웹문서분류체계의 설계

Design for the System of Web Document Classification

남 영 준 (전주대학교 문헌정보학과)

Nam Young-Joon (Dept. of Lib. & Inf. Sci., JeouJu Univ.)

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.

서 론

현대 도서관은 인터넷이라는 전혀 예기치 못한 거대한 데이터베이스의 출현으로 정보봉사범위를 인터넷까지 확대할 것인지 또한 이를 도서관 장서데이터베이스외의 통합된 데이터베이스로 간주할 지에 대해서 적정한 해결책을 제시하지 못하고 있다. 또한, 도서관에 소장된 자료와 인터넷에 등재된(up loading) 자료는 각각의 특성 때문에 검색방법을 별도로 개발해야 하는 문제가 제기되기 때문에 기존의 참고봉사 방법과 검색방법으로는 이 두 개의 데이터베이스를 운영하기 어려운 실정이다. 즉, 기존의 도서관들은 전혀 새로운 차원의 검색이론과 방법을 개발할 필요성이 증대된 것이다. 왜냐하면, 인터넷에 존재하는 자료들을 검색할 수 있는 일부 웹검색엔진들이 웹문서의 효율적인 검색을 위해 자체적으로 개발한 분류 체계를 운영하고 있기 때문이다. 또한, 웹검색엔진의 분류체계들은 각기 다른 구조를 갖고 있으며 그 구조화된 형태는 문헌분류체계의 원리를 원용하고 있기 때문이다. 따라서 도서관에서 사용하는 메뉴체계와 각기 서로 다른 메뉴체계를 웹검색엔진들이 사용함에 따라 이용자들은 디지털 검색에 많은 어려움과 혼란이 야기되고 있다.

이에 본 연구에서는 디지털 시대에 사용자를 위한 표준화된 역동적 웹문서 분류체계를 제시한다.

1. 문헌분류체계분석

분류의 목적은 크게 다음 두 가지로 대별될 수 있다. 하나는 학문분류로서 학문 자체의 분류 및 사물이나 개념의 상호간의 관계를 발견하는 수단으로 사용하는 것이다. 다른 하나는 문헌분류로서 자료(문헌)의 효과적 이용을 위해 체계적으로 정보를 배치하는 도구로서 활용하는 것이다. 일반적으로 문헌정보학분야에서 분류라 함은 후자의 경우에 속한다. 그 가운데에서 국내 대학 도서관계에서 가장 많이 사용되는 DDC와 KDC의 분류체계에 대해서는 많은 논란과 비판이 지속되어 왔던 것은 주지의 사실이며, 현재의 관점에서는 특히 DDC의 기본 골격자체에 문제점이 더욱 크게 부각되고

있다.(정필모) 본 연구에서는 도서관에서 널리 사용되고 있는 KDC가 실제로 어느 정도 균형있게 배열되어 있는지를 판단하였다. 분석대상으로는 전북 지역의 대학도서관 3곳을 선정하였다.

1.1 대학교 도서관 장서분석

분석대상으로 선정한 대학도서관은 오리지널 캐털로깅을 수행하는 기관이다. 즉, 다른 기관과의 연계성을 고려하지 않았기 때문에 장서구성에 독립적인 성격을 갖고 있다. 그럼에도 불구하고 장서분포도에 큰 차이를 보이지 않고 있다. 또한, 대학도서관의 특성상 학술적인 자료가 우선임에도 불구하고 류번호에 배정된 장서구성도의 편차가 매우 높다. 류번호의 장서배정의 균형을 이루기 위해서는 각 류별당 10%내외의 소장분포형태가 이상적이나 10%내외의 분포도를 보이는 것은 순수과학을 제외하고는 거의 없는 실정이다. 도서관 분류체계를 분석한 결과를 웹문서의 메뉴체계로 활용하는데 도출될 수 있는 문제점을 지적하면 다음과 같다.

1) 체제의 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 메뉴로 데이터를 찾기에는 불편함이 초래될 수 있다. 따라서 인터넷의 자원을 기존 도서관의 십진분류체계로 분류하기에는 너무 고정적인 형태라 할 수 있다.

또한, 일반적으로 도서관 장서에 부여된 분류번호(KDC)가 배정되면, 이의 수정은 현실적으로 불가능하다. 왜냐하면, 인터넷에 등재된 많은 웹문서들의 주제는 대부분 새로운 분야의 것이 있기 때문이다.

