웹문서분류체계의 분석 및 새로운 설계 *

Analysis and Design for the System of Korean Web Document Classification

남 영 준(Young-Joon Nam) **

목 차

초 록

인터넷에 존재하는 웹문서와 사이트들은 충분히 학술적 가치를 갖고 있기 때문에 중요한 정보원으로 간주된다. 도서관은 이 새로운 정보원을 대상으로 도서관 이용자를 위한 새로운 검색기법과 관리기법을 개발할 필요가 증대되었다. 왜냐하면 현재 웹검색엔진에서 제공하는 분류체계는 도서관학적 관점에서 개발되지도 않았으며 또한 웹검색엔진간 분류체계의 설계원칙도 없기 때문이다. 본 논문에서는 이점에 착안하여 웹문서를 효율적으로 검색할 수 있는 실험적인 새로운 웹문서분류체계를 설계하였다. 설계는 해당 분류항목과 연관된 웹문서의 수와 접속비율에 근거하였으며, 설계의 수준은 1차적으로 류·강항목까지 제한하였다.

ABSTRACT

Because of a rapid increase of information available through web site, a user often falls into confusion of which web sites should be visited for his information needs. If a web site search engine can classify web sites according to their subject or topics, it can help the user to determine which web sites are worth accessing and thus to easily acquire relevant information. In this study, I propose new classifying system with a two level hierarchy and 57 items.

* 이 연구는 1997년도 전주대학교 인문과학연구소의 지원에 의해 이루어졌음.
** 전주대학교 문헌정보학과
접수일자 1998년 9월 2일

1. 서 론

전통적으로 도서관에서 분류라 함은 그 관점을 문헌분류로 제한하여 도서관에서 입수한 공통개념들을 추출하고, 기호를 부여하고, 도서번호를 조정하는 것을 의미하였다. 이러한 분류작업은 많은 시간과 노력을 필요로 하기 때문에 현대와 같이 문헌자료를 비롯한 각종 정보들이 기하급수적으로 급증하는 시대에 오히려 장서구입에 소요되는 비용보다 이를 정리하는데 소요되는 인건비가 높은 실정이다.

한편, 사서들은 인터넷이라는 전혀 예기치 못한 거대한 데이터베이스의 출현으로 정보봉사범위를 인터넷까지 확대할 것인지 또한 이를 도서관 장서데이터베이스외의 통합된 데이터베이스로 간주할 지에 대해서도 적정한 해결책을 제시하지 못하고 있다. 통합된 데이터베이스나 혹은 별개의 데이터베이스로 간주하더라도 도서관 정보봉사범위에 인터넷 자료들은 반드시 포함되어야 할 것이다. 이러한 당위성은 도서관에 소장된 자료와 인터넷에 등재된(up loading) 자료는 각각의 특성 때문에 검색방법을 새롭게 개발해야 한다는 문제가 제기되며, 기존의 참고봉사방법과 검색방법으로는 이 두 개의 데이터베이스를 한꺼번에 운영할 수 없는 상황에 이르렀다. 즉, 도서관들은 전혀 새로운 차원의 검색이론과 방법을 개발할 필요성에 직면한 것이다. 왜냐하면, 인터넷에 존재하는 자료들을 검색할 수 있는 일부 웹검색엔진들이 웹문서의 효율적인 검색을 위해 자체적으로 개발한 분류 체계를 별도로 운영하고 있기 때문이다(Chen, H; 1998). 또한, 웹검색엔진의 분류체계들은 각기 다른 구조를 갖고 있으며, 구조의 원리는 문헌분류체계에 적용된 전개원칙을 일부 수용하고 있다. 따라서 현재 웹상에서 서비스를 제공하고 있는 웹검색엔진들은 도서관에서 사용하고 있는 분류표와는 다른 메뉴체계를 사용하고, 또한 각각의 검색엔진메뉴체계도 서로 다른 분류구조를 갖고 있기 때문에 이용자들은 디지털 검색에 많은 어려움과 혼란이 야기되고 있다.

이에 본 연구는 현재 도서관에서 사용하고 있는 분류체계의 적절성을 측정하고, 인터넷상에서 제공되는 웹검색엔진의 분류체계의 적절성도 측정한다. 측정결과에 근거하여 향후 디지털 시대에 문헌분류와 웹문서(web document) 분류에 호환성을 갖는 역동적 분류체계(dynamic classification system)를 제시한다. 측정방법은 크게 두가지 방법을 사용한다. 첫째, 도서관분류체계에 따라 정리된 문헌자료의 양을 조사한다. 둘째, 국내 웹검색엔진의 문서분류체계를 분석하여 각 항목에 연결된 웹문서의 수와 각 항목의 접속이용비율을 측정한다.

2. 문헌분류체계의 분석

분류의 목적은 크게 다음 두가지로 대별될 수 있다. 하나는 학문분류로서 학문 자체의 분류 및 사물이나 개념상호간의 관계를 발견하는 수단으로 사용하는 것이다. 다른 하나는 문헌분류로서 자료(문헌)의 효과적 이용을 위해 체계적으로 정보를 배치하는 도

구로서 활용하는 것이다. 일반적으로 문헌정보학분야에서 분류라 함은 후자의 경우에 속한다. 문헌분류표 가운데 지금까지 사용되는 분류법으로 듀이십진분류표(DDC), 미의회도서관분류표(LCC) 및 콜론분류표(CC)를 들 수 있다. 그 가운데 현대의 일반적인 문헌분류법에 있어서 DDC는 '십진식 전개'라고 하는 독창성과 이에 따른 분류작업의 편리성 및 실용성 때문에 현재 세계적으로 널리 사용되고 있다. 그러나 DDC는 신학대학인 앰허스트 대학도서관의 장서를 분류하기 위해서 편찬된 것으로서, 그 류강목 배열이 인문과학 위주로 고안되었으며, 또한 미국과 유럽위주로 전개된 특징을 갖고 있다. 그후 학술과 문화의 발전에 따라 DDC는 여러 학자들에 의해서 계속적으로 연구되고 개정되어 1998년 현재 21판까지 발행되었다. 개정판이 발행될 때마다 그 개정의 내용은 주로 세목의 세부전개와 특정한 몇 개 강목의 개정에 불과한 것이었으며, 당초의 류강목의 기본골격은 그대로 유지되고 있다. 따라서 그 동안 DDC의 분류체계에 대해서는 많은 논란과 비판이 지속되어 왔던 것은 주지의 사실이며, 현재의 관점에서는 특히 DDC의 기본 골격자체에 문제점이 더욱 크게 부각되고 있다(정필모). 이러한 사실은 대부분 문헌정보학자들이 모두 지적하고 있으나, DCC의 편이성 때문에 이를 모본으로 하여 개발한 한국십진분류표(KDC)를 국내 대부분 도서관들이 사용하고 있다.

이렇게 십진분류표가 도서관장서의 분류를 위해 널리 사용되고 있음에도 불구하고 이를 웹문서분류체계에 적용하지 못하는 것은 다음과 같은 한국십진분류표의 특성에 기인한다.

① 체계의 고정성

KDC는 류 및 강·목·세목 모든 항목이 철저하게 10진 전개를 원칙으로 하고 있다. 이러한 원칙은 전개의 단순성으로 인해 사서와 같이 분류표를 운영관리하는 관리자 입장에서는 편리할 수 있으나, 새로운 학문의 분화가 많은 현대사회에 검색자(이용자)입장에서는 해당 분류시스템으로 데이터를 찾기에는 불편함이 초래될 수 있다.

일반적으로 도서관 장서를 분류함에 있어 기존분류체계의 최소 수정으로 자관특성에 맞게 활용될 수 있으나, 인터넷에 등재된 자료를 분류함에 있어서 기존분류체계를 활용하는 것은 어려운 실정이다. 왜냐하면, 인터넷에 등재된 많은 웹문서들의 주제는 대부분 새로운 주제분야의 것이 많기 때문이다. 예를 들면, '홈페이지'라는 개념은 자서전이나 혹은 컴퓨터의 하부주제로 분류될 수도 있지만 KDC의 주제전개표에 '홈페이지'개념이 없기 때문에 세목이하에 배열되어야 한다. 한편, 현재 웹문서분류체계에서는 '홈페이지'라는 개념은 주요한 개념으로 간주하여 대부분 류·강수준에 배열하고 있다. 즉, 이 분야는 웹검색엔진의 분류체계에서 매우 일반적이고 실용적인 주제로서 류·강수준에 배정되어 있으나, 문헌분류체계에서는 세세목이하의 수준에 배정된다. 따라서 인터넷 자원을 기존 도서관의 십진분류체계로 분류하기에는 너무 고정적인 형태라 할 수 있다.

