네트웍 데이터베이스에서의 주제별 디렉토리와 키워드 검색엔진의 검색효율에 관한 탐색적 연구

An Exploratory Study of Performances between a Subject Directory and Keyword Search Engine in the Network Databases

이 명 희(Myeong-Hee Lee)

목 차

초 록

본 연구는 주제별 디렉토리인 Yahoo와 키워드 검색엔진인 Alta Vista가 대학도서관 이용자들에 의해 제기된 탐색질문에 대해 얼마나 적합한 문헌을 탐색해 내는지 알아보기 위하여 탐색적 연구의 형태로 진행되었다. 탐색결과는 검색된 문헌의 양, 검색된 적합문헌의 양, 재현율, 정확률의 측정기준에 의해 평가되었다. 특히 Alta Vista는 특정적이고 전문적인 용어의 탐색에 적합한 반면 Yahoo는 일반적이며 추상적인 용어의 탐색에 적합한 것으로 드러났다.

ABSTRACT

The study measured whether two search engines retrieve different Web documents for 6 queries. Two different search engines, Alta Vista in terms of keyword search engines and Yahoo in terms of subject directory engines were measured using as criteria, total number of documents retrieved, total number of relevant documents retrieved, recall and precision ratios. In addition, Alta Vista was suitable for specific and technical terms, while Yahoo was effective for general and plain terms. However, more elaborate research needs to be tested in terms of query characteristics.

상명대학교 문헌정보학과 전임강사
접수일자 1997년 5월 19일

1. 서 론

인터넷의 잠재력은 인터넷상의 정보자료의 급증과 더불어 이용자들이 큰 노력을 들이지 않고도 효율적이고 효과적인 정보접근이 가능하게 됨으로써 그 위력을 발휘하고 있다. 특히 인터넷상에 Web이 등장하게 됨으로써 정보제공은 몇몇 기관을 넘어서서 개인으로까지 확대되었으며 하이퍼텍스트와 하이퍼미디어 기법을 이용한 대량의 자료검색이 가능하게 되었다. Web의 가장 큰 특징은 하이퍼텍스트 기술을 사용하여 문자 뿐 아니라 그림, 음성, 비디오 등 멀티미디어 자료를 전송하게 되었고, 하이퍼링크의 기술을 통하여 전세계에 있는 컴퓨터에 거미줄같이 연결함으로써 전세계의 Web서버에 손쉽게 이동할 수 있다는 점이다. 도서관에서도 과거 참고사서나 열람사서에 의해 직접 제공되던 많은 참고질문들이 Web을 통하여 제공되고 때로는 최종이용자들이 Web에 직접 접근하는 등의 접근성을 높이게 되었다. 그러나 인터넷상의 정보는 아직까지 비체계적이며 무조직적이어서 이용자들이 원하는 정보를 정확하게 찾는 것은 상용 데이터베이스 검색에서보다 더욱 어렵고 과거 그 어느 때보다 정보량의 과다에 비해 많은 탐색의 실패를 초래하고 있다 (Janes and Rosenfeld, 1996).

최근 들어 이용자들이 원하는 정보를 손쉽게 찾을 수 있도록 도와주는 각종 검색엔진들이 개발되어 정보제공업무에 도움을 주고 있다. 검색엔진은 일종의 정보중개인 역할을 하는 탐색도구로서 수천개의 링크를 가진 강력한 온라인 데이터베이스이다. 여기서 제기되는 질문들은 특히 이들 검색엔진은 각각 어떤 특징을 가지고 있으며 검색업무 수행에 어떤 차이를 가지고 있는가? 그들의 데이터베이스 구축방법과 색인방법, 탐색능력, 출력형태, 검색효율 등은 어떠한가? 등이며 이들 질문에 답하기 위해 여러 연구가 시도되었다. 그러나 Web의 역사가 짧은 것과 마찬가지로 검색엔진의 성능을 평가한 문헌 또한 극히 초보적인 단계에 머무르고 있다.

다양한 검색엔진을 그들의 동작방식에 따라 또는 색인구축대상에 따라 나누면 주제별 디렉토리 검색엔진과 키워드 검색엔진으로 크게 구별된다. 또한 이들은 정보구축 범위에 의해 각각 초록 검색과 전문검색으로 나뉘기도 한다. 본 연구에서는 주제별 디렉토리의 대표적인 검색엔진인 Yahoo와 대표적인 키워드 검색엔진인 Alta Vista가 질문의 유형에 따라 어떻게 다른 종류의 정보를 검색해 내는지 밝혀봄으로써 검색엔진의 특성과 질문의 유형에 따른 정보제공에 도움을 주기 위하여 시도되었다.

2. 배경이론 및 선행연구 개관

검색엔진을 분류하는 기준이 공식적으로, 학문적으로 명확하게 정립된 것은 없다. 대부분 인터넷 이용자에 의해 구분되

는데, 무엇을 분류의 기준으로 삼는지에 따라 구분은 달라지며, 동일한 내용도 사람에 따라 각기 달리 불리우고 있다. 여기서는 주제별 디렉토리 검색엔진과 키워드 검색엔진과의 관계, 초록검색과 전문검색과의 관계를 알아보고, 검색엔진의 효율측정에 관해 현재 수행된 선행연구와 본 연구에서 실험대상으로 하는 검색엔진인 Yahoo와 Alta Vista의 특성에 관해 알아보고자 한다.

