로그 분석을 통한 네이버 이용자의 디렉토리 접근 행태에 관한 연구*

Directory Access Behavior of the NAVER users via Log Analysis

배희진(Hee Jin Bae)** · 이준호(Joon Ho Lee)*** · 박소연(Soyeon Park)****

<목 차>

초 록

대다수의 웹 검색 포탈들은 인터넷 상의 정보들을 주제별로 분류한 디렉토리 서비스를 제공하며, 이러한 디렉토리들에 대한 접근 과정을 기록한 로그는 이용자의 실제 디렉토리 접근 행위를 사실적으로 반영한다. 본 연구는 디렉토리 서비스 이용자들의 다양한 디렉토리 접근 행태를 파악하기 위해 네이버에서 생성된 디렉토리 접근 로그들을 분석하였다. "세션"을 한 명의 이용자가 단일한 정보 요구를 지니고 디렉토리들을 접근한 일련의 과정으로 정의한 후, 본 연구에서는 전체 세션들을 정보 획득까지의 경로에 따라 여섯 가지 유형으로 분류하는 방법론을 개발하였다. 또한 세션 유형별 디렉토리 접근 빈도, 세션 내 항해 길이, 세션 내 주제에 대한 분석 결과를 제시하였다. 본 연구의 결과는 보다 효과적인 디렉토리 서비스 구축을 위한 근거로서 활용될 것으로 기대된다.

주제어 : 디렉토리 서비스, 로그 분석, 세션 유형, 디렉토리 접근 행태

Abstract

Most web portals provide a web directory service which selects and classifies web sites according to their subject matter. In order to investigate the directory access behavior of general Korean web users, this study analyzes directory access logs of NAVER, a major Korean web search engine. This study suggests a methodology to classify the total sessions into six different session types. This study also discusses directory access behaviors of the NAVER users by examining the distribution of sessions according to session types, the lengths of navigation within a session, and the most frequently visited categories. It is expected that this study could contribute to the development of more effective web directory services.

Key Words : directory service, log analysis, session type, directory access behavior

* 본 연구는 숭실대학교 교내연구비 지원으로 이루어졌음

** 숭실대학교 컴퓨터학부 대학원(jinybae@irlab.ssu.ac.kr)
*** 숭실대학교 컴퓨터학부 부교수(joonho@comp.ssu.ac.kr)
**** 덕성여자대학교 문헌정보학과 조교수(sypark@duksung.ac.kr)

· 접수일 : 2004. 1. 27 · 최초심사일 : 2004. 3. 3 · 최종심사일 : 2004. 3. 9

1. 서 론

대다수의 웹 검색 포탈들은 웹 상의 수많은 정보들 중에서 비교적 양질의 정보를 선택하여 주제별로 분류한 디렉토리 서비스를 제공한다. 디렉토리는 기본적으로 상하, 연관 관계와 같은 주제어의 상호 관계를 나타내고, 일반적으로 단계별 계층 구조로 구성된다[1]. 이용자들은 최상위 계층의 디렉토리로부터 또는 질의를 입력하여 검색된 디렉토리부터 항해를 시작하여 원하는 정보를 발견할 수 있다.

디렉토리 서비스는 이용자가 특정한 주제 분야에 어떠한 정보들이 존재하는지 알고 싶을 경우, 또는 원하는 정보의 검색에 적합한 키워드들이 모호할 경우에 유용하다. 그러나 이용자는 원하는 정보를 발견할 때까지 디렉토리 구조 내에서 여러 단계의 디렉토리들을 항해하며, 특히 원하는 정보가 속한 디렉토리가 불명확한 경우 디렉토리 항해에 많은 시간과 노력을 소비할 수 있다. 따라서 디렉토리 서비스의 이용자가 원하는 정보를 효율적으로 발견할 수 있도록 합리적이고 체계적으로 디렉토리를 구축하는 것이 중요하다.

디렉토리 구축과 관련된 국내 선행 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; 남영준, 1998; 신동민, 2001), 그리고 디렉토리 설계에 전통적 분류 이론과 체계를 적용한 연구(최재황, 1998)로 구분될 수 있다. 그리고 국외 선행 연구로서 Srikant & Yang[2]은 디렉토리 서비스 이용자들이 남긴 디렉토리 접근 로그들의 분석을 통하여 디렉토리 내의 웹사이트의 위치가 이용자가 기대하였던 위치와 불일치하는 경우 이를 찾아내는 알고리즘을 개발하였다.

한편, 실제 디렉토리 서비스 이용자들의 디렉토리 접근 행태에 대한 분석은 보다 합리적이고 체계적인 디렉토리 구축을 위한 근거로서 활용될 수 있다. 그러나 국내외 선행 연구들 중에서 실제 디렉토리 서비스 이용자들을 대상으로 이들의 디렉토리 접근 행태를 분석한 연구는 찾아보기 어려운 실정이다.

이에 본 연구는 이용자와 디렉토리 서비스 시스템 사이의 모든 상호 작용을 기록, 저장한 디렉토리 접근 로그를 이용하여 실제 디렉토리 서비스 이용자들의 디렉토리 접근 행태를 분석하고자 한다. 이를 위하여 디렉토리 접근 로그 분석을 위한 세션 정의 방법과 로그 정제 방법 등을 제시하였다. 또한 본 연구에서는 네이버 디렉토리 서비스로부터 일주일 동안 생성

1) 신동민, "인터넷 검색엔진의 디렉토리 구성에 관한 연구," 정보관리학회지, 제18권, 제2호 (2001, 6), p.147.
2) R. Srikant, and Y. Yang, "Mining Web Logs to Improve Website Organization," *Proceedings of the Tenth International World Wide Web Conference*, (2001), pp.430-437.

된 디렉토리 접근 로그들을 분석하여, 정보 획득까지의 경로에 따라 모든 세션들을 여섯 개의 유형으로 분류하는 방법론을 새롭게 개발하였다. 마지막으로 본 연구는 세션 수, 세션 내 항해 길이, 그리고 세션 주제에 대한 분석 결과를 제시하였다.