② 장서의 불균형성

KDC를 사용하고 있는 도서관에서는 류항목을 비롯한 모든 항목의 배열에 있어 각 항목에 속해있는 장서구성의 균형이 심하게 편중되어 있다. 이러한 불균형적인 구성에도 불구하고 도서관들은 KDC의 기본 골격을 혁신적으로 개편하지 못하고 있다. 왜냐하면, 이미 정리된 장서를 새로운 분류체계로 재분류해야한다는 것은 막대한 인적 경제적 부담이 수반될 수 있기 때문이다.

이러한 불균형적인 장서구성현상을 확인하기 위해서 3개대학을 임의로 추출하여 각 대학의 류별 장서배정을 분석한 결과 3개대학이 모두 '사회과학' 분야의 장서가 가장 많이 배정되었으며, '어학' 분야에 가장 적은 장서가 배정되었다〈표 1〉. 10개의 류분야에는 10%내외의 장서가 배정되어야 이상적이나 본 조사에서는 최대 분포비(사회과학류)와 최소분포비(어학류)와의 차이가 25.50%에 달해 사회과학분야의 장서가 어학분야의 장서에 비해 약 70배의 장서배정을 확인할 수 있었다. 또한, 사회과학의 경우 강체계를 비교할 경우, 최대로 배정된 분야와 최소한 배정된 분야간의 편차가 9.55%(320강과 390강)에 달해 강체계간 장서불균형도 확인할 수 있었다.

③ 계층간 불균형성

KDC를 비롯한 열거식 분류표는 개념의 열거와 고정배열을 기본으로 하고 있다. 따라서 개념의 열거를 확정짓는 시기에 존재하지 않았던 개념이나 주제는 KDC의 항목으로 채택되지 않고 있다. 이후 새롭게 생성되는 개념과 주제에 대해서는 세목과 세세목이하의 분류기호를 부여하여 새로운 주제나 개념을 확장수용하고 있다. 따라서 새로운 지식이나 주제가 실시간별로 생성되는 인터넷의 특성을 기존 열거식 분류체계로 수용하기는 현실적으로 매우 어렵다. 이러한 예의 대표적인 것으로는 전기공학, 전자공학 및 컴퓨터공학을 들 수 있다. 컴퓨터 공학은 초기 KDC가 구조화될 때 존재하지 않은 주제였

〈표 1〉 사회과학/어학 강분야의 장서배정비율

사회과학	장서분포율	어학	장서분포율
300	1.05	700	0.64
310	0.46	710	0.82
320	9.78	720	0.27
330	3.08	730	0.56
340	2.21	740	1.08
350	1.03	750	0.12
360	2.89	760	0.10
370	4.64	770	0.02
380	0.50	780	0.00
390	0.23	790	0.07
계	25.87	계	0.37

으나 이후 이와 관련된 자료들이 오히려 상위항목인 전기공학에 비해 정보생산성이 높아져 장서수에 있어서도 우위를 점하게 되었다. 특히, 컴퓨터공학은 현재에도 계속 세분화되고 생산성도 높기 때문에 이러한 계층간의 불균형성이 점차 심화될 것으로 판단된다.

이는 지식의 발달에 따라 학문이 세분되어 발전되기 때문에 일반적으로 母學問보다 子學問이 발전하는 것이 일반적인 현상이기 때문에 향후 계층간 불균형성 문제가 야기될 것이다. 따라서 기존의 문헌분류체계를 웹문서분류체계로 적용하는 것은 불합리하다고 할 수 있다.

3. 웹문서분류체계의 분석

인터넷을 통해 얻을 수 있는 서비스 종류는 매우 다양하다. 특히, 정보를 입수하는 방법으로 구분하면 첫 번째는 텍스트 형태의 데이터를 배치방식으로 입수할 수 있는 것으로서 원격접속방식(telnet)과 파일전송프로토콜(ftp : file transfer protocol), 아키, 웨이즈, 고퍼 등이 있다. 이 서비스는 인터넷망을 이용하여 텍스트 형태의 정보를 직선식 메뉴체계를 이용하여 사용자가 원하는 정보를 얻는 방식을 의미한다. 두 번째는 멀티미디어 형태의 데이터를 온라인 실시간방식으로 입수할 수 있는 것으로서 웹(World Wide Web)이 있다. 이 서비스는 인터넷을 통해 멀티미디어자료를 열람하고, 실시간으로 음악이나 동영상같은 자료도 열람할 수 있으며 필요에 따라서는 해당 데이터를 자신의 컴퓨터로 복사할 수 있다.

통신망과 관련기술이 발전됨에 따라 도서관에서 제공하였던 텍스트 위주의 배치방식에서 인터넷을 통해 실시간 방식으로 멀티미디어자료를 제공할 수 있게 되었다. 인터넷상에서 실시간 검색방식과 멀티미디어데이터의 입수를 가능하도록 한 것은 웹브라우저와 웹검색엔진(web search engine)이다. 특히, 웹검색엔진은 인터넷이라는 거대한 데이터베이스에 연결되어있는 모든 형태의 데이터를 검색할 수 있도록 하는 새로운 검색도구이다 (남영준 1).

3. 1 정보탐정

정보탐정(http://www.idetect.com/)은 한국통신 멀티미디어 연구소에서 개발한 웹검색엔진이다. 이 엔진은 1994년 국내에서 자체개발된 한국어 상용 정보검색엔진이며, 1996년 3월에 서비스를 개시하였다. 그 후 1997년 6월 대폭 개량된 최신 버젼의 정보검색엔진을 탑재하고 각종 서비스를 추가한 정보탐정 II를 개발하여 서비스하고 있다. 정보탐정은 언어에 종속되지 않고 전세계 문서를 모두 수용할 수 있으면서도 빠르고 정확한 검색이 가능한 검색엔진을 사용하고 있다. 특징으로는 웹문서 링크에 있어 로봇을 사용하고 있으며, 로봇은 다중 프로세스간의 통신을 통한 병렬적인 문서수집을 수행한다. 문서 수집 로봇은 특히 문서들간의 URL링크에 역링크를 내부적으로 생성하여 문서들간의 양방향 지도(map)를 만들고 있다. 연결된 문서에 대해 최소의 수작업을 통하여 일

부 웹문서 연결과정을 처리하고 있다.

3. 1. 1 정보탐정의 분류체계

정보탐정은 1998년도 5월 현재 다음과 같이 12개의 류항목과 32개의 강항목으로 분류체계가 운영되고 있다. 또한, 목분야와 세목분야, 세세목분야로 주제크기와 중요도에 따라 체계가 전개되고 있으며, 현재도 계속해서 목이하분야가 추가되고 있다. 다음은 정보탐정의 분류체계가운데 류강분야까지 분류항목들을 조사한 것이며 각 분류항목에 연결된 웹문서와 사이트이 수를 조사하였다 〈표 2〉.

〈표 2〉에서 분류번호란은 주제에 대응하는 KDC분류번호이며, 괄호안의 숫자는 98년 5월 현재 해당 주제아래에 링크된 웹문서와 사이트들를 합한 개수이다. 류주제와 링크된 문서의 수와 강주제에 해당하는 링크의 수가 차이가 나는 것은 웹의 특성상 목, 세목이하의 메뉴체계에 해당하는 웹문서의 수가 실제 류강항목에 연결된 문서의 수에 포함되어 있지 않기 때문이다.

3. 1. 2 정보탐정의 분류체계 분석

정보탐정의 류수준의 개념을 KDC의 분류항목과 비교하면 유사한 주제개념이 부여된 것은 '언론·매체'를 비롯하여, '경제·산업', '영화·음악·연예', '건강·병원', '생활·주택', '정치·행정·법' 등을 들 수 있다. 이 가운데 KDC의 류수준과 정보탐정의 류수준이 일치하는 것은 거의 없으며, 이를 강수준까지 확대하여 류수준의 주제와 강수준의 주제가 일치하는 것은 '경제·산업' 분야를 비롯하여 '영화·음악·연예', '정치·행정·법' 세분야뿐이다.