2. 1 주제별 디렉토리와 키워드 검색엔진의 특성

주제별 디렉토리 검색엔진은 특정 주제에 해당하는 각종 정보를 목록으로 제공하기 때문에 디렉토리 서버, 주제별 검색엔진, 주제 카탈로그, 메뉴검색, subject-oriented searching으로 불리며, 이들은 전통적인 도서관에서의 계층적인 분류개념과 유사한 의미를 가지고 있다. 최근에 몇몇의 주제 디렉토리와 비슷한 검색엔진들이 개발되어 자료의 주제접근을 가능하게 해주고 있다. Yahoo, The Argus Clearing House, The Whole Internet Catalog, Einet Galaxy, Magellan, The WWW Virtual Library등은 이 분야의 대표적인 검색엔진으로서 많은 양의 자료를 찾게 해주는 계층적인 주제 트리구조를 가지고 있다. 특히 The WWW Virtual Library는 미국회도서관 분류체계를 채택하고 있어서 같은 맥락에서 정보를 찾을 수 있게 도와주고 있다. 그러나 이들 검색엔진은 그들의 메뉴의 깊이나 다양성, 자료의 설명 등에 있어서 각각 다른 양상을 보여주고 있으며 제공된 이들 자료에 대한 거시적인 평가는 아직 이루어지지 않았고 이들 검색도구의 어떤 것도 대규모의 자료로 확대되지 않았다.

검색엔진의 또 다른 한 유형은 키워드 검색엔진으로서, Lycos, Alta Vista, HotBot, InfoSeek 등의 키워드 검색엔진은 키워드 검색을 주로 하고 있으며, 인터넷에 있는 홈페이지의 내용과 url을 자체 데이터베이스로 구축해 둔 것이다. 키워드는 하이퍼링크된 url, title tag, 본문자체의 내용으로서 키워드 검색엔진의 종류는 로봇의 정보수집 분류방식의 차이때문에 다양하다. 키워드 검색에서 중요한 것은 Web 사이트의 수(보유 데이타의 양), 출력량을 제한하는 filtering 기능, 컴퓨터 하드웨어 능력, 소프트웨어 능력으로서 이에 따라 검색데이터의 질에는 차이가 있다.

정보전문가들은 일관된 효율적인 접근을 제공하기 위해 개발된, 또는 복잡한 탐색전략을 가능케하는 도구를 사용하면서 정보탐색을 하는데 익숙해져 있다. 이러한 모습가운데 가장 중심이 되는 것은 불리안 탐색, 용어절단, 인접탐색, 타이틀, 저자 등의 특별한 필드에 탐색을 국한시키는 제한탐색 등이다. Web환경에서 전문텍스트를 가진 Web문헌에 대한 접근을 제공하는 검색엔진의 증가와 이용은 특수한 Web페이지나 Web사이트의 탐색을 어떤 면에서는 쉽게 만들었지만 또한 정보검색의 전문탐색에 관한 어려움을 야기하고

있다. 더구나 이들 시스템 중에서 불리안 탐색과 용어절단보다 정교한 테크닉을 사용하여 탐색하는 것을 허용하는 시스템은 거의 없으며, 상용 데이터베이스 검색시스템에서 이용할 수 있는 통제언어나 시소러스 등을 거의 가지고 있지 않다. 이러한 여건은 네트웍 환경에서 탐색을 더욱 정교하고 확고하게 하는 것을 어렵게 만든다(Janes & Rosenfeld, 1996).

키워드 검색에서의 디스크립션은 로봇에 의해 작성된 서술이므로 내용이 크게 중요하지 않으며 하이퍼링크는 사이트가 아니라 문장의 Web페이지를 보여 준다. 따라서 키워드 검색에서는 방대한 검색정보를 어떻게 줄이는가 하는 것이 중요하고 이를 위해 다양한 검색옵션이 발달하게 되었다. 키워드 검색엔진에서는 정보를 더욱 정확하게 찾아낼 수 있도록 AND, OR, 큰따옴표 등 여러 가지 검색옵션을 지정해 줄 수 있으며 검색결과에 대한 신뢰도 점수나 가중치를 보여주기도 한다. 신뢰도 점수나 가중치란 해당 검색결과가 얼마나 정확한지 검색시스템 자체에서 알려주는 점수로서 이들 점수가 높을수록 정확한 검색결과라고 각 검색엔진은 밝히고 있다.

