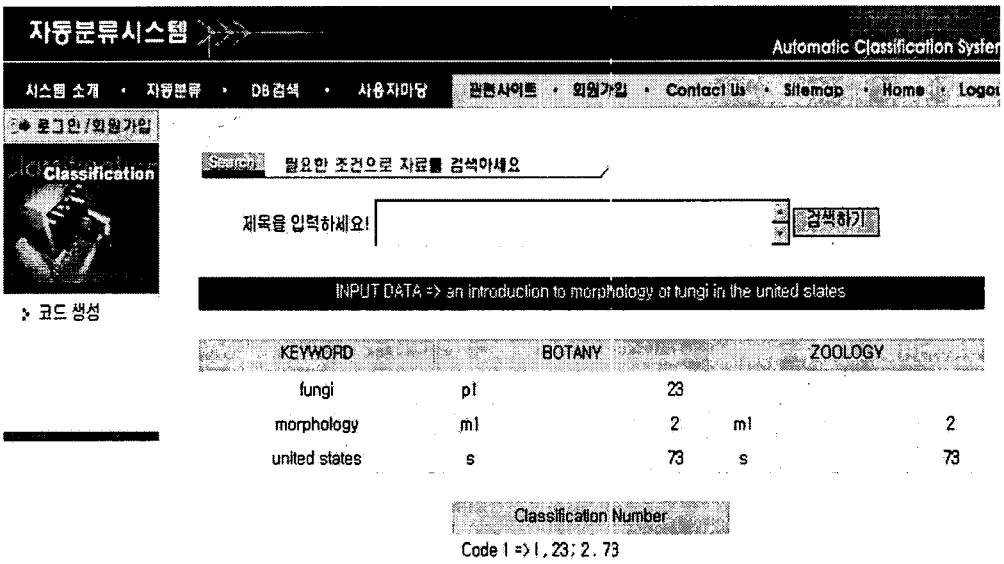
3. 주제인지 및 분류

탐색된 키워드를 근거로 주제분야별 키워드의 출현빈도를 계산하여 출현빈도가 높은 주제를 해당주제로 결정한다. 여기서는 식물학분야만을 대상으로 하고 있어 주제인지가 쉽지만

모든 주제를 대상으로 한 종합적인 분류시스템에서도 이러한 출현빈도 계산으로 주제인지가 가능하다. 탐색과 동시에 주제가 결정되면 화면상에 주제가 바로 나타나고, 주제인지가 되지 않으면 분류자로 하여금 주제를 입력하도록 컴퓨터가 요구하는 때도 있다. 예를 들어 식물학이나 동물학, 또는 생물학의 주제에 같은 빈도로 용어가 출현하였을 경우, 분류자가 최종적으로 판단한 후 하나의 주제를 선택해서 주제어를 입력해야만 세부 분류가 가능하게 된다.

이렇게 주제인지가 되고 나면 각 주제분야별 분류공식 즉, 자동분류 원리에 따라 분류기호를 조합해 낸다. <그림 6>은 문헌의 표제가 "An introduction to morphology of fungi in the United States" 인 경우를 예를 들어 나타낸 것이다. 이때 입력데이터는 'INPUT DATA'다음에 표시가 되며, 용어의 분석내용이 함께 제시되어 있어 분류내용을 확인할 수가 있다. 여기서는 키워드의 출현빈도에 따라 주제가 I(식물학)로 결정됨으로써 I/P1/M1/S와 같은 형태로 조합이 되며, 분류기호는 I,23;2.73과 같이 출력된 것이다. 그리고 키워드는 morphology(M1: 2), fungi(P1: 23), United States(S: 73)와 같이 분석이 되며, 이 가운데 'morphology'와 'fungi'에 의해 주제는 I로 인지된 것이다.

