WWW 탐색도구의 색인 및 탐색 기능 평가에 관한 연구

A Comparative Study of WWW Search Engine Performance

정 영 미 (Young-Mee Chung) *
김 성 은 (Seong-Eun Kim)**

목 차

초 록

WWW 탐색도구들은 인터넷 정보자원의 탐색에 있어서 매우 중요한 역할을 하고 있다. 본 연구에서는 주요한 WWW 탐색도구들의 성능을 평가할 목적으로 먼저 각 탐색도구의 색인 데이터베이스 특성, 탐색 기능, 적합성 순위 부여 방법 등을 비교한 후, 탐색실험을 통하여 검색효율, 중복탐색의 정도, 탐색결과의 유사도 등을 측정하였다. 탐색실험 결과 탐색질문의 유형에 관계없이 Alta Vista, HotBot, Open Text Index가 비교적 좋은 검색효율을 보였으며, 대부분의 탐색도구가 질문의 유형에 따라 검색효율에 있어서 차이를 보였다. 동일한 사이트를 중복하여 탐색하는 탐색의 중복도는 Magellan, WebCrawler, Yahoo!를 제외한 나머지 탐색도구들에서 모두 높게 나타났다. 탐색결과의 유사도를 측정한 결과 대부분의 탐색도구들이 매우 낮은 유사도를 보였다.

ABSTRACT

The importance of WWW search services is increasing as Internet information resources explode. An evaluation of current 9 search services was first conducted by comparing descriptively the features concerning indexing, searching, and ranking of search results. Secondly, a couple of search queries were used to evaluate search performance of those services by the measures of retrieval effectiveness, the degree of overlap in searching sites, and the degree of similarity between services. In this experiment, Alta Vista, HotBot and Open Text Index showed better results for the retrieval effectiveness. The level of similarity among the 9 search services was extremely low.

* 연세대학교 문헌정보학과 교수
** 연세대학교 대학원 문헌정보학과
접수일자 1997년 3월 5일

1. 서 론

인터넷상의 정보자료가 급증하면서 이에 대한 효율적이고 효과적인 접근을 위하여 여러 가지 도구들이 개발되어 사용되고 있다. 최근 들어 이용자들이 필요로 하는 정보를 손쉽게 찾을 수 있도록 도와주는 각종 색인 및 탐색도구들이 다양하게 제공되고 있으며 이에 대한 연구는 국외는 물론 국내에서도 활발하게 진행 중이다. 특히 이러한 연구의 일환으로 현재 인터넷상에서 제공되고 있는 여러 탐색도구들의 성능을 평가해 보고자 하는 시도가 많이 이루어지고 있다. 대부분의 연구에서는 각 탐색도구들이 제공하는 탐색기능이나 데이터베이스의 구성방법 등을 성능 평가의 기준으로 삼아 이들을 비교하는 방식을 택하거나(Sullivan, 1997) 몇 개의 탐색질문을 이용하여 탐색을 수행한 후 예상 결과와의 비교를 통해 그 평가를 시도하고 있다(Courtois, et al., 1995; Zorn, et al., 1996). 그러나 이러한 방법들은 각 탐색도구가 실제로 적합한 정보를 얼마만큼 검색해내는지를 평가하기에는 부족하다. 따라서 본 연구에서는 먼저 각 탐색도구의 색인 및 탐색 기능과 검색된 문서의 순위부여 방법을 비교한 후, 각 탐색도구의 검색 성능을 평가하였다. 평가기준으로는 검색효율과 중복검색의 정도를 측정하였으며, 검색효율 척도로는 정확률과 상대재현율을 사용하였다. 또한 유사계수 공식을 사용하여 탐색도구간의 유사도를 측정하였다. 본 연구는 현 시점에서 각 탐색도구의 성능을 평가하는 것 뿐만 아니라 앞으로의 지속적인 평가를 위한 기준과 방법을 제시하는 데에 그 목적이 있다.

평가 대상으로 선정된 탐색도구는 World Wide Web(WWW 또는 웹) 서비스에서 최근 가장 많이 이용되고 있는 Alta Vista, Excite, HotBot, InfoSeek, Lycos, Magellan, Open Text Index, WebCrawler, Yahoo! 등 9개이며 이들의 URL은 다음과 같다.

- Alta Vista
  (http://altavista.digital.com)
- Excite
  (http://www.excite.com)
- InfoSeek
  (http://www.infoseek.com)
- HotBot
  (http://www.hotbot.com)
- Lycos
  (http://www.lycos.com)
- Magellan
  (http://www.mckinley.com)
- Open Text Index
  (http://index.opentext.net)
- WebCrawler
  (http://webcrawler.com)
- Yahoo!
  (http://www.yahoo.com)

2. 색인 데이터베이스의 특성

각 탐색도구들이 구축하는 색인 데이터

베이스는 제공하는 탐색기능이나 서비스의 목적에 따라 그 구성이나 기타 특성에 있어서 차이가 있다. Nicholson(1996)은 Lycos, Alta Vista, Excite, Open Text Index, Yahoo!, Magellan 등 6개의 탐색도구를 대상으로 한 색인 및 초록의 비교연구에서 각 데이터베이스의 구축방법을 평가하는 기준으로는 사이트(site) 선정방법(수작업/로봇프로그램), 사이트 선정기준, 선정을 위한 분석에 사용되는 인터넷 정보자원의 유형, 사이트 선정을 위한 인터넷 탐색의 범위, 한 사이트가 데이터베이스에 포함되어 있는 기간, 데이터베이스 각 항목의 갱신 빈도, 데이터베이스의 규모와 성장 속도 등을 제시하였고, 색인기법의 평가 기준으로는 각 사이트에서 색인되는 부분과 이러한 부분들의 대표성 정도, 통제어휘의 사용여부와 이용자가 이를 사용할 수 있는지의 여부, 키워드색인 수행방법, 이용자가 색인된 용어들을 탐색할 수 있는 방법 등을 제시한 바 있다.

인터넷 정보자원 데이터베이스의 질을 평가하는 척도는 일반적인 서지 데이터베이스를 평가하는 척도와는 많은 차이가 있는데 이는 주로 인터넷 상의 정보가 갖는 유동성에 기인한다(Notess, 1996). 그리고 인터넷 정보자원 중에서도 하이퍼링크로 복잡하게 연결되어 있는 웹문서들을 탐색이 가능한 색인 데이터베이스로 구축하고 유지하는 것은 쉽지 않으며, ftp 정보자원에 대한 색인인 Veronica나 고퍼(gopher) 정보자원에 대한 색인인 Archie에서 사용하는 비교적 간단한 방식으로는 해결하기 어렵다는 것이 지적된 바 있다(McMurdo, 1995).

다음은 각 색인 데이터베이스의 특성을 비교하기 위하여 데이터베이스의 규모와 포함되는 정보자원의 유형, 전문(full text) 색인 여부, 그리고 최신성 유지 등의 측면을 중심으로 하여 각 탐색도구를 살펴 본 것이다.

(1) Alta Vista

Alta Vista는 여러 개의 로봇프로그램을 이용해 웹을 탐색하여 색인하는 프로그램으로서 1995년 12월 일반에게 공개된 이래 현재 476,000개의 서버에서 3,100만여개의 웹문서 전문과 14,000개 Usenet 뉴스그룹의 400만개 뉴스기사 전문을 대상으로 하여 색인 데이터베이스를 구축하고 있다. 웹문서의 경우 브라우저에 의해 보여지는 텍스트의 전문을 색인할 뿐만 아니라 html 파일 전체도 색인해 준다. 모든 단어를 색인하므로 재현율은 높일 수 있으나 중요 단어의 식별 과정이 없어 탐색시 정확률을 떨어뜨리는 결과를 초래한다.

Alta Vista 데이터베이스의 필드 중 유용한 것은 'Date' 필드로서 웹문서가 가장 최근에 갱신된 날짜 정보를 담고 있다. 이러한 정보는 웹문서에 명시되어 있지 않더라도 로봇이 탐색하는 과정에서 서버의 파일 날짜 등을 통하여 확보하는 것으로 보인다(Notess, 1996). Alta Vista는 다른 탐색도구들과는 달리 http 프로토콜에 의해 접근이 가능한 웹정보자원만을 데이터

베이스에 포함시키고 있으며 telnet, ftp, 고퍼 등에 대한 접근은 제공하지 않는다. 한글이 지원되며 데이터베이스의 갱신은 매일 300만개씩의 웹문서를 대상으로 수행되고 있다(Halttunen, 1996).

(2) Excite

데이터베이스는 '개념기반' 키워드색인 방식이 특징으로서 로봇프로그램이 소프트웨어적 알고리듬을 이용해 웹문서를 분석하여 주제를 찾아낸후 적합한 키워드를 선정함으로써 전문 색인이 가져올 수 있는 정확률 저하의 문제를 보완한다(Nicholson, 1996). 새로운 웹문서의 탐색과 함께 일주일에 한번씩 데이터베이스에 등록되어 있는 사이트들을 재방문해 갱신작업을 수행함으로써 최신성을 유지하고 있다.

Excite는 크게 6개의 서비스, 즉 Excite Search, Excite Reviews, Excite City.Net, ExciteSeeing, Tours, Excite Live!, Excite Reference로 구성되어 있는데 이 중 가장 핵심적인 기능이라 할 수 있는 'Excite Search'는 탐색 대상 데이터베이스로 World Wide Web/News Tracker/Excite Web Site Reviews/Usenet Newsgroups/Usenet Classifieds 중에서 하나를 선택할 수 있다. 각 데이터베이스의 색인 대상은 'World Wide Web'의 경우 5,000만개의 웹문서 전문을, 'News Tracker'는 New York Times, Forbes, Entertainment Weekly 등 각종 신문 및 잡지의 기사 전문을, 그리고 'Web Site Reviews'는 16개의 주제분야에서 전문가들이 선정하여 평가한 60,000여개의 웹사이트 전문을 색인한다. 그밖에 'Usenet Newsgroups'는 인터넷상에 있는 10,000개 Usenet 뉴스그룹의 기사 100만여건의 내용을 색인하며, 'Usenet Classifieds'는 Usenet의 최근 2주간 광고들을 분류, 색인한다(Halttunen, 1996).

(3) HotBot

Inktomi社가 개발한 NOW(Network of Workstations) 병렬 컴퓨팅 기술, 즉 LAN으로 연결된 워크스테이션들을 병렬로 연결하여 슈퍼컴퓨터급의 성능을 내는 기술을 기반으로하여 'Slurp the Web Hound'라는 로봇프로그램이 일주일에 한번씩 웹사이트들을 방문함으로써 데이터베이스를 구축한다. NOW 기술을 통해 하루에 1,000만개의 웹문서를 다운로딩받을 수 있으며, 현재 5,400만개의 URL과 Usenet 뉴스그룹을 대상으로 한 데이터베이스가 구축되어 있다. 이 중 3,600만개의 웹문서가 실제 색인되어 있고, 데이터베이스의 갱신은 하루에 700만건씩 수행되고 있다.

(4) InfoSeek

InfoSeek는 인터넷 탐색에 요금을 부과했던 최초의 서비스이다. 현재는 무료로 서비스되며 500만여개의 웹문서, 고퍼,

ftp 등의 정보자원 전문이 색인 대상이며 Usenet 뉴스그룹, 미국기업 디렉토리, e-mail 주소 등에 대한 데이터베이스도 구축되어 있다. 이외에도 다양한 데이터베이스 서비스를 제공하는 것이 InfoSeek의 특징인데 Hoover Company Profiles, MDX Health Digest, Corptech Directory of Technology Companies, Microcomputer Abstracts, CSA Biomedical Database, CSA Computer and Engineering Database, CSA Worldwide Market Research Database, Softbase, EDGE Newletters 등이 이용가능하다(Suoniemi, 1996). 데이터베이스의 갱신은 한달에 20회 정도씩 각 사이트와 웹문서를 방문하여 수행한다(Suoniemi, 1996).

