국내 웹 디렉토리들의 커버리지 및 커버리지 중복성 분석*

Analyzing Coverage and Coverage Overlap of Korean Web Directories

배희진(Hee-Jin Bae)**, 이진숙 (Jin-Suk Lee)***,
이준호(Joon-Ho Lee)****, 박소연(So-Yeon Park)*****

초 록

본 연구에서는 국내 주요 웹 검색 포탈인 네이버, 야후 코리아, 엠파스가 제공하는 웹 디렉토리들의 커버리지 및 커버리지 중복성을 분석하였다. 이를 위하여 본 연구는 웹 디렉토리에 등록된 사이트들의 수집 방법을 개발하고, 대분류 매핑, 중복 분류 및 참조 링크 고려와 같은 커버리지 및 커버리지 중복성 분석에 필요한 방법론을 제시하였다. 조사 결과, 참조 링크의 허용 여부가 웹 디렉토리의 커버리지에 매우 큰 영향을 미치며, 국내 웹 디렉토리들 사이의 커버리지 중복성이 매우 낮은 것으로 나타났다. 본 연구는 국내 웹 디렉토리들에 대한 이해를 넓히고, 웹 디렉토리들의 커버리지 및 커버리지 중복성 분석에 필요한 방법론을 제시함으로써, 웹 디렉토리에 관한 연구에 기여할 것으로 기대된다.

ABSTRACT

This study examines coverage and coverage overlap of the three major Korean web directories, Naver, Yahoo Korea, and Empas. This study also suggests a methodology for collecting and processing web sites provided by these web directories. A method for mapping main categories was developed. Each directory provided registered web pages in a slightly different way. Reference links had a significant influence on the coverage of each web directory. The overlap of pages among three directories was quite low. It is expected that this study could contribute to the field of web research by providing insights to how directories provide web pages and suggesting a methodology for the analysis of directory coverage.

키워드: 웹 검색 포탈, 웹 디렉토리 수집, 웹 디렉토리 평가, 커버리지, 커버리지 중복성, 참조 링크, web search portal, web directory crawling, web directory evaluation, coverage, coverage overlap, reference link

---

* 본 연구는 숭실대학교 교내연구비 지원으로 이루어졌음
** 숭실대학교 컴퓨터학부 대학원 박사과정 (jinybae@irlab.ssu.ac.kr)
*** 숭실대학교 컴퓨터학부 대학원 석사과정 (gufoo@irlab.soongsil.ac.kr)
**** 숭실대학교 컴퓨터학부 부교수(joonho@computing.soongsil.ac.kr)
***** 덕성여자대학교 문헌정보학과 조교수(sypark@duksung.ac.kr)
■ 논문 접수일 : 2004. 2. 25
■ 게재 확정일 : 2004. 3. 18

1 연구의 목적

1990년대 중반 이후 인터넷의 사용과 보급이 폭발적으로 증가함에 따라 인터넷을 통한 정보의 접근을 지원하기 위해 웹 검색 서비스들이 활성화되었다. 이러한 웹 검색 서비스들의 성능 개선을 위해서는 기존 웹 검색 서비스들의 비교 및 평가가 선행되어야 한다. 지금까지의 연구들은 웹 검색 서비스들의 성능을 비교 및 평가하기 위하여 검색 엔진의 커버리지, 갱신 주기, 안정성, 중복 페이지 검출, 스팸 및 포르노 페이지 검출, 응답 속도, 연결성 등의 다양한 척도들을 사용하였다(Dong & Su, 1997; Heydon & Najork, 1999; Najork & Heydon, 2001). 이러한 평가 척도 중 커버리지는 특정한 검색 서비스나 저널이 제공하는 주제의 범위, 자료의 범위를 의미하며, 검색엔진의 커버리지는 검색엔진이 수집하여 제공하는 사이트의 수와 다양성으로 측정될 수 있다. 검색엔진의 커버리지는 검색 서비스가 제공하는 자료가 얼마나 포괄적인가를 나타내는 지표로서, 웹 이용자들이 검색 엔진을 선택하는 데 영향을 미치는 중요한 요소이므로, 많은 검색 서비스들이 커버리지를 향상시키기 위하여 노력하고 있다.

국외의 경우, 웹 검색 엔진들의 커버리지를 분석한 대표적인 연구는 Lawrence & Giles(1998, 1999)에 의해 수행되었다. 즉, 1998년 연구에서 NEC에 근무하는 연구자들이 선택한 575개의 질의를 6개의 검색 엔진들에서 실행한 후, 이들의 검색 결과에 근거하여 검색 엔진들의 커버리지와 검색 엔진들 사이의 커버리지 중복성을 조사하였다. 또한 1999년 2월 한 달 동안 실험을 통하여 전체 웹 서버를 280만으로 추정하였다. 그리고 이들 중 2500개의 웹 서버를 임의로 추출한 후, 이들이 제공하는 웹 페이지들의 수를 조사함으로써 전체 웹의 규모를 추정하였다. 이들은 이러한 연구들을 통하여 각각의 검색 엔진은 웹의 일부분만을 수집, 색인하며, 검색 엔진들 사이의 커버리지 중복성이 낮다는 사실을 발견하였다.