2) 장서의 불균형

KDC는 류항목간의 배열과 혹은 목이하의 항목간의 배열에 있어 해당 항목에 속해있는 장서의 양이 심하게 편중되어 있는 현상을 보이고 있다. 이를 확인하기 위해서 각 대학별로 가장 높은 장서구성비를 보이고 있는 사회과학분야(300)의 강부문과 가장 낮은 장서구성비를 보이고 있는 어학분야(700)의 강부문을 비교·분석하면 다음과 같다. 균형있는 분류체계일 경우에는 각 류별장서구성비가 10%내외이지만, 본 조사에서는 최대 분포비와 최소분포비의 차이가 25.50%라는 극심한 편차가 있었다. 특히, 강체계 자체적으로도 사회과학의 경우 그 편차가 9.55%(320강과 390강)를 확인할 수 있었다.

3) 계층간 불균형성

KDC는 거의 대부분 DDC의 분류체계를 수용하고 있다. DDC는 판이 개정되어도 기본적인 구조의 변화는 현실적으로 어렵다는 특징을 갖고 있다 즉, 새로운 학문분야나 주제영역이 발전되어도 류나 강, 목수준의 분류번호부여는 현실적으로 어렵게 되었다. 즉, 최신학문이지만 DDC의 초창기 개발시기에는 없었던 주제의 경우는 열거식 분류표의 특성상 주제분야 삽입이 어려운 것이다. 따라서 학문 발전의 순서는 늦었지만 현재 학문의 활발하게 연구가 이루어지는 분야는 세목이하의 분류번호가 배정되는 문제점을 내포하고 있다.

2. 웹검색엔진(web search engine)의 종류

인터넷에 존재하는 데이터들은 컴퓨터와 통신망등과 같은 물리적 요소와 HTML(Hyper Text Markingup Language)이라는 기술언어의 준칙과 그림파일로서는 gif파일등을 준용한다는 등의 기초적인 소프트웨어적인 기준만을 만족한다면, 주제나 형식에 상관없이 어떠한 내용의 웹문서도 웹에 등재될 수 있다. 따라서 웹은 분류와 조직에 있어 특정한 기준이 없으며 각 사이트의 개발자의 의도에 따라 다양한 형태를 보이고 있다. 이러한 웹검색엔진을 검

색범위와 특성에 따라 다음과 같이 4가지로 크게 구분한다.(남영준 1)

① 주제 검색엔진

주제 검색엔진은 전세계에 산만하게 흩어져 있는 웹문서들을 회사의 특성에 따라 준비한 분류체계에 맞추어 계단식 주제별 검색이 가능하도록 설계된 검색엔진이다. 대표적인 것으로는 '야후(KOREA)'와 '정보탐정', '심마니', 'Naver'등이 있다.

② 키워드 검색엔진

키워드 검색엔진은 별도의 분류표를 준비하지 않고 검색창에 직접 사용자가 원하는 정보를 입력하는 방식을 택한 검색엔진이다. 즉, 분류체계에 웹문서들을 링크한 것이 아니라 웹문서에 출현하는 주요 단어(keyword)에 웹문서들을 링크한 것이다. 특히, 이 방식은 특정 주제어가 정해져 있지 않고 사용자가 입력한 자연어로 검색을 시작하며, 자연어검색에서 야기될 수 있는 문제를 補整할 수 있도록 검색어 확장검색옵션도 제공한다.[1] 또한, 정도율(precision ratio)을 높일 수 있는 방법으로 검색대상을 제목과 본문으로 구분하여 검색할 수 있는 옵션도 제공한다.(남영준2) 단점으로는 검색어를 올바로 선정하지 않으면 불필요한 정보가 너무 많이 검색되거나 혹은 전혀 검색되지 않는 등 검색효율이 극단적으로 나타날 수 있다. 대표적인 검색엔진으로는 '알타비스타(KOREA)', 'Lycos', 'InfoSeek', 'WebCrawler', 'Excite', 'Magellan', 'OpenText', '까치네' 등이 있다.

③ 메타 검색 엔진

메타 검색엔진은 키워드 검색엔진의 일종이지만 각각의 웹검색엔진들이 갖고 있는 웹문서들을 대상으로 검색이 이루어지는 超키워드(super keyword) 검색엔진이라 할 수 있다. 대표적인 검색엔진으로는 'ALL-in -One'과 '미스 다찾니'가 있다.