한편 표제만으로 분류가 되지 않는 경우, 내용목차나 본문의 키워드 등을 인위적으로 추출하여 입력하여도 분류가 되지 않을 때는 분류데이터베이스에 이러한 용어를 분류자가 새로이 분석하여 추가 등록해 주어야 한다.



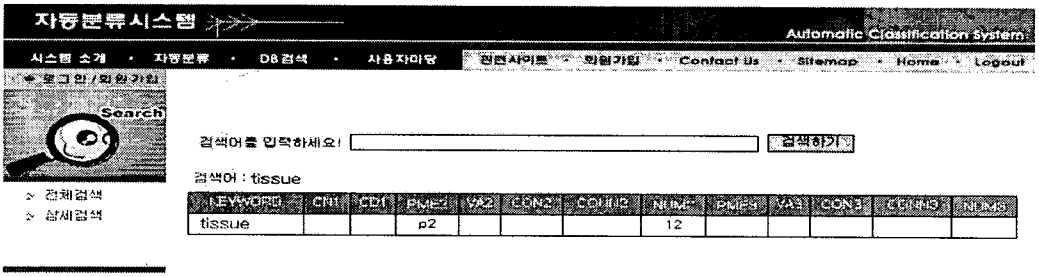
<그림 6> 자동분류시스템에 의한 분류기호의 출력 예

4. 분류데이터베이스의 운용 및 갱신

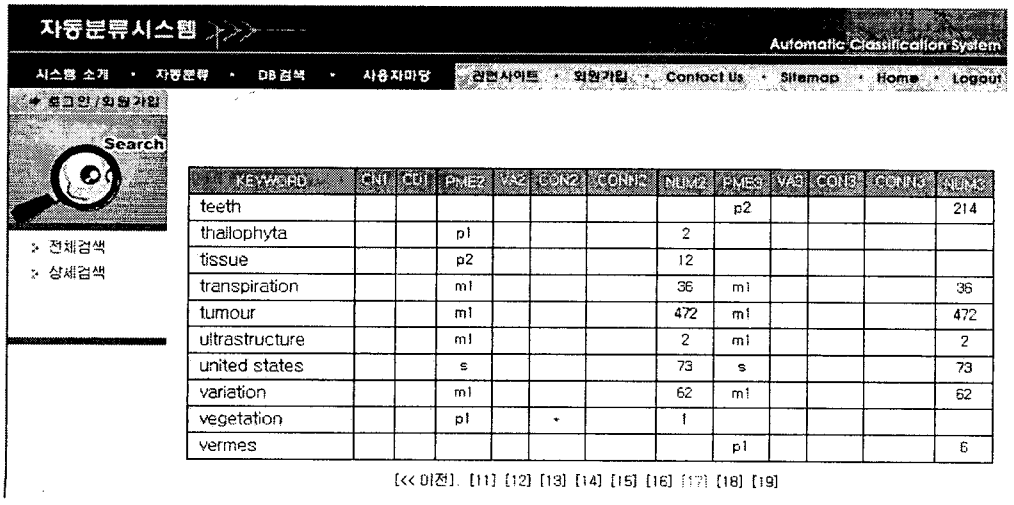
분류데이터베이스는 언제든지 필요한 용어에 대해 수정, 추가, 삭제, 검색 및 색인이 가능하다. 예를 들어 'tissue'란 용어를 수정하고자 한다면 자동분류시스템의 'DB검색' 항목을 눌러 Search화면으로 가서, '상세검색'을 누른 다음 'tissue'를 입력하고 enter키를 누르면 <그림 7>과 같이 나타난다. 이때 용어를 수정하고 enter키를 누르면 된다.

새로운 용어가 출현한 경우나 기존의 용어가 수집 및 분석과정에서 제외되어 데이터베이스에 수록되어 있지 않을 때는 [F6]키를 사용하여 용어를 추가할 수 있으며, 데이터베이스에 잘못 입력된 용어나 코드를 삭제하고자 할 때는 [F4]키를 사용하여 삭제할 수 있다. 이러한 분류데이터베이스의 관리는 분류시스템 관리자에 한해 제한적으로 허용한다.