실제적으로 방송(326.7)분야의 경우는 언론·매체(070)와 관련된 분야이면서 KDC상에서 서로 다른 분류번호가 배정된 것은 '방송'은 신문방송학이라는 학문적 분류를 우선한 것이고, 정보탐정에서는 '방송' 자체를 고려한 분류로서 실용적인 분류가 우선한 것이다. 이는 웹분류체계의 원칙이 실용적인 것에 있는 것을 반증하는 한 예이다. 한편, '홈페이지'라는 개념은 KDC관점에서는 자서전, 홍보, 컴퓨터등과 관련된 주제이지만 실제적으로는 컴퓨터에 관련된 주제가운데 파생된 하나의 개념이라 할 수 있다. 또한 강수준에 배열되어 있는 '개인과 단체'는 주제적으로 개인홈페이지와 단체홈페이지에 해당하기 때문에 이를 KDC로 분류하는 것은 더욱 어려운 상황이다. 또한, '개인과 단체'라는 항목은 강수준에 배열된 디스크립터이면서 다른 영역에서도 나타나는 항목이다. 즉, '개인과 단체'는 KDC개념으로는 보조기호표적인 성격이 매우 강하다. 이러한 보조기호표적인 주제 혹은 항목으로는 지역을 나타내는 용어인 '서울'을 비롯하여 많은 개념어들을 들 수 있다. 특히, 열거된 분류체계가운데 '경제/산업' 류항목에 가장 많은 웹문서가 링크되어 있으며, '교육/출판', '홈페이지' 순으로 링크된 자료의 수가 많았다.

〈표 2〉 정보탐정 분류체계일부

류		강	
주 제	분 류 번 호	주 제	분 류 번 호
컴퓨터	004	소프트웨어	004(40)
인터넷	566(867)	인터넷	004.575/566(126)
언론	070(362)	신문	070(195)
매체	070	잡지	070(129)
		방송	326.7(38)
경제	320(1155)	기업	324(305)
산업	323	산업	323(745)
		취업	336.24(44)
취미	691(338)	취미	691(91)
레저		스포츠	692(130)
		여행	980.2(110)
학 문	(573)	과학기술	500(382)
		인문사회	001.3/300(191)
교육	370(1,089)	대학	377(858)
출판	012	영어	740.7(48)
영화	688(318)	연예계	688(40)
음악	670	영화	688(124)
연예	689	음악	670(111)
예술	600(355)	화랑	660.69(65)
문화	900	문학	800(66)
출판	012	행사	606.3(47)
건강	517(501)	건강	517(53)
병원	510	질병	(53)
		병원	510(195)
생활	617(366)	쇼핑	326.17(183)
주택		생활정보	335(111)
		주택	(73)
정치	340(318)	행정	350(140)
행정	350	정치	340(78)
법	360	법률	360(89)
홈 페이지	(971)	개인	(737)
		단체	(227)

3. 2 심마니

심마니(http://simmany.chollian.net/)는 한글 워드프로세서를 개발한 한글과 컴퓨터(社)에서 개발한 검색엔진이다. 초기에는 자체적으로 운영하였으나 회사 내부사정으로

현재 심마니의 관리를 데이콤에서 운영하고 있다. 대부분의 검색엔진이 영어로 구성되어 검색 옵션과 방법이 영어권에 편리하도록 개발한 것에 비해 심마니는 한글검색을 가장 효과적으로 가능하게 해주는 최초의 국내 웹 검색엔진이다. 심마니는 조사나 어미 등 불용어가 붙어있는 키워드로 검색하거나 혹은 키워드만으로 검색어가 입력되어도 해당 키워드를 갖고 있는 웹문서들이 모두 출력된다.

3. 2. 1 심마니의 분류체계

심마니[1]는 1998년 5월 현재 개인 홈페이지를 비롯하여 16개의 류항목을 1차메뉴체계로 제공하고 있으며 강항목수준으로는 207개 항목을 제공하고 있다〈표 3〉. 특징적인 것은 항목에 해당 항목에 링크되어 있는 웹문서의 수와 하위항목에 대한 정보를 제공하는 것이다. 심마니는 해당 분류체계에 링크된 정보를 제공하여 이용자에게 검색에 도움이 되도록 한다. 예를 들면, 〈예 1〉은 심마니의 류항목인 '교육, 대학'의 사이트를 나타낸 것이다. '교육, 대학'이라는 류수준에는 17개의 강항목이 배열되어 있으며 강항목가운데 '가상학교'가 있음을 나타낸다. 각괄호안의 숫자는 해당 분류항에 연결된 정보의 양이며 아래 예에서 첫 번째 0은 카테고리의 개수를 나타내며, 두 번째자리수인 29는 카테고리와 연결되어 있지 않은 웹문서의 개수이며, 세 번째 자리수 29는 카테고리와 각각의 웹문서의 총수를 나타낸다.

/첫 화면/교육, 대학
* 하위 분류 개수 : 17
가상학교 [0/29/29]

〈예 1〉심마니의 분류체계일부

3. 2. 2 심마니의 분류체계 분석

심마니의 류수준과 KDC의 류수준과의 일치도를 조사하면 '사회/문화'를 비롯하여 4개의 분류항목(류수준)이 일치하며 특히, '어린이와 청소년'과 같이 KDC로는 분류번호배정이 모호한 것을 류수준에 배열하고 있다. 특히, 강항목수준은 다른 웹검색엔진에 비해 비교적 많이 배정되었으며, 강항목가운데는 분류번호배정이 모호한 것도 다른 검색엔진에 비해 많은 편이다. 강수준에 배열된 분류항의 개념은 대부분 단일개념으로 이루어져있으며, 다른 검색엔진과 같이 류항목의 개념은 복합개념으로 이루어졌다. 그러나 '사회(300)/문화(900)' 항목을 제외하고는 복합개념에 전혀 상이한 개념끼리 합쳐진 항목은 없으며, 강수준에 배정된 분류항도 대부분 류수준의 분류항의 주제와 동일한 주제이다. 심마니의 분류체계는 개념색인에 근사한 것으로 기존 웹검색엔진가운데 문헌분류체계에 가장 유사하지만, 이용도가 높은 분류항에 속하는 '신문, 잡지' 등이 '산업/경제' 류에 속하도록 한 것과 '컴퓨터/인터넷'에 신문을 다시 한번 배정한 것, 류수준에 '뉴스/언론매체'를 별도로 배정한 사실은 심마니도 다른 웹검색엔진과 같이 반복성과 순환성을 염두에 두어 분류체계가 체계적이지

1) 대괄호의 숫자는 해당 류개념어에 배정된 웹문서의 갯수이다.

〈표 3〉 심마니의 분류체계 일부

類		綱	
주제	분류번호	주제	분류번호
개인홈페이지	[14591]	국내	[14166]
		국외	[424]
교육	370	가상학교	[25]
대학	377[5066]	기타	[36]
		단체,기관	370.6[87]
		대학교(대학원)	377[3530]
		도서관	027[16]
		야학	[2]
		외국어	701[88]
		유아(아동)	[26]
		유학	377.7[137]
		직업전문학교	336.24071[12]
		초,중,고등학교	375.2, 376[606]
		편입	373.3[6]
		학술정보	[30]
		학습자료	[140]
		학원	[291]
		행사,대회	[1]
		회사(업체)	[33]
사회	300	기념일	[9]
문화	900[1305]	기타	[11]
		노동,인권	336[9]
		단체,기관	339/060[206]
		도서관	026[11]
		문화비평	[1]
		문화재	601.5[45]
		사회과학	300[31]
		사회복지	338[79]
		생활정보서비스	[239]
		언어	700[20]
		여성	337[18]
		잡지	070[16]
		정신과학,초현상	187[17]
		족보	999.11[6]
		지역(지방)	[412]
		철학	100[26]
		컬럼	[1]
		풍습	380[30]
		한국의 사회,문화	331.5[15]
		행사,대회	[45]
		환경	539.9[58]

못한 것을 나타낸다. 또한, 개인홈페이지를 별도의 류로 배정하여 이 분야에 링크된 웹문서가 다른 류보다 월등히 많은 것도 특징이다.

3. 3 야후(한국판)

야후(한국판)는 1997년부터 본격적인 서비스를 개시하였다. 야후는 기본적으로 전세계

에서 가장 널리 알려진 Yahoo(US)판과 동일한 메뉴체계를 유지하였으나 최근에 Yahoo(US)측의 일방적인 분류체계의 변경에 따라 약간의 차이가 있다. 그러나 검색에 관련된 모든 기법은 Yahoo(US)와 동일한 서비스를 제공하고 있다.