④ 통합 검색 엔진

통합검색엔진은 메타 검색엔진과 유사하게 한 번의 검색으로 많은 검색 엔진에서 정보를 얻는 엔진이다. 그러나 엄밀하게 구분하면 통합 검색 엔진은 검색엔진이라기 보다는 서지데이터베이스에서 위치(holding & location)정보만을 제공하는 순수 디렉토리라 할 수 있다. 즉, 자체 검색 엔진을 사용하지 않고 각 회사의 웹검색엔진을 사용한다. 검색의 방법은 연결된 웹검색엔진 가운데 가장 우선적인 것을 검색하고, 다음 웹검색엔진을 검색하는 방법을 취함으로서 검색성능이 전적으로 연결된 검색엔진의 성능에 좌우된다. 대표적인 엔진으로는 SavvySearch가 있다.

2.1 정보탐정

정보탐정은 1998년도 5월 현재 12개의 류분야와 32개의 강분야로 주제를 구분하고 있다. 목분야와 세목분야, 세세목분야로 주제가 전개되고 있으며, 계속해서 목이하분야가 추가되고 있다. 정보탐정의 류개념을 KDC분류체계에 근거하여 분석하면 동일 주제개념이 부여된 것은 '언론·매체'를 비롯하여, '경제·산업', '영화·음악·연예', '건강·병원', '생활·주택', '정치·행정·법'등이다. 이 가운데 KDC의 류수준에 해당하는 것은 한 분야도 없다. 이를 강개념까지 확대하여 류수준의 주제와 강수준의 주제가 일치하는 것은 '경제·산업'분야와 '영화·음악·연예', '정치·행정·법' 3분야뿐이다. 실제적으로 방송(326.7)분야의 경우는 언론·매체(070)와 관련된 분야이면서 KDC상에서 서로 다른 분류번호가 배정된 것은 '방송'은 신문방

1) 예를 들면, 정보탐정의 경우 사용자가 검색어로서 '호출기'를 입력하면, 유의어인 '삐삐'나 '휴대용 무선호출기', '페이저', '비퍼'등 유사동의어로 검색범위를 확장한다.

	정보 산업	언론	경제	산업	레저	스포츠	취미/오락	교육	예술	건강	법/법률	과학	종교	사회/문화	사건	지역 정보
카테고리	226	187	90	305	293	175	26	837	240	248	22	129	119	343	3	88
사이트	5768	2741	3210	2,926	1696	1186	630	7023	3,736	1,544	610	866	592	2788	105	2562
총수	5994	2928	3,300	3,231	1989	1361	656	7860	3,976	1,792	632	995	711	3131	108	2650
백분율	14.5	7	8	7.8	4.8	3.3	1.6	19	9.6	4.3	1.5	2.4	1.7	7.6	0.3	6.4

<표 2> 새로운 웹문서분류체계의 문서배정율

송학이라는 학문적 분류를 우선한 것이고, 웹상에서는 '방송'자체로서 실용적인 분류가 우선한 것이다. 이는 웹분류체계의 원칙이 실용적인 것에 있는 것을 반증하는 한 예이다. 한편, '홈페이지'라는 개념은 KDC관점에서는 홍보에 해당하는 것이지만 실제적으로는 컴퓨터에 관련된 주제가운데 파생된 하나의 개념이라 할 수 있다. 강수준으로 있는 '개인과 단체'는 주제적으로 개인홈페이지와 단체홈페이지에 해당하기 때문에 이를 KDC로 분류하는 것은 더욱 어려운 상황이다. 또한, 개인과 단체라는 항목은 강수준에 배열된 디스크립터이면서 다른 영역에서도 나타나는 항목이다. 즉, '개인과 단체' 항목은 KDC에서 활용되고 있는 보조기호표적인 성격이 매우 강하다. 정보탐정은 한 항목에 최대 3개이상의 개념을 복합적으로 사용하고 있다.

2.2 심마니 (http://simmany.chollian.net/)

심마니는 1998년 5월 현재 개인 홈페이지를 비롯하여 16개의 류항목을 1차메뉴체계로 제공하고 있으며 강항목수준으로는 163개항목을 제공하고 있다. 특징적인 것은 류항목에 류주제항목에 연결되어 있는 웹문서와 하위항목에 대한 정보를 제공하고 있는 점이다. 심마니의 경우 류수준에 있는 16개 항목가운데 KDC의 류수준과 일치하는 것은 '과학'을 비롯하여 '역사', '예술', '종교' 를 들 수 있다. 이를 강수준까지 확대하면 '사회·문화'과 '건강·병원', '오락·취미', '법률'를 비롯하여 대부분이 포함된다. 심마니도 다른 웹검색엔진과 같이 메뉴체계상에서 하나의 항목에 두 개의 개념을 부여하여 해당 주제 영역을 복합적으로 설정하고 있다. 심마니에서도 '단체·기관'과 같이 메뉴체계에서 여러 항목의 하부메뉴로 제공하는 개념들이 존재한다. 또한, 상위 체계와 하위체계의 개념 또는 항목이 동일한 경우도 발생한다. <그림 1>