한편, 국내 선행 연구의 경우 웹 검색 엔진이나 디렉토리들의 커버리지와 커버리지 중복성을 분석한 연구는 미진한 편이다. 디렉토리 구축과 관련된 연구들은 크게 특정한 주제 분야나 영역의 디렉토리 구조를 분석하고 개선 방안을 제시하는 연구(김영보, 1997; 오동근, 황재영, 배영활, 2001; 이란주, 성기주, 양정하, 2001; 정연경, 2001; 최희윤, 1998; 한상길, 2001)와 다수 디렉토리들의 구조를 비교, 분석하여 디렉토리 구축의 지침을 제시하는 연구(곽철완, 2001; 신동민, 2001), 디렉토리 설계에 전통적 분류 이론과 체계를 적용한 연구(최재황, 1998), 한국십진분류표와 주요 웹 검색엔진을 통합한 분류체계를 개발로 한 연구(남영준, 1998; 남영준, 최승순, 2002)로 구분될 수 있다.

국내외 선행 연구 중 실제 서비스되고 있는 웹 디렉토리들을 대상으로 웹 디렉토리들의 커버리지 및 커버리지 중복성을 분석한 연구는 찾아보기 어려운 실정이다. 이에 본 연구에서는 국내 주요 웹 검색 포탈들이

제공하는 웹 디렉토리들의 커버리지 및 커버리지 중복성을 분석하고자 한다. 즉, 네이버, 엠파스, 야후 코리아의 웹 디렉토리에 등록된 사이트들의 수를 조사하고, 서로 다른 웹 디렉토리들에 동시에 등록된 사이트들의 분석을 통하여 웹 디렉토리들 사이의 커버리지 중복성을 파악하고자 한다. 이를 위하여 본 연구는 웹 디렉토리에 등록된 사이트들의 수집 방법을 개발하고, 대분류 매핑, 중복 분류 및 참조 링크 고려와 같은 커버리지 및 커버리지 중복성 분석에 필요한 방법론을 제시하였다.

현재 웹 검색 포탈들이 제공하고 있는 웹 디렉토리들의 구축 현황을 보여주는 본 연구의 결과는 보다 합리적이고 체계적인 웹 디렉토리 구축을 위한 근거로서 활용될 수 있을 것이다. 또한 본 연구의 결과는 이용자들이 웹 디렉토리들을 선택하고 활용하는 전략을 세우는 데에도 영향을 미칠 것으로 기대된다.

2 선행 연구

디렉토리 구축과 관련된 국내 선행 연구를 구체적으로 살펴 보면 첫째, 특정한 주제 분야나 영역의 디렉토리 구조를 분석하고 개선 방안을 제시하는 연구를 들 수 있다. 곧 김영보(1997)는 컴퓨터, 인터넷 분야를, 오동근, 황재영, 배영활(2001)은 군사학 분야를, 이란주, 성기주, 양정하(2001)는 여성학 분야를, 정연경(2001)은 인터넷 서점을, 최희윤(1998)은 물리학 분야를, 그리고 한상길(2001)은 산업 분야를 중심으로 디렉토리 구조를 검토하고 개선 방안을 제시하였다. 둘째, 신동민(2001)은 다수의 디렉토리 구조의 비교, 분석을 통하여 디렉토리 구성, 주제 선정, 인터페이스 등과 관련된 지침을 제시하였다. 곽철완(2001)은 세 쇼핑몰의 분류체계 비교, 분석을 통하여 인터넷 쇼핑몰 분류체계 구축을 위한 지침을 제시하였다.

셋째, 남영준(1998)은 정보탐정, 심마니, 야후 코리아, 네이버의 디렉토리의 웹문서분류체계와 한국십진분류표의 비교를 통하여 실험적인 웹 문서 분류체계를 설계하였다. 또한 남영준, 최승순(2002)은 한국십진분류표와 심마니, 야후코리아, 한미르의 분류체계의 비교, 분석을 통하여 전통적인 분류체계와 웹문서분류체계를 통합할 수 있는 분류체계를 개발하였다. 이들의 통합 분류체계는 한국십진분류표의 구조와 원칙을 수용하며, 각 검색엔진에서 공통적으로 채택하고 있는 개념과 구조를 반영하였다. 곧 남영준, 최승순은 류와 강항목까지의 계층을 개발하였으며, 항목의 자모순으로 배열하였다. 또한 인터넷 자료의 특수성을 고려하여 한국십진분류표에 사용되지 않은 개념 가운데 필요한 부분은 웹 검색엔진에서 사용되는 체계와 용어를 사용하되, 웹 검색엔진 분류체계의 비교를 통하여 일치도와 중복도를 기준으로 하였다. 본 연구는 네이버, 엠파스, 야후 코리아가 제공하는 웹 디렉토리들의 상이한 대분류들을 포괄하기 위하여 남영준, 최승순의 통합 분류체계 중 류항목 부분을 참고하였다.

3 연구 방법

웹 사이트들에 대한 효과적인 접근을 지원하기 위하여 국내외의 대다수 웹 검색 포탈들은 웹 사이트들을 주제별로 분류한 디렉토리 서비스를 제공하고 있다. 디렉토리는 기본적으로 상하, 연관 관계와 같은 주제어의 상호 관계를 나타내고, 일반적으로 단계별 계층 구조로 구성된다(신동민, 2001).