(5) Lycos

Lycos는 http, ftp, 고퍼 사이트 등을 색인 대상으로 하며 telnet, mailto, wais, Usenet 뉴스그룹 등은 포함하지 않고 있다. 로봇프로그램은 확률적인 기법을 이용해 데이터베이스에 포함시킬 URL을 확보하는데 한 서버 내의 모든 파일들을 일일이 검사하는 것이 아니라 서버 단위로 이동하면서 URL을 찾아나간다(McMurdo, 1995). 선정기준은 사이트의 인기도, 즉 해당 사이트로의 링크수와 경로명의 길이 등으로서 링크수가 많고 경로명이 짧을수록 선호된다. 이용자가 직접 데이터베이스에 자신의 URL을 등록할 수 있으며 이를 삭제할 수도 있다(Suoniemi, 1996).

현재 인터넷상에 있는 모든 웹문서의 90%인 1,600만여개를 데이터베이스의 색인대상으로 한다고 말하고 있는데 실제로는 360만개의 문서만을 완전하게 색인하였으며 700만개의 웹문서는 부분적으로 색인되어 있다(Courtois, 1996). 이 밖에도 소리, 이미지, 비디오 등에 대한 500만여개의 이진파일도 포함한다. Lycos의 데이터베이스는 2주에 한번씩 전체 데이터베이스를 갱신하고 있으며 주당 약 300,000여개의 페이지가 추가되고 있다(Suoniemi, 1996). Lycos의 색인은 전문색인이 아니라 통계적 분석에 의거한 개념기반 색인이며(Singh, 1996), 탐색시 Alta Vista와 마찬가지로 한글이 지원된다.

(6) Magellan

1995년부터 서비스에 들어간 Magellan은 색인자와 함께 로봇프로그램이 웹사이트를 선정하고 색인하여 데이터베이스를 구축하고 있다. 색인자는 키워드와 사이트의 예상이용자를 나타내는 디스크립터를 선정하고 적절한 주제카테고리에 해당 사이트를 할당하는 일을 한다(Suoniemi, 1996). 사이트의 선정기준은 페이지의 구성, 내용의 흡인력, 정보의 최신성 및 깊이 등이다(Nicholson, 1996). 데이터베이스는 150만여개의 웹사이트, ftp 및 고퍼 서버, Usenet 뉴스그룹, 그리고 telnet 서비스 등을 포함하고 있으며 전문을 색인대상으로 한다. 전문 데이터베이스와는 별

도로 키워드 및 통제어휘 데이터베이스가 구축되어 있다. 그리고 리뷰와 평가를 거친 사이트 40,000여개가 데이터베이스에 포함되어 있는데 여기에는 뉴스그룹과 메일링리스트도 포함되며(Singh, 1996), 이를 위해 매주 수천 개의 새로운 사이트를 방문하고 있다.

(7) Open Text Index

1,900만여개의 하이퍼링크와 100만여개에 달하는 웹문서 전문이 색인대상이며(Courtois, 1996), ftp와 고퍼 등도 데이터베이스에 포함된다. Open Text 5라는 독자적인 소프트웨어에 의해 각 웹문서의 단어들을 색인하여 데이터베이스에 저장한다. Open Text Index의 강점은 그 이름에서도 알 수 있듯이 색인 자체의 특성에 있다(Zorn, et al., 1996). 40개 이상의 파일 유형을 색인하고 있으며 각 웹문서의 모든 단어와 구절을 색인하는 전문색인이므로 탐색문의 불용어도 무시되지 않는다. 그리고 용어가 출현한 필드를 인식함으로써 필드제한 탐색이 가능하며 부분적으로 일본어, 스페인어, 포르투칼어 등 다국어를 지원하고 있다. 하루에 약 50,000개의 웹문서가 추가되거나 갱신되고 있다.

(8) WebCrawler

로봇프로그램을 이용하여 구축되는 WebCrawler의 데이터베이스에는 현재 220,000여개의 웹문서 전문이 색인되어 있고, 아직 색인되지 않은 360만여개 웹문서의 리스트를 확보하고 있다 (Kimmel, 1996). 본문 텍스트 뿐만 아니라 html 파일 자체도 색인대상이며(단, URL은 제외) 웹문서외에 ftp, 고퍼 등의 인터넷 정보도 포함하고 있다. 웹로봇인 WebCrawler가 "breadth-first" 방법, 즉 파일단위의 탐색이 아닌 사이트단위 탐색을 통해 문서를 수집하고 색인한다. 따라서 다른 탐색도구들과는 달리 하루에 같은 사이트를 지나치게 여러번 방문하는 일은 생기지 않는다. 데이터베이스의 갱신은 2주에 한번씩 수행되고 있다(Sullivan, 1997).

(9) Yahoo!

Yahoo!는 웹사이트를 비롯하여 Usenet 뉴스그룹, e-mail 주소 등을 색인하여 데이터베이스를 구축한다. 각 웹사이트에 대해서는 URL, 표제, 코멘트 등이 색인되며 현재 65,000여개의 항목이 데이터베이스에 포함되어 있다(Courtois, et al., 1995). 원래는 이러한 과정을 10명의 색인자가 전부 수작업으로 수행하였으나 인터넷 정보가 급증함에 따라 1995년 9월 로봇프로그램을 등장시켰다(Ross and Hutheesing, 1995). 색인자는 전체 사이트를 검토하여 적절하다고 판단되면 사이트의 홈페이지만을 데이터베이스의 적절한 주제명 아래 포함시킨다. 데이터베이스의 갱신은 이용자가 문제를 제기할 때에만 수행된다(Nicholson, 1996).

3. 탐색 기능

각 탐색도구들은 초보자들도 쉽게 이용할 수 있는 기본적인 키워드 탐색 외에도 전문적인 정보검색사나 사서들에게 적합한 발전된 형태의 탐색 지원기능을 다양하게 제공하고 있다. Zorn 등(1996)은 WWW 탐색의 특성상 정확성과 효율성을 높이기 위해서는 지나치게 많은 양의 탐색결과를 제한하기 위해 복합 불논리(Boolean) 구문(중첩 논리), 사이트의 중복검색 감지 능력, KWIC 색인, 필드지정에 의한 제한탐색, 인접탐색 및 구(phrase) 탐색, 적합성 순위별 결과 출력, 탐색결 과집합의 처리, 절단탐색 등의 발전된 기능들이 제공되어야 한다고 지적하고 있다. 실제로 각 탐색도구의 데이터베이스의 규모, 색인의 깊이와 함께 이와 같이 다양하게 제공되는 탐색기능에 따라 탐색의 난이도와 탐색결과가 달라지게 된다. 다음은 각 탐색도구가 제공하는 주요한 탐색기능의 개요이다.

(1) Alta Vista

탐색문의 구성방법이나 기타 특징들이 상업용 온라인 DB 서비스와 유사한 것이 특징이며, 'Simple Query'와 'Advanced Query'를 구분하여 탐색 서비스를 제공한다. 먼저 'Simple Query'에서는 불논리 연산기호(Boolean operators)의 지정없이 기본적으로 OR 연산을 수행하며, 용어앞에 + 기호를 붙여 반드시 포함되어야 할 탐색어를 표시해 줌으로써 AND 연산을 수행할 수 있다. 이중인용부호를 사용하여 구탐색을 할 수 있으며 기호를 사용한 우측절단탐색도 가능하다. 탐색어를 소문자로 표기하면 소문자와 대문자를 모두 검색해낼 수 있으나 대문자로 탐색어를 표기하면 완전일치 검색이 된다. 탐색할 필드도 지정할 수 있는데 필드명과 탐색어를 콜론(:)으로 연결하여 표현한다. 'Advanced Query'의 경우에는 'Simple Query'에서 제공하는 기능 외에 AND(&), OR(|), NOT(~), NEAR(!, 10단어 이내 인접탐색) 등의 기호를 사용하여 불논리에 의한 탐색을 할 수 있다. 괄호를 사용한 중첩 탐색문 입력이 가능하며 탐색문을 입력하는 칸(Selection criteria text box)과 탐색결과에 순위를 부여하기 위한 기준으로 사용할 용어를 입력하는 칸(Results ranking criteria text box)이 따로 제공된다. 또한 'Start date'와 'End date' 옵션을 이용하여 색인된 시점을 기준으로 한 제한탐색이 가능하다.

(2) Excite

Excite의 탐색엔진은 단순한 키워드 탐색만을 수행하는 다른 탐색엔진과는 달리 ICE(Intelligent Concept Extraction)라는 기술을 사용한다. 이는 색인 과정에서 해당 단어와 관련된 단어들을 학습해 두었다가 탐색시 탐색문에 포함되지 않은 용어라도 해당 주제와 관련있는 경우 이 용

어들을 탐색어로 확장하여 사용하는 것이다. 이용자가 탐색어로 여러 개의 용어를 사용하면 그 용어들간의 관계를 파악하여 관련 개념을 표현하는 다른 용어들도 탐색에 사용한다. 이러한 관련어의 학습 및 탐색어의 확장은 웹문서 내용의 통계적 분석을 통해 이루어진다.

탐색문은 자연어 질문, 단어의 단순 나열, 불논리 탐색문 등 세가지가 모두 가능하며 불논리 탐색문에서 사용되는 논리연산기호는 AND(+), OR, AND NOT(-) 등이다. 또한 연산의 우선순위 지정 및 중첩된 탐색문 표현을 위해 괄호를 사용할 수 있고 자동으로 우측절단 탐색이 수행된다. 탐색어의 대소문자를 구별하므로 인명과 같은 고유명사를 탐색어로 사용하는 경우에는 각 단어의 첫글자를 대문자로 표기하여야 원하는 탐색 결과를 얻을 수 있다. 그리고 탐색결과의 표제 옆에 나오는 'More Like This' 앵커는 해당 문서를 예로 삼아 다시 새로운 탐색을 시도하여 유사한 문서들을 더 볼 수 있도록 한다.

위와 같은 'Excite Search'는 포괄적이고 망라적으로 웹사이트를 탐색하고자 할 때 적합한 방식이며 좀더 선별된 웹사이트들만을 살펴보고자 할 때는 'Excite Reviews' 서비스를 이용한다. 'Excite Reviews'에서는 16개 주제분야에서 전문가들이 선정한 웹사이트 160,000여개를 1~4등급으로 평가하여 제공하며, 이용자는 계층적 주제디렉토리를 통해서 접근하거나 탐색대상을 'Excite Reviews'로 제한하여 탐색을 수행함으로써 원하는 정보를 얻을 수 있다. 그리고 전형적인 키워드 탐색외에도 개념탐색이 가능하여 하나의 주제에 대해서 동의어 등을 이용하여 탐색할 수 있으나(Courtois, et al., 1995) 현재 이 기능은 제공되지 않고 있다.