2. 연구 방법

본 연구에서는 웹 이용자들의 디렉토리 접근 행태를 파악하기 위해 디렉토리 접근 로그를 분석하였다. 분석 대상이 된 로그는 2003년 6월 25일부터 7월 2일까지 일주일동안 네이버의 디렉토리 서비스에서 생성되었다. 네이버는 대중성이나 인지도면에서 국내 주요 검색 포탈로 인정받고 있다. 즉, 네이버는 2002년 12월 한국인터넷기업협회와 한국기자협회가 선정한 올해의 인터넷기업 대상을 수상하였으며, 한국생산성 본부에서 실시한 2003년 1/4분기 국가고객만족도(NCSI) 조사에서 검색 포탈 서비스 부분 1위를 차지하였다. 네이버는 디렉토리 검색, 웹문서 검색, 백과사전 검색, 지식iN검색, 뉴스 검색, 이미지 검색 등을 개별적으로 지원하고 있으며, 또한 이들 검색 결과들을 통합하여 보여 주는 통합검색을 제공하고 있다. 본 연구에서는 네이버의 디렉토리 서비스에서 생성된 로그를 분석함으로써 국내 웹 이용자들의 전반적인 디렉토리 접근 행태를 파악하고자 하였다. 다음에서는 디렉토리 접근 로그 분석에 필수적인 세션 정의 방법, 로그 정제 방법에 대하여 기술한다.

2.1 세션 정의

일반적으로 "세션"은 한 명의 이용자가 단일한 정보 요구를 지니고 처음 검색을 시작하여 검색을 종료하기까지의 일련의 과정으로 정의된다. 로그 분석을 이용한 선행 연구들은 세션에 관한 일반적인 정의에는 동의하나, 검색 질의들을 세션으로 구분함에 있어서 상이한 방법을 적용하고 있으며, 익사이트의 질의 로그를 분석한 Spink et al.[3]의 연구, 알타비스타의 로그를 분석한 Silverstein et al.[4]의 연구와 로이터, 알타비스타, 익사이트 로그를 분석한 He

3) A. Spink et al., "Searching the Web : The Public and Their Queries," *Journal of the American Society for Information Science and Technology*, Vol.52, No.3(2001), pp.226-234.

4) C. Silverstein et al., "Analysis of a Very Large Web Search Engine Query Log." *SIGIR Forum*, Vol.33, No.1(1999), pp.6-12.

& Goker[5]의 연구가 그 대표적인 예이다.

Spink et al.은 세션의 정의를 위하여 익사이트 서버가 할당한 이용자 식별자를 이용하였다. 곧 임의의 컴퓨터가 브라우져를 통하여 익사이트에 검색을 요청할 때, 익사이트는 쿠키를 생성하여 검색을 요청한 컴퓨터에게 쿠키의 소멸 시간을 지정하지 않은 상태로 전달한다. 이후부터 이 쿠키는 이용자 식별자로 이용되며, 검색을 요청한 컴퓨터의 브라우져들이 모두 종료될 때 소멸된다. 이러한 세션 설정 방법은 공공장소에 위치한 하나의 컴퓨터를 다수의 이용자가 이용할 경우 또는 다수의 컴퓨터들이 하나의 프록시로 설정되어 있는 경우, 세션 수가 과소평가되는 문제점을 지니고 있다.

He & Goker는 로이터와 익사이트 로그를 대상으로 세션의 시간 간격을 1분부터 50분까지 달리한 분석을 수행한 후, 10분에서 15분을 최적의 세션 간격으로 제시하였다. 그러나 이들의 세션 정의 방법은 비교적 소규모의 웹 트랜잭션 로그에서만 사용되었고, 이들이 제시한 결론은 엄격한 통계 분석에 기초한 것이 아니라는 문제점을 지니고 있다.

Silverstein et al.은 이용자가 특정한 검색 목적을 다른 검색 목적으로 전환하게 되는 경우 시간적 공백이 발생한다는 점에 착안하여, 검색을 요청한 컴퓨터에 쿠키를 전달할 때, 이 쿠키가 5분 후에 소멸되도록 지정하였다. 따라서 이들은 5분 내에 새로운 질의가 입력될 경우 기존 세션을 연장하고, 5분 동안 질의를 입력하지 않으면 새로운 세션을 정의하는 방법을 제시하였다. 이준호, 박소연, 권혁성[6]의 2003년 연구에서는 각각의 방법의 장단점을 비교, 평가한 후 Silverstein et al.이 제안한 방법을 사용하여 검색 질의에 대한 로그들을 세션으로 분리하였다.

한편, 본 연구의 분석 대상은 질의 로그가 아닌 디렉토리 접근 로그이기 때문에, 세션을 한 명의 이용자가 단일한 정보 요구를 지니고 처음 디렉토리를 접근하여 디렉토리 접근을 종료하기까지의 일련의 과정으로 재정의한다. 세션을 이와 같이 정의할 경우, 디렉토리 접근 로그들을 세션으로 분리하기 위하여 Silverstein et al.의 세션 정의 방법을 적용할 수 있다. 즉, 본 연구에서는 이용자가 5분 동안 디렉토리 접근을 수행하지 않으면 새로운 세션을 생성하고, 5분 이내에 디렉토리 접근을 수행하면 기존 세션을 연장하였다.

---

5) D. He, and A. Goker, "Detecting Session Boundaries from Web User Logs," *Proceedings of the 22nd Annual Colloquium on Information Retrieval Research*, (2000), pp.57-66.

6) 이준호, 박소연, 권혁성, "질의 로그 분석을 통한 네이버 이용자의 검색 행태 연구," 정보관리학회지, 제20권, 제2호(2003, 6), pp.27-41.

2.2 로그 정제

디렉토리 접근 로그들로부터 정상적인 세션에 포함시키기 어려운 다수의 로그들을 발견하였으며, 이러한 디렉토리 접근 로그들이 생성되는 이유는 다음과 같이 크게 세 가지로 구분될 수 있다. 첫째, 이러한 로그들은 웹 문서들을 수집하는 웹 로봇 프로그램이 네이버 디렉토리를 접근하거나, 또는 네이버 이외의 타 웹 검색 엔진의 검색 결과를 통하여 네이버 디렉토리를 접근한 경우에 생성될 수 있으며, 본 연구에서는 이러한 디렉토리 접근 로그들을 분석 대상에서 제외하였다.