본 연구에서는 국내 주요 웹 검색 포탈들인 네이버, 엠파스, 야후 코리아를 대상으로 웹 디렉토리들의 커버리지 및 커버리지 중복성을 분석하고자 한다. 디렉토리 서비스를 제공하는 국내 여러 포탈 서비스들 중에서 네이버, 엠파스, 야후 코리아를 선택한 이유는 첫째, 이들의 인지도와 대중성 때문이며, 둘째, 이들이 다른 서비스들보다 적극적으로 디렉토리 서비스를 지원하고 있기 때문이다. 즉, 2003년 국가 고객 만족 지수(NCSI) (http://www.ncsi.or.kr) 검색 포탈 부문에서 네이버와 엠파스는 공동 1위였고, 야후 코리아가 3위를 차지하였다. 또한 웹 사이트 평가 및 트래픽 분석업체인 인터넷 매트릭스 (http://www.metrixcorp.com)에 의하면 국내 이용자들이 2004년 2월 1개월 동안 가장 많이 방문한 검색 포탈들이 네이버, 야후 코리아, 엠파스, 한미르 순으로 나타났다. 또한 다음, 드림위즈, 네이트, 하나포스, 네띠앙 등과 같은 다른 포탈 서비스들의 경우 디렉토리 서비스를 제공은 하나 초기 페이지에서 제공하지 않고 있다. 따라서, 디렉토리 서비스를 초기 페이지에서 적극적으로 제공하고 인지도도 높은 네이버, 엠파스, 야후코리아의 웹 디렉토리들의 분석을 통하여 국내 웹 검색 포탈들의 전반적인 웹 사이트 수집 및 처리 현황을 파악할 수 있을 것으로 사료된다.

네이버, 엠파스, 야후 코리아가 제공하는 웹 디렉토리들의 커버리지 및 커버리지 중복성을 분석하기 위하여 2003년 4월 11일부터 15일 동안 네이버, 엠파스, 야후 코리아의 웹 디렉토리에 등록된 사이트들을 수집하였다. 사이트의 수집 당시 네트워크 혼잡으로 인한 수집 속도의 저하, 웹 디렉토리 서버의 장애 등 외적인 요소로 인하여 네이버에 등록된 웹 사이트들의 수집은 2003년 4월 26일에 완료되었고, 엠파스와 야후 코리아에 등록된 웹 사이트들의 수집은 각각 4월 14일과 4월 19일에 완료되었다.

3.1 웹 사이트 수집

웹 검색 포탈들이 제공하는 웹 디렉토리들의 커버리지 및 커버리지 중복성을 분석하기 위해서는 분석 대상인 웹 디렉토리들에 등록된 웹 사이트들의 수집이 수행되어야 한다. 본 연구에서 웹 디렉토리에 등록된 웹 사이트들을 자동으로 수집하기 위해 사용한 방법은 준비 과정과 수집 과정으로 구성되어 있다. 준비 과정에서는 웹 디렉토리의 대분류 URL들을 대분류 URL 큐에 삽입한다. 예를 들어, 네이버의 웹 디렉토리는 14개의 대분류들로 구성되어 있으며, 네이버 웹 디렉토리에 등록된 웹 사이트들을 수집

하기 위해서는 이들의 URL들을 대분류 URL 큐에 저장한다. 수집 과정은 다음과 같은 3단계로 구성되어 있다.

단계 1: 대분류 URL 큐로부터 하나의 대분류 URL을 삭제하고, 이 대분류 URL을 디렉토리 URL 큐에 삽입한다. 이때 대분류 URL 큐가 비어 있을 경우, 웹 사이트 수집이 종료된다.

단계 2: 디렉토리 URL 큐로부터 하나의 URL을 삭제하고, 이 URL에 해당하는 웹 페이지를 수집한다. 이때 디렉토리 URL 큐가 비어 있을 경우, 단계 1을 다시 수행한다.

단계 3: 수집된 웹 페이지를 분석하여 URL들을 추출하고, 이들 중에서 디렉토리 URL들은 디렉토리 URL 큐에 삽입하고, 웹 사이트 URL들은 웹 사이트 URL 파일에 저장한다. 그리고 단계 2를 다시 수행한다.

3.2 통합 대분류

<표 1>은 네이버, 엠파스, 야후 코리아가 제공하는 웹 디렉토리들의 대분류들을 보여준다. 세 검색 포탈들이 제공하는 웹 디렉토리들은 모두 14개의 대분류들로 구성되어 있다[1]. 이러한 웹 디렉토리들의 분류 체계는 DDC(Dewey Decimal Classification)나 한국십진분류표와 같은 전통적이고 이론적인 문헌 분류 체계에 근거한 것이 아니라 이용자들의 관심도나 편의성, 시대성을 반영하는 것이라고 볼 수 있다. 이용자의 관심사와 시대성은 주관적이고 유동적이므로, 각 웹 디렉토리들의 전개방법과 항목에는 차이를 보이고 있다(남영준, 최승순, 2002). 곧 세 웹 디렉토리들의 대분류들 중 일부 대분류명은 완전히 일치하지만, 나머지 대분류명들에 있어서는 약간의 차이가 있음을 알 수 있다. 예를 들어, 세 검색 포탈들은 모두 "뉴스," "건강, 의학," "컴퓨터, 인터넷"이라는 대분류를 공통으로 제공한다. 그러나 네이버와 야후 코리아가 "비즈니스와 경제"라는 대분류를 제공하는데 비해 엠파스는 이와 유사한 "경제, 재테크"라는 대분류를 제공한다.