2. 2 초록데이터베이스와 전문데이터베이스의 특성

우수한 초록은 문헌의 요점이 되는 주제내용의 기본 표현수단으로서 문헌의 실제적 대용물이다. 초록에 포함된 대부분의 단어는 주제내용을 정확하게 전달하여야 하며 우수한 초록작성자는 주제의 개념을 정확하게 나타내기 위하여 저자가 본문에서 직접 사용하지 않더라도 적절한 단어를 할당하기도 한다. 표제 및 부표제는 주제내용에 대한 중요한 단서를 제공하고 있으나 몇 가지 장애요소를 지니고 있다. 덜 학술적인 문헌이나 심지어는 연구논문에 있어서도 표제가 지나치게 불명료하거나 일반화되어 있기도 하고 자극적인 언어로 되어 있어서 표제가 취급된 주제를 제대로 표현하지 못하는 경우도 있다. 이러한 결함에도 불구하고 표제는 색인작성의 기본단위이며 주제내용을 판단하기 위하여 최초로 주의를 기울이는 곳이다.

통제어를 이용한 초록데이터베이스 탐색방식에서는 원문에 실려있지 않은 단어가 초록자에 의해 디스크립터로 사용될 수 있으며 특정의 용어도 일반적인 디스크립터로 전환되어 입력되거나 원문에서 사용된 원래의 용어 자체가 누락될 수 있기 때문에 탐색어 선정에 상당한 주의를 기울여야 하는 단점이 있지만 전문 데이터베이스는 원문의 모든 용어가 대부분 디스크립터가 될 수 있다는 장점을 가지고 있다. 그러나 접근점이 많으면 많을수록 그만큼 출력건수도 많아지기 때문에 전문데이터베이스의 탐색은 재현율이 높은 반면에 정확률이 떨어지며, 부적합 정보의 출력건수가 많으면 많을수록 이용자의 비용부담은 늘어날 뿐 아니라 적합정보의 선택에 있어서도 혼란을 일으키게 된다는 단점이 있다.

Tenopir(1985)는 전문 데이터베이스인 HBRO(Harvard Business Review Online)를 대상으로 31개의 탐색주제를 선정하였다. 정기간행물의 전문에서 추출한 주제어나 자연어를 사용한 온라인 탐색은 높은 재현율과 낮은 정확률을 나타낼 것이며, 이와 같은 방법을 사용한 탐색은 표제, 초록, 통제어 및 이들의 조합을 이용한 탐색보다 출력자료가 월등하게 많을 것이다라는 두 가설을 검증하였다. 전문, 초록, 통제어, 서명, 서지사항의 조합 및 서지사항과 전문의 조합으로 구성하여 검증한 결과 이들은 모두 입증된 것으로 나타났다. 그러나 그녀는 전문탐색은 비교적 정확한 탐색결과를 제공하지만 대부분의 경우에 있어서 전문탐색 한가지만으로는 정확률이 높은 정보를 제공하지 못하므로 전문과 통제어 뿐 아니라 초록을 함께 사용하는 탐색방법이 필요하다고 주장하고 있다.

남영준(1987)은 HBRO를 대상으로 전문 데이터베이스 탐색방식과 서지데이터베이스 탐색방식을 데이터 항목으로 선정하여 두 탐색방식에서의 재현율, 정확률 및 비용대 효과를 비교하였다. 이 연구에서 그는 표제는 탐색의 접근점으로 충분치 못하며 탐색범위를 전문까지 확대할 경우에는 표제, 초록 및 통제어로 범위를 정했을 경우에 비하여 출력건수와 적합정보의 수가 3배 이상 증가한 것을 발견하였다. 그리고 전문 데이터베이스 탐색방식이 서지데이터베이스 탐색방식에 비하여 재현율은 높고 정확률은 낮게 나타났으나 비용적인 측면을 고려하면 서지데이터베이스의 탐색방식보다 경제성이 높다고 주장하였다.

Web에 있는 문서는 html이라는 특수한 형식으로 작성되어 있는데, 문서가 있는 위치를 의미하는 url, 해당문서의 이름을 의미하는 title, 문서의 앞부분임을 의미하는 header, 문서의 본문 등 다양한 부분으로 이루어져 있다. 검색엔진이 사용하고 있는 로봇은 이들 html문서의 내용 모두를 대상으로 정보를 색인화하고 데이터베이스를 구축하기도 하며 Web문서의 내용을 특징적으로 축약한 초록만을 참조하여 데이터베이스를 구축하기도 한다. 전자의 경우를 전문검색엔진이라 하고 후자의 경우를 초록형 검색엔진이라고 한다. Yahoo와 WWW Virtual Library는 초록을 대상으로 검색을 진행하는 초록형 검색엔진이며, Alta Vista, OpenText, InfoSeek 등 대부분의 검색엔진은 전문검색엔진에 해당한다. 초록형 검색엔진은 Web페이지들의 내용을 설명해 주는 간단한 초록을 색인화한 것으로, 초록의 내용, 제목, 주제별 카테고리의 사이트와 일치하는 키워드를 쉽고 빠르게 찾아주며 사용자가 원하는 내용에 근접하는 정보를 찾아준다는 장점이 있지만 Web페이지 내의 모든 단어를 색인화하는 전문검색엔진에 비해 검색가능한 키워드의 갯수가 적고 자기에게 필요한 내용이 들어있는데도 자칫 키워드로 선별되지 못한 단어를 사용하면 찾을 수 없다는 문제점도 있다. 또한 초록형 검색엔진은 해당 Web페이지의 중요한 키워

드나 초록을 색인화하므로 그 성격상 전문검색엔진이 제공하는 몇몇 연산자를 사용할 수 없다. 즉 전문검색엔진은 Web페이지의 모든 단어를 색인화하기 때문에 서로 가까운 거리나 몇몇 단어사이에 있는 정보를 찾게 해주는 기능, 특정 키워드의 바로 뒤에 또 다른 단어가 오는 것만을 찾아주는 기능을 제공하지만 초록형 검색엔진은 그 특성상 이러한 기능을 지원하지 못한다.