3. 3. 1 야후(한국판)의 분류체계

야후는 14개의 류항목을 갖고 있으며 52개 강항목을 갖고 있다〈표 4〉. 또한, 전체분류항의 총수는 국내 웹검색엔진가운데 최대 4,177개의 분류항을 갖고 있다. 이와 같이 상대적으로 많은 분류항을 갖고 있는 것은 검색자의 편의성과 웹정보의 균형성을 유지하기 위한 것이라 판단된다. 특히, 야후는 분류항의 반복성과 순환성을 특징으로 하고 있다. 반복성은 어느 체계에 있어도 필요에 따라서는 몇번이고 반복해서 다른 체계에 배열되는 성질이고, 순환성은 @기호를 활용하여 다른 분류항과의 링크를 통해 해당 정보를 입수할 수 있도록 하는 성질이다. 〈예 2〉는 '예술과 인문' 류에 하위항목으로 전개된 예이다. 인덴션은 분류의 수준을 의미하며, 일부 분류항은 생략하였다.

반복성을 나타내는 것은 '교육' 이라는 분류항으로써 초기에는 목수준에 있으나 필요에 따라 세목수준에도 출현한다. 또한, '기관, 단체' 등과 같은 여러 분류항도 각각의 류수준에 반복해서 출현하며 체계수준도 일정하지 않는 특징을 보여주고 있다. 한편, 순환성은 분류항 뒤에 @표시를 하여 이를 계속해서 추적할 경우 최하위 해당정보(leaf node)에 수렴하도록 하였다. 예를 들면, '무용@'

```
공연예술
  교육
    무용@
  극, 연극
  교육
    전문대학, 종합대학교
  극작가@
  기관, 단체
기관, 단체
  무용@
  연극@
  음악@
무용, 댄스
  교육
    전문대학, 종합대학교
  기관, 단체
  무용가
    전통무용가@
  볼룸댄스
```

〈예 2〉 야후의 분류체계 일부

은 '교육과 기관', '단체' 와 같이 다른 수준에서 동시에 출현하나 두 개 항목의 최종적으로 검색결과를 보여주는 것은 동일하도록 안내의 의미를 갖고 있다. 그러나 '건축' 과 같은 분류항은 @표시가 부여되어 있으나 무한순환(loop)으로 수렴하여, 각각의 카테고리에서 별도의 웹문서를 보여주고 있다.

3. 3. 2 야후(한국판)의 분류체계 분석

야후의 류수준을 KDC의 분류항목과 비교하면 유사한 수준의 분류항이 부여된 것은 '사회과학' 를 비롯하여, '사회/문화', '엔터테인먼트', '자연과학', '지역정보' 가 해당한다. 이를 강수준까지 확대하여 류수준의 주제와 강수준의 주제가 일치하는 것은 '언어학' 만 일치할 뿐이다. 그밖에 항목들은 대부

〈표 4〉 야후(한국판)의 분류체계 일부

류		강	
주제	분류번호	주제	분류번호
예술 인문	600 001.3(13875-4)	사진 전통예술 디자인 박물관/화랑 문학	660(119) 600.9(163) 658( ) 606.9(146) 800(@)
비지니스 경제	320(31422-5)	회사 취업, 채용 투자/재테크 항목별 광고	324.4(26782) 336.24(899) 327.8(845) 326.14(503)
컴퓨터 인터넷	004.075 004.575/566 (33732-14)	인터넷 WWW 통신/네트워크 하드웨어	004.575/556(7168) 004.575(@) 567/004.575(9242) 566(10718)
교육	370(5143-11)	대학교 전문대학 초/중등교육 입시정보	377.6(@) 377.5(@) 375/376(607) ( )
엔터테인먼트	689(5838-8)	콜사이트 영화 음악 유머/재미	(671) 688(1635) 670(1382) 817(124)
정부	350.2(1770)	행정부 정치 법	350.2( ) 340(414) 360(332)
건강 의학	510(1934-8)	의학 병원 질병과증상	510(531) 517.16(194) (76)
뉴스 미디어	070(1745-12)	이벤트 잡지 텔레비전 신문	(55) 070(197) 070(796) 070(158)
레크리에이션 스포츠	691(3262-10)	스포츠 게임 여행, 관광 자동차	692(617) 691(335) 980.2(927) 556(505)
참고자료	018(1446-2)	도서관 사전류 전화번호	026(1258) 030(25) 326.447(30)
지역정보	900(35474-1)	국가별 한국	(8184) 911(11425)
자연과학	400(19759-9)	컴퓨터공학 생물학 천문학 공학	566.1(3168) 470(480) 440(172) 530(2285)
사회과학	300(5076-6)	인류학과고고학 사회학 언어학 경제학	471/902.5(848) 331(208) 700(832) 320(2465)
사회 문화	300/900(13562-2)	사람들 개인홈페이지 어린이 종교	(7295) ( ) ( ) 200(1637)

분 강과 목수준이 배정되었으며 세목이하로 전개되어 배정된 것은 '정부/공공기관'을 비롯하여 류수준에는 1개와 강수준에는 13개의 분류항이 배정되어 있다. 특히, '서울'과 같은 분류항(강)은 지리구분에 배정된 것으로 특별히 분류번호가 부여되기 애매한 것도 8항목에 달하고 있다.

이는 기존의 분류체계로 야후에 연결된 인터넷자원의 분류와 체계성 유지에 많은 어려움이 있음을 나타내는 것이다. 특히, '엔터테인먼트(689)'와 같은 류체계에 '유머/재미(817)'등이 배정되어 류강항목의 기호연계성도 높지 않음을 알 수 있다. 따라서 유사한 정보의 군집이 이루어지지 않아 검색에 어려움이 초래될 수 있다. 또한, 야후에서 제공하는 분류체계가운데 가장 많은 웹문서들이 배정되어 있는 것은 개인홈페이지이며, 다른 웹문서의 류수준에는 없는 '지역정보'류에 일련의 웹문서들이 배정되는 것도 중요한 특징이다.

3. 4 네이버

네이버(http://www.naver.com/)도 확장검색기능을 갖고 있으며, 특징적인 것은 일반도서관 검색방법과 유사하게 키워드로 검색을 수행할 수 있도록 조치하고 있다. 또한, 검색대상을 분류체계를 대상으로 하거나 혹은 분류만을, 혹은 웹만을 대상으로 검색을 수행한다. 웹을 검색대상으로 할 경우 검색대상을 웹문서의 제목만을 검색대상으로 설정할 수도 있으며, 검색대상을 홈페이지(사이트)만으로도 검색을 할 수 있도록 하고 있다. 이는 도서관에서 출판지와 서명 혹은 분류번호, 전문을 대상으로 조건별 검색이 이루어지는 것과 유사하다.

3. 4. 1 네이버의 분류체계

네이버는 14개 류항목과 44개 강항목으로 구성되어 있다〈표 5〉. 네이버에는 웹문서가 3,000,514건이 링크되어 있으며 각 분류항에는 해당 항목에 링크된 자료의 수가 제시되고 있으며 다른 검색엔진과 같이 분류항이 순환하는(@) 특성을 갖고 있다. 그러나 이 순환기호는 항상 동일한 최하위 사이트를 지적하고, 해당항은 다른 체계에도 반드시 출현하는 성질을 갖고 있다. 〈예 3〉의 네이버 분류체계의 괄호 안의 숫자는 항상 해당 분류항에 연결된 웹사이트의 수를 의미한다. 특징적인 것은 해당 항목에 있는 하위카테고리에 대한 정보를 제공하지 않는 것이다.