따라서, 네이버, 엠파스, 야후 코리아가 제공하는 웹 디렉토리들의 커버리지 및 커버리지 중복성을 대분류별로 분석하기 위해서는 상이한 대분류들을 포괄할 수 있는 별도의 분류 체계가 필요하다. 이를 위하여 본 연구에서는 남영준과 최승순이(2002) 제안한 통합분류체계를 참고하고, 대분류들간의 일치성과 중복성을 고려하여 세 검색 포탈들이 제공하는 웹 디렉토리들의 대분류들을 포괄할 수 있는 총 10개의 통합 대분류들을 <표 1>과 같이 제시하였다. 또한 남영준과

1) 이들 중 야후 코리아의 대분류들은 사회과학이 인문과학과 통합되어 있다는 점을 제외하고는 야후(Yahoo, U.S.A.)의 대분류들과 동일하였다.

최승순의 통합분류체계가 2001년 9월 당시의 야후, 한미르, 심마니의 웹 디렉토리의 분류 체계의 비교, 분석에 근거한 만큼, 현재의 연구에 직접적으로 적용되기 어려운 경우, 세 디렉토리 서비스들간의 대분류 항목의 일치성과 중복성에 따라 통합 대분류를 개발하였다. 곧, 네이버의 "학문, 과학", 엠파스의 "학문", 야후 코리아의 "인문, 사회과학"과 "자연과학"은 모두 "학문"이라는 대분류로 통합되었다. 또한 동일한 개념이 별도의 단일 대분류 항목으로 분리되어 있는 경우와 다른 개념과 통합되어 대분류를 구성하는 경우, 포괄적인 대분류를 채택하였다.

3.3 중복 분류와 참조 링크

웹 디렉토리들로부터 웹 사이트들을 수집한 결과, 동일한 웹 사이트가 다수의 디렉토리에 등록되어 있는 경우를 발견하였다. 이러한 중복 분류는 웹 사이트의 특성과 부합하는 디렉토리가 두개 이상 존재할 경우, 이 웹 사이트를 다수의 디렉토리들에 중복해서 등록하기 때문에 발생한다. 예를 들어, 네이버 '지도검색' 사이트는 두개의 디렉토리들 '네이버 홈>교육, 참고자료>참고자료>지도'와 '네이버 홈>컴퓨터, 인터넷>인터넷>포털>네이버'에 중복 등록되어 있고, 네이버 '메일'

〈표 1〉 통합 대분류로의 매핑

네이버	엠파스	야후 코리아	통합 대분류
뉴스, 미디어	뉴스, 미디어	뉴스와 미디어	뉴스, 언론
건강, 의학	건강, 의학	건강의학	의학
컴퓨터, 인터넷	컴퓨터, 인터넷	컴퓨터와 인터넷	컴퓨터, 인터넷
지역정보	대한민국, 세계	지역정보	지역정보
교육, 참고자료	교육, 학교 사전, 참고자료	교육 참고자료	교육, 참고자료
학문, 과학	학문	인문, 사회과학 자연과학	학문
비즈니스, 경제 쇼핑	경제, 재테크 기업, 쇼핑몰	비즈니스 경제	경영, 경제
가정, 여성 사회, 문화	생활, 취미 정부기관, 사회	사회와 문화 정부	사회, 정부
엔터테인먼트, 예술	문화예술, 종교 연예, 오락	예술 엔터테인먼트	예술, 엔터테인먼트
게임 스포츠 레크레이션	여행, 스포츠	여가생활과 스포츠	여가, 스포츠

사이트는 '네이버 홈>컴퓨터, 인터넷>인터넷>전자우편>무료메일계정'과 '네이버 홈>컴퓨터, 인터넷>인터넷>전자우편'에 중복 등록되어 있다.

한편, 대부분의 웹 검색 디렉토리들은 참조 링크를 포함하고 있으며, 일반적으로 국내 웹 검색 포탈들은 참조 링크를 표시하기 위해 '@'를 사용한다. 이러한 참조 링크는 다른 디렉토리에 사이트가 등록되어 있지만 현재 디렉토리와의 연관성 때문에 연결되어진 범주를 나타내며, 구체적으로 이미 분류체계 내에 존재하는 디렉토리를 다른 디렉토리의 하위 디렉토리로 등록할 때 사용된다. 예를 들어, 엠파스의 '연예, 오락' 디렉토리에는 '음악@'이라는 하위 디렉토리가 존재하며, 이 디렉토리는 '엠파스 홈>문화예술, 종교>예술>음악' 디렉토리로 참조 링크되어 있다.

웹 디렉토리를 구성하는 각각의 대분류에 대하여 커버리지 및 커버리지 중복성을 조사할 경우, 이러한 중복 분류와 참조 링크의 배제 및 허용 여부는 조사 결과에 커다란 영향을 미칠 수 있다. 특히 중복 분류의 배제는 다수의 대분류에 등록된 웹 사이트가 하나의 대분류에만 출현하도록 만드는 문제를 야기한다. 본 연구는 커버리지 및 커버리지 중복성을 대분류별로 분석하고자 하므로, 하나의 웹 사이트가 다수의 대분류들로 중복 분류되는 것을 허용한다. 단, 특정 대분류 내에서 다수의 하위 디렉토리들에 중복 분류된 웹 사이트는 그 대분류에 한번 등록된 것으로 간주한다.