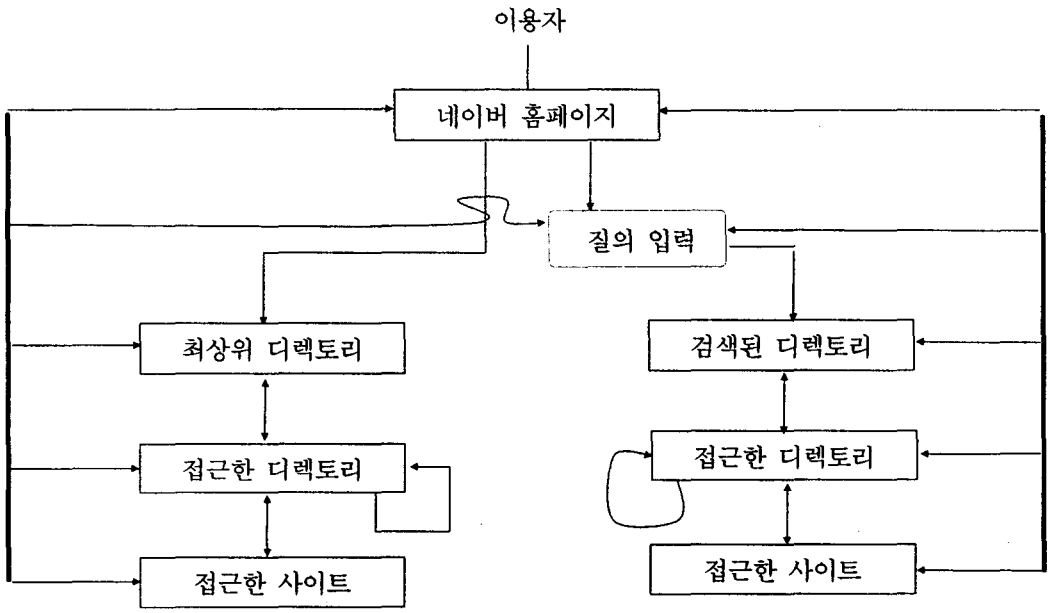
둘째, 2.1절에서 설명된 것처럼 본 연구에서 사용한 세션 정의 방법이 현재까지 알려진 최선의 방법일 지라도, 이 방법 역시 문제점이 있음을 발견하였다. 본 연구에서는 세션이 생성된 후, 이용자가 5분 안에 디렉토리 접근을 수행하는 동안은 세션이 연장된다. 그러나 5분의 휴식이나 공백 이후, 이전 디렉토리 결과를 기반으로 디렉토리 접근을 반복 수행하는 경우에는 새로운 세션이 생성되며, 이러한 세션을 비정상적으로 간주하여 로그 파일로부터 제거하였다.

셋째, 이용자는 네트워크의 속도 저하, 디렉토리 시스템의 과부하 등의 이유로 결과 화면 생성이 다소 지연되면 새로 고침 버튼을 클릭하거나 재접근을 수행하는 경향이 있다. 이때 이용자가 결과 화면을 받지 않은 페이지에 대해서도 로그가 생성되기 때문에, 동일한 디렉토리 접근 로그가 연속해서 파일에 기록된다. 따라서 본 연구에서는 이러한 경우에 두 번째 이후의 중복되는 디렉토리 접근 로그를 제거하였다.

3. 세션 유형 분석 결과

네이버 디렉토리 서비스는 웹 사이트들을 14개의 대분류로 구분하고, 이 14개의 대분류 각각은 다수의 하위 주제 범주들로 세분화된다. 네이버 디렉토리의 초기 화면은 14개 대분류뿐만 아니라, 사용자들이 빈번하게 접근하는 70개 하위 주제 범주들로 구성되어 있으며, 본 논문에서는 이들 84개의 주제 범주들을 최상위 디렉토리로 명명한다. <그림 1>은 이용자가 네이버 홈페이지로부터 디렉토리들에 접근하여 원하는 사이트를 발견하기까지 선택할 수 있는 모든 경로들을 보여준다. 실선 직사각형은 이용자가 접근한 웹 페이지이고, 타원은 검색 창을 통한 이용자의 질의 입력을 표시하며, 화살표는 웹 페이지들의 접근 순서, 즉 화살표 시

작 부분의 웹 페이지로부터 화살표 끝 부분의 웹 페이지가 접근되었음을 의미한다. 예를 들어, 이용자는 네이버 홈페이지의 최상위 디렉토리들 중에서 하나를 클릭하고 순환적으로 디렉토리들을 항해한 후, 최종적으로 원하는 사이트에 접근할 수 있다. 이외에도 정보 획득까지 이용자가 선택할 수 있는 다양한 경로들이 존재하며, 본 장에서는 정보 획득까지의 경로들에 따라 세션들을 여섯 가지 유형들로 구분하고, 이러한 유형들에 대하여 기술한다.



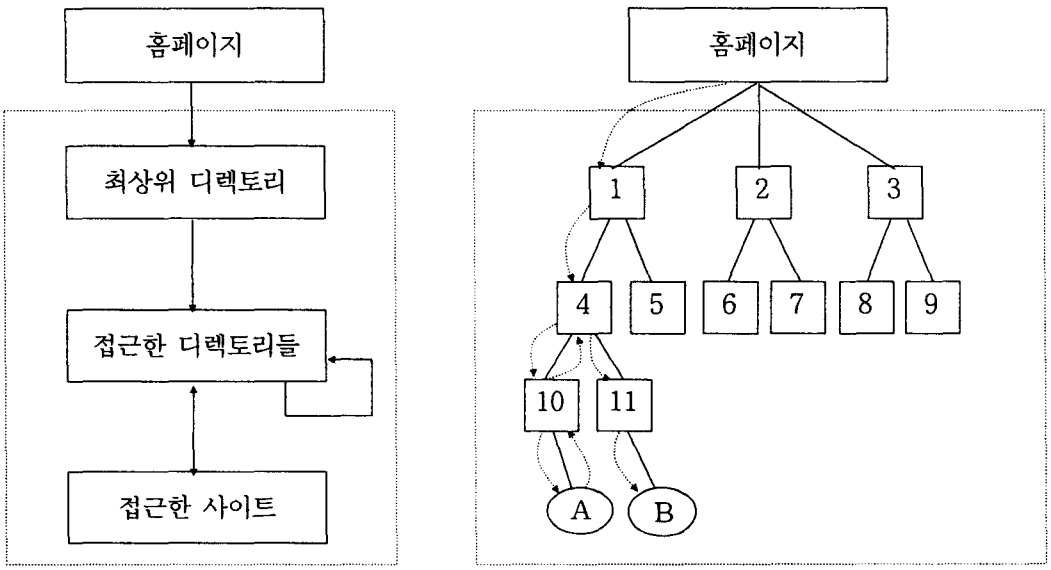
<그림 1> 정보 획득까지의 접근 경로들

3.1 기본 항해 세션

기본 항해 세션은 이용자가 네이버 홈페이지에 열거된 최상위 디렉토리들 중의 하나를 클릭할 경우에 시작되며, 세션이 시작된 후에는 네이버 홈페이지로 재접근하지 않는 세션이다. <그림 2>의 (a)는 모든 접근 경로들 중에서 기본 항해 세션에 포함되는 경로들을 보여준다. 이용자는 네이버 홈페이지에 열거된 최상위 디렉토리로부터 순환적으로 디렉토리들을 항해함으로써 원하는 정보를 발견할 수 있으며, 이때 생성되는 일련의 디렉토리 및 사이트 접근 로그들이 기본 항해 세션에 포함된다.