2. 3 선행연구 개관

현재 인터넷상에 약 300여개의 Web 검색엔진이 사용되고 있는데 이들은 1994년까지 존재하지 않았으며 이들을 취급하고 있는 문헌은 더욱 짧은 역사를 가지고 있다. 따라서 Web 검색엔진을 비교 평가한 연구는 소수에 불과하며 또한 이들이 기초연구의 성격을 가지고 있다기 보다는 기술적인 서술에 머무르고 있는 경우가 대부분이다(Shirky, 1995; Taubes, 1995).

Courtois et al.(1995)은 Yahoo, Lycos, Opentext, CUI 등을 포함한 10개의 검색엔진의 성능에 대해 3개의 질문을 가지고 비교하였다. 검색엔진의 유연성, 강력한 탐색 인터페이스, 빠른 응답시간의 측면에서 OpenText가 가장 효율적이며, 또한 초보이용자에게는 WebCrawler가 가장 쉬운 인터페이스를 제공한다고 주장하였다.

Leighton(1995)은 교과과정상 검색엔진의 성능을 검사하였는데, 사용된 평가기준은 정확률이었다. Leighton은 InfoSeek, Lycos, WWW Worm, WebCrawler등 4개의 검색엔진을 대상으로 대학도서관에서 실제 사용하고 있는 8개의 참고질문에 대해 성능을 측정하였다. 그는 Lycos와 InfoSeek는 비슷한 정도의 정확률을 가진데 비해 WebCrawler의 정확률은 극히 낮음을 발견하였다. 또한 WWW Worm은 높은 정확률을 가진 적어도 1-2개의 문헌을 검색해내므로 검색양은 적으나 히트율이 높은 문헌을 검색할 때에 WWW Worm이 적합하다고 주장하였다.

Kimmel(1996)은 WWW Worm, Lycos, Open Text등 7개의 검색엔진을 대상으로 단일 단어로 구성된 문장에 대한 이들의 성능을 테스트하였다. 테스트 결과 키워드 검색엔진에 있어서 Lycos가 가장 강력한 검색엔진이라고 주장하였다.

Leonard(1996)는 온라인 탐색과 서비스를 평가하는 전문회사인 C/NET의 Web 사이트에 대해 19개 검색엔진의 성능을 비교 연구하였다. 공공도서관에서 제기된 참고질문인 15개의 탐색질문에 대해 결과의 정확성, 사용의 편리성, 발달된 옵션의 제공여부에서 평가가 이루어졌다. 그 결과, Alta Vista의 검색효율성이 가장 높고, All-in-One Search Page와 Internet Sleuth가 통합검색엔진 중 최고의 검색효율을 나타낸다고 주장하였다.

Chu와 Rosenthal(1996)은 로봇에 의해 운영되는 검색엔진인 Alta Vista, Lycos,

Excite를 대상으로 해서 그들의 탐색능력(불리안 논리, 용어절단, 제한탐색, 단어와 구탐색)과 검색효율(정확률과 응답시간)을 측정하였다. 대학도서관의 참고업무중에 실제 발생된 10개의 질문을 대상으로 수행된 이 연구에서 적합성의 판정은 두 연구자 자신에 의해 이루어졌다. 그들은 Alta Vista와 Excite는 각 질문당 10개 이상의 문헌을 검색해 내었으나 Lycos는 수록범위가 가장 넓은데도 불구하고 어떤 질문에 대해서는 하나의 문헌도 검색해내지 못하는 것을 발견하였다. 탐색성능이나 검색효율에 있어서 Alta Vista가 가장 성능이 뛰어남을 밝히고 높은 정확률을 요구하는 이용자에게는 Alta Vista가 가장 권고할 만하다고 주장하였다. 또한 검출해낸 250여개의 논문중 오버랩되어 검출된 것은 거의 없었다고 지적하고 각 검색엔진의 구축은 거대한 Web 시스템의 각각 다른 지적구조를 나타낸다고 주장하였다. 그들은 추후의 Web 검색엔진의 평가방법으로 1) Web 색인의 구성(범위, 갱신빈도, 색인된 Web 페이지의 부분), 2) 탐색능력(불리안 연산자, 구탐색, 용어절단, 제한기능 등), 3) 검색효율(정확률, 재현율, 응답시간), 4) 결과의 선택여부 (결과 선택의 숫자와 결과의 내용), 5) 이용자 노력(인터페이스, 문헌의 구성내용)을 제안하였다.