3.4 중복성 척도

네이버, 엠파스, 야후 코리아가 제공하는 웹 디렉토리들의 커버리지 중복성 분석을 위해서, 서로 다른 2개의 웹 디렉토리들 사이의 중복성을 측정할 수 있는 척도와 서로 다른 3개의 웹 디렉토리들 사이의 중복성을 측정할 수 있는 척도가 필요하다. 본 연구에서는 색인의 중복성을 측정하기 위하여 개발되어 널리 사용되고 있는 Rolling(1981)의 공식을 근거로 하여 2개의 웹 디렉토리들 A, B 사이의 중복성을 측정하는 함수 *Overlap*2 (*A,B*)와 3개의 웹 디렉토리들 A, B, C 사이의 중복성을 측정하는 함수 *Overlap*3(*A,B,C*)를 다음과 같이 정의한다.

$$Overlap2\ (A,B) = \frac{2\times N(A\cap B)}{N(A)+N(B)}$$

$$Overlap3(A,B,C) = \frac{3\times N(A\cap B\cap C)}{N(A)+N(B)+N(C)}$$

여기에서 N(S)는 집합 S에 포함된 사이트들의 수이다. 예를 들어 N(A∩B)는 집합 *A*와 집합 *B*의 교집합에 포함된 사이트들의 수를 의미한다.

함수 *Overlap*2와 함수 *Overlap*3는 입력 집합들 사이에 중복이 전혀 없으면 중복도 0을 생성하고, 입력 집합들이 동일한 사이트들을 포함할 경우 중복도 1을 생성한다. 즉, 이 함수들은 0부터 1사이의 값을 생성하고, 입력 집합들 사이에 중복이 많을수록 큰 값을 생성한다. 예를 들어, 3개의 입력 집합 A, B, C에 포함된 사이트들의 수가 각각 1000개, 800개, 600개이고, 이들 사이의 중복도가 0.6

이라고 가정하자. 이러한 사실은 평균적으로 집합 *A, B, C*에 포함된 사이트들 중의 480개 사이트가 중복 출현함을 의미한다.

4 커버리지 분석

네이버, 엠파스, 야후 코리아의 웹 디렉토리들에 등록된 유일 웹 사이트들의 수는 각각 157,194개, 163,445개, 190,858개이다. 즉, 야후 코리아의 웹 디렉토리가 국내 웹 사이트들에 대한 가장 높은 커버리지를 제공하고 있으며, 야후 코리아의 웹 디렉토리 커버리지는 네이버 1.21배, 엠파스의 1.17배에 해당한다. 다음에서는 참조 링크를 배제한 경우와 참조 링크를 허용한 경우에 대하여, 네이버, 엠파스, 야후 코리아가 제공하는 웹 디렉토리들의 커버리지를 2.2절에서 정의된 통합 대분류별로 기술한다.

4.1 참조 링크 배제

<표 2>는 참조 링크를 배제한 경우, 네이버, 엠파스, 야후 코리아가 제공하는 웹 디렉토리들의 커버리지를 통합 대분류별로 보여준다. 예를 들어 참조 링크를 배제한 경우, 통합 대분류들 중 하나인 "뉴스"에 대한 네이버, 엠파스, 야후 코리아의 커버리지는 각각 1,970개, 3,440개 2,070개이다. <표 2>로부터 "지역 정보" 대분류가 웹 디렉토리들의 커버리지에 결정적인 영향을 미치고 있음을 알 수 있다. 즉, "지역 정보"에 등록된 사이트들을 제외할 경우, 네이버, 엠파스, 야후 코리아에 등록된 사이트들의 수가 유사하다. 그리고 "지역 정보" 카테고리를 제외할 경우, "경영, 경제," "예술, 엔터테인먼트" 카테고리 순으로 가장 많은 사이트들이 등록되어 있다.

〈표 2〉 국내 웹 디렉토리들의 통합 대분류별 커버리지(참조 링크 배제)

	네이버		엠파스		야후 코리아	
뉴스, 언론	1,970	(1.1%)	3,440	(2.1%)	2,070	(0.9%)
의학	3,380	(1.9%)	4,706	(2.9%)	2,787	(1.3%)
컴퓨터, 인터넷	3,559	(2.0%)	4,621	(2.8%)	4,518	(2.0%)
지역정보	17,691	(9.9%)	2,266	(1.4%)	68,606	(30.7%)
교육, 참고자료	16,064	(9.0%)	16,582	(10.1%)	12,665	(5.7%)
학문	7,064	(4.0%)	7,445	(4.5%)	10,594	(4.7%)
경영, 경제	81,591	(45.8%)	55,789	(33.9%)	64,894	(29.1%)
사회, 정부	15,041	(8.5%)	22,573	(13.7%)	19,248	(8.6%)
예술, 엔터테인먼트	19,648	(11.0%)	37,619	(22.9%)	30,731	(13.8%)
여가, 스포츠	12,157	(6.8%)	9,294	(5.7%)	7,067	(3.2%)
총계	178,165	(100.0%)	164,335	(100.0%)	223,180	(100.0%)

한편, 네이버, 엠파스, 야후 코리아의 웹 디렉토리들에 등록된 유일 웹 사이트들의 수는 각각 157,194개, 163,445개, 190,858개이다. <표 2>의 총계는 이러한 유일 웹 사이트들의 수보다 크며, 이는 하나의 웹 사이트가 다수의 통합 대분류들로 중복 분류되었기 때문이다. 즉, 네이버, 엠파스, 야후 코리아는 하나의 웹 사이트를 평균적으로 각각 1.13개, 1.01개, 1.17개의 통합 대분류들로 중복 분류함을 알 수 있다.