<그림 2>의 (b)는 기본 항해 세션의 예를 보여준다. (b)에서 정사각형과 타원은 각각 이

용자가 접근한 디렉토리와 사이트를 표시하고, 정사각형 내의 번호와 타원 내의 알파벳은 각각 디렉토리 이름과 사이트 이름이며, 점선 화살표는 이용자의 디렉토리 접근 순서를 나타낸다. 이 예에서 세션은 이용자가 네이버 홈페이지에 열거된 최상위 디렉토리들 중에서 디렉토리 1을 클릭할 때 시작된다. 이후에 이용자는 디렉토리들 4, 10과 사이트 A, 그리고 디렉토리들 10, 4, 11을 거쳐 최종적으로 사이트 B에 접근한다.



(a) 접근 경로 (b) 예제

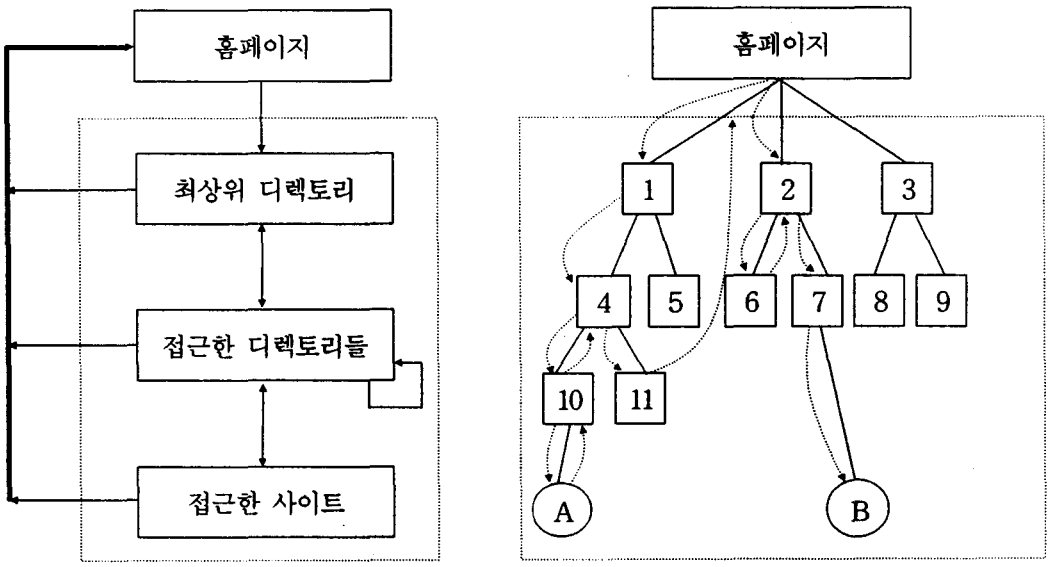
<그림 2> 기본 항해 세션

3.2 반복 항해 세션

반복 항해 세션은 기본 항해 세션에 포함된 경로들과 더불어 네이버 홈페이지에 재접근하는 경로를 포함한다. <그림 3>의 (a)는 반복 항해 세션에 포함되는 경로들을 보여준다. 이용자는 네이버 홈페이지에 열거된 최상위 디렉토리로부터 순환적으로 디렉토리들을 항해하거나 홈페이지에 재접근함으로써 원하는 정보를 발견할 수 있다. 이때 생성되는 일련의 디렉토리 및 사이트 접근 로그들이 반복 항해 세션에 포함된다.

<그림 3>의 (b)는 반복 항해 세션의 예를 보여준다. 반복 항해 세션은 이용자가 네이버 홈페이지에 열거된 최상위 디렉토리들 중에서 디렉토리 1을 클릭할 때 시작된다. 이후에 이

용자는 디렉토리들 4, 10과 사이트 A, 디렉토리들 10, 4, 11을 순차적으로 접근하고, 네이버 홈페이지에 재접근한다. 그리고 디렉토리들 2, 6, 2, 7을 거쳐 최종적으로 사이트 B에 접근한다.