한편 용어의 검색은 크게 분류데이터베이스에서 전체검색과 상세검색으로 나뉘어진다. 전체검색은 <그림 8>과 같이 a에서 z까지 전체용어를 10개씩 묶어 화면 하단에 아라비아숫자로 표시하여 탐색하도록 하였다.



<그림 7> 자동분류시스템의 분류데이터베이스 수정 예



<그림 8> 자동분류시스템의 분류데이터베이스 검색 예

5. 자동분류시스템에 의한 분류결과

K대학 도서관의 식물학관련 영어문헌 약 380권을 대상으로 실험하였다. <그림 9>는 식물학 문헌의 자동분류 결과의 일부 예이다. 실험결과 표제만으로 분류가 가능한 문헌이 83% 이며, 분류데이터베이스의 키워드부족으로 인하여 새로운 키워드를 삽입한 후 분류한 문헌이 17% 이다. 따라서 분류데이터베이스의 규모가 커지면 대부분의 문헌은 표제만으로 분류가 가능할 것이다.

분류기호	표 제
I	An introduction to plant biology
I	How to identify plants
I	Research methods in plant science
I,11	Introduction to plant cell culture
I,11;3;122	Cell physiology of higher plant
I,11;33E	Ion transport in plant cells and tissues
I,11:18	Electron microscopy of plant cells
I,12	Plant tissue culture in liquid systems
I,13	Methods of studying root systems
I,13;3	Physiology of the plant root system
I,13;4	Epidemiology and management of root diseases
I,13;7	Limitations to plant root growth
I,15	Progress in leaf protein research
I,18;2	Principles of seed pathology
I,18;4	Plant stems; physiology and functional morphology
I,23	Fungi in vegetation science
I,23,11	Introduction to cell culture of fungi
I,23;2	Morphology of plants and fungi
I,23;4	The fungi which cause plant disease
I,23;44	Poisonous plants and fungi: an illustrated guide
I,23;567	The biology of symbiotic fungi
I,23;66	Evolutionary biology of the fungi
I,23;675	Genetics of sexuality in higher fungi
I,23;7	Biotechnology of fungi for improving plant growth
I,23.73	Fungi on plants in the United States
I,237:195	Introduction to smut and rust fungi slide set
I,2372	Rust fungi of cereals, grasses and bamboos
I,2372;2	Ultrastructure of rust fungi

I,2372;2.73	Morphology of rust fungi in the United States
I,27	How to know the lichens
I,8,15;33	Metabolism of dicotyledon leaves
I;115	An introduction to plant taxonomy
I;33;85	Protein metabolism in the plant
I;33;150	Nitrogen metabolism in plants
I;331;140	Carbon dioxide assimilation in a higher plant
I;62	Plants variation and classification
I;62:17	Collecting plant genetic diversity; technical guidelines
I;643	Plant breeding reviews
I;643;85	New approaches to breeding for improved plant protein
I;643.8T	Plant breeding in New Zealand
I;66;b	Heavy metal tolerance-plants; evolutionary aspects
I;67	A color atlas of plant propagation and conservation
I;7;122	Calcium in plant growth
I;7;86	Plant hormones and their role in plant growth and development
I:17	Plant collecting and documentation field notebook
I:188	Staining technique in botany

<그림 9> 식물학 문헌의 자동분류 예

V. 결 론

본 연구에서는 문헌분류 자동화를 실현하기 위해서 식물학문헌을 대상으로 분류데이터베이스를 구축하고, CC에 의한 자동 분류원리를 유도하여 문헌의 표제나 키워드를 입력함으로써 자동적인 주제인지 및 분류기호가 생성될 수 있는지에 대하여 실험을 통해 입증하고자 하였다.

연구의 결론을 요약하면 다음과 같다.