4.2 참조 링크 허용

<표 3>은 참조 링크를 허용한 경우, 네이버, 엠파스, 야후 코리아가 제공하는 웹 디렉토리들의 커버리지를 통합 대분류별로 보여준다. 예를 들어 참조 링크를 허용한 경우, 통합 대분류들 중 하나인 "뉴스 ,언론"에 대한 네이버, 엠파스, 야후 코리아의 커버리지는 각각 39,341개, 45,390개 105,424개이다. <표 2>와 <표 3>의 총계를 비교함으로써 세 검색 포탈 모두 많은 수의 참조 링크를 이용하고 있으며, 특히 엠파스와 야후 코리아가 네이버에 비하여 매우 많은 수의 참조 링크를 사용함을 알 수 있다.

한편, 네이버, 엠파스, 야후 코리아의 웹 디렉토리들에 등록된 유일 웹 사이트들의 수는 각각 157,194개, 163,445개, 190,858개이다. <표 3>의 총계는 이러한 유일 웹 사이트들의 수보다 매우 크며, 이는 하나의 웹 사이트가 중복 분류 및 참조 링크로 인하여 다수의 통합 대분류에 소속되어 있기 때문이다. 즉, 네이버, 엠파스, 야후 코리아의 웹 디렉토리들에서 하나의 웹 사이트는 평균적으로 3.80개, 8.13개, 7.18개의 통합 대분류에 소속됨을 알 수 있다.

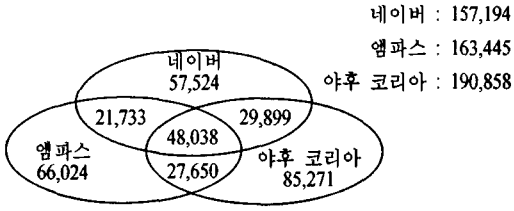
〈표 3〉 국내 웹 디렉토리들의 통합 대분류별 커버리지(참조 링크 허용)

	네이버		엠파스		야후 코리아	
뉴스, 언론	39,341	(6.6%)	45,390	(3.4%)	105,424	(7.7%)
의학	12,829	(2.2%)	19,979	(1.5%)	103,434	(7.6%)
컴퓨터, 인터넷	47,012	(7.9%)	42,928	(3.2%)	49,360	(3.6%)
지역정보	28,536	(4.8%)	116,924	(8.8%)	98,661	(7.2%)
교육, 참고자료	54,927	(9.2%)	214,756	(16.2%)	169,816	(12.4%)
학문	80,220	(13.4%)	137,290	(10.3%)	266,470	(19.4%)
경영, 경제	115,926	(19.4%)	252,941	(19.0%)	102,138	(7.4%)
사회, 정부	129,105	(21.6%)	263,801	(19.9%)	256,973	(18.7%)
예술, 엔터테인먼트	42,839	(7.2%)	111,082	(8.4%)	99,767	(7.3%)
여가, 스포츠	46,094	(7.7%)	123,897	(9.3%)	119,208	(8.7%)
총계	596,829	(100.0%)	1,328,988	(100.0%)	1,371,251	(100.0%)

5 커버리지 중복성 분석

네이버, 엠파스, 야후 코리아의 웹 디렉토리들에 등록된 유일 웹 사이트들의 수는 각각 157,194개, 163,445개, 190,858개이며, <그림 1>은 세 웹 디렉토리들의 커버리지 중복성을 보여준다. 예를 들어 네이버에 등록된 157,194개의 웹 사이트들 중에서 엠파스와 야후 코리에 등록되지 않은 웹 사이트들의 수는 57,524개이고, 네이버, 엠파스, 야후 코리아 모두에 등록된 웹 사이트들의 수는 48,038개이다.

<그림 1>로부터 네이버, 엠파스, 야후 코리아 사이의 중복도는 0.2818임을 알 수 있으며, 이는 평균적으로 세 검색 포탈들 중 하나의 웹 디렉토리에 등록된 사이트들의 28%를 다른 두 검색 포탈들의 웹 디렉토리에서 발견할 수 있음을 의미한다. 한편, 네이버와 엠파스 사이의 중복도는 0.4352, 네이버와 야후 코리아 사이의 중복도는 0.4478, 그리고 엠파스와 야후 코리아 사이의 중복도는 0.4273임을 알 수 있다. 이러한 결과는 국내 웹 디렉토리들의 커버리지 중복성이 비교적 낮음을 의미한다. 다음에서는 참조 링크를 배제한 경우와 참조 링크를 허용한 경우에 대하여, 네이버, 엠파스, 야후 코리아가 제공하는 웹 디렉토리들의 커버리지 중복성을 통합 대분류별로 기술한다.



〈그림 1〉 국내 웹 디렉토리들의 커버리지 중복성

5.1 참조 링크 배제

<표 4>는 참조 링크를 배제한 경우, 네이버, 엠파스, 야후 코리아가 제공하는 웹 디렉토리들의 커버리지 중복성을 통합 대분류별로 보여준다. 'N'은 네이버, 'E'는 엠파스, 'Y'는 야후 코리아의 웹 디렉토리에 등록된 사이트들의 집합이고, 'N∩E'는 네이버와 엠파스의 웹 디렉토리 모두에 등록된 사이트들의 집합이다. 예를 들어, <표 4>로부터 네이버, 엠파스, 야후 코리아의 뉴스 분야에 공통적으로 등록되어 있는 사이트들의 수가 431개임을 알 수 있다.