(a) 접근 경로 (b) 예제

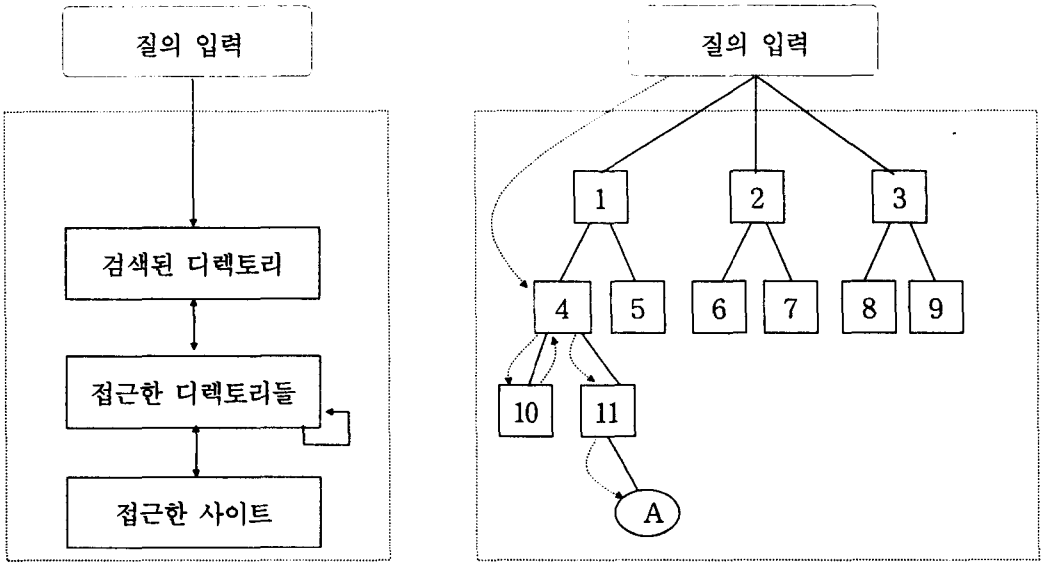
<그림 3> 반복 항해 세션

3.3 기본 검색 세션

기본 검색 세션은 이용자가 검색 창에 질의를 입력하여 검색을 수행한 후, 검색 결과로서 출력된 디렉토리들 중의 하나를 클릭할 경우에 시작되며, 세션이 시작된 이후에는 검색 창에 질의를 재입력하지 않는 세션이다. <그림 4>의 (a)는 기본 검색 세션에 포함되는 경로들을 나타낸다. 이용자는 검색된 디렉토리로부터 순환적으로 디렉토리들을 항해함으로써 원하는 정보를 발견할 수 있으며, 이때 생성되는 일련의 디렉토리 및 사이트 접근 로그들이 기본 검색 세션에 포함된다.

<그림 4>의 (b)는 기본 검색 세션의 예를 보여준다. 세션은 이용자가 검색 창에 질의를 입력하여 검색된 디렉토리들 중에서 디렉토리 4를 클릭할 때 시작된다. 기본 검색 세션에서 이용자가 최초로 클릭하는 디렉토리는 검색 결과로부터 출력된 디렉토리들 중 하나이므로, 네이버 디렉토리의 최상위 디렉토리가 아닐 수도 있다. 이후에 이용자는 디렉토리들 10, 4,

11을 거쳐 최종적으로 사이트 A에 접근한다.



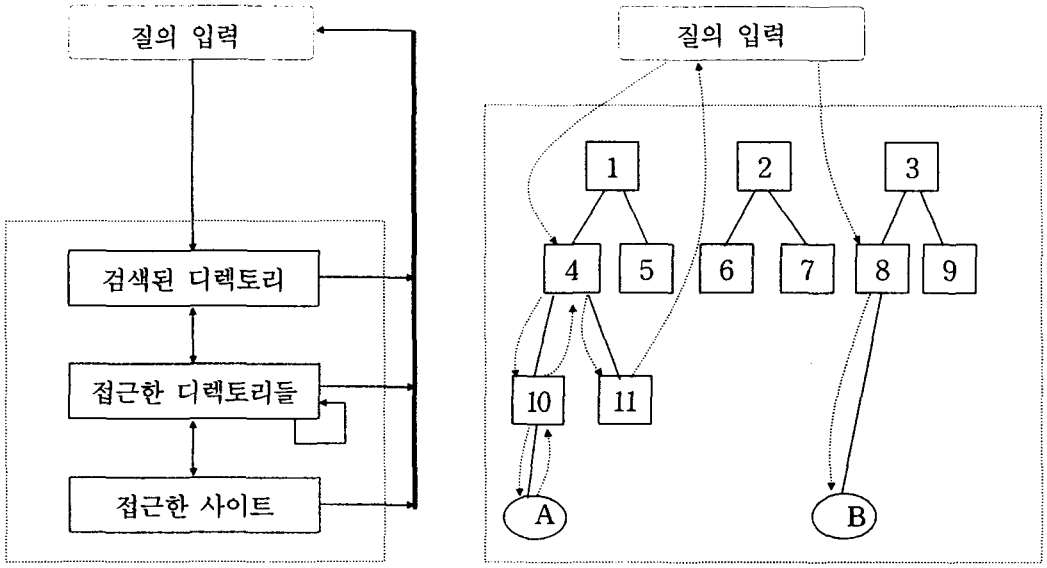
(a) 접근 경로 (b) 예제

<그림 4> 기본 검색 세션

3.4 반복 검색 세션

반복 검색 세션은 기본 검색 세션에 포함된 경로들과 더불어 검색 창에 질의를 재입력하는 경로를 포함한다. <그림 5>의 (a)는 반복 항해 세션에 포함되는 경로들을 나타낸다. 이용자는 검색된 디렉토리로부터 순환적으로 디렉토리들을 항해하거나 검색 창에 질의를 재입력함으로써 원하는 정보를 발견할 수 있다. 이때 생성되는 일련의 디렉토리 및 사이트 접근 로그들이 반복 검색 세션에 포함된다.

<그림 5>의 (b)는 반복 검색 세션의 예를 보여준다. 세션은 이용자가 검색 창에 질의를 입력하여 검색 결과로서 출력된 디렉토리들 중에서 디렉토리 4를 클릭할 때 시작된다. 이후에 이용자는 디렉토리 10과 사이트 A 그리고 디렉토리들 10, 4, 11에 접근한다. 또한 이용자는 검색 창에 질의를 재입력하여 검색된 디렉토리 8을 거쳐 최종적으로 사이트 B에 접근한다.