자연과학
- 공학 (227)
- 교육 (20)
- 기관, 단체 (8)
- 기업@
- 농학 (33)
- 뉴스 (7)

〈예 3〉 네이버의 분류체계 일부

3. 4. 2 네이버의 분류체계 분석

네이버의 14개 류항목가운데 KDC의 수준과 일치하는 것은 '사회과학'을 포함하여 6

〈표 5〉 네이버의 분류체계 일부

류		강	
주제	분류번호(문서갯수)	주제	분류번호(분서갯수)
건강/의학	517/510(481-4)	건강관리	517(9)
		병원	517.16(40)
		의학	510(239)
교육	370(290-4)	대학	377.6(85)
		시험, 자격증	517.14(39)
		유학, 어학연수	377.7(42)
뉴스/미디어	070(362-3)	신문	070(68)
		잡지	070(81)
		텔레비젼	326.76(45)
레크리에이션	691(674-3)	스포츠	692(364)
		게임	691(136)
		여행	980.2(44)
		야외활동, 레져	691(41)
사업/경제	320(4814-2)	기업	324(4302)
		취업정보	326.49(133)
		온라인쇼핑	326.173(95)
사회과학	300(239-3)	경제학	320(59)
		사회학	331(48)
		언어학	700(58)
사회/문화	300/900(807-2)	결혼	332.22(53)
		기관/단체	(57)
		어린이	(26)
연예/오락	689(3502-6)	연예인	(121)
		영화	688(211)
		음악	670(405)
		유머, 우스개	807(29)
인문/예술	001.3/600(498-1)	디자인	658(124)
		인문과학	001.3(155)
		박물관, 갤러리	069.8(53)
자연과학	400(642-1)	공학	530(176)
		컴퓨터과학	566(77)
		생물학	470(77)
정부/공공기관	350.2/360.6(530)	정치	340(122)
		한국정부	(248)
		국제기구	(10)
지역정보	900(285-2)	대한민국	911(@)
		서울	()
		국가	(269)
참고자료	018(147-1)	도서관	026(84)
		사람찾기 사전	(19)
컴퓨터	004(901-6)	인터넷	004.575/556(329)
		S/W	004(102)
		O/S	005.43(91)
		통신, 네트워크	028.65(73)

개 항목으로 다른 검색엔진에 비해 가장 높은 일치율을 보이고 있다. 특히, '참고자료' 항을 류수준으로 한 것은 도서관학적인 발상으로 판단되며, 하위분류항도 '도서관'과

'사람찾기 사전'으로 구성되어 있다. 이에 해당하는 웹문서의 개수가 다른 류강수준에 비해 낮음에도 불구하고 배열한 것도 이를 반영한 것이라 판단한다. 특히, 네이버는 항목의 중복성을 최대한 배제한 검색엔진이다. 류강항목가운데 동일한 항목어를 사용한 것은 '건강/의학' 항목의 강항목에서 '의학'이 반복된 것을 제외하고는 거의 없는 수준이다. 또한 세목이하에서도 다른 검색엔진에 비해 분류항의 반복이 최소화되어 있다. 웹문서의 균형성도 다른 엔진에 비해 편차가 적은 것을 보여주고 있다. 특히, 네이버의 류에 배정된 항목은 대부분 다른 검색엔진의 류에 출현한 것과 동일하다. 그 가운데 '사업/경제'류에 가장 많은 웹문서가 배정되어 있다.

3. 5 웹분류체계와 문헌분류체계의 비교

웹문서분류라는 입장에서 웹분류체계와 한국십진분류표와의 차이점을 살펴보면 다음과 같다.

1) 분류표제항의 차이

KDC의 경우 표제항은 단일개념으로 요약되며, 복합주제개념의 사용이 최대한 제한되는 성질을 보이고 있다. 또한, 류항목에 근접할수록 단일개념으로 표현이 되며 세세목과 같이 하위 항목으로 세분될수록 복합항목이 사용된다. 즉, 류항목에 배열된 분류항목은 모두 단일어로 이루어진 반면, 세세목으로 전개될수록 복합어나 복합개념이 사용된다. 예를 들면, 류항목에 '철학', '종교'등과 같이 단일어로 이루어진 반면 강항목(060)의 '일반학회, 단체, 협회, 기관, 연구기관'은 여러개의 복수어로 하나의 강항목을 차지하고 있다. 또한, '상업윤리(195.3261)'와 같이 세세목으로 갈수록 복합개념이 사용되고 있다. 웹분류체계는 분류수준에 상관없이 어느 위치에서도 복합개념 혹은 복합어가 하나의 분류항으로 사용된다. 이를 특별히 구분할 경우, KDC와는 달리 류수준에 가까울수록 복수개념과 복합어를 사용하고 있다. 예를 들면, 정보탐정의 경우 류수준에 '예술(600), 문화(900), 출판(012)'과 같이 기존 분류표의 개념으로는 전혀 다른 주제가 하나의 류에 복합적으로 있으며, 분류항의 구성도 복합어로 이루어졌다. 강수준에는 32개 강항목중 복합어나 복합개념으로 사용된 분류항은 없다.

2) 계층의 차이

KDC의 경우는 일반 문헌분류체계와 마찬가지로 논리적 분류에 근거하고 있다. 따라서 가장 복잡한 것에서 단순한 것으로 유사성의 정도에 따른 분류가 이루어진다. 즉, 개념적으로 큰 개념에서 작은 개념으로 세분화되는 방식으로 주제가 전개된다. KDC분류는 철학적 지식분류에 근거하고 있기 때문에 해당류에 해당하는 주제를 다루고 있는 정보의 생산성이 매우 낮더라도 해당 분류체계를 재구조화하지 않는다. 특히, 그 주제가 류수준에 가까워질수록 해당 주제는 변화가 없다. 예를 들면, DDC의 경우 초기 류항목을 개발하였을 때의 구조와 현재 구조와는 거의 차이가 없으나, 세부항목에는 분류항의 조정이 있다.

웹분류체계는 실용성에 주안점을 두고, 현

재 인터넷상에서 생산되고 있는 정보를 근거로 분류체계가 구축되고 있다. 이는 역동적 분류체계를 의미하며 새로운 지식의 분화와 통합을 인정하면서 기존 학문간 개념간 경계가 모호해지는 것을 분류에 받아들이고 있는 것이다. 따라서 계층구조면에서 이 두 개의 특성을 고려한다면, KDC의 경우는 상대적으로 매우 논리적이며 전개의 원칙이 있다. 한편, 웹문서분류체계는 상대적으로 비논리적이며, 전개의 원칙은 현재 관련분야의 정보를 필요로 하는 사람들의 관심의 크기가 분류기준이 된다.

4. 웹문서분류체계의 새로운 설계

분류는 크게 철학적 관점의 분류와 실용적 관점의 분류로 구분된다. 전자는 지식의 분류로 표현한다면, 후자는 문헌분류를 들 수 있다. 도서관에서의 문헌분류는 기본적으로 지식의 분류원칙을 따르고 있지만, 문헌분류는 분류대상이 도서관내에 소장된 자료라는 제한된 분야의 분류라는 측면에서 지식의 분류와는 차이가 있다. 즉, 문헌분류는 도서나 기타 자료를 그 기억된 지식 또는 정보의 내용에 따라서 분류하는 특성을 지니고 있다. 한편, 지식은 실용적 지식(practical knowledge), 지적 지식(intellectual knowledge), 오락적 지식 (small-talk and paste knowledge), 영적 지식 (spiritual knowledge) 및 불필요한 지식(unwanted knowledge)등으로 유형을 구분할 수 있다(Machlup). 이러한 지식가운데 전통적으로 문헌분류는 분류대상으로 사회적으로 인정받는 지식가운데 도서관에 보관되어 이용될 만한 가치가 있는 자료로 한정하고 있다. 도서관들은 이러한 선정자체를 수서라는 업무영역으로서 설정하고 사서의 중요한 임무가운데 하나로 간주하고 있다. 한편, 웹상에 등재된 자료들은 사서와 같은 정보처리전문가의 통제가 전혀 없이 누구라도 인터넷과 통신에 대한 기초적인 지식을 가진 이들이 올린 데이터가 대부분이다. 그 등재된 지식이 어떠한 유형의 지식인지에 대해서는 어떠한 제약도 없기 때문에 기존의 문헌분류체계로 이러한 지식들을 구조화한다는 것은 어려운 작업이다.

그러나 웹에 등재된 자료가운데 상당한 양의 데이터(웹문서)가 학술적 가치를 갖고 있으며, 기존 도서관에서 보관되어 일반 이용자에게 필요한 자료라고 판단되어 이의 학술적 활용가치와 방안에 대해 다양한 연구가 이루어지고 있다(Bruce, H).

4. 1 설계원칙

본 연구에서는 인터넷의 특성으로 대변될 수 있는 실시간적 정보제공형태와 주제와 내용의 무제한성을 설계원칙의 기준으로 한다. 즉, 지식의 세분화와 통합화를 통해 끊임없이 발전하고 쇠퇴하는 주제분야의 역동성을 고려하여 현재 웹상에서 가장 중요한 관심사가 되는 주제를 분류체계에 실시간적으로 최대한 반영하는 것이다.

기존의 문헌분류체계는 10개의 류분류와 100개의 강분류, 1,000개의 목분류로 구분하

고 있으며, 그 기준은 대체적으로 학문의 발전단계를 따르고 있다. 이 분류체계는 문헌자료의 분류적 측면에서는 의의를 갖고 있으나, 웹과 같이 역동적으로 이용자의 주제관심이 다양하고, 급속하게 변화하는 성질의 데이터를 대상으로 분류원칙을 적용하는 것에는 많은 무리가 있다.