<표 5>는 <표 4>를 근거로 계산된 국내 웹 디렉토리들의 통합 대분류별 커버리지 중복성을 보여준다. <표 5>로부터 참조 링크를 배제한 경우, 네이버와 엠파스, 엠파스와 야후 코리아, 네이버와 야후 코리아 사이의 통합 대분류별 중복도들의 평균은 각각 0.2589, 0.2565, 0.2660임을 알 수 있다. 이러한 결과는 국내 웹 디렉토리들의 통합 대분류별 커버리지 중복성이 매우 낮음을 의미한다. 한편, 네이버와 엠파스, 엠파스와 야후 코리아, 네이버와 야후 코리아 사이의 중복도들인 0.4352, 0.4273, 0.4478과 비교하여, <표 5>의 통합 대분류별 중복도들이 매우 낮음을 알 수 있다. 이러한 사실은 국내 웹 검

〈표 4〉 국내 웹 디렉토리들의 통합 대분류별 커버리지 중복성(참조 링크 배제)

통합 대분류	N	E	Y	N∩E	E∩Y	N∩Y	N∩E∩Y
뉴스, 언론	1,970	3,440	2,070	712	1,100	633	431
의학	3,380	4,706	2,787	1,253	1,551	825	651
컴퓨터, 인터넷	3,559	4,621	4,518	1,067	1,187	903	525
지역정보	17,691	2,266	68,606	367	652	5,287	215
교육, 참고자료	16,064	16,582	12,665	5,989	3,437	3,483	1,868
학문	7,064	7,445	10,594	1,511	2,625	1,962	804
경영, 경제	81,591	55,789	64,894	35,048	25,243	34,958	19,058
사회, 정부	15,041	22,573	19,248	4,855	3,502	3,071	1,486
예술, 엔터테인먼트	19,648	37,619	30,731	6,008	9,407	7,464	3,831
여가, 스포츠	12,157	9,294	7,067	1,768	1,482	2,118	731

〈표 5〉 국내 웹 디렉토리들의 통합 대분류별 중복도(참조 링크 배제)

통합 대분류	N&E	E&Y	N&Y	N&E&Y
뉴스, 언론	0.2632	0.3993	0.3134	0.1729
의학	0.3099	0.4140	0.2676	0.1796
컴퓨터, 인터넷	0.2609	0.2598	0.2236	0.1240
지역정보	0.0368	0.0184	0.1225	0.0508
교육, 참고자료	0.3669	0.2350	0.2425	0.1237
학문	0.2083	0.2910	0.2222	0.0961
경영, 경제	0.5102	0.4183	0.4773	0.2827
사회, 정부	0.2581	0.1675	0.1791	0.0784
예술, 엔터테인먼트	0.2098	0.2753	0.2963	0.1306
여가, 스포츠	0.1648	0.1812	0.2204	0.0769
평균	0.2589	0.2660	0.2565	0.1316

색 포탈들이 동일한 웹 사이트를 서로 다른 대분류에 중복 등록하는 경우가 많음을 시사한다.

5.2 참조 링크 허용

<표 6>은 참조 링크를 허용한 경우, 네이버, 엠파스, 야후 코리아가 제공하는 웹 디렉토리들의 커버리지 중복성을 통합 대분류별로 보여준다. 그리고 <표 7>은 <표 6>을 기반으로 계산된 국내 웹 디렉토리들의 통합 대분류별 중복도를 보여준다. <표 5>와 <표 7>의 통합 대분류별 중복성의 비교를 통하여, 참조 링크의 허용이 커버리지 중복성에 크게 영향을 미치지 않음을 알 수 있다.

〈표 6〉 국내 웹 디렉토리들의 통합 대분류별 커버리지 중복성(참조 링크 허용)

통합 대분류	N	E	Y	N∩E	E∩Y	N∩Y	N∩E∩Y
뉴스, 언론	39,341	45,390	105,424	4,656	9,386	12,044	2,540
의학	12,829	19,979	103,434	5,576	7,940	6,781	4,052
컴퓨터, 인터넷	47,012	42,928	49,360	13,228	12,523	15,919	7,527
지역정보	28,536	116,924	98,661	12,025	34,131	10,443	5,869
교육, 참고자료	54,927	214,756	169,816	24,286	60,910	28,186	16,539
학문	80,220	137,290	266,470	31,171	53,816	37,272	19,918
경영, 경제	115,926	252,941	102,138	49,859	39,826	46,831	26,148
사회, 정부	129,105	263,801	256,973	51,287	89,180	52,968	29,133
예술,엔터테인먼트	42,839	111,082	99,767	14,102	26,487	17,774	9,013
여가, 스포츠	46,094	123,897	119,208	13,631	43,215	16,758	8,785

〈표 7〉 국내 웹 디렉토리들의 통합 대분류별 중복도(참조 링크 허용)

통합 대분류	N&E	E&Y	N&Y	N&E&Y
뉴스, 언론	0.1099	0.1245	0.1664	0.0401
의학	0.3399	0.1287	0.1166	0.0892
컴퓨터, 인터넷	0.2942	0.2714	0.3304	0.1621
지역정보	0.1653	0.3166	0.1642	0.0721
교육, 참고자료	0.1801	0.3168	0.2508	0.1129
학문	0.2866	0.2666	0.2150	0.1235
경영, 경제	0.2703	0.2243	0.4295	0.1666
사회, 정부	0.2611	0.3425	0.2744	0.1345
예술, 엔터테인먼트	0.1832	0.2512	0.2493	0.1066
여가, 스포츠	0.1604	0.3555	0.2028	0.0911
평균	0.2251	0.2598	0.2399	0.1099

6 결 론

본 연구에서는 국내 주요 웹 검색 포탈인 네이버, 야후 코리아, 엠파스가 제공하는 웹 디렉토리들의 커버리지와 커버리지 중복성 현황을 분석하였다. 또한 웹 검색 포탈들의 웹 디렉토리에 등록된 사이트들의 수집 방법을 개발하고, 최상위 주제 범주 매핑, 중복 분류, 참조 링크 등 디렉토리 커버리지 분석에 필요한 방법론을 제시하였다.