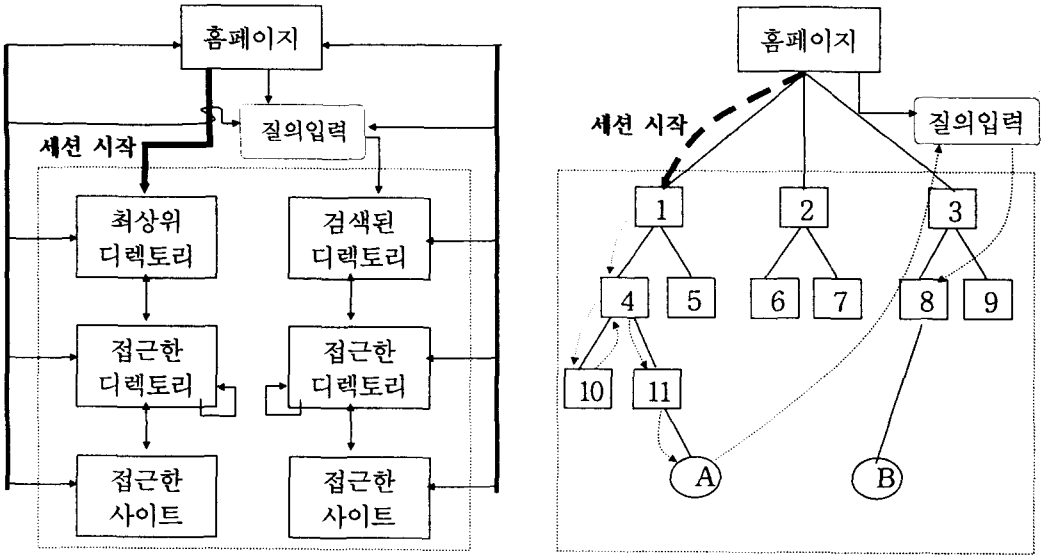
(a) 접근 경로 (b) 예제

<그림 5> 반복 검색 세션

3.5 교차 항해 세션

<그림 6>의 (a)는 교차 항해 세션에 해당하는 경로들을 보여준다. 교차 항해 세션은 네이버 홈페이지에 열거된 최상위 디렉토리들 중의 하나를 클릭함으로써 시작되며, 반복 항해 세션 및 반복 검색 세션에 포함된 모든 경로들을 포함한다. <그림 6>의 (a)를 <그림 1>과 비교해 보면 네이버 홈페이지로부터 정보 획득까지의 모든 접근 경로들이 교차 항해 세션에 포함될 수 있음을 알 수 있다.

<그림 6>의 (b)는 교차 항해 세션의 예를 보여준다. 세션은 이용자가 네이버 홈페이지에 열거된 최상위 디렉토리들 중에서 디렉토리 1을 클릭할 때 시작된다. 이후에 이용자는 디렉토리들 4, 10, 4, 11과 사이트 A에 접근한다. 또한 이용자는 검색 창에 질의를 입력하여 검색된 디렉토리 8을 거쳐 최종적으로 사이트 B에 접근한다.



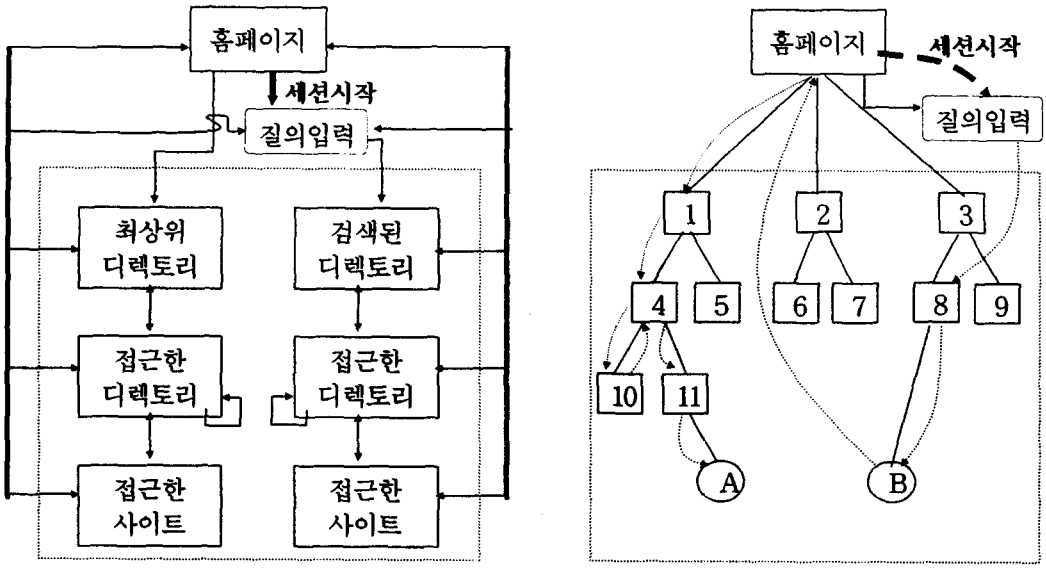
(a) 접근 경로 (b) 예제

<그림 6> 교차 항해 세션

3.6 교차 검색 세션

<그림 7>의 (a)는 교차 검색 세션에 포함되는 경로들을 보여준다. 교차 검색 세션은 이용자가 검색 창에 질의를 입력하여 검색을 수행한 후, 검색 결과로서 출력된 디렉토리들 중의 하나를 클릭할 경우에 시작되며, 반복 항해 세션 및 검색 항해 세션에 포함된 모든 경로들을 포함한다. 한편, <그림 7>의 (a)와 <그림 6>의 (a)의 비교로부터 교차 검색 세션은 교차 항해 세션이 포함할 수 있는 모든 경로들을 포함할 수 있으며, 교차 검색 세션과 교차 항해 세션은 세션을 시작하는 방법에서 차이가 있음을 알 수 있다.

<그림 7>의 (b)는 교차 검색 세션의 예를 보여준다. 세션은 이용자가 검색 창에 질의를 입력하여 검색된 디렉토리들 중에서 디렉토리 8을 클릭할 때 시작된다. 이후에 이용자는 사이트 B에 접근한다. 또한 이용자는 홈페이지에 열거된 최상위 디렉토리들 중에서 디렉토리 1을 클릭한 후, 디렉토리들 4, 10, 4, 11을 거쳐 최종적으로 사이트 A에 접근함으로써 세션을 종료한다.



(a) 접근 경로 (b) 예제

<그림 7> 교차 검색 세션

4. 세션 분석 결과

4.1 세션 수 및 항해 길이 분석

본 연구에서는 이용자의 디렉토리 접근 행태를 파악하기 위해 네이버의 디렉토리 서비스로부터 2003년 6월 26일부터 7월 2일까지 일주일 동안 생성된 로그들을 분석하였다. <표 1>은 전체 세션을 세션 유형별로 구분한 후, 각 세션 유형별 세션 수와 항해 길이에 대한 기술통계를 보여준다. 또한 <표 2>는 세션 유형별 평균 항해 길이를 나타낸다. 여기에서 특정한 세션의 항해 길이는 그 세션에 포함된 로그들의 수, 즉 이용자가 하나의 정보 요구를 지니고 디렉토리 또는 사이트를 클릭한 횟수로서 정의된다.