Zorn(1996)은 Alta Vista, Lycos, InfoSeek, OpenText 등을 대상으로 복합탐색(advanced search)의 특성을 살펴보았다. 검색엔진의 복합탐색기능을 사용하여 3개의 탐색질문이 각각 탐색되었는데, 검색엔진의 색인과 평가과정은 시스템마다 다르기 때문에 어떤 단일의 검색엔진도 최고의 효율성을 가질 수 없다는 결론에 도달하였다. Alta Vista와 Lycos가 URL을 포함하였다는 견지에서 가장 포괄적인 색인을 구성하고 있었지만 Alta Vista, OpenText, InfoSeek가 적합성과 검색의 정확성에 있어서 높이 평가되었으며, 고도의 복합탐색기능과 보다 복잡한 기능에 대한 우수한 이용자 인터페이스와 documentation을 제공해 주는 점에서는 OpenText가 우수하다고 지적하였다. Web탐색의 특성상 정확성과 효율성을 높이기 위해서는 지나치게 많은 양의 탐색결과를 제한하기 위한 복합 불논리 구문, 사이트의 중복검색 감지능력, KWIC색인, 필드지정에 의한 제한탐색, 인접탐색 및 구탐색, 적합성 순위별 출력결과, 탐색결과 집합의 처리, 절단탐색 등의 발전된 기능이 제공되어야 한다고 주장하였다. 전문적인 탐색자가 좋은 탐색결과를 가지기 위해서는 온라인 상용 데이터베이스에서 한 탐색자가 여러 데이터베이스를 탐색하여 검색된 자료의 포괄성과 검색의 정확성을 유지하는 것처럼 Web탐색에서도 같은 과정이 필요하며 이를 위해서 전문탐색자는 인터넷 탐색서비스의 복합탐색기능의 특성에 익숙해야 한다고 제안하였다.

Pollack와 Hockley(1997)는 초보적인 최종이용자의 일상생활에서의 정보탐색과 그들의 인터넷 검색엔진 사용에 대한 접근방법 등을 탐색적 연구로 수행하였다.

최종이용자를 일상적으로 컴퓨터를 사용하는 집단과 전혀 컴퓨터 사용경험이 없는 두 집단으로 나누어 그들 자신의 정보탐색을 위해 Yahoo, Lycos, WebCrawler 등의 검색엔진을 사용하도록 요구하였다. 최종이용자들은 컴퓨터 사용경험의 유무에 상관없이 정보탐색에 많은 어려움을 가졌는데 그들은 인터넷 정보자료의 성격을 잘 이해하지 못하였으며 탐색용어의 선정 및 탐색식의 구성에 많은 어려움을 가진 것으로 드러났다. 또한 최종이용자들이 일반적으로 직면하는 어려움은 자연언어로 표현된 이용자의 요구를 컴퓨터 언어로 변환하는 데의 어려움, 탐색요구의 지나친 상세화 또는 생략, 스펠링 에러, 세계관에 대한 지식의 결여로 지적되었다. 이들은 현재 사용되고 있는 인터넷 검색엔진이 지나치게 어렵고 복잡하므로 초보이용자에게는 복합탐색(advanced search)보다 단순탐색(simple search)의 기능이 강화되어야 하며 이용자의 불분명한 또는 잘못된 요구에 지능적으로 대처하는 스펠링에러 체커 등과 같은 검색기능의 강화가 요청되었다. 그들은 또한 초보이용자들에게는 직접 탐색식을 작성하게 하는 것보다 계층적인 분류체계를 선택하게 하는 Yahoo가 가장 인기있는 검색엔진임을 밝혀내었다.

정영미와 김성은(1997)은 Alta Vista, Lycos, HotBot 등 9개의 검색엔진을 대상으로 그들의 색인 및 탐색기능과 검색된 문헌의 순위부여방법을 비교한 후 검색성능을 평가하였다. 탐색실험을 위해 Zorn(1996)등이 사용했던 탐색질문 가운데 2개를 사용하였으며, 각 검색엔진이 검색해낸 문헌가운데 적합성 순위가 상위 15개인 문헌으로 제한하여 문헌 자체에서 제공된 적합성 점수를 가지고 판단하였다. 평가기준으로는 검색효율과 중복검색의 정도를 측정하였으며 검색효율 척도로는 정확률과 상대 재현율을 사용하였다. 탐색실험 결과 탐색질문의 유형에 관계없이 Alta Vista, HotBot, OpenText가 비교적 좋은 검색효율을 보였으나 대부분의 검색엔진이 질문의 성격과 작성된 탐색문에 따라 탐색결과에 있어 많은 차이를 보인다는 것을 발견하였다. 각 검색엔진마다 재현율은 매우 낮은 것으로 나타났으며, 다이스 유사계수공식을 사용하여 검색엔진간의 유사도를 측정한 결과 탐색도구간 유사도는 매우 낮다는 것을 발견하였다.