(3) HotBot

이용자의 탐색 능력을 차별화하여 기본탐색과 Modify/Date/Location/Media Type을 구분하여 제공한다. 이중 'Date', 'Location', 'Media Type' 옵션들은 이전의 'Expert' 옵션으로 함께 제공되 던 것을 각각 독립시킨 것이다. 기본탐색에서는 우선 탐색대상을 웹과 Usenet 뉴스기사 중에서 선택할 수 있으며, 탐색어의 구문은 all of the words/any of the words/the exact phrase/the person/links to this URL/the Boolean expression 중에서 선택할 수 있다. 그리고 탐색결과의 출력양식에 있어서는 한번에 보여 주는 탐색결과의 건수와 해당 사이트에 대해 제공되는 정보(full descriptions/brief description/URLs only) 등을 지정할 수 있다. 'all of the words' 탐색은 불논리 연산의 AND 연산과 동일하며 'any of the words'는 OR 연산과 같은 결과를 낳는다. 구탐색은 메뉴에서 'the exact phrase' 탐색을 선택하거나 용어들을 이중인용부호로 묶어 표현함으로써 수행한다. 이러한 구탐색은 용어들의 순서도 그대로 적용된다는 점에서 AND 연산과는 다르다. 'the Boolean expression'을 선택

하는 경우에는 AND (&), OR(|), NOT(!) 등을 연산기호로 사용하며 괄호의 사용도 가능하다. 그러나 중첩탐색문이나 절단탐색(부분문자열 탐색)은 지원하지 않으며, 인접탐색은 두 단어 이상으로 된 인명을 탐색하는 경우에만 제한적으로 수행되고 있다. 그리고 탐색어의 대소문자는 구분하지 않지만 대소문자가 섞여 있는 경우에는 완전일치 탐색을 수행한다.

(4) InfoSeek

InfoSeek는 이용자의 각기 다른 탐색요구에 부응하기 위해 'Ultrasmart'와 'Ultraseek'의 두가지 탐색 서비스를 제공한다. 'Ultrasmart'는 이용자가 자신의 정보요구에 대해 구체적인 생각을 갖고 있지 못할 때 포괄적인 탐색을 수행할 수 있도록 한다. 이것은 탐색과 일람(browsing) 기능을 결합한 것이라고도 말할 수 있는데, 탐색결과 적합한 웹사이트들을 검색해낼 뿐만 아니라 이용자의 질문에 적합한 주제디렉토리상의 관련 주제명과 뉴스기사의 리스트 등도 함께 제공하여 이용자가 일람할 수 있도록 한다.

'Ultraseek'는 검색된 웹사이트의 리스트만을 바로 원할때 사용할 수 있는 탐색서비스로서 'Ultrasmart'의 속도, 정확성, 최신성, 포괄성 등을 모두 제공함과 동시에 이용자에게 효율적인 형식으로 탐색결과를 제공하는 서비스라고 할 수 있다. 즉 'Ultrasmart'가 초보적인 이용자들이 자신의 탐색요구를 좀더 정확하게 파악할 수 있도록 인도하는 서비스라면, 'Ultraseek'는 자신이 찾고자 하는 정보가 무엇인지 정확하게 알고 있고 빠른 탐색을 원하는 이용자에게 적합한 서비스이 다.

또한 12개의 주제카테고리별 리뷰사이트를 일람할 수 있도록 주제별 디렉토리를 제공한다. 각 주제카테고리 페이지 상단에는 붉은색 화살표로 현재 선택된 주제명을 보여 주며, 하단에는 탐색문을 입력할 수 있는 칸이 있어 해당 주제카테고리로 한정되는 탐색이나 전체 웹을 대상으로 하는 탐색을 바로 수행할 수 있다.

탐색문의 구문을 살펴보면 우선 불논리 연산기호는 지원되지 않고 있는데 이것은 InfoSeek가 가능한 한 간단한 탐색문을 입력하도록 요구하기 때문이다. 대신 탐색어 앞에 + 기호를 사용하여 AND 연산과 같은 결과를 가져올 수 있으며 - 기호로는 NOT 연산의 효과를 얻을 수 있다. 구로 된 탐색어는 이중인용부호나 하이픈을 이용하여 표현함으로써 구탐색이 가능하다.

또한 탐색어 앞에 소문자로 된 필드명(link, site, url, title)과 콜론을 사용함으로써 필드지정 탐색을 수행할 수 있으며 절단탐색 기능은 제공되지 않는다. 키워드를 이용한 탐색외에도 자연어 질문 그대로를 탐색문으로 사용할 수도 있는데 탐색 결과는 동일하다(Singh, 1996).

(5) Lycos

Lycos는 우선 'Customize Your

Search' 메뉴를 통해 탐색대상, 탐색연산, 출력양식 등에 대한 옵션을 지정할 수 있다. 탐색 대상은 The Web/Sounds/Pictures/Sites By Subject/Top 5% Sites 중에서 선택한다. 탐색 연산은 'match any term(OR)', 'match all terms(AND)', 그리고 불논리 연산을 약간 수정하여 탐색어들 중 일부만 매치되는 경우에 검색하는 'match 2/3/4/5/6/7 terms' 옵션을 지정할 수 있다. Lycos는 괄호를 사용한 중첩 탐색문을 사용할 수 없으므로 포괄적인 탐색 결과를 얻기 위해서는 여러가지 동의어나 관련 용어를 혼합하여 탐색하여야 한다.

또한 탐색연산과 관련하여 지정할 수 있는 옵션으로서 loose/fair/good/close/strong match가 있는데 이는 검색될 문서의 적합성 최소 점수를 지정하는 옵션이다. 이외에도 페이지당 출력되는 탐색결과 건수를 지정할 수 있으며 출력 양식도 detailed/summary/standard results 중에서 지정할 수 있다. 또한 용어의 뒤에 $를 붙임으로써 우측절단탐색의 결과를 얻을 수 있으나 필드제한 탐색과 인접탐색 및 구탐색 기능은 제공하지 않고 있다.

Lycos는 'a2z Sites by Subjects'라는 주제별 디렉토리에서 총 16개의 대주제로 사이트들을 분류하여 놓았으며, 각 사이트에 대한 한두 줄의 요약을 제시하고 있다. 또한 'Find Related Site' 앵커를 선택하면 해당 사이트의 용어를 이용하여 새로운 탐색을 시도하며 그 결과 관련 웹문서로 연결시켜 준다.

(6) Magellan

크게 키워드 탐색과 주제카테고리별 일람의 두 가지 기능을 제공한다. 먼저 키워드 탐색은 그 탐색대상을 'rated and reviewed sites only'와 'entire database' 중에서 선택할 수 있다. 전자는 Magellan의 편집진들이 평가해 놓은 사이트들만을 대상으로 하여 탐색을 수행하는 경우이며, 후자는 전체 데이터베이스를 대상으로 하는 탐색이다. 탐색어의 대소문자는 구분하지 않으며, 탐색어에 사용된 불용어는 모두 무시된다. 그리고 탐색어 앞에 + 기호를 붙여 AND 연산과 같은 탐색결과를 낼 수 있으며 - 기호는 NOT 연산과 같은 결과를 가져온다. 절단탐색은 지원하지 않지만 대신 입력된 탐색어를 키워드 데이터베이스에서 찾아 내어 그 전후에 오는 키워드를 관련어로 간주하여 탐색에 사용한다(Nicholson, 1996).

탐색어 입력시 Boolean operators/Mininum Rating/Duraion of Options 등을 지정할 수 있다. 불논리 연산은 자동설정 기능이 OR 연산이며 AND 연산을 선택할 수 있다. 'Minimum Rating'은 검색해 내는 사이트들의 적합성 점수에 제한을 두는 것이며, 'Duration of Options'를 통해서는 한 번 설정한 탐색옵션을 저장할 수 있다. Magellan은 키워드 탐색 외에도 주제카테고리의 계층적 일람 기능을 제공한다. 이는 전문가들이 리뷰한 사이트들을 주제별로 모아 놓은 것으로서 사이트의 선정 및 평가는 사이트가 제공

하는 정보의 깊이(포괄성 및 최신성), 이용의 용이성(사이트의 구성), 특징적 매력(참신성, 흥미성)의 세가지 측면에서 이루어진다.

(7) Open Text Index

탐색메뉴는 크게 Simple Search, Power Search/Current Events/Newsgroups/Email/Other Languages로 구분되어 있다. 'Simple Search'에서는 단어탐색과 구탐색을 지정하여 탐색을 수행할 수 있다. 'Power Search'는 웹페이지의 탐색할 부분(anywhere/summary/title/first heading/URL)을 지정하는 옵션과 인접탐색이 지원되는 불논리 연산기호(AND/OR/BUT NOT/NEAR/FOLLOWED BY)를 사용할 수 있도록 옵션을 제공한다. 이러한 연산기호는 사용된 순서대로 적용된다. 탐색어의 대소문자나 구두점 등은 무시되며 절단탐색이 불가능하므로 망라적인 탐색을 하기 위해서는 탐색어로 단수형과 복수형, 동사원형과 과거형, 명사형 등 가능한 각종 변형을 모두 사용하여 OR 연산을 해야 한다. 그리고 또 다른 문제는 괄호를 사용하지 못하므로 중첩 탐색문을 표현하지 못하기 때문에 맨 마지막에 있는 탐색어에 대한 연산이 잘못 해석될 우려가 있다는 점이다.

특징적인 기능으로는 최대 5개의 용어를 이용한 초기 탐색이 수행된 후 탐색어를 하나씩 추가하여 새로운 탐색문으로 재탐색을 시도함으로써 이용자의 정보요구를 좀더 정확하게 반영할 수 있도록 하는 적합성 피드백 과정이 있다.

(8) WebCrawler

단순한 탐색 인터페이스를 제공한 이용의 용이성이 특징이다. 'Search' 메뉴에서는 초보자를 위한 자연어 탐색과 전문가를 위한 불논리 탐색구문을 모두 지원한다. 자연어 탐색은 기본적으로 AND 연산을 수행한다. 불논리 탐색구문에서 이용할 수 있는 불논리 연산기호는 AND/OR/NOT/NEAR/ADJ이며, NEAR/n은 앞 뒤 n단어 이내로 인접해 출현하는 용어들을 탐색어로 사용할 때, ADJ는 두 단어로 된 구가 탐색어일 때 사용한다. ADJ와 같은 결과를 얻기 위해 두 단어를 이중인용부호로 묶어 입력할 수도 있다. 그밖에 연산의 우선순위 지정을 위해 괄호를 사용할 수 있고 * 기호를 사용한 우측절단탐색이 가능하다.

'Browse' 메뉴에서는 웹사이트를 전문가들이 리뷰하여 15개 주제로 모아 놓은 것을 일람할 수 있도록 하고 있다. 일람과정에서 'Spidey Search' 버튼을 누르면 해당 주제카테고리나 리뷰한 사이트와 관련된 웹문서을 탐색한다. 이는 WebCrawler의 편집진이 최적의 검색결과를 내도록 선정한 키워드로 작성된 유사 탐색문을 이용하는 것이다.

(9) Yahoo!

Yahoo!의 가장 큰 장점은 계층적으로 잘 조직되어 있는 주제별 디렉토리라고 할 수 있다. 특히 탐색에 있어서도 이와 같이 조직된 주제별 디렉토리의 특징을 이용하여 탐색을 한정할 수 있다. 먼저 초기 화면에서 'options'를 선택하면 기본탐색에 대한 옵션을 지정할 수 있다. 우선 탐색 대상을 Yahoo!/Usenet/Email addresss 중에서 선택할 수 있으며, 'Yahoo!'를 선택한 경우에는 다시 세분된 탐색대상(Yahoo Categories/Web Sites /Today's News/Net Events)옵션과 탐색방법(Intelligent default/An exact phrase match/Matches on all words(AND)/ Matches on any word(OR)/A person's name)옵션을 제공한다. 그밖에 날짜(1 day/3 days/1 week/1 month/3 months/6 months/3 years)와 페이지당 출력되는 검색결과의 건수(10/25/50/100) 등도 지정할 수 있다.