〈표 6〉은 문헌분류체계와 웹검색엔진 분류체계의 주제적 차이와 이용자의 관심도의 차이를 보여주고 있다. 즉, 학문적 성격과 논리적 성격이 강조된 전자와 실제 이용자 관심도가 강조된 후자의 분류체계차이를 확인할 수 있다. 특히, 아래 표와 같이 일부 웹검색엔진에서는 웹분류체계의 류수준에 있는 항목과 문헌분류체계 류수준 항목이 전혀 일치하지 않고 있음을 확인할 수 있다.

이러한 차이점과 특성을 고려하여 인터넷 자원을 효과적으로 분류할 수 있는 웹문서분류체계의 설계 원칙을 제시하면 다음과 같다.

① 류강분야위주로 설계한다. 실제로 웹검색엔진의 세목·세세목부분은 계속해서 증가하는 추세를 보이며 또한 하이퍼 링크[2] 라는 특징을 갖기 때문에 세목·세세목 부분은 고정된 분류체계로 구축하기가 어렵기 때문에 세목이하 설계는 최소화한다.

② 류분야를 비롯한 하부전개에 최소 분야만을 제시하되, 고정된 전개(예를 들면, 10진)원칙을 따르지는 않는다. 고정된 원칙을 제시할 경우, 주제별로 점차 다양해지는 웹문서들을 주제별로 범주화하기가 현실적으로 불가능하기 때문이다. 특히, 듀이십진분류의 초기설계는 그 당시 소장한 자료를 중심으로 설계가 이루어지고 향후 발전될 학문분야나 기술부문을 완전하게 예측하지 못했기 때문에 장서분포비가 불균형적으로 나타나게 되었다. 이에 비해 웹문서는 주제별 특성상 향후 미래사회에는 주제적으로 전혀 예측할 수 없는 분야와 문헌들이 크게 증가할 것이기 때문에 10진과 같은 고정된 분류체계로는 웹문서들을 효과적으로 분류할 수 없게 될 것이다. 따라서 새로 설계될 웹분류체계는 10진으로 전개하지 않는다.

③ 류분야와 강분야의 구분은 학술적인 관점과 사용빈도를 고려하여 설정한다. 류분야는 학술적인 분류원칙에 따라 실제 해당 항목에 링크된 웹문서의 숫자보다는 하위 항목에 링크된 웹문서의 숫자를 고려한다. 인터넷의 사용빈도는 해당 항목의 접속빈도를 의미하며 인터넷분류체계구축에 매우 중요한 역할을 하지만 실제 정확한 수치데이터(접속횟수)는 각 회사별로 광고수익에 절대적 영향을 미치기 때문에 이를 공개하지 않

〈표 6〉 문헌분류와 웹분류체계 비교 예

DDC의 류항목	라이코스의 류항목
총류	경영
철학	교육
종교	오락
사회과학	참고 및 참조
언어	정부
순수과학	뉴스
기술과학	스포츠
예술	여행
문학	날씨
역사	웹자원(기타)

2) 하이퍼 링크(Hyper link) : 동일한 항목명이 서로 다른 세목 혹은 세세목 체계에 출현한다.

〈표 7〉 네이버의 상위접속사이트

주 제 영 역	접 속 건 수	백 분 율	비 고
/뉴스/신문	698	2.6	
/오락	494	1.8	
/오락/연예인(Entertainers)	471	1.7	
/오락/연예인/탤렌트	448	1.6	
/뉴스	392	1.4	
/비즈니스/고용(취업)	382	1.4	
/오락/사람	381	1.2	
/컴퓨터	321	1.1	
/컴퓨터/인터넷	276	1.0	
/오락/유머, 우스개	267	1.0	
/오락/음악/가수/한국팝송	260	1.0	
/비즈니스/재정 및 투자	247	0.9	
/오락/영화 및 필름	234	0.9	
/비즈니스/회사/매체/신문/중앙일보	214	0.8	
/비즈니스/고용/직업	213	0.8	
/비즈니스/회사/매체/신문/조선일보	212	0.8	
/비즈니스	212	0.8	
/컴퓨터/멀티미디어/비디오/MPEG/MP3	209	0.8	
/오락/음악/	204	0.8	
평 균	323	1.2	

는다. 그러나 〈표 7〉은 웹서치엔진의 하나인 '네이버'에서 평일(98년 3월)을 기준으로 접속빈도횟수와 접속분포를 공개한 자료이다.

이 자료는 1998년 3월 30일 현재 이용자들이 접속한 건수를 근거한 것으로서 접속건수 대비 5%이상인 주제영역만 나열한 것이다. 전체 접속건수는 26,795건에 달하고 있으며, 상위 10개의 주제영역이 전체 접속건수 대비 15.3%(4,130건)에 해당한다.

위 표의 대쉬(/)는 주제전개의 표시이다. 예를 들면, "/오락/영화 및 필름"은 "오락"이라는 류항목 밑에 "영화 및 필름"이라는 강항목이 있음을 나타내는 것이다.

26,795번의 접속가운데 상위에 해당하는 19개의 주제를 기준으로 환산하면 "뉴스/신문"이 가장 높은 숫자를 나타내지만 류 및 강분야만으로 환산할 경우 "오락" 분야와 "비즈니스" 주제에 접속한 숫자가 절대적이다. 즉, 이용자들은 인터넷에서 메뉴체계를 이용하여 접근하는 가장 인기있는 주제로는 오락과 비즈니스라 할 수 있다. 이러한 수치는 이용자별 조사가 이루어지지 않았기 때문에 대학교와 같은 연구기관에 그대로 적용되는 것은 무리가 있으나, 현실적인 관점에서 판단한다면 오히려 현실적인 수치로 활용할 수 있다.

위와 같은 접속빈도외에 본 논문의 분석대상으로 선정한 4개의 웹검색엔진에서 제공

〈표 8〉 검색엔진의 웹문서배정순위(류수준)

종류\순위	1	2	3	4	5	6	7	8	9	10
정보탐정	경제 /산업	교육 /출판	홈페이지	컴퓨터 /인터넷	학문	건강 /병원	생활 /주택	언론 /매체	예술/문화 /출판	취미 /레저
심마니	개인 홈페이지	산업 /경제	교육 대학	컴퓨터 /인터넷	오락 /취미	사회 /문화	건강 /병원	종교	어린이 /청소년	예술
야후	지역정보	컴퓨터 /인터넷	비지니스 /경제	자연과학	예술 /인문	사회 /문화	엔터테인 먼트	교육	사회과학	레크레이 션/스포츠
네이버	사업 /경제	연예 /오락	컴퓨터	사회 /문화	레크 레이션	자연과학	정부/ 공공기관	인문 /예술	건강 /의학	뉴스 /미디어

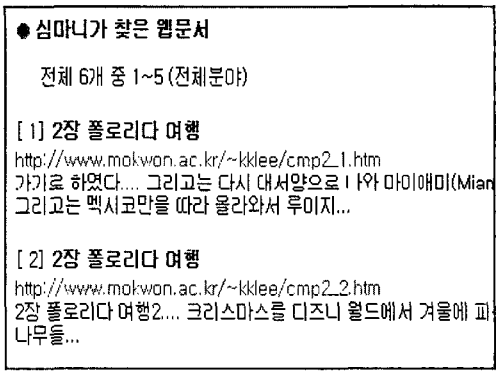
하고 있는 문서분류체계를 링크된 웹문서(사이트포함)의 수로서 상위 10개류를 선정하면 〈표 8〉과 같다.

이 표를 분석한 결과 4개의 검색엔진은 공히 '경제/산업'에 관련된 정보의 양이 가장 많았으며 '컴퓨터와 인터넷'도 유사한 순위를 확인할 수 있었다. 4개의 검색엔진가운데 류가 다른 검색엔진의 류와 한 번도 일치하지 않는 것으로 정보탐정의 '생활/주택'과 '언론/매체', 심마니의 '종교', '어린이/청소년'이 있다. 또한, 야후에서는 '지역정보'가 가장 많은 웹문서가 링크되어 있으나 타검색엔진의 류수준에는 나타나지 않고 있다. 네이버는 모든 류항목이 다른 검색엔진의 류항목과 대부분 일치하여 가장 보편적인 분류체계를 갖고 있다.