이상에서 살펴본 것처럼 각 연구의 결과는 연구마다 다른 것으로 나타났으며 일치하는 것같지 않다. 또한 각 연구에서 사용된 방법론과 평가기준 또한 각각 달라서 Web 이용자들이 그들의 특수한 탐색요구에 사용할 적절한 검색엔진을 선택할 수 있도록 적절한 방법론을 개발하는 것이 필요하다. 이를 위해서는 현재 사용되고 있는 각 검색엔진의 독특한 특성과 성격을 파악하는 것이 필요하다.

2. 4 실험대상 데이터베이스의 특성

각 검색엔진이 구축·제공하는 색인데이터베이스의 성격은 제공하는 탐색기능

이나 서비스의 목적에 따라 그 구성요소나 특성 등에 차이가 있다. 특히 이 연구에서는 사이트의 선정방법과 선정기준, 색인방법의 차이 등에서 현격한 차이를 가지고 있는 주제별 디렉토리 검색엔진과 키워드 검색엔진의 대표인 Yahoo와 Alta Vista를 대상으로 그들의 색인데이터베이스의 구축범위, 색인방법, 데이터베이스 규모, 탐색기능, 출력형태 등 다양한 방면을 살펴보고자 한다.

2. 4. 1 Yahoo

Yahoo 데이터베이스는 1994년부터 서비스를 시작하여 14개 주제별 디렉토리와 초록형 검색을 제공하고 있다. Yahoo는 Web 사이트를 비롯하여 유즈넷 뉴스그룹, 전자우편 주소에 있는 내용을 대상으로 색인하여 데이터베이스를 구축한다. 각 Web 사이트에 대해서는 url, title, abstract의 내용등이 색인되며 현재 약 65,000여개의 항목이 데이터베이스에 포함되어 있다(Courtois et al, 1995). 원래 이러한 과정은 색인자에 의한 수작업과정으로 진행되었으며 각 초록은 저자 자신에 의해 만들어진 저자초록이다. 따라서 초록의 내용은 저자에 의해 주어진 것이므로 내용설명이 충실한 것으로 알려져 있다. 색인자는 전체 사이트를 검토하여 적절하다고 판단되면 사이트의 홈페이지만을 데이터베이스의 적절한 주제명 아래에 포함시킨다. 그러나 최근에 정보가 급증하게 됨에 따라 1995년 9월부터 로봇 프로그램을 등장시켜 키워드색인을 병행하고 있다.

Yahoo 데이터베이스의 내용은 수천가지에 이르는 상세한 주제별로 분류된 특징을 가지고 있다. 그리고 하나의 사이트 성격이 주제로 분류하기에 중복되는 부분이 있더라도 모두 Yahoo 시스템 내부적으로 연결하기 때문에 한층 정확한 분류가 가능하다. 자신이 원하는 사이트를 찾을 때 꼭 자신이 선정한 키워드를 포함하지 않더라도 성격상 해당되는 경우가 많이 있으며 주제별 분류는 이때 도움이 된다.

Yahoo에서 정보를 찾는 방법은 주제별 디렉토리와 키워드 검색의 두가지가 있는데, 주제별 디렉토리를 통한 검색은 찾고자 하는 내용을 명확하게 키워드로 뽑아내기 힘들거나 특정분야와 관련있는 것일 때 이용하면 좋은 것으로 알려져 있다. 찾고자 하는 정보가 어떤 분야에 속하는지를 판단한 다음 14가지 분야 중 한 곳을 클릭하고 이어서 나타나는 소분류 중 다시 하나를 선택하여 점차 검색의 범위를 좁혀가면 된다. Yahoo의 키워드는 분류카테고리 타이틀, 사이트의 타이틀, 사이트의 디스크립션 안의 내용들이다.

Yahoo의 키워드 검색은 세 곳에서 제공되는데 Yahoo의 홈페이지, 14개 목록의 소분류화면, Yahoo Search이다. Yahoo의 홈페이지와 14개 목록의 소분류화면에서 제공하는 키워드 검색기능은 단순히 "search"라고 부르며, Yahoo 홈페이지와 키워드 입력상자 오른쪽에 있는 "options" 항목을 클릭했을 때 나타나는 검색환경을 "Yahoo search"라고 부른다. 홈페이지에

있는 검색어 입력상자를 통해 검색하게 되면 Yahoo의 카테고리와 Yahoo 사이트의 데이터베이스에 있는 내용을 대상으로 정보를 찾게 된다. 이곳에서 검색을 할 때 여러개의 단어를 입력하면 이들이 모두 들어있는 정보 즉, AND 조건으로 검색을 진행한다. 검색결과의 출력은 Yahoo 카테고리, Yahoo 사이트, Alta Vista 등 세곳에서 검색결과를 출력해 준다. 소분류화면에서의 키워드 검색은 홈페이지 검색과 거의 동일하나 하나의 차이는 검색의 범위를 해당 소분류에 들어있는 정보만을 대상으로 할 것인지 Yahoo의 전체 데이터베이스를 대상으로 할 것인지 지정해 줄 수가 있다.