기본탐색의 옵션을 지정하는 화면에서 다시 'Advanced Search Syntax'를 선택하면 정확률 향상을 위해 보다 정교한 탐색문을 작성할 수 있도록 한다. 탐색문 자체에 탐색 조건을 표현할 수 있는 방법으로서 + 기호를 이용한 AND 연산과 - 기호를 이용한 NOT 연산이 가능하다. 이중 인용부호로는 구탐색어를 표현할 수 있으며 * 기호를 사용해 우측절단 탐색을 수행할 수 있다. 또한 탐색어 입력시 표제, URL 등 웹문서상의 특정한 부분만을 탐색하도록 지정할 수 있다.

4. 적합성 순위 부여 알고리듬

Brandt(1996)는 '적합성'이라는 개념이 정보검색 과정에서 탐색주제의 정의, 탐색, 평가의 세 단계에서 적용된다고 보았다. 먼저 주제정의와 관련해서는 인터넷 정보자원이 급증하고 있고 대부분의 탐색도구들이 가능한 한 모든 정보자원에 대한 접근을 제공하기 위해 노력하고 있기 때문에 이용자가 자신의 정보요구에 부합하는 탐색을 하기 위해서는 적절한 시작점을 결정해야만 한다는 것이다. 즉 주제지향적 접근방법을 통해 일단 탐색의 범위를 좁혀야 한다는 뜻이다. 대부분의 탐색도구에서 제공하는 계층적 주제디렉토리나 주제카테고리의 일람 등이 바로 이러한 시도중 하나라고 할 수 있다. 두번째로 탐색단계에서의 적합성이란 이용자가 자신의 정보요구에 얼마나 적합한 탐색문을 작성하고 탐색연산을 수행하는지를 의미한다.

마지막으로 평가단계에서의 적합성이란 탐색결과의 적합성을 의미한다. 즉, 탐색결과가 이용자의 정보요구와 얼마나 관련이 있고 어느 정도 이용될 수 있는지를 나타내는 것이다. 이용자가 적합성 판정을 하기 위해서는 탐색결과를 하나씩 일일이 살펴 보아야 하는데 대부분의 WWW 탐색도구의 탐색결과가 수백, 수천건씩 되므로 효과적인 적합성 판정을 위해서는 일정한 기준에 의해 순위를 부여하여 이용자에게 출력해 주는 것이 필요하다. 그리고 적합성 판정이라는 것이 주관적인 것

이기 때문에 이용자가 판정을 내리는 데 사용할 정보가 정확하고 충분하게 제공되어야 한다. 대부분의 탐색도구에서 탐색어의 출현빈도나 출현위치, 텍스트의 길이 등을 이용해 탐색결과에 순위를 매겨 출력하고 있다. 다음은 각 탐색도구의 적합성 순위 부여 알고리듬과 출력양식의 특징을 기술한 것이다.

(1) Alta Vista

탐색 결과 검색된 문서는 이용자의 정보요구에 적합한 순서로 출력된다. 'Simple Query'에서는 탐색어가 웹문서의 앞부분(웹문서의 표제나 Usenet 뉴스기사의 헤더)에서 출현하는 경우, 문서내에서 출현한 탐색어들간의 거리가 서로 가까운 경우, 그리고 탐색어가 한 문서에서 여러번 출현하는 경우에 이 탐색어는 높은 가중치를 부여 받게 되며 따라서 관련된 문서는 높은 순위로 검색된다.

'Advanced Query'에서는 순위부여 기준이 되는 용어를 입력하지 않으면 순위없이 무작위로 탐색결과가 출력되며, 기준용어가 입력된 경우는 다시 두가지로 나누어진다. 즉, 해당 용어가 탐색어로 입력된 용어인 경우에는 그 용어에 높은 가중치를 주어 그 용어가 출현한 문서가 높은 순위로 출력되고, 해당 용어가 탐색어로 사용되지 않은 경우에는 일차로 검색된 문서들 중 이 용어를 포함하지 않은 문서들은 모두 제거하고 나머지 문서들을 순위에 따라 출력한다.

출력양식은 standard form/compact form/detailed form의 세가지 중에서 선택할 수 있는데, 'standard form'에서는 표제, 앞에 오는 25개 단어, URL, 최근 색인날짜, 파일 크기 등에 대한 정보를 출력하고 'compact form'에서는 표제, 날짜, 그리고 웹페이지의 첫번째 줄을 보여 준다. 'detailed form'은 'standard form'과 동일하다.

탐색결과 한번에 10건씩의 문서를 출력하며 각 문서의 적합성 점수는 출력되지 않는다. 또한 각 문서마다 일치하는 탐색어의 수를 보여 주지 않기 때문에 OR 연산인 경우에는 이용자가 적합성을 판단하기가 용이하지 않다.

(2) Excite

이용자가 탐색문으로 표현한 정보요구와 각 웹사이트의 정보를 비교하여 자동으로 적합성 점수를 산출한다. 적합성 점수는 100%를 만점으로 하는 백분율 점수이며, 탐색결과는 적합성 점수가 큰 것부터 출력된다. 탐색결과의 출력은 이러한 적합성 순위별 방식 외에도 같은 사이트에서 여러 문서가 검색되는 경우가 많기 때문에 사이트별로 모아 출력하는 방식을 택할 수도 있다.

각 사이트에 대해 탐색결과로 제공하는 내용은 탐색대상에 따라 약간씩 다르다. World Wide Web을 대상으로 한 탐색에서는 한번에 10건씩의 결과를 보여 주며 각 사이트에 대한 적합성 점수, 표제와

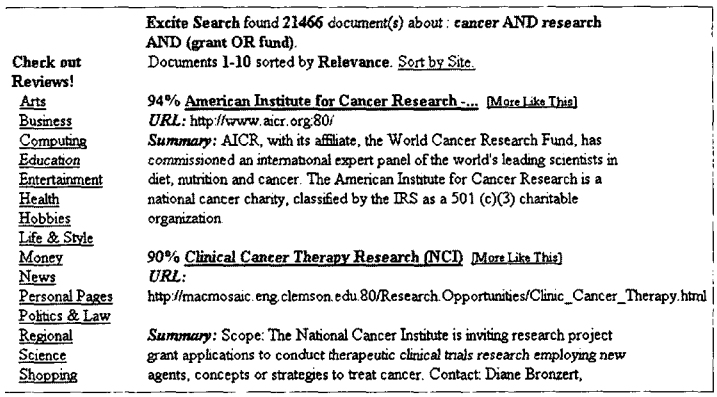

그림 1. Excite의 탐색결과 출력 양식

URL, 요약을 제공한다. 요약은 로봇프로그램에 의해 생성되는데 개념기반 색인기법에 따라 적합한 주제와 키워드가 결정되면 이 키워드가 출현한 줄을 찾아낸 후 이들을 연결해 구성하는 것이다(Nicholson, 1996). 그림 1은 Excite의 탐색결과 출력된 문서의 내용이다.

(3) HotBot

탐색결과는 적합성 순위에 따라 출력되며 100%를 만점으로 한다. 이러한 적합성 점수는 탐색어의 출현 양상에 의해 결정되는데 탐색어의 출현빈도가 높을수록, 탐색어가 공통어인 경우보다 흔히 사용되지 않는 용어일 경우, 탐색어가 표제나 키워드에 출현하는 경우에 해당 문서의 적합성 점수는 높아진다. 또한 웹문서의 텍스트 길이가 짧을수록 높은 순위가 매겨진다.

그리고 URL 주소만 다를 뿐 내용이 같은 웹문서는 'alternates'로 간주하여 하나의 표제 아래 모아서 출력한다. 출력양식은 적합성 점수, 표제, 웹문서 전문의 첫번째 문장을 이용한 요약문, URL, 파일크기, 색인 날짜 등으로 구성된다.

(4) InfoSeek

검색 결과의 순위는 백분율로 표현된 적합성 점수에 의해 결정된다. 기본적으로 탐색어와 일치하는 용어의 수가 많을수록 높은 순위로 검색되며, 그밖에도 탐색어가 웹문서의 첫머리 혹은 표제 부분에 출현하는 경우, 탐색어의 출현빈도가 높은 경우에도 해당 웹문서는 높은 순위로 검색된다. 그리고 InfoSeek 데이터베이스내에서의 출현빈도가 낮은 용어, 즉 문헌분리능력이 높은 용어는 높은 가중치를 부여받는데 이러한 높은 가중치를 갖는 탐색어를 포함한 웹문서 또한 높은 순위로 검색된다.

탐색결과 출력양식은 먼저 각 탐색어에 대한 검색건수 및 탐색연산을 거친 최종

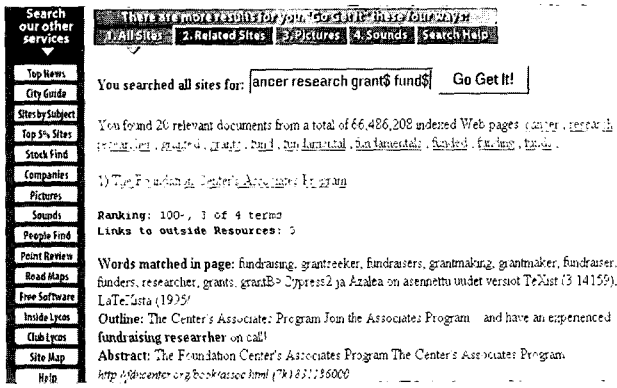

그림 2. Lycos의 탐색결과 출력 양식

적인 검색건수를 보여 준 후 검색된 각 문서에 대한 정보를 제공한다. 이러한 정보에는 표제, 요약, URL, 적합성 점수, 파일 크기 등이 포함되며 체크 표시가 되어 있는 웹사이트는 'select' 사이트로서 InfoSeek가 추천하는 사이트이다.

(5) Lycos

검색된 웹문서의 적합성 점수는 100%를 만점으로 하는 백분율 점수로 표현되는데, 기본적으로 일치하는 탐색어의 수가 많을수록 높은 순위를 차지하며, 웹문서내에서의 탐색어의 위치에 의해서도 점수가 달라져서 텍스트 본문에 출현하는 경우보다 표제나 제목에 출현하는 경우가 더 높은 적합성 점수를 얻게 된다. 그리고 탐색어와 출현 용어가 일치할수록, 즉 변형이 적을수록 해당 웹페이지가 높은 순위로 검색되는데 예를 들어 'glow'라는 탐색어를 사용하였을 때 'glows'가 출현한 웹페이지가 'glowworm'이 출현한 웹페이지보다 높은 순위로 검색된다(McMurdo, 1995). Lycos에서는 구탐색어를 표현할 수는 없으나 탐색어들이 인접해서 나타나는 경우 해당 문서는 높은 순위로 검색된다. 이밖에 탐색어의 출현빈도도 적합성 점수에 영향을 미치며, 사이트의 상대적 이용도나 인기도 고려된다.

출력양식을 보면 우선 각 탐색어와 일치되는 단어의 모든 변형을 열거하고 총 검색건수를 보여준다. 그 다음 표제, 백분율 값으로 표현된 적합성 점수, 일치하는 용어 수와 인접도, 외부로의 링크 수, 일치하는 탐색어 리스트, 상위 10개의 키워드로 구성된 개요(outline), 불용어를 제외한 첫 100개의 중요 단어들로 작성된 요약, URL, 최근 갱신날짜, 파일 크기 등의 정보를 제공한다. 100개의 중요 단어를 결정하는 기준은 역문헌빈도 가중치이다(McMurdo, 1995). 그런데 개요에 대한 설명이 부족하여 이용자가 이를 이해하기가 쉽지 않고, 요약 작성시 웹문서의 길이에 관계 없이 100개의 단어만을 사용하므로

길이가 짧은 문서는 대부분의 내용이 요약에 포함되나 길이가 긴 페이지의 경우에는 극히 일부분만이 요약으로 표현되므로 이용자가 적합성을 판정할 때 문제 가 발생할 수 있다(Nicholson, 1996). 그림 2는 Lycos의 탐색결과 출력된 문서의 내용이다.