④ 인터넷상의 웹문서에 대한 다음과 같은 특성을 인정한다.

i) 웹문서수의 생동성 : 웹문서의 수는 생동성을 갖고 있기 때문에 시간대별로 또는 일자별로 차이가 현저하게 발생하기 때문에 웹문서구성비를 조사하기 위해서는 고정된 웹문서의 수를 확보해야 하는 어려움이 있다. 예를 들면, '풍습'라는 검색어를 심마니를 통해 검색할 경우, 98년 5월에 30개의 웹문서가 있으나, 98년 8월 현재 24개의 웹문서만이 링크되어 웹문서의 수가 일정하지 않은 점이 있다.

ii) 웹문서의 중복성 : 웹검색엔진은 하나의 웹문서일지라도 하나의 파일에 여러 개의 하이퍼 링크의 주소가 부여되었다면 이를 기계색인시에 여러개의 문서로 간주한다. 따라서 검색된 웹문서의 숫자가 정확치 않을 확률이 있다. 예를 들면, '키웨스트'라는 검색어로 검색을 할 경우, 다음의 〈그림 1〉과 같이 첫 번째 문서와 두 번째 문서는 동일한 내용의 자료이지만 웹검색결과서로 다른 문서로 제시되어 2개의 문서로 구분되어 있으나 실제로는 동일한 문서이다. 즉, 동일한 문서내에 여러개의 웹링크가 되어있는 경우에는 실제 문서의 수는 1건이지만 링크된 숫자에 따라 복수개의 문서로 중복계수된다.


● 심마니가 찾은 웹문서

전체 6개 중 1~5 (전체분야)

[1] 2장 폴로리다 여행
http://www.mokwon.ac.kr/~kklee/cmp2_1.htm
가기로 하였다.... 그리고는 다시 대서양으로 ! 나와 마이애미(Mian
그리고는 멕시코만을 따라 올라와서 루이지...

[2] 2장 폴로리다 여행
http://www.mokwon.ac.kr/~kklee/cmp2_2.htm
2장 폴로리다 여행2.... 크리스마스를 디즈니 월드에서 겨울에 피
나무들...

〈그림 1〉 웹문서중복의 예

4. 2 웹검색엔진용 분류체계설계

이상과 같은 설계원칙과 특성을 고려하고, KDC의 구조와 장서분포도, 웹검색엔진의 분류체계를 비롯한 여러 조사결과에 따라 류·강수준의 가장 기본적인 분류체계를 제시한다.

새로운 분류체계는 류개념의 수를 11개로 선정하였으며, 강개념으로는 46항목을 선정하였다. 새롭게 설계된 분류체계의 정보함유도를 측정하기 위해 야후(한국판)를 활용하여, 해당 주제어에 링크된 문서의 숫자를 조사하였다. 괄호안의 숫자는 새로이 설계한 해당 분류항을 검색어로 입력하여 나타난 숫자이다. 앞의 숫자는 관련 웹사이트(카테고리)의 숫자이며, 뒤의 숫자는 입력어(분류항)를 색인어로 한 관련 웹문서의 수이다. 앞의 류개념의 숫자와 강개념의 숫자가 차이가 나는 것은 류주제어를 색인어로 하는 웹

〈표 9〉 신분류체계 (류·강수준)

류주제	강주제
전산학(1-18)	컴퓨터(70-1807)
	인터넷(104-2818)
	정보통신(3-475)
언론(0-115)	신문(32-622)
	잡지(76-340)
	방송(14-50)
	광고/홍보(29-910)
경제(23-471)	금융(11-246)
	부동산(27-171)
	결혼(10-144)
	기업(13-989)
레저(4-125)	여행(134-1,366)
	스포츠(49-369)
	쇼핑(20-270)
	음식점(33-96)
	오락(17-879)
	만화(26-817)
	자동차(21-624)

〈표 9〉 신분류체계의 일부

류 주 제	강 주 제
교육(320-2184)	초등학교(108-169)
	중고등학교(0-9)
	대학교(289-2572)
	학원(23-349)
	도서관(81-565)
	유학(6-262)
	취업(49-473)
문화(58-909)	문학(8-150)
	연극(1-95)
	영화(27-1083)
	음악(22-1197)
	연예인(2-157)
	미술(3-175)
건강(99-337)	병원(17-322)
	한의원(1-19)
	약국(1-30)
	치과(0-74)
법률(4-242)	특허(3-85)
과학(86-788)	순수과학(0-1)
	기술과학(11-154)
	인문과학(4-3)
종교(87-49)	가톨릭(4-100)
	기독교(6-265)
	불교(5-102)
정부(152-115)	통일(2-51)
	국방(0-10)
	대한민국정부(892)
	정치(442)

문서/사이트와 강개념을 색인어로 하는 웹문서/사이트와의 관계를 고려하지 않기 때문이다.

새로운 웹문서 분류체계가운데 1차적으로 류수준의 설계는 KDC분류체계에서 가장 많은 장서분포율을 보이고 있는 '사회과학', '기술과학' 및 '문학'을 주요 세분전개분야로 간주하여 이에 해당하는 세부항목을 조사하였다. 한편, 웹검색엔진에서 제공하고 있는 분류체계를 분석한 결과, 가장 많은 정보가 링크된 류수준으로는 문헌분류체계상의 '경제'와 '문학' 분야에 해당하는 '문화' 류와 '과학' 분야로서 상대적으로 많은 웹문서들이 배정되어 있다. 이들 분야를 웹분류체계의 주요분석대상으로 선정하였다. 특히, 링크정보함유도가 10위내에 있는 것은 몇 개의 류

항목을 제외하고는 4개의 검색엔진에서 제공하는 주제분야들이 대부분 일치하고 있었다. 그 가운데 본 연구에서는 '언론', '문화', '교육' 및 '과학' 등의 항목을 1차 류항목으로 선정하였다. 단, '언론'의 경우는 정보탐정의 경우만 출현하고, 타검색엔진의 류수준의 분류체계에는 출현하지 않았지만 앞에서 조사한 네이버의 류항목중 최대의 접속 횟수가 '언론(뉴스)' 항목인 것을 중시하여 이를 류수준으로 배정하였다. 종교의 경우 류수준에 배열한 것은 향후 종교관련정보가 급속하게 증가할 것으로 판단되기 때문이다. 그 판단의 근거는 현재 인터넷과 함께 통신서비스로서 폭넓게 활용되고 있는 PC통신의 메뉴체계에서 소프트웨어를 포함한 관련 자료의 정보생산성이 매우 높은 분야이기 때문이다. '법률' 항의 경우도 전문데이터베이스의 주서비스 주제가 법률분야라는(정영미) 점을 중시하고 이 분야에 대한 이용빈도도 급증할 것이라는 판단에 근거하였다. 이밖에 제시된 류항목은 10위권내에 들지 않은 류주제와 강수준에 링크된 웹문서의 양을 고려하여 선정하였다. 이러한 원칙과 방법을 분석한 결과 다음과 같이 새로운 웹문서분류체계를 제시하였다.

4. 3 새로운 웹문서 분류체계의 분석

앞에서 제시한 웹문서분류체계의 류수준에는 학문적 성격을 상대적으로 많이 반영하였으며, 강개념이하 세부주제로 전개될수록 실용적인 주제가 전개되도록 설계하였다. 새로운 분류체계의 균형성을 조사하면 〈표 10〉과 같다.

류수준의 균형성이 유지되기 위해서는 대체적으로 각 류개념이 11개로 구분되어 있기 때문에 한 개의 류에 정보함유비율이 9%내외를 유지해야 한다. 그러나 〈표 10〉에 따르면 류수준의 경우 전산학항목은 0.3%의 정보 함유비율을 보이고 있으며, 교육항목은 40.5%의 정보함유비율을 보이고 있으며, 대부분의 분류항(류수준)이 9%내외의 정보함유도를 보이지 못하고 있다. 이를 강수준으로 확대하면 법률항이 0.4%로 최소로 배정되어 있으며, 전산학항목이 21.9%로 최대의 정보함유비율을 보이고 있다. 류강수준을 종합하여 웹문서분포율을 분석하면 '법률'과 '종교' 항목이 상대적으로 작은 정보함유도를 보이고 있고, '교육' 항목이 최대의 정보함유도를 나타내고 있다. 이러한 현상만으로 새로운 웹문서분류체계를 분석할 경우, 기존의 분류체계와 같이 한 주제분야에 편중된 체계로 판단할 수 있다. 이와 같은 현상이 나타나

〈표 10〉 새로운 분류체계의 균형성

	전산학	언론	경제	레저	교육	문화	건강	법률	과학	종교	정부	합계
류주제	0.3	1.9	8	2	40.5	15.6	7	4	14.1	2.2	4.3	100
강주제	21.9	8.6	6.7	19.5	20.5	12.1	2.0	0.4	0.7	2.1	5.8	100
계	22.2	10.5	14.7	21.5	61	27.7	9	4.4	14.8	4.3	10.1	200

는 것은 국내 웹검색엔진의 색인방식에 기인하며, 각 검색엔진들이 웹문서계수 방식의 차이에도 그 원인이 있다.