(1) CC의 패싯원리를 이용하여 문헌의 표제만으로 분류기호를 자동생성할 수 있다.

(2) 분류하고자 하는 문헌에 대한 자동적인 주제인지는 원통형과 행렬식의 원리를 응용한 데이터베이스설계와 개개 용어에 대해 주제분야를 명시하여 둠으로써 가능하다.

(3) 분류데이터베이스의 설계를 위해 식물학분야의 용어 약 670개를 수집하여 CC의 카테고리별로 분석하였으며, 패싯 분류공식에 따라 식물학분야의 기본 카테고리별 조합의 경우

수는 개념이 중복 없이 출현할 때 128개로 나타났다.

(4) 입력데이터에 의한 주제의 자동적인 인지는 탐색된 용어의 주제분야별 출현빈도 측정으로 인지할 수 있다.

(5) 식물학 관련 380권의 영어문헌을 대상으로 실험한 결과 표제만으로 분류가 가능한 문헌이 83%이며, 분류데이터베이스의 키워드부족으로 인하여 새로운 키워드를 삽입한 후 분류한 문헌이 17%로 나타났다.

한편, 향후 자동분류가 실제로 적용되기 위해서는 각 주제별로 분류공식에 따라 분류데이터베이스를 구축하고, 조합의 원리를 정립해야 하며, CC의 5개 기본 카테고리뿐만 아니라 공통보조기호에 의한 분류도 함께 연구되어져야 할 것이다.

참 고 문 헌

이경호. 『콜론분류법에 바탕한 자동분류시스템의 개발에 관한 연구: 농학 및 의학전문도서관을 사례로』. 성균관대학교 박사학위논문, 1992.

이경호. "문헌정보학 문헌을 위한 자동분류시스템의 개발", 『慶北大學校 文獻情報學科 創立二十周年紀念論文集』. 1994. pp. 365-422.

丸山昭二郎. "分類作業の一致率",《情報の科學と技術》Vol. 37, No.5(1987). pp. 198-199.

Cosgrove, S.J. and Weiman, J.M. "Expert System Technology Applied to Item Classification", *Library Hi Tech*, Vol. 10, No.1(1992). pp. 33-40.

Dewey, Melvil. *Dewey Decimal Classification and Relative Index*. 21th ed. Albany : Forest Press, 1996.

Endres-Niggemeyer, Brigitte. "Knowledge Based Classification Systems: Basic Issues, a Toy System and Further Prospects", *International Classification*, Vol. 16, No.3 (1989). pp. 146-156.

Ishikawa, Tetsuya. "The Man-Machine Interface Aspect of an Automatic Classification Numbering System in a Computerized Library System", *Journal of Information Processing*, Vol. 11, No.3(1988). pp. 199-205.

Ranganathan, S.R. *Colon Classification, Vol.1(Schedules for Classification)*. 7th ed. revised and edited by M.A. Gopinath. Bangalore : Sarada Ranganathan Endowment for

Library Science, 1987.

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

Sharif, Carolyn A.Y. *Developing an Expert System for Classification of Books Using Micro-based Expert System Shells.* Yorkshire : British Library Research and Development Department, 1988.

Valkonen, Pekka and Nykanen, Olli. "An Expert System for Patent Classification", *World Patent Information*, Vol. 13, No.3(1991). pp. 143-148.

Vashista, Rama N. "Automatic Classification: Some Latest Development", *Indian Librarian*, Vol. 32, No.2(1997). pp. 73-83.

Venkataraman, S. and Nelameghan, A. "Formation of Isolate Number by Computer Using the Devices of Colon Classification", *Library Science with a Slant to Documentation*, Vol. 6, No.1(1969). pp. 141-190.

Venkataraman, S. and Nelameghan, A. "Preparation of Schedule-on-Tape for Synthesis of Class Number by Computer", *Library Science with a Slant to Documentation*, Vol. 6, No.1(1969). pp. 130-140.