과학
  과학
  단체·기관
    단체·기관
    공공기관
    동호회
    연구소
  순수과학√
    순수과학√
    물리학
      물리학
      단체·기관

<그림 1> 심마니 분류체계의 전개예

2.3 웹분류체계의 특징

웹문서분류체계는 기본적으로 다음과 같은 특징을 지니고 있다.

① 학문적 특성보다는 현실적인 특성이 우선하고 있다.

② 분류의 기본적인 원칙은 개념분류가 아니며, 단순히 키워드 분류에 머무르고 있다. 예를 들면, 자연과학이라고 분류된 문헌과 순수과학이라고 분류된 문헌의 종류가 다르다. 단

지, 해당 문헌들은 개념적 군집화보다 해당 용어가 웹문서에 있고 이 단어가 색인어로 선정되었기 때문에 해당 항목에 링크된다.

③ 호환성이 없다. 예를 들면, 열거식 분류표상에서 KDC로 분류된 자료와 DDC로 분류된 자료간에는 최소한의 필터링 작업으로 분류번호의 군집화가 가능하지만, 웹문서 분류체계에서는 검색엔진의 특성이 강하기 때문에 분류의 호환이 현실적으로 어려움이 있다.

④ 한 문헌에 대해 여러개의 분류항목이 부여된다. 도서관의 장서에는 하나의 분류개념이 부여되는 것이 일반적이나 웹문서의 경우는 복수의 개념 부여가 일반적이다.

3. 웹분류체계 설계

본 연구에서 제시할 웹분류체계는 다음과 같이 메뉴체계설계 원칙을 설정하였다.

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

② 류분야를 비롯한 하부전개에 최소 분야만을 제시하되, 고정된 전개(예를 들면, 10진) 원칙을 따르지는 않는다.

③ 류분야와 강분야의 구분은 학술적인 관점과 사용빈도를 고려하여 설정한다.

3.1 웹검색엔진용 분류체계설계

앞에서 입수한 결과에 근거하여 다음과 같이 웹검색엔진용 분류체계를 설계하였다. 본 연구에서 제시한 분류체계는 류강수준의 가장 기본적인 분류체계를 제시하였다. 설계의 기본 원칙은 다음과 같다.

① 류항목에 배열된 주제는 학문적 개념을 택한다.

② 전개원칙은 학문적 분류체계에서 나타난 원칙으로 학문의 상위개념에서 하위개념으로 전개하는 분화원칙을 따른다.

③ 세부적으로 전개될수록 항목은 학문적 개념보다 실용적 개념을 택한다.

이상과 같은 원칙에 따라 류개념을 16개로 선정하였으며 강개념로는 122개 항목을 선정하였다. 설계의 타당성을 검증하기 위해 야후(한국판)를 대상으로 해당 류항목과 강항목에 배정된 디스크립터에 링크된 카테고리수와 엡문서의 수를 계수하였다. 계수한 결과는 다음 <표 2>와 같다.

위 표에서는 41,314개의 웹문서 및 카테고리를 대상으로 분석이 이루어졌다. 류·강 항목상에서는 최고 19%(교육)과 최저 0.3%(사건)로 편차가 크나 이러한 현상이 나타나는 것은 다음과 같은 웹검색의 특징에 기인한다.

개념색인이기 보다는 용어색인방식을 택하고 있기 때문에 해당 용어와 일치하지 않은 문서의 경우는 링크가 이루어지지 않기 때문에 이상과 같은 편차를 보이고 있다.

한 주제에 대해 세부적인 전개가 이루어지지 않을 경우 범용적인 디스크립터에 링크된 웹문서의 수가 상대적으로 적다. 즉, '과학'이란 용어에 링크된 문서보다는 '수학'과 '물리'와 같이 구체적인 용어를 항목으로 결정하는 것이 훨씬 많은 웹문서들이 링크된다.