① 웹문서분류방법 : 국내뿐만 아니라 외국의 웹검색엔진 색인은 거의 모든 과정을 기계색인방식에 의존한다. 이는 기존 도서관에서 이루어지는 색인방식이 사서의 판단에 따라 개념적으로 이루어지는 분류방식이라면 기계색인은 웹문서의 표제나 내용중에 나타난 용어를 대상으로 이루어지는 방식으로서 수작업색인에 비해 훨씬 단순하고 간단한 색인과정이다. 즉, 웹문서색인은 해당 웹문서가나 사이트에 나타난 단어에 의존할 뿐이며, 개념이나 의미분석과정이 전혀 이루어지지 않고 있다. 웹문서에서 동일한 의미의 서로 다른 용어로 기록될 경우 각 문헌들은 서로 다른 자료로 간주한다. 예를 들면, 웹문서 가운데 "입학전 아동발달에 TV의 영향"이라는 제목을 갖는 웹문서와 "입학전 아동발달에 텔레비존의 영향"이라는 웹문서는 '텔레비존'과 'TV'는 서로 다른 용어로 색인이 이루어진다.

② 문서링크방법 : 웹의 특성가운데 문서표현방식이 HTML로 이루어져 분류항은 항상 하나의 상위분류항만을 갖고 있지 않는다. 예를 들면, 야후(한국판)에서 "처음:비지니스와 경제:컨벤션과 회의"라는 계층구조와 "처음:비지니스와 경제:회사:컨벤션, 무역쇼"라는 복수의 계층구조를 유지할 수 있는 것은 분류항끼리 공간연결이 가능하기 때문이다. '비즈니스와 경제' 분류항 밑에 '컨벤션과 회의'라는 분류항과 '컨벤션, 무역쇼'라는 분류항이 서로 유사하지만 서로 수준이 다른 상위항을 갖고 있어도 이용자입장에서는 커다란 문제가 발생하지 않는 것이다. 따라서 링크된 웹문서의 개수는 일정부분 중복이 가능한 것이다. 도서관에 "철학과 도시공학"이라는 문헌자료가 입수되었다면 이 자료에는 반드시 해당 분류번호 하나가 부여되며, 복권이 구입되어도 동일한 분류번호가 부여된다. 그러나 웹상에서 분류는 '철학'이라는 항목과 '도시공학'이라는 항목에 각각 해당 웹문서가 배정될 수 있으며 이는 2권의 동일한 자료를 각기 다른 주제에 배열하는 것이다.

이러한 웹문서 분류체계와 특성을 보완하기 위해서는 반드시 전거파일의 도입이 절대적으로 필요하다. 이는 기계색인의 단순성을 보완할 수 있는 유일한 방법으로서 실제적으로 검색과정에서 전거파일을 이용하여 조정된 후의 균형비율을 조사하면 〈표 11〉과 같다.

즉, 류항목을 대상으로 전거파일의 효용성을 조사한 결과 기계색인만으로 단순검색이 이루어졌을 때보다 항목간 균형편차가 40.2%에서 36.2%로 감소하였다.

예를 들면, 전산학의 경우 0.3%에서 2.3%

〈표 11〉 전거파일로 조정된 균형비율

	전산학	언론	경제	레저	교육	문화	건강	법률	과학	종교	정부
류항목	2.3%	3.3%	6.8%	2.3%	38.5%	13.7%	6.9%	4.9%	11.8%	3.4%	6.0%

로 증가한 것은 '전산학'만으로 단순검색을 시도한 경우에 1개의 카테고리와 18개의 웹문서만이 출력되나, 전거어인 '전자계산학'을 사용하였을 경우에는 2개의 카테고리 50개의 웹문서를, 또하나의 전거어인 '컴퓨터공학'을 사용하였을 경우에는 2개의 카테고리와 116개의 웹문서가 출력되었다.

한편 가장 높은 정보함유비율을 보이고 있는 '교육' 부분을 세부항목으로 분리하지 않은 것은 앞에 조사한 바와 같이 교육관련 웹문서의 수는 많으나 실제 이용빈도는 상위순위에 나타나지 않아 불필요한 정보함유율이 높다고 판단하였기 때문이다.

5. 결 론

현재 사용되고 있는 웹검색엔진의 분류체계는 정보사서와 같은 전문인의 판단이 거의 배제되고 기계에 의해 추출된 정보를 자체적으로 개발한 분류에 적용하여 이용자들에게 제공하고 있다. 한편, 기존의 문헌분류체계는 인터넷상에 있는 새롭고 다양한 주제를 수용하기에는 분류체계의 경직성과 고정전개원칙 때문에 현실적으로 어려운 실정이다. 따라서 웹검색엔진에 사용될 이용자중심의 분류체계는 기계에 의존한 용어색인보다는 개념색인이 필요하며 웹문서 분류체계의 논리성과 체계성의 확보가 절대적으로 필요하다. 이에 따라 본 연구에서는 다음과 같이 분류체계설계원칙에 따라 새로운 웹문서분류체계를 제시하였다.

① 전개원칙은 주제범위의 상위개념에서 하위개념으로 전개하는 분화원칙을 따른다.

② 분류항 선택의 기준은 해당 분류항이 갖는 정보함유비율로 한다.

③ 세부적으로 전개될수록 항목은 학문적 개념보다 실용적 개념을 택한다.

이러한 원칙에 따라 류수준에 '전산학'을 비롯한 11개 항목과 46개 강항목으로 구성된 웹문서분류체계를 설계하였다. 제시된 류항목을 기준으로 균형성을 조사한 결과 새로운 웹문서분류체계도 정보함유비율에 있어 높은 편차를 보이고 있다. 이러한 편차는 웹문서색인방식과 링크방식, 웹분류체계의 특성에 기인한 것으로 판단되었다. 이를 평준화하기 위해서는 별도의 전거파일이 필요하였으며 전거파일을 도입하여 류수준의 균형성을 조사한 바에 따르면 일반기계색인방식으로 처리하였을 때보다도 4%정도 편차를 줄일 수 있었다. 즉, 정련된 전거파일을 이용할 경우 링크된 문헌의 편중성은 보완될 수 있음을 확인할 수 있었으며, 이는 새롭게 제시한 웹문서분류체계가 실제 웹검색엔진의 분류체계로 사용될 수 있음을 의미한다고 판단한다.

현재, 국내외에서 개발되고 있는 웹검색엔진의 분류체계에 있어 전거파일의 개발과 분류항의 선정에 많은 연구가 필요하다. 특히, 웹문서분류체계에 류강수준이외에 세목이하까지 연구범위를 확대하여 실용화될 수 있는 연구가 절대적으로 필요할 것이며, 이에 대한 연구가 지속적으로 이루어져야 한다.

참 고 문 헌

Bruce, H. (1998). "User Satisfaction with Information Seeking on the Internet" Journal of the American Society for Information Science. 49 (6): pp. 541-556.

Chen, H etc. (1998). "Internet Browsing and Searching:User Evaluations of Category Map and Concept Space Techniques" Journal of the American Society for Information Science. 49(7) : pp. 582-603.

Chen, H., Schatz, B.R., Orwig, R. (1996). Internet categorization and search: A machine learning approach. Journal of Visual Communications and Image Representation, 7(1), pp. 88-102.

Machlup, F. The production and distribution of knowledge in the United States. Princeton: Princeton University Press. 1962. (정영미. 지식구조론. 1997. pp.30-32에서 재인용)

Van Rijsbergen, C.J.(1977). "A theoretical Basis for the Use Co-occurrence Data in Information Retrieval," Journal of Documentation, 33(2), pp. 106-119.

남영준1. 디지털 정보검색론. 전주대학교 출판부. 1998.

남영준2. 인터넷으로 떠나는 세계여행. 전주대학교 출판부. 1998. pp.75-76.

정영미. 정보검색론. 구미무역출판부. 1993.

정필모. 국제백진분류법연구 〈인문학분야편〉. 중앙대학교 출판부. 1995. pp.1-5.

포항공과대학교 정보통신연구소. 웹검색서비스용 자동 문서분류시스템 연구. 한국통신 연구개발원. 1997.