(6) Magellan

검색된 사이트는 그 적합성 정도에 따라 순위가 매겨지는데 우선 리뷰 사이트가 일반 사이트 보다 높은 순위로 검색된다. 이용자가 탐색어로 사용한 용어들이 모두 포함되어 있는 사이트가 그렇지 못한 사이트보다 순위가 높고, 용어의 출현위치(표제, 키워드, 기술부, URL)에 따라서도 점수가 달라진다. 그리고 탐색어의 출현빈도가 높을수록 해당 사이트는 적합한 것으로 판단된다.

검색결과는 10건씩 출력되며 출력내용은 총 검색건수와 함께 표제, 요약, URL로 구성되는데 적합성 점수는 포함하지 않는다. 리뷰 사이트에 대해서는 'Review' 앵커를 제공하여 리뷰 전문과 평가정보를 볼 수 있도록 하고 있다. 그리고 각 검색결과 화면 상단에는 관련된 주제 카테고리명의 리스트가 제공되는데 그 중 하나를 선택하면 이용자의 탐색문과 관련있는 리뷰 사이트의 리스트를 살펴볼 수 있다.

(7) Open Text Index

적합성 순위는 탐색어의 출현빈도에 의해 매겨지며 검색결과의 출력양식에는 웹문서의 첫 100여개 단어를 이용해 자동으로 생성한 요약도 포함된다. 그러나 적합성 점수의 구체적인 생성 방법과 적합성 점수의 최고점 등에 대한 설명이 부족하다. 10건씩 출력되는 문서의 출력양식은 상세한 편으로서 표제, URL, 웹페이지의 첫 100개 단어들로 구성된 요약문, 적합성 점수, 파일 크기로 구성된다.

(8) WebCrawler

기본적으로 일치하는 탐색어의 수가 많을수록 해당 웹문서는 높은 순위로 검색된다. 그밖에 적합성 점수에 영향을 미치는 요인은 크게 두가지로서 탐색어의 출현빈도와 탐색어의 고유성이다. 탐색어의 고유성이란 탐색어가 특정한 문서에서만 출현하는 용어인가 아니면 여러 다른 문서에서도 많이 출현하는 용어인가를 의미하는 것으로서 해당 용어의 문헌빈도(document frequency)를 고려하는 것이다. 즉 탐색어의 문헌빈도가 낮을수록 적합성 점수는 높아진다. 검색된 각 웹문서의 적합성은 백분율 점수로 표현된다.

25건씩 출력되는 문서의 출력양식은 간략형식과 상세형식의 두가지가 있는데 간략형식은 웹문서의 표제만 보여 주는 반면 상세형식은 표제와 함께 요약을 보여주면서 'Similar Pages' 앵커를 통해 유사

한 웹문서들로 링크시켜 준다. 한번에 출력하는 문서의 건수는 따로 지정할 수 있다.

(9) Yahoo!

적합성 순위를 결정하는 요인으로는 일치하는 탐색어의 수, 출현위치에 따른 가중치, 주제카테고리의 포괄성 등이 있는데, 일치하는 키워드 수가 많을수록, 탐색어가 URL이나 표제에 출현하는 경우, 그리고 주제카테고리가 Yahoo! 주제 트리(tree)의 윗쪽에 위치할수록 해당 웹문서는 높은 순위로 검색된다.

탐색대상을 'Yahoo!'의 'Web Sites'로 선택하여 탐색을 수행하면 우선 해당 탐색문에 적합한 주제카테고리의 리스트를 제시한 다음 각 사이트들을 주제카테고리별로 모아 적합성 순위별로 출력한다. 여러 탐색도구 중 출력양식이 가장 간단하며 웹사이트 검색결과의 경우 검색된 사이트의 주제카테고리와 표제, 그리고 짧은 기술부로 구성된다.

5. 탐색 실험

5. 1 실험용 탐색질문과 적합성 판단 기준

본 연구에서 탐색도구의 검색효율 평가를 위해 사용한 탐색질문은 Zorn 등(1996)이 4대 탐색 도구의 탐색기능 비교를 위해 사용했던 3개의 탐색질문 가운데 2개이며, 이 탐색질문들을 상용 온라인 정보서비스에서 사용할 경우에는 일반적으로 다음과 같은 탐색식을 구성하게 된다.

① Cancer research grant or funding opportunities
=〉 cancer and research and (grant* or fund*)

② XI International conference on AIDS
=〉 (XI or 11th or eleventh) and international and conference and AIDS

이 실험은 1996년 12월에 수행되었으며 실험시 각 탐색도구의 특성에 비추어 가장 좋은 검색결과를 낼 수 있도록 적절한 탐색문과 탐색기능을 사용하였다. 그러나 각 탐색도구의 색인 데이터베이스에는 매일 많은 양의 새로운 웹사이트가 추가되고 있으며 또한 탐색기능도 계속 개선되고 있으므로 앞으로 수행될 다른 탐색 실험의 결과는 이 실험결과와 차이가 있을 수 있다.

적합성 판정은 탐색결과 순위에 따라 출력된 간단한 내용만으로는 어려운 경우가 많으므로 출력물에 연결된 링크를 따라 해당되는 사이트로 가서 문서 자체를 살펴 봄으로써 적합 여부를 판단하였다. 단, 링크로 연결된 정보가 이용이 불가능한 경우(예: 해당 웹문서 "Not Found") 에는 해당 탐색도구의 최신성 유지라는 측면에서 부정적이므로 그 문서는 부적합

한 것으로 간주하였다.

첫번째 탐색질문인 "Cancer research grant or funding opportunities"는 암 분야 연구자의 입장에서 자신의 연구에 재정적 도움을 줄 수 있는 기회를 발견하는 데에 탐색 목적을 두고, 직접 기금 지원을 알리는 내용의 문서 뿐만 아니라 이러한 재정적 지원이 가능한 재단이나 연구소, 기타 기관에 대한 소개 내용을 담고 있는 사이트도 모두 적합한 것으로 간주하였다. 그러나 구체적으로 암연구를 지원한다는 것이 반드시 명시되어 있어야만 한다. 그리고 기한이 많이 지난 내용이나 이미 어떤 기금으로 수행된 프로젝트의 연구 결과에 대한 문서는 부적합한 것으로 간주하지만 연결되어 있는 하이퍼링크를 따라간 결과 새로운 기금 지원에 대한 내용이 나오는 경우에는 적합한 것으로 간주하였다. 단, 하이퍼링크를 따라 가는 것은 3번으로 제한하였다.

두번째 탐색질문인 "XI International conference on AIDS"에 대해서는 에이즈에 관한 제11회 국제학술회의에 관련된 내용은 그 정보 유형에 관계없이 모두 적합한 것으로 간주하였다. 여기에는 이 학술회의의 홈페이지를 비롯하여, 뉴스레터, 발표된 각 논문에 대한 초록 및 핸드북 등이 포함된다. 하이퍼링크만을 제공하는 웹문서가 검색되었을 경우 링크를 따라간 결과 사이트의 주된 내용이 이 회의에 관한 것이면 해당 웹문서는 적합한 것으로 간주한다. 그러나 주된 내용이 이 회의에 대한 것일 때에만 적합한 것으로 판단하였고, 탐색결과 출력된 리스트의 표제에는 " XI International conference on AIDS"라는 문구가 포함되어 있으나 실제 링크를 따라가 본 결과 이 회의에 대한 내용이 사이트의 주된 내용이 아닌 경우나 이 회의에 대한 내용이 거의 포함되어 있지 않은 경우에는 모두 부적합한 것으로 간주하였다. 첫번째 탐색질문과 마찬가지로 하이퍼링크를 따라 가는 것은 3번으로 제한하였다.

적합성 평가의 대상은 각 탐색도구가 검색해낸 문서들 가운데 적합성 순위가 상위인 15개의 문서들로 제한하였다. 그 이유는 첫째, 이미 다른 연구(Tillman, 1995)에서도 지적된 바와 같이 순위가 15-20위 이하로 내려가면 질문과 관련있는 적합문서의 수가 급격하게 감소하기 때문이며, 둘째, 적합문서가 탐색결과 화면 한두 개 이내에 출력되지 않는다는 것은 탐색도구의 적합성 순위 결정방식에 문제가 있는 것이며 이는 이미 해당 탐색도구의 성능이 우수하지 않음을 반영하기 때문이다.

5. 2 탐색도구별 탐색문 작성

본 연구에서는 웹문서를 대상으로 하여 각 탐색도구의 효율을 평가하므로 탐색대상을 모두 웹으로 한정하여 실험을 수행하였다.

(1) Alta Vista

'Advanced Query'를 사용하였으며 출력양식은 상세형식을 선택하였다. 탐색질문 1에서는 날짜를 1996년 1월 1일 이후로 한정하였다.

〈탐색문 1〉

① *Selection Criteria: cancer and research and (grant* or fund*)*

② *Results Ranking Criteria: cancer research grant* fund**

〈탐색문 2〉

① *Selection Criteria: (xi or 11th or eleventh) and international and conference and aids*

② *Results Ranking Criteria: xi international conference aids*

(2) Excite

〈탐색문 1〉

Cancer AND research AND (grant OR fund)

〈탐색문 2〉

(XI OR 11th OR eleventh) AND international AND conference AND AIDS

(3) HotBot

불논리 연산식을 이용하여 탐색문을 구성하고 출력양식은 'full description'을 선택하였으며 탐색질문 1에서는 날짜를 1996년 1월 1일 이후로 한정하였다.

〈탐색문 1〉

cancer and research and (grant or fund)

〈탐색문 2〉

(xi or 11th or eleventh) and international and conference and aids

(4) InfoSeek

Ultraseek를 이용한다.

〈 탐색문 1 〉

+cancer +research grant fund

〈 탐색문 2 〉

xi +"international conference" +aids

(5) Lycos

탐색질문 1의 경우 탐색옵션은 'match 3 terms', 'good match'를 지정하고 출력옵션은 '20 results per page', 'detailed results'를 지정하여 수행한다. 탐색질문 2의 경우는 탐색질문 1과 같은 옵션으로는 검색건수가 너무 적었으므로 15개 이상의 검색결과를 얻기위해 탐색옵션은 'match any term', 'loose match'를 사용하였고 출력옵션은 탐색질문 1과 같다.

〈탐색문 1〉

cancer research grant$ fund$

〈탐색문 2〉

xi 11th eleventh international conference aids

(6) Magellan

탐색대상을 'entire database'로 하며 minimum rating은 'all matching sites'로 지정해 한정하지 않았다. 그리고 탐색질문 1은 OR 연산, 탐색질문 2는 AND 연산을 수행하였다.

〈탐색문 1〉
+cancer +research grant fund
〈 탐색문 2 〉
xi international conference aids

(7) Open Text Index

'Power Search'를 이용하였으며 Within옵션을 'Anywhere'로 지정하여 탐색을 수행하였다.

〈탐색문 1〉
grant Or fund And cancer And Research
〈탐색문 2〉
xi Or 11th Or eleventh And international And conference And aids

(8) WebCrawler

탐색질문 2의 경우에 '(xi or 11th or eleventh) and international and conference and aids'를 사용하면 엉뚱한 결과를 초래하였기 때문에 '11th'과 'eleventh'는 탐색어에서 제외시켰다.

〈탐색문 1〉
cancer and research and (grant or fund*)*
〈탐색문 2〉
xi and international and conference and aids

(9) Yahoo!