3.2 웹검색엔진용 분류체계의 제한 및 분석

앞에서 제시한 메뉴체계는 학문적 성격이 상대적으로 많이 반영된 메뉴체계이다. 강항목이하 목, 세목, 세세목분야까지의 확대가 이루어진 후의 결과는 실제적인 성격이 많이 가미될 것이다. 새로운 웹분류체계의 검증은 현재 웹검색엔진에 있는 자료들을 대상으로 함에 따라 다음과 같은 제한점이 있다.

① 웹문서 자료가 실시간으로 변화되기 때문

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

에 완벽한 근거자료의 제시에 어려움이 있다.

② 웹문서의 수가 많기 때문에 웹검색엔진에서 제공하는 웹문서의 수는 회사자체 계수과정에 불완전성에도 불구하고 분류체계에 나타난 수치데이터를 원용하고 있다.

③ 엔진별 공인 접속횟수는 분류체계 설계에 중요한 기준이 되고 있지만 광고수입과 같은 요인 때문에 모든 검색엔진을 대상으로 정확한 정보를 입수할 수 없었다. 단, 일부회사에 한해 접속비율을 공개했으나 일정기간의 자료이기보다는 특정일 하루만의 데이터를 공개하여 적용자료로 활용할 수가 없었다.

결 론

도서관은 웹문서처리에 대해 새로운 해석과 기술을 개발해야 할만큼 웹문서의 수와 이용자가 증가하였다. 따라서 웹에 대한 새로운 접근이 필요하나 현재 웹검색엔진의 분류체계는 도서관학적 관점보다는 현실적 상황 체계로 구분하여 도서관 이용자들의 어려움을 야기하고 있다. 따라서 본 연구에서는 새로운 웹문서체계를 제시하였으며, 새로운 웹문서분류체계에 따라 실제로 웹문서의 분포도를 측정하였다. 그 결과는 다음과 같았다.

첫째, 인터넷상의 웹문서를 분류하기 위해 도서관에서 사용하는 기존의 문헌분류체계로 적용하는 것은 바람직하지 않다.

둘째, 웹문서의 색인은 웹문서에 나타난 용어를 기준으로 하기 때문에 분류 체계의 항목에 링크된 것은 일차적으로 개념색인보다는 단순 용어색인에 해당한다.

셋째, 넓은 의미를 갖는 용어가 디스크립터로 사용된 경우는 세부적으로 용어가 전개되지 않을 경우 상대적으로 링크된 문헌이 적었다.

넷째, 웹문서분류체계로 검색이 효율적으로 이루어지기 위해서는 반드시 전거사전이 구축되어야 한다.

<<참고문헌>>

• 정필모. 국제백진분류법연구 <인문학분야편>. 중앙대학교 출판부. 1995. pp.1-5.

• 남영준 1. 인터넷으로 떠나는 세계여행. 전주대학교 출판부. 1998. pp.75-76.

• 남영준 2. 디지털정보검색론. 전주대학교 출판부. 1998. pp.42-60.

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

• H. Chen *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) : 582-603.

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

류항목	강 항 목
정보산업	컴퓨터, 정보통신, 인터넷, 인트라넷, 반도체
언론	신문, 잡지, 방송, 방송국, 광고/홍보, 출판
경제	금융, 보험, 주식(증권), 채권,세금, 수/출입, 부동산, 생활,창업, 기업, 국제경제
산업	농업, 제조업, 에너지, 금속(철강), 기계, 섬유산업, 토목/건축, 운송, 마케팅산업, 전기/전자산업, 상업,유통, 서비스업
레저	여행, 도로, 관광
스포츠	구기스포츠, 무술, 종합경기, 육상, 동계스포츠, 수상스프츠,자동차, 오토바이, 체육시설
취미/오락	오락, 취미, 동호회
교육	유치원, 초등학교, 중고등학교, 대학교, 대학원, 특수학교,학원, 유학, 취업, 입대
예술	문학,연극,영화,영화제, 비디오, 공연장, 공연예술, 음악,악기, 연예인, 운동선수,미술,패션
건강	질병, 병원(의학), 한의원, 약국, 치과병원, 건강관리, 미용, 성인, 유/아동
법/법률	헌법, 법률, 재판, 소송
과학	자연과학, 응용과학, 인문과학, 사회과학
종교	가톨릭(천주교),기독교,불교, 유교, 민간신앙, 사이비종교
정치	선거 외교 국회 국방 통일
정부	행정부 사법부
사회/문화	인구문제, 사회운동, 성문제, 결혼, 장례, 제사,사회복지,연금, 환경공해, 도서관, 박물관
사건	사고, 재해
지역정보	아시아, 북미, 중남미, 유럽, 오세아니아/극지, 한국