본 연구의 조사 결과, 첫째, 유일 웹 사이트들만을 대상으로 측정한 경우, 야후 코리아의 커버리지가 네이버나 엠파스보다 더 큰 것으로 나타났다. 웹 디렉토리들의 커버리지를 대분류별로 측정한 결과 "지역정보" 대분류가 야후 코리아의 커버리지에 결정적인 영향을 미치고 있음을 알 수 있었다. 그러나 "지역 정보" 대분류를 제외할 경우, 세 웹 디렉토리들은 유사한 수준의 커버리지를 제공하였다. 참조 링크를 배제한 경우, 웹 사이트들이 가장 많이 등록된 대분류들은 "경영, 경제," "지역정보", "엔터테인먼트, 예술" 순으로 나타났다. 한편 세 디렉토리들 중 야후 코리아가 중복분류를 가장 많이 사용하는 것으로 나타났다.

둘째, 참조 링크의 사용이 대분류별 커버리지에 매우 큰 영향을 끼치고 있음을 알 수 있다. 야후 코리아와 엠파스가 네이버에 비하여 많은 수의 참조 링크를 사용하고 있었다. 셋째, 국내 웹 디렉토리들 사이의 커버리지 중복성이 상당히 낮은 것으로 나타났다. 따라서 이용자들은 복수의 웹 디렉토리 서비스들을 사용하는 것이 보다 효율적일 것이다. 넷째, 이용자들은 참조링크, 중복 분류 등의 영향으로 동일한 사이트가 동일한 대분류 또는 상이한 대분류에서 여러 번 출현할 수 있음을 염두에 두고 웹 디렉토리를 접근하는 것이 효과적일 것이다. 다섯째, 상이한 대분류명의 사용, 상이한 참조 링크의 비율, 상이한 중복 분류의 비율 등이 이용자들에게 혼돈을 줄 수 있으므로, 웹 디렉토리들 사이의 일관성 있는 대분류명, 분류 시스템, 참조 링크의 사용 등이 요청된다.

한편, 본 연구의 수행 결과 향후 연구가 요구되는 사항들은 다음과 같다. 첫째, 디렉토리들이 공통적으로 제공하는 페이지들을 대상으로 분류의 일치성이나 일관성에 관한 연구를 수행할 수 있을 것이다. 둘째, 본 연구에서 제시한 웹 디렉토리들 커버리지 및 커버리지 중복성 분석 방법론에 대한 검증과 보완 작업이 요구된다.

참 고 문 헌

곽철완. 2001. 인터넷 쇼핑몰의 상품 분류체계에 대한 연구. 『정보관리학회지』, 18(4): 210-215.

김영보. 1997. 『인터넷 탐색엔진의 분류체계에 관한 연구 : 컴퓨터, 인터넷 분야를 중심으로』. 석사학위논문, 성균관대학교.

남영준. 1998. 웹 문서 분류체계의 분석 및 새로운 설계. 『한국문헌정보학회지』, 32(3): 207-230.

남영준, 최승순. 2002. 한국십진분류체계와 웹문서의 통합분류체계 개발. 『국회도서관보』, 39(1): 25-43.

신동민. 2001. 인터넷 검색엔진의 디렉토리 구성에 관한 연구. 『정보관리학회지』, 18(2): 143-163.

오동근, 황재영, 배영활. 2001. 군사학 분야 웹 문서 분류체계의 설계. 『한국도서관 정보학회지』, 32(2): 323-347.

이란주, 성기주, 양정하. 2001, 여성학분야 인

터넷 자원의 분류체계에 관한 연구. 『한국도서관 정보학회지』, 32(3): 397-417.

정연경. 2001. 인터넷 서점의 주제별 분류체계 설계에 관한 연구. 『한국문헌정보학회지』, 35(3): 17-34.

최재황. 1998. 인터넷 학술정보자원의 디렉토리 서비스 설계에 있어서 DDC분류 체계의 활용에 관한 연구. 『정보관리학회지』, 15(2): 47-67.

최희윤. 1998. 인터넷 정보서비스의 분류체계에 대한 비교연구 : 물리학을 중심으로. 『정보관리학회지』, 15(3): 45-57.

한상길. 2001. 산업분야 인터넷 자원의 분류체계에 관한 연구. 『정보관리학회지』, 18(3): 285-309.

Dong, X., and Su, L. T. 1997. "Search engines on the world wide web and information retrieval from the internet: A review and evaluation." *Online & CDROM Review,* 21(2): 67-82.

Heydon A., and Najork, M. 1999. "Mercator: a scalable, extensible web crawler." *World Wide Web,* 2(4): 219-229.

Lawrence, S., and Giles, C. L. 1999. "Accessibility of information on the web". *Nature,* 400: 107-109.

Lawrence, S., and Giles, C. L. 1998. "Searching the World Wide Web," *Science,* 280: 98-100.

Najork, M., and Heydon, A. 2001. "High-Performance Web Crawling," *SRC Research Report,* 173, Compaq Systems Research Center.

Rolling, L. 1981. "Indexing consistency, quality and efficiency", *Information Processing and Management,* 17: 69-76.