탐색방법 옵션은 'Matches on any word(OR)'로, 탐색대상 옵션은 'All of the Above'로, 한번에 출력하는 검색결과의 건수는 20건으로 지정하여 탐색을 수행하였다. 그리고 다른 탐색도구들과의 비교를 위해 웹사이트에 대한 검색결과만을 대상으로 하여 15건을 선정하였다. 탐색질문 1에 대해서는 날짜를 최근 6개월로 한정하였다.

〈탐색문 1〉
+cancer research grant fund**
〈탐색문 2〉
xi 11th eleventh international +conference +aids

6. 탐색실험 결과 분석

순위									
1		94	99	70	100		4093	86	
2		90	98	70	73		4455	85	
3		89	98	70	69		4229	84	
4		89	98	70	67		3630	84	
5		89	98	70	66		3630	84	
6		89	98	70	65		3541	84	
7		89	98	69	64		3318	84	
8		89	97	68	64		3314	83	
9		88	97	68	63		3230	83	
10		88	97	68	63		3089	83	
11		88	97	68	60		3048	83	
12		88	97	67	54		3048	83	
13		88	97	66	53		3031	83	
14		88	97	66	52		2954	83	
15		88	97	66	52		2753	82	
	Alta Vista (6)	Excite (4)	HotBot (9)	InfoSeek (3)	Lycos (6)	Magellan (5)	Open Text Index(8)	Web Crawler(3)	Yahoo! (5)

적합문서, ( ) 안은 적합문서 총수

그림 3. 탐색질문 1에 대한 적합문서의 분포와 적합성 점수

순위									
1		91	99	84	100		5326	96	
2		91	99	83	100		5020	95	
3		91	98	83	95		4722	95	
4		91	97	83	94		4510	92	
5		91	96	83	92		4408	90	
6		91	96	83	91		3133	90	
7		90	95	83	81		3024	90	
8		90	95	83	45		2835	89	
9		89	95	83	30		2548	89	
10		89	95	82	30		2236	89	
11		89	95	81	30		2132	89	
12		89	95	81	30		2132	89	
13		89	95	81	30		2131	88	
14		89	95	81	30		2048	88	
15		89	95	79	30		1925	87	
	Alta Vista (15)	Excite (7)	HotBot (15)	InfoSeek (14)	Lycos (5)	Magellan (4)	Open Text Index(11)	Web Crawler(4)	Yahoo! (2)

적합문서, ( ) 안은 적합문서 총수

그림 4. 탐색질문 2에 대한 적합문서의 분포와 적합성 점수

6. 1 적합문서의 분포와 적합성 점수

그림 3과 그림 4는 각 탐색질문에 대해 탐색도구별로 검색된 적합문서들의 분포 및 각 문서의 적합성 점수를 보여준다. 적합성 점수는 출력결과에 점수가 포함되는 탐색도구만을 비교하였으며 Open Text Index를 제외한 나머지 탐색도구들의 적

표 1. 탐색질문 1에 대한 정확률 및 재현율

척도 \ 탐색도구	AltaVista	Excite	HotBot	InfoSeek	Lycos	Magellan	Open Text Index	Web Crawler	Yahoo!
정확률	.4667	.2667	.6	.2	.4	.3333	.5333	.2	.3333
재현률	.1277	.0851	.1702	.0638	.1277	.1064	.1489	.0638	.1064

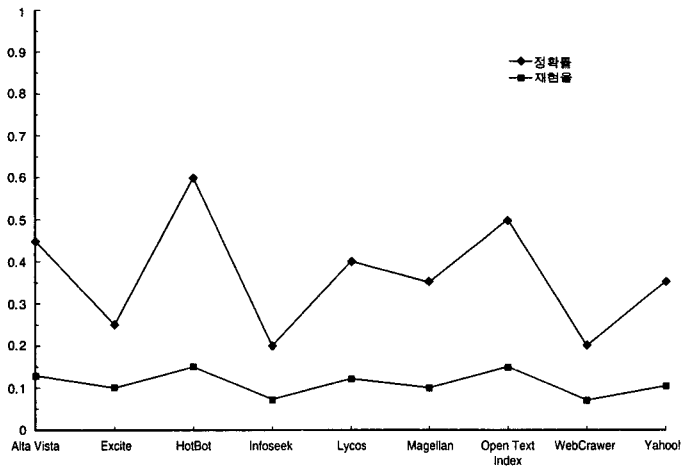

그림 5. 탐색질문 1에 대한 정확률 및 재현율

합성 점수는 백분율 값이다. 탐색질문 1에 대해 가장 많은 수의 적합문서를 검색해낸 탐색도구는 9건을 검색한 HotBot이었으며 상위에 집중적으로 적합문서를 검색해 낸 탐색도구는 HotBot과 Alta Vista로 나타났다. 그리고 다른 탐색도구들은 그 적합성 점수의 변화 폭이 작은데 비해 Lycos나 Open Text Index는 두 질문 모두에서 점수가 큰 폭으로 하락하고 있음을 볼 수 있다. 탐색질문 2에 대해서는 Alta Vista와 HotBot, InfoSeek 등에서 상위 15건 대부분이 적합한 문서인 반면에 Yahoo!의 경우에는 적합문서 수가 극히 적었다. 탐색질문 2의 경우 대부분 상위권에서 적합문서가 검색되었으나 Excite의 경우에는 그렇지 못하였다.

6. 2 검색효율 평가

각 탐색도구의 검색효율을 측정하기 위해 정확률과 재현율을 척도로 사용하였다. 정확률은 각 탐색도구가 검색한 적합문서의 수를 전체 검색건수로 나눈 값이며 이 실험에서 전체 검색건수는 15건으로 모두 일정하다. 이 실험에서 산출한 재현율은 상대재현율로서 9개의 탐색도구를 이 용

하여 검색한 적합문서 총수에 대한 각 탐색도구가 검색한 적합문서 수의 비율로 계산한다. 이때 검색된 적합문서 총수에서 중복되는 내용(파일명까지 정확히 일치하는 웹문서)은 제외한다. 실험결과 검색된 적합문서 총수는 탐색질문 1의 경우 47건, 탐색질문 2의 경우는 62건으로 나타났다. 그리고 재현율 계산시 하나의 탐색도구에서 정확히 일치하는 웹문서가 두번 이상 검색된 경우는 1건으로 간주하여 계산하였다.

탐색질문 1에 대해서는 표 1과 같이 모든 탐색도구들의 정확률과 재현율이 모두 높지 않은 것으로 나타났다. 정확률이 .5를 넘는 탐색도구는 HotBot과 Open Text Index로 나타났는데 각각 .6과 .5333의 정확률을 나타냈다. 재현율은 모든 탐색도구가 매우 저조한 것으로 나타나 탐색도구 간에 공통적으로 검색된 웹사이트가 적었음을 알 수 있다. 각 탐색도구들이 비슷한 성능을 보였 으며 가장 높은 재현율을 나타낸 탐색도구 HotBot은 .1702, 가장 낮은 재현율을 보인 InfoSeek와 WebCrawler는 .0638을 나타냈다. 그림 5는 표 1의 값을 그래프로 나타낸 것이다.

표 2. 탐색질문 2에 대한 정확률 및 재현율

척도 \ 탐색도구	AltaVista	Excite	HotBot	InfoSeek	Lycos	Magellan	Open Text Index	Web Crawler	Yahoo!
정확률	.1	.4667	.1	.9333	.3333	.2667	.7333	.2667	.1333
재현률	.2419	.4667	.2097	.2258	.0483	.0645	.1613	.0645	.0161

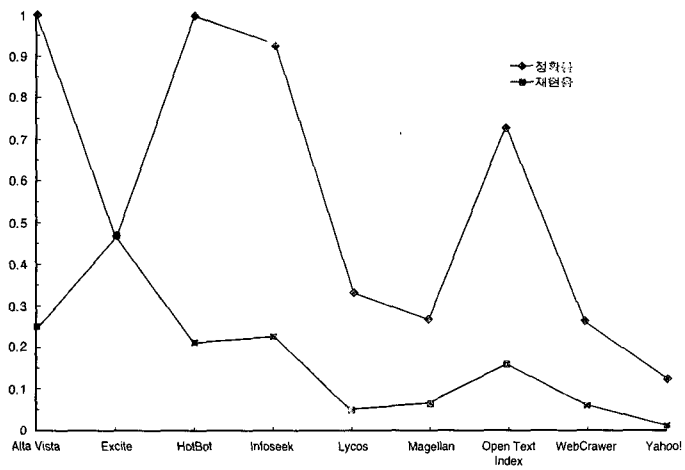

그림 6. 탐색질문 2에 대한 정확률 및 재현율

탐색질문 2는 표 2에서와 같이 탐색질문 1에 비해 정확률과 재현율 값이 각 탐색도구에 따라 큰 편차를 보이고 있다. 정확률의 경우 질문의 특성상 100%의 정확률을 보인 탐색도구가 Alta Vista, HotBot 두 개인 반면, 가장 낮은 정확률을 보인 Yahoo!는 .1333으로 매우 낮은 것으로 나타났다. 재현율은 Excite가 .4667로 가장 높았고 Lycos가 .0483으로 가장 낮았다. 그림 6은 표 2의 값을 그림으로 나타낸 것이다.

그림 7과 그림 8은 탐색질문의 특성에 따라 검색효율이 달라지는지를 알아 보기

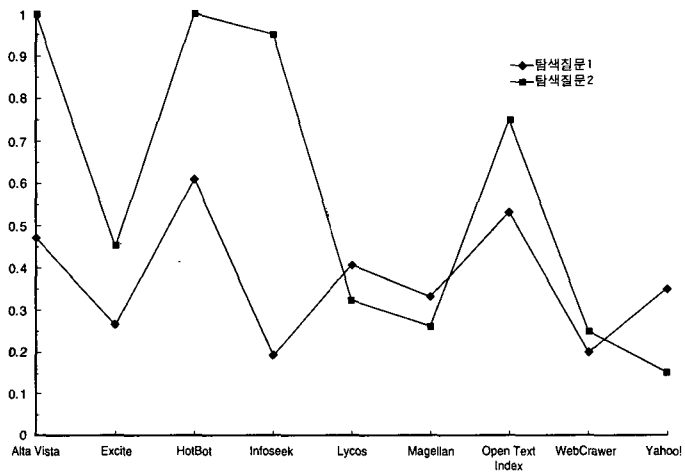

그림 7. 탐색질문 1과 탐색질문 2의 정확률 비교

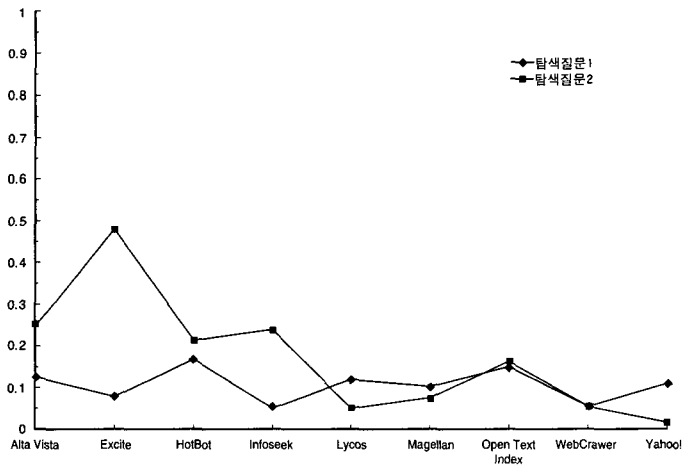

그림 8. 탐색질문 1과 탐색질문 2의 재현율 비교

위한 것이다. 그림 7은 탐색질문 1과 탐색질문 2의 정확률을 비교한 것으로서 대부분의 탐색도구가 질문 1에 비해 질문 2에서 훨씬 높은 정확률을 나타냈다. 이것은 탐색질문 2가 탐색질문 1에 비해 특정적이고 구체적이며 질문 자체가 사이트의 표제나 제목 부분에 나타날 수 있을 만큼 정보요구를 명확하게 표현하고 있다는 점에 기인하는 것으로 보인다. 9개 탐색도구의 평균정확률을 산출한 결과 탐색질문 1에 대한 값은 .3704, 탐색질문 2에 대한 값은 .5333으로 나타났다. 그림 8에서 보는 바와 같이 재현율 역시 대부분의 탐색도구가 질문 1보다는 질문 2에 대해 높거나 유사한 값을 보였으나 그 차이는 정확률에서만큼 크지 않았다. 그러나 탐색도구별로 볼 때는 정확률의 비교에서와 비슷한 패턴을 보여서 Lycos, Magellan, Yahoo!의 경우 질문 1의 재현율이 질문 2보다 오히려 높게 나타났다.