<표 1>로부터 다음과 같은 이용자의 디렉토리 접근 행태를 알 수 있다. 첫째, 홈페이지에 열거된 최상위 디렉토리들 중의 하나를 클릭함으로써 시작되는 항해 세션들의 수는 45.4%이고, 검색 결과로서 출력된 디렉토리들 중의 하나를 클릭함으로써 시작되는 검색 세션들의 수는 54.6%를 차지한다. 따라서 최상위 디렉토리부터 세션을 시작하는 이용자보다 검색 결과로

서 출력된 디렉토리부터 세션을 시작하는 이용자가 좀 더 많음을 알 수 있다. 둘째, 전체 세션들 중에서 기본 항해 세션은 28.3%를 차지하며, 기본 검색 세션은 44.5%를 차지한다. 따라서 약 73%의 이용자들이 복잡한 디렉토리 접근 방법보다 단순한 디렉토리 접근 방법을 선호함을 알 수 있다. 또한, <표 2>에 의하면 항해 세션의 평균 항해 길이가 검색 세션의 평균 항해 길이 보다 2배 정도 긴 것을 알 수 있다. 이로부터 이용자는 항해 세션보다 검색 세션을 통해 원하는 정보에 보다 신속하게 접근하고 있음을 알 수 있다.

<표 1> 세션 유형별 세션 수 및 항해 길이 총계

세션 유형	세션 수(%)	항해 길이 총계(%)
기본 항해 세션	944,955 (28.3%)	3,589,938 (22.9%)
반복 항해 세션	544,599 (16.3%)	5,771,447 (36.8%)
교차 항해 세션	24,633 (0.8%)	276,441 (1.7%)
기본 검색 세션	1,482,855 (44.5%)	3,674,692 (23.4%)
반복 검색 세션	326,491 (9.8%)	2,253,444 (14.4%)
교차 검색 세션	10,942 (0.3%)	124,399 (0.8%)
총 계	3,334,475 (100%)	15,690,361 (100%)

<표 2> 세션 유형별 평균 항해 길이

세션 유형	평균 항해 길이	세션 유형	평균 항해 길이
기본 항해 세션	3.8	기본 검색 세션	2.5
반복 항해 세션	10.6	반복 검색 세션	6.9
교차 항해 세션	11.2	교차 검색 세션	11.4
항해 세션	6.4	검색 세션	3.3

4.2 세션 주제 분석

네이버 디렉토리 서비스에서 웹 사이트들은 14개의 대분류들로 구분된다. 본 연구에서는 네이버 디렉토리 서비스 이용자들의 주제별 관심도를 파악하기 위하여, 항해 및 검색 세션들의 주제를 조사하였다. 여기에서 주어진 세션의 주제는 14개 대분류들 중의 하나로서, 세션

내에서 이용자가 최초로 접근한 디렉토리가 소속된 대분류로 정의된다.

<표 3>은 항해 및 검색 세션 내에서 주제들의 분포를 보여주며, 이로부터 다음과 같은 이용자의 디렉토리 접근 행태를 알 수 있다. 첫째, 항해 세션을 통하여 이용자가 가장 많이 접근하는 주제는 "뉴스, 미디어"이고, 검색 세션을 통하여 이용자가 가장 많이 접근하는 주제는 "비즈니스, 경제"임을 알 수 있다. 둘째, 항해 세션들의 경우 전체의 과반수 이상이 "뉴스, 미디어" 주제에 집중되어 있는 반면, 검색 세션들은 다수의 주제들에 비교적 균등하게 분포되어 있음을 알 수 있다. 셋째, "비즈니스, 경제," "엔터테인먼트," "쇼핑," "게임" 등과 같은 주제들의 경우, 이용자들이 홈페이지로부터 항해를 시작하기보다 검색을 통하여 더 빈번하게 접근하는 것을 알 수 있다.

<표 3> 항해 및 검색 세션 내에서 주제들의 분포

대분류	항해 세션 수	검색 세션 수
뉴스, 미디어	**813,328 (54%)**	59,451 (3%)
엔터테인먼트	132,825 (9%)	316,934 (17%)
비즈니스, 경제	66,329 (6%)	**389,346 (21%)**
컴퓨터, 인터넷	84,728 (6%)	186,245 (10%)
쇼핑	64,897 (4%)	171,185 (9%)
게임	142,215 (9%)	291,521 (16%)
가정, 여성	39,498 (3%)	30,950 (2%)
레크리에이션	56,020 (4%)	60,912 (3%)
사회, 문화	13,279 (1%)	94,963 (5%)
스포츠	70,929 (5%)	28,462 (2%)
학문, 과학	2,954 (0%)	34,780 (2%)
건강, 의학	16,848 (1%)	28,851 (2%)
교육, 참고	7,072 (0%)	112,347 (6%)
지역정보	3,265 (0%)	14,341 (1%)

3장에서도 언급되었듯이 네이버 디렉토리의 초기 화면은 14개의 대분류뿐만 아니라, 사용자들이 빈번하게 접근하는 70개 하위 주제 범주들로 구성되어 있다. 예를 들면, 네이버 홈페이지에는 대분류 항목인 "뉴스, 미디어"와 더불어 이의 하위 주제 범주들인 "최신 뉴스", "신문", "방송", "스포츠신문", "날씨"가 함께 열거되어 있다. 본 연구에서는 네이버 디렉토리 서

비스 이용자들의 주제별 관심도를 보다 세밀히 파악하기 위하여, 항해 세션 내에서 84개 최상위 주제 범주들의 분포를 조사하였다.

<표 4>는 84개 최상위 디렉토리들 중에서 항해 세션의 분포율이 높은 상위 10개의 디렉토리들을 보여주며, 이 표로부터 다음과 같은 이용자의 디렉토리 접근 행태를 알 수 있다. 첫째, 이용자가 가장 빈번히 접근하는 최상위 디렉토리는 "스포츠 신문"으로서, 항해 세션의 분포율이 매우 높음을 알 수 있다. 둘째, "게임", "스포츠"를 제외한 나머지 8개 디렉토리는 대분류가 아닌 하위 단계 주제 범주들이다. 이러한 결과는 14개 대분류들과 더불어 하위 단계 주제 범주들을 디렉토리 초기 페이지에 함께 제공하는 방식이 효율적임을 시사한다.