Yahoo의 홈페이지나 소분류화면에서 "options"를 클릭하면 좀더 자세하고 구체적인 검색형태를 지정할 수 있는 Yahoo Search가 나타난다. Yahoo Search에서 검색할 수 있는 대상은 Yahoo 데이터베이스, 유즈넷 뉴스, 전자우편 주소로서 Yahoo 데이터베이스는 Yahoo 카테고리와 Yahoo 사이트를 가리킨다. 유즈넷 검색은 데자뉴스의 협찬아래 유즈넷뉴스에 있는 기사내용을 검색할 수 있게 만든 것이며 전자우편 주소검색은 인터넷 전자우편 주소를 찾아주는 기능을 가지고 있다.

Yahoo Search에서 Web을 검색대상으로 설정한 후 정보를 찾게 되면 다음과 같이 세 개로 나누어진 검색결과가 출력된다; Yahoo 카테고리, Yahoo 사이트, Alta Vista Web페이지. Yahoo 카테고리에서는 Yahoo의 14개 분류목록 및 그에 속한 소분류의 제목에 있는 내용을 대상으로 하여 정보를 검색해 낸다. Yahoo 사이트는 Yahoo의 소분류화면에서 해당분야의 특정적인 사이트를 따로 모아 놓은 것으로 즉, 소분류 화면의 아래쪽에 위치한 내용중 Web 사이트의 제목과 그에 대한 설명을 50단어의 정도로 표시해 놓은 부분을 가리킨다. 또한 Yahoo 카테고리와 Yahoo 사이트에서 미처 검색해내지 못한 정보를 Alta Vista를 이용하여 검색한 결과를 보여준다. Yahoo Search에서 사용할 수 있는 검색옵션으로는 검색대상 정보의 기간, 각 키워드들의 연산조건, 검색된 결과가 출력된 갯수 등을 지정할 수 있다. 기간의 지정에 있어서 기본값은 3년이며, 하루, 일주일, 한달 등으로 변경할 수 있다. 불리안 연산자 AND나 OR의 조건을 지정해 줄 수 있는데, 검색어 입력상자 아래의 옵션에서 "Match on any word"는 공백을 주어서 입력한 키워드를 불리안 연산자의 OR 조건으로 검색하는 것이며 "Match on all words"는 AND 조건에 해당한다. "consider keys to be" 항목에는 "substring"과 "complete words"라는 두 개의 선택조건이 있는데 "substrings"는 입력한 키워드 중 일부분이라도 있는 것을 찾아주며 "complete words"는 완전하게 찾고자하는 단어가 있는 경우만 찾아준다. 또한 "Advanced search syntax" 항목에서는 "+"부호를 사용함으로써 AND 연산자와 같은 기능을, "-"부호를 사용함으로써 OR 연산자와 같은 기능을 수행한다.

2. 4. 2 Alta Vista

Alta Vista는 여러개의 로봇프로그램을 이용해 Web을 탐색하여 색인하는 프로그램으로서 1995년 12월 일반에게 공개된 이래 현재 476,000개의 서버에서 3,100여만개의 Web문서 전문과 14,000개 유즈넷 뉴스그룹의 400만개 뉴스기사 전문을 대상으로하여 색인 데이터베이스를 구축하고 있다. Web문서의 경우 브라우저에 의해 보여지는 텍스트의 전문을 색인할 뿐 아니라 html화일 전체도 색인해 준다. Alta Vista는 키워드형 검색엔진으로서 주제디렉토리는 제공하고 있지 않으며 전문검색을 대상으로 하며 검색대상은 Web과 유즈넷, 뉴스기사의 내용이다. 영어는 물론 한국어, 일어 등 2바이트 문자권의 언어로 작성된 정보까지 검색할 수 있으며 빠른 접속속도를 제공할 뿐 아니라 검색결과도 빠른 것으로 나타나고 있다(Zorn et al., 1996).

탐색에서 사용자의 기호와 수준에 따라 두가지의 검색환경인 단순검색(simple search)과 복합검색(advanced search)을 제공하고 있다. 단순검색은 복잡한 검색식을 사용하지 않고 간편하게 검색을 진행할 때 사용하는 기본형태의 검색으로서 여기서는 +, -같은 특수문자를 사용할 수 있으나 불리언 연산자는 사용할 수 없다. 그러나 복합검색은 AND, OR, NOT과 NEAR의 사용이 가능하다. 또한 복합검색은 특정한 기간내에 만들어진 정보만을 검색할 수 있도록 설정이 가능하며 출력에 있어서 특정한 단어를 포함하는 정보가 먼저 출력되도록 조정할 수 있는 우선순위 지정이 가능하나 단순검색에서는 이것이 불가능하며 적중도가 높은 순위에 따라 자동출력되도록 하고 있다.