질문 1과 질문 2에 대한 탐색에서 모두 전체적으로 매우 낮은 재현율이 나타 났는데 이는 상대재현율을 이용해 비교를 하였으므로 각 탐색도구가 서로 상이한 검색결과를 냈다는 것을 의미하기도 한다. 탐색질문 1에 대한 평균재현율은 .1111, 탐색질문 2에 대한 값은 .1665였다.

6. 3 탐색도구간 유사성

각 탐색도구의 검색 패턴을 비교하기 위해 다이스계수를 이용한 유사도 행렬을 작성하였다. 다이스계수는 코사인계수와 함께 가장 많이 사용되는 유사계수로서 유사계수 S(X,Y) 를 구하는 공식은 다음과 같다.

$$S(X,Y) = \frac{2 \cdot |X \cap Y|}{|X| + |Y|}$$

(|X| :탐색도구 x가 검색한 문서 수,
|Y| :탐색도구 y가 검색한 문서 수,
$|X \cap Y|$:탐색도구 x와 y가 공통적으로 검색한 문서 수)

각 탐색도구들이 검색해 낸 사이트가 매우 다양하여 두 탐색도구간에 파일명까지 정확히 일치하는 사이트를 동시에 검색해낸 경우는 거의 없었기 때문에 유사계수를 계산하는데 있어서 전체 URL뿐만

	Alta Vista	Excite	HotBot	InfoSeek	Lycos	Magellan	Open Text Index	WebCrawler	Yahoo!
Alta Vista	1	.0556	0	.1111	.1111	.0952	.1587	0	0
Excite		1	0	.0588	0	.35	.1639	0	0
HotBot			1	0	0	0	0	0	.0164
InfoSeek				1	.0588	.1	0	0	0
Lycos					1	.05	0	0	0
Magellan						1	0	0	.1053
Open Text Index							1	0	0
WebCrawler								1	0
Yahoo!									1

그림 9. 탐색질문 1에 대한 유사도 행렬

아니라 도메인명을 기준으로 하여 같은 도메인에 속하는 문서를 검색한 경우도 X ∩Y 집합에 포함시켰다.

그러나 도메인명을 기준으로 하였기 때문에 하나의 탐색도구를 통해 같은 도메인의 여러파일이 검색되는 경우는 모두 하나로 취급하여야 하는 문제가 생긴다. 예를 들어 탐색질문 1에 대한 Excite와

	Alta Vista	Excite	HotBot	InfoSeek	Lycos	Magellan	Open Text Index	WebCrawler	Yahoo!
Alta Vista	1	0	.0147	.0444	0	0	0	0	0
Excite		1	.0588	.1429	.1429	.0455	0	.0357	.0714
HotBot			1	.16	.16	.0227	.7272	.2	.0889
InfoSeek				1	.5455	0	0	0	.2727
Lycos					1	0	0	0	0
Magellan						1	0	0	0
Open Text Index							1	.3636	.1852
WebCrawler								1	.0588
Yahoo!									1

그림 10. 탐색질문 2에 대한 유사도 행렬

Magellan의 유사계수를 계산하는 경우 먼저 Excite에서 9, 14위가 같은 도메인에서 검색되었고 Magellan은 각각 1, 2, 3위와 5, 12위가 각각 같은 도메인에서 검색되었음을 알 수 있다. 그리고 Excite와 Magellan에서 공통적으로 검색된 도메인은 두 개인데 하나는 각 탐색도구에서 한 번씩 검색되었고 다른 도메인은 Excite에서 2회, Magellan에서 4회씩 각각 검색되었다. 따라서 유사도 계산시 이원벡터가 아닌 가중치벡터에 적용되는 다이스계수 공식을 이용하였다.

탐색질문 1에 대해서는 질문이 특정적이지 않기 때문에 파일명까지 일치하는 URL을 검색해낸 경우는 2건뿐이었다. 그림 9는 탐색질문 1에 대한 각 탐색도구간 유사도를 행렬로 표현한 것이다. 0이 아닌 유사도를 보이는 탐색도구들도 그 유사계수 값이 극히 낮아 각 탐색도구들의 탐색 양상이 판이함을 알 수 있다. 가장 높은 유사도(.35)를 보인 탐색도구는 Magellan과 Excite였으며 평균유사도는 .0399이다.

탐색질문 2에 대해서는 그림 10에서와 같이 조금이라도 유사성을 나타낸 경우가 탐색질문 1보다 많았으며 .5 이상의 유사도를 보인 경우도 2건이나 나타났다. .7272의 가장 높은 유사도를 보인 탐색도구는 HotBot과 Open Text Index였으며 평균유사도는 .0928이다.

6. 4 중복탐색의 정도

탐색도구의 중복탐색은 두가지 의미를 갖는다. 첫째는 동일한 도메인이나 사이트에 질문에 관련된 여러개의 문서가 각기 다른 파일명으로 저장되어 있는 경우 각 파일을 검색하기 위해 동일한 도메인이 중복하여 탐색되는 것을 말한다. 둘째는

동일한 사이트나 문서가 오류에 의해 중복되어 검색되는 경우이다. 앞의 유사도 행렬 작성에서 이미 언급했듯이 각 탐색도구는 특정한 탐색질문에 대해 하나의 도메인 혹은 사이트에서 여러 관련파일들을 검색하는 경우가 빈번하였다. 동일한 도메인이나 사이트의 중복탐색은 로봇프로그램이 적절한 히스토리 파일을 관리하지 않는 경우에 발생하는 것으로 지적된 바 있다(McMurdo, 1995). 일반적으로 이용자가 하나의 도메인 혹은 상위 경로명에 접근하게 되면 제공되는 링크를 따라 여러 파일에의 접근이 가능하기 때문에 위와 같은 탐색방법은 효과적이지 못하다. 그리고 어떤 경우에는 같은 사이트나 문서임에도 불구하고 URL 표기의 차이 등으로 인해 두번 이상 검색되거나 또는 같은 URL인데도 두번 이상 검색되는 경우가 발생했다. 예를 들면 하나의 IP 주소에 대해 여러개의 DNS alias가 할당되어 있는 경우 로봇프로그램이 이를 인식하지 못하면 이러한 중복탐색 결과가 나타나게 되는 것이다(McMurdo, 1995). 웹 탐색은 대부분의 경우 그 검색결과가 매우 많은 것이 특징이므로 이러한 중복탐색은 탐색기능의 효율을 저하시키는 요인 중의 하나라고 할 수 있다. Hotbot의 경우에 이러한 탐색결과가 발생하면 같은 내용임을 표시해주는 'alternate' 지시기호를 달아 주고 있으나 검색결과의 순위에는 변화가 없기 때문에 효율성 제고라는 측면에서는 별로 효과적이지 못하다. 그림 11과 그림 12는 각 탐색질문에 대해 동일한 사이트나 도메인을 중복탐색한 양상을 탐색도구별로 비교한 것이다.

탐색질문 1에 대한 분석 결과 탐색의

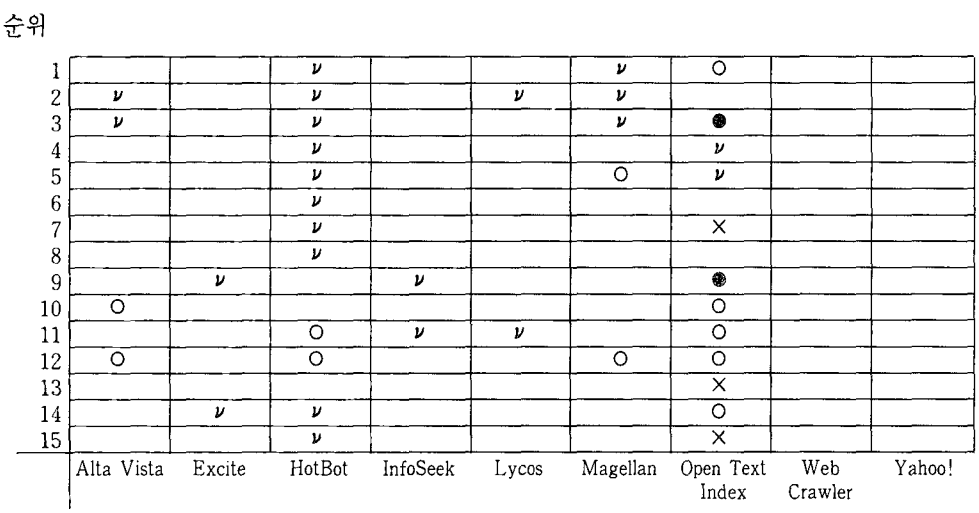

그림 11. 탐색질문 1에 대한 중복탐색

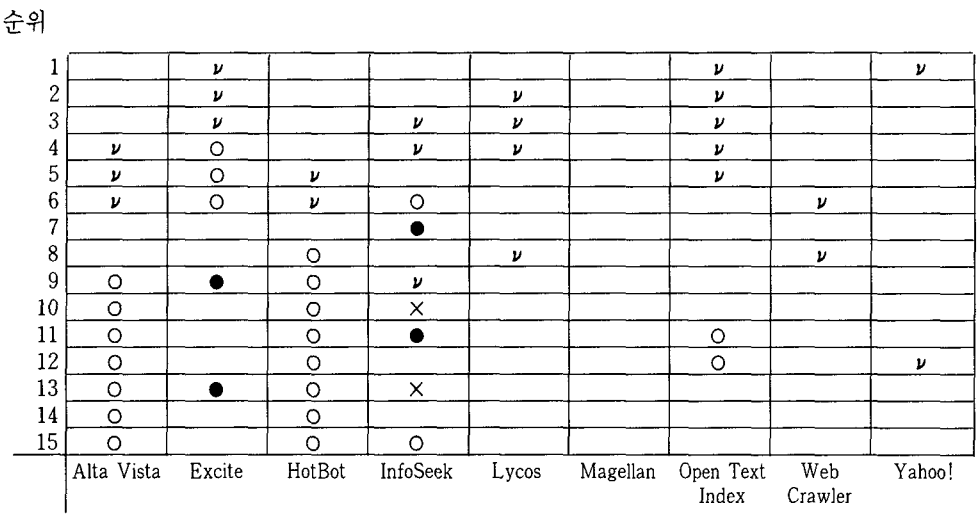

그림 12. 탐색질문 2에 대한 중복탐색

중복도가 가장 심한 탐색도구는 HotBot인 것으로 나타 났다. 15건의 검색결과 중 10건이 동일한 도메인에서 검색되었다. Open Text Index의 경우에도 4개의 도메인으로부터 대부분의 파일들이 검색되었음을 알 수 있다. 반면 WebCrawler와 Yahoo!는 15건 모두가 각기 다른 도메인으로부터 검색되었다. WebCrawler와 Yahoo!의 이러한 현상은 이미 색인 데이터베이스 부분에서 언급했듯이 파일단위가 아닌 사이트단위로 데이터베이스를 구축하기 때문인 것으로 보인다.