<표 4> 항해 세션의 분포율이 높은 상위 10개의 최상위 디렉토리들

최상위 디렉토리	세션 수
스포츠신문	530,172 (35%)
신문	142,654 (9%)
한게임	86,383 (6%)
최신뉴스	53,117 (4%)
게임	44,771 (3%)
스포츠	34,547 (2%)
채팅	33,596 (2%)
연예인	31,051 (2%)
방송	28,934 (2%)
음악	27,270 (2%)

5. 결 론

이용자와 검색 서비스 시스템간의 모든 과정을 기록하고 저장한 로그는 이용자의 실제 행태를 사실적으로 반영하므로, 웹 이용자들의 이용 행태 연구를 위한 합리적이고 객관적인 방법으로 활용되고 있다. 따라서, 본 연구에서는 2003년 6월 25일부터 7월 2일까지 일주일동안 네이버의 디렉토리 서비스에서 생성된 로그를 분석하여, 다양한 이용자들의 디렉토리 접근 행태를 조사하였다.

디렉토리 접근 로그들의 분석을 통하여 나타난 다음과 같은 조사 결과는 보다 효과적인

디렉토리 구축을 위한 근거로서 활용될 수 있다. 첫째, 세션 내에서 이용자가 항해하는 길이가 비교적 짧으므로, 디렉토리 구조에 지나치게 많은 계층을 제공하는 것은 바람직하지 않다. 둘째, 최상위 디렉토리부터 항해를 시작하는 이용자보다 검색 결과로서 출력된 디렉토리부터 세션을 시작하는 이용자가 더 많은 것으로 나타났다. 셋째, 약 73%의 이용자들이 복잡한 세션 유형보다는 단순한 세션 유형을 선호하는 것으로 나타났다. 따라서 이용자가 원하는 정보에 최대한 신속하게 도달할 수 있도록 지원하는 것이 필요하다. 넷째, 이용자는 항해 세션보다 검색 세션을 통해 원하는 정보에 보다 신속하고 간단히 접근한다. 다섯째, "비즈니스, 경제"나 "쇼핑"과 같은 주제의 경우 이용자가 항해보다는 검색을 선호하는 것으로 나타났다. 이는 이용자가 이러한 분야의 디렉토리에 접근하여 항해하는 것이 용이하지 않음을 시사한다. 따라서 이들 주제 범주의 디렉토리 구조에 대한 재검토와 이용자들을 대상으로 이러한 접근 행태에 대한 심층조사를 수행하는 것이 필요하다. 마지막으로, 이용자의 관심도가 가장 높은 최상위 주제 범주는 각각 "뉴스, 미디어"와 "스포츠 신문"으로 나타났다. 따라서 이러한 주제 범주에 대해 디렉토리 서비스를 강화하는 것이 필요하다.

한편, 본 연구의 분석 자료 외에 향후 연구가 요구되는 사항은 다음과 같다. 첫째, 디렉토리 접근 로그를 통해 이용자의 디렉토리 항해 경로를 추적할 수 있다. 이러한 항해 경로의 분석을 통해 디렉토리 구조 및 분류의 문제점을 파악하고, 디렉토리 구조의 합리성 및 분류의 정확성을 개선할 수 있다. 둘째, 본 연구에서는 디렉토리 이용자들의 디렉토리 접근 행태를 계량적인 방법으로 분석하였다. 따라서 이용자들의 디렉토리 서비스에 대한 만족도나 디렉토리 서비스에 대한 개선 사항, 특정한 방식으로 행동하는 이유 등을 분석하기 위해서는 실험, 면접, 관찰 등의 질적인 방법을 통한 보완작업이 요구된다. 셋째, 본 연구에서는 일주일 간의 디렉토리 접근 로그를 통하여 이용자들의 디렉토리 접근 행태를 조사하였다. 그러나 이용자들이 관심을 가지고 접근하는 주제 범주들의 경우 시간이 지남에 따라 변화할 수 있으므로, 향후 연구에서는 보다 장기간의 로그 분석을 통하여 이용자들의 관심사의 변화를 추적하고 이를 디렉토리 구축에 반영할 수 있다.

참 고 문 헌

곽철완. "인터넷 쇼핑몰의 상품 분류체계에 대한 연구." 정보관리학회지, 제18권, 제4호(2001, 12), pp.210-215.

김영보. 인터넷 탐색엔진의 분류체계에 관한 연구 : 컴퓨터, 인터넷 분야를 중심으로. 석사학위 논문, 성균관대학교, 1997.

남영준. "웹 문서 분류체계의 분석 및 새로운 설계." 한국문헌정보학회지, 제32권, 제3호(1998, 9), pp.207-230.

신동민. "인터넷 검색엔진의 디렉토리 구성에 관한 연구." 정보관리학회지, 제18권, 제2호(2001, 6), pp.143-163.

오동근, 황재영, 배영활. "군사학 분야 웹 문서 분류체계의 설계." 한국도서관·정보학회지, 제32권, 제2호(2001, 6), pp.323-347.

이란주, 성기주, 양정하. "여성학분야 인터넷 자원의 분류체계에 관한 연구." 한국도서관·정보학회지, 제32권, 제3호(2001, 9), pp.397-417.

이준호, 박소연, 권혁성. "질의 로그 분석을 통한 네이버 이용자의 검색 행태 연구." 정보관리학회지, 제20권, 제2호(2003, 6), pp.27-41.

정연경. "인터넷 서점의 주제별 분류체계 설계에 관한 연구." 한국문헌정보학회지, 제35권, 제3호(2001, 9), pp.17-34.

최재황. "인터넷 학술정보자원의 디렉토리 서비스 설계에 있어서 DDC분류 체계의 활용에 관한 연구." 정보관리학회지, 제15권, 제2호(1998, 6), pp.47-67.

최희윤. "인터넷 정보서비스의 분류체계에 대한 비교연구 : 물리학을 중심으로." 정보관리학회지, 제15권, 제3호(1998, 9), pp.45-57.

한상길. "산업분야 인터넷 자원의 분류체계에 관한 연구." 정보관리학회지, 제18권, 제3호(2001, 9), pp.285-309.

He, D. and A. Goker. "Detecting Session Boundaries from Web User Logs." *Proceedings of the 22nd Annual Colloquium on Information Retrieval Research*, (2000), pp.57-66.

Silverstein, C., M. Henzinger, H. Marais, and M. Moricz. "Analysis of a very large web search engine query log." In *SIGIR Forum*, Vol.33, No.1(1999), pp.6-12.

Spink, A. et al. "Searching the Web : The Public and Their Queries." *Journal of the American Society for Information Science and Technology*, Vol.52, No.3(2001), pp.226-234.

Srikant, R., and Y. Yang. "Mining Web Logs to Improve Website Organization." *Proceedings of the Tenth International World Wide Web Conference*, (2001). pp.430-437.