동일한 문서를 두번 이상 검색한 경우는 Alta Vista, HotBot, Open Text Index에서 나타났는데, Alta Vista의 경우 10위와 12로 검색된 사이트가 동일한 표제를 갖고 있었고, Open Text Index에서는 3위와 9위로 검색된 사이트가 마지막 파일명만 다를 뿐 내용이 정확히 일치하였으며 4위와 5위는 파일명의 대소문자에만 차이가 있을뿐 동일한 문서를 가리키는 것으로 나타났다. HotBot의 경우에는 11위로 "Welcome to the Cancer Research Institute's World Wide Web Home Page!"라는 표제를 갖는 사이트가 http://www.cancerresearch.org/라는 URL로 검색되었는데 이와 동일한 문서를 가리키는 또 다른 URL인 http://www.cancerresearch.org/index.html이 12 위에 'alternate'로 표시되어 검색되었다.

탐색질문 2에 대한 분석 결과 Alta Vista, HotBot에서 가장 심한 중복탐색 결과가 나타났는데 Alta Vista는 총 15건 중 7건이, HotBot은 8건이 같은 도메인에서 검색되었다. 그밖에 Open Text Index는 15건 중 5건이 같은 도메인에서 검색되는 중복도를 나타냈고 Exctie, InfoSeek 등도 상당한 중복도를 보여 주고 있다. 반면

에 Magellan은 모두 다른 도메인에서 파일들이 검색되었고 WebCrawler와 Yahoo!도 중복도가 낮은 것으로 나타났다.

동일한 문서를 두 번 이상 검색한 경우는 Excite, HotBot, Lycos, Open Text Index, Yahoo! 에서 나타났다. Excite에서는 1, 2, 3위가 모두 같은 문서였으나 요약부분에서 기술하고 있는 내용이 각기 달랐으며 URL도 찾지 못했다("Not Found"). 4~6위 사이트도 URL이 모두 같았으나 찾지 못하는 경우였다. HotBot에서는 5위의 "XI International Conference on AIDS Welcome Page"가 http://www.interchg.ubc.ca/aids11/의 URL로 검색되었는데 이에 대한 'alternate'로 6위에 http://www. interchg.ubc.ca/aids11/index.html의 URL이 제시되었으며, 14위로 검색된 웹문서인 "Special service - XI international conference on aids"도 14위에서는 URL이 http://www.ansa.it/aids96/daily/00000217.HTM으로 검색되었으나 'alternate'로 검색된 15위에서는 http://www. ansa.it/aids96/commun/00000218.HTM의 URL로 검색되었음을 알 수 있었다. 그밖에 Lycos에서는 2, 3, 4위로 검색된 문서들의 내용은 모두 동일하나 URL 표기 상에 약간의 차이가 있었으며, Open Text Index에서는 파일명의 확장자만 .HTM과 .htm으로 다른 경우였다. 그리고 Yahoo!의 경우에는 http://www. interchg.ubc.ca/aids11/aids96.html의 URL을 갖는 웹문서인 "XI International Conference on AIDS"가 1위와 12위로 검색되었다.

7. 결 론

본 연구에서는 주요 탐색도구들의 색인, 탐색, 적합성 순위부여 기능을 비교한 후 실제 탐색 실험을 통해 각 탐색도구의 검색효율과 중복탐색도를 측정함으로써 이들의 탐색성능을 평가하였다. 탐색실험 결과 탐색질문의 유형에 관계없이 Alta Vista, HotBot, Open Text Index가 비교적 좋은 검색효율을 보였으나 대부분의 탐색도구가 질문의 성격과 작성된 탐색문에 따라 탐색결과에 있어 많은 차이를 보이는 것으로 나타났다. 그리고 재현율에 있어서는 그 값이 전체적으로 매우 낮게 나타났는데 척도가 상대재현율이었음을 감안할 때 각 탐색도구가 서로 상이한 정보를 검색해 냈다는 것을 알 수 있다. 이것은 다이스계수를 이용한 유사도행렬을 통해서도 확인되었는데 유사성의 기준을 도메인명으로 완화했음에도 불구하고 각 탐색도구간 유사도가 매우 낮았으며 탐색질문에 따라서도 탐색도구간의 분포가 다른 것으로 나타났다.

동일한 도메인이나 사이트의 중복탐색에 관한 분석에서는 로봇프로그램이 색인데이터베이스를 구축할 때 파일단위(depth-first)가 아닌 사이트 혹은 서버 단위(breadth-first) 방식을 사용하는 WebCrawler나 Yahoo!와 같은 탐색도구

가 다른 탐색도구들에 비해 낮은 중복탐색도를 보였다. 또한 오류 등에 의한 동일한 문서의 중복탐색도 여러 경우가 발견되었다.

실험 결과에서도 나타났듯이 각 탐색도구들은 각각 중점을 두고 있는 서비스가 다르고 색인 데이터베이스의 구성 방법 등에 차이가 있기 때문에, 탐색자는 자신의 정보요구에 가장 적합한 탐색도구들을 선택하여 가장 효과적인 탐색을 수행할 필요가 있다. 대부분의 탐색도구에서는 이를 위해 다양한 방법으로 도움말을 제공하고 있으나 이들은 대부분이 탐색 초보자들을 위한 것이 었고 사서나 정보검색사들에게 필요한 정보는 부족한 경우가 많았다. 그리고 데이터베이스 규모에 대한 설명이나 이용횟수 등에 대한 내용은 다른 탐색도구와의 경쟁을 의식해서인 듯 확실하지 않은 정보를 제공하는 경우도 발견할 수 있었다. 따라서 각 탐색도구에 대한 정확하고 충분한 정보제공이 요구되며, 이용자는 색인 데이터베이스가 자신의 탐색목적에 적합한지, 탐색엔진은 어떤 탐색자를 위해 만들어진 것인지, 그리고 데이터베이스나 탐색엔진이 어떠한 방향으로 어떠한 방식에 의해 개선되고 있는지 등을 기준으로 하여 적절한 탐색도구를 선택하고 자신의 정보요구에 적합한 탐색을 수행하여야 할 것이다. 현재까지는 하나의 탐색도구가 완벽한 서비스를 제공하지 못하고 있기 때문에 여러 탐색도구들을 함께 사용하는 것이 바람직하다고 할 수 있다. Nicholson(1996)은 Alta Vista의 데이터베이스와 Excite의 색인기법, Open Text Index의 탐색엔진, Yahoo!의 디렉토리 구조, Magellan과 Lycos의 출력양식이 결합하면 이상적인 인터넷 탐색도구가 될 것으로 보았다. 또 이러한 맥락의 연구로서 워싱턴대학에서 실험적으로 개발하고 있는 탐색도구인 MetaCrawler를 들 수 있다(Ross, 1995). MetaCrawler는 입력된 탐색요구에 대해 Lycos, WebCrawler, InfoSeek, Open Text Index, Yahoo!, Galaxy 등의 6개 탐색도구를 동시에 실행시킴으로써 각 탐색도구들의 장점을 반영한 검색 결과를 얻을 수 있도록 하고 있다.

또한 인터넷 정보의 양이 많아지면서 그 처리 과정에 사람의 개입이 점점 힘들어지므로 좀더 지능적인 로봇프로그램을 만들어야 할 필요성이 대두되고 있다. 특히 색인대상이 될 인터넷 정보 자원의 선정, 효과적인 색인, 웹문서의 요약, 웹사이트/문서의 분류 등의 작업을 보다 지능적으로 처리할 필요가 있다. 그러나 웹로봇이라는 것이 일종의 프로그램이기 때문에 그 코드가 잘못 구현되어 있거나 예상치 못한 네트워크 환경에서의 오동작으로 인해 정보검색에 혼란을 가져올 수도 있으며, 로봇프로그램이 더욱 지능화된다고 하더라도 네트워크 자원과 서버의 부하가 커지고 데이터베이스 갱신에 따른 오버헤드 역시 증가하기 때문에 ALIWEB이나 Harvest와 같이 데이터베이스 구축시 로봇프로그램과는 다른 성격의 방법을 사용하기도 한다(Koster, 1995).

인터넷 정보자원과 이의 활용이 급증하고 있는 현시점에서 절실하게 요구되는 것은 효율적이며 동시에 효과적인 탐색도구의 등장이라고 할 수 있다. 이를 위해 기존의 탐색도구들에 대한 다각적인 평가를 통해 탐색도구의 설계자들에게는 이상적인 탐색도구의 모형을 제시하고 이용자들에게는 최적의 탐색도구 선택을 위한 기준을 제시할 수 있는 연구가 계속되어야 할 것이다.

참고문헌

Brandt, D. Scott. 1996. "Relevancy and Searching the Internet." *Computers in Libraries* 16(8): 35-39.

Courtois, Martin P., et al. 1995. "Cool Tools for Searching the Web: A Performance Evaluation." *Online* 19(6): 14-32.

Courtois, Martin P. 1996. "Cool Tools for Web Searching: An Update." *Online* 20(3): 29-36.

Halttunen, Kai. 1996. "Alta Vista by DIgital - Search Services: Analytical Form." http://www.uta.fi/~likaha/desire/altavista3.htm.

Halttunen, Kai. 1996. "Excite - Search Services: Analytical Form." http://www.uta.fi/~likaha/desire/excite.htm.

Halttunen, Kai. 1996. "Hotbot - Search Services: Analytical Form." http://www.uta.fi/~likaha/desire/hotbot.htm.

Halttunen, Kai. 1996. "Open Text Index - Search Services: Analytical Form." http://www.uta.fi/~likaha/desire/opentext3.htm.

Kimmel, Stacey. 1996. "Robot-generated Databases on the World Wide Web." *Database* 19(1): 40-49.

Koster, Martijn. 1996. "Robots in the Web: Threat or Treat?." http://info.webcrawler.com/mak/projects/robots/threat-or-treat.htm.

McMurdo, George 1995. "How the Internet was Indexed." *Journal of Information Science* 21(6): 479-489.

Nicholson, Scott. 1996. "Indexing and Abstracting on the World Wide Web: An Examination of Six Web Databases." http://www.telepath.com/santa/lis/iapaper.html.

Notess, Greg R. 1996. "Searching the Web with Alta Vista." *Database* 19(3): 86-88.

Ross, Philip E. and Hutheesing, Nikhil. 1995. "Along Came the Spiders." *Forbes* October 23: 210-216.

Singh, Amarendra, et al. 1996. "All-Out Search." *PC Magazine Online*, http://www.pcmagazine.com/iu/srchsite/__open.htm.

Sullivan, Danny. 1997. " A Webmaster's Guide to Search Engines and Directories." http://calafia.com/webmasters/.

Suoniemi, Anne. 1996. " InfoSeek Guide, InfoSeek Professional - Outline Search Services: Analytical Form." http://www.uta.fi/tmansu/infoseek.htm.

Suoniemi, Anne. 1996. "Lycos, The Catalog of the Internet - Outline Search Services: Analytical Form." http://www.uta.fi/tmansu/lycos.htm.

Suoniemi, Anne. 1996. " Magellan Internet Guide - Outline Search Services: Analytical Form." http://www.uta.fi/tmansu/magellan.htm.

Tillman, H. N., ed. 1995. Internet Tools for the Profession: A Guide for Special Librarians. Washington, D. C.: Special Libraries Association.

Zorn, Peggy, et al. 1996. "Advanced Web Searching: Tricks of the Trade." *Online* 20(3): 14-28.