단순검색과 복합검색에서 공동으로 사용할 수 있는 기본요소는 단어(words), 구(phrases), 큰 따옴표, 대문자, 악센트, "*" 등이다. 또한 Alta Vista의 검색에서는 Web페이지의 타이틀이나 유즈넷 뉴스그룹의 내용에서 정확한 정보를 찾아내기 위한 특수키워드를 사용할 수 있다. 이러한 키워드는 host, title, link, url, text 등 다양한데 이렇게 지정해 줌으로써 Alta Vista는 이곳에 있는 정보를 대상으로 검색을 진행한다. 예를 들면 "title:search"하면 Alta Vista는 제목에 "search"가 있는 것만 검색해 줌으로써 출력범위를 제한할 수 있다.

출력옵션의 지정은 compact form, standard form, detailed form 등 3개분야이며, 검색결과가 출력되는 주화면에서 검색을 위해 사용한 키워드, 각 키워드별 검색된 문헌 수, 자체 점수매기기 시스템에 의해 산출된 결과가 현재 페이지에 몇개 출력되었는지의 갯수, 개별 검색결과가 나타난다.

3. 연구계획

본 연구에서는 주제별 디렉토리 검색엔진인 Yahoo와 키워드 검색엔진인 Alta Vista가 주어진 탐색질문에 각각 얼마나

적합한 문헌을 검색해내는가를 알아보기 위하여 시도되었다. 실제 대학도서관 이용자들에 의해 제기된 질문에 대해 두 검색엔진이 얼마나 적합한 문헌을 검색하는지를 알아보기 위해 실험연구 방법이 진행되었으며 적합성의 판정은 각 질문의 제공자들에 의해 이루어졌다.

3. 1 변인의 설정

이 연구에서 사용된 독립변인은 검색엔진이며, 종속변인은 1) 검색된 문헌의 양, 2) 검색된 적합문헌의 양, 3) 재현율, 4) 정확률이다.

3. 1. 1 독립변인

독립변인은 검색엔진으로서, 여기서 사용된 검색엔진은 주제별 디렉토리 검색엔진인 Yahoo와 키워드 검색엔진인 Alta Vista이다.

3. 1. 2 종속변인

1) 검색된 문헌의 양

검색된 문헌의 양은 하나의 검색엔진이 최적의 효과를 거두기 위해 탐색식을 작성하여 검색해낸 문헌의 전체양을 의미한다.

2) 검색된 적합문헌의 양

검색된 적합문헌의 양은 탐색효율의 측정에 있어서 기본이 되는 가장 중요한 요소이다. 여기서는 각 검색엔진의 출력문헌 중 출력순위 20개로 제한된 상황에서 적합한 문헌으로 판단된 것을 검색된 적합문헌의 양으로 본다. 이것은 정확률과 재현율을 측정하기 위한 전제요소이다.

3) 재현율

실제 어느 검색엔진의 재현율을 측정하기 위해서는 특정의 탐색질문에 대한 전체 시스템 내에서의 적합한 문헌의 전체양을 알아야 하는데, 실제로 이것을 파악한다는 것은 불가능하다. 여기서는 재현율 대신에 상대 재현율을 사용하였다. 상대재현율은 두 검색엔진을 이용하여 출력한 적합문헌양에 대해 각 검색엔진이 출력한 적합문헌양의 비율로 계산한다.

4) 정확률

정확률은 한 검색엔진이 하나의 탐색질문에 대해 출력한 검색문헌 중, 출력순위 20건에 대해 적합하다고 판단된 검색문헌의 비율이다.

3. 2 탐색질문의 구성

탐색질문은 대학도서관에서 실제 이용자들에 의해 제기된 참고질문들이다. 이들은 최종이용자로서 다양한 유형의 질문을 제기하였다.

질문 1: 사회 내에서의 자원봉사자의 실태에 관한 자료를 제시하시오.

질문 2: 세계 통신업계에 관한 신기술 관련정보, 각국 통신정책, 대표적 통신사업체 등 통신분야에 관련된

정보를 찾으시오.

질문 3: 미국의 1996년도 동절기 기상 전망에 관한 자료를 찾으시오.

질문 4: 대만의 1996년도 무역관계 통계를 찾으시오.

질문 5:영국 런던대학 산하 교육연구소에 관한 안내정보를 찾으시오.

질문 6: NAFTA에 관한 정보를 찾으시오.

3. 3 탐색과정 및 탐색식 작성

Yahoo의 탐색과정에 있어서는 먼저 14개의 주제별 디렉토리를 선택하고 다시 소분류 중 하나를 선택하여 검색의 범위를 좁힌다. 소분류화면에서 "options"를 클릭하여 "Yahoo search"를 찾아내고 Yahoo Search에서 Web을 검색대상으로 설정하였다. 검색옵션 중 기간의 지정에 있어서 기본값 3년을 지정하고 불리안 연산자 AND 조건인 "Match on all words"를, "Consider keys to be"항목에서는 "complete words"를 선택하였다. 또한 "Advanced search syntax"에서 AND 연산자의 기능을 가진 "+"부호와 OR 연산자의 기능을 가진 "-"부호를 사용하여 탐색식을 구성하였다.

Alta Vista에서는 복합검색(advanced search)을 사용하여 AND, OR, NOT, NEAR 등의 불리안 연산자를 사용하였으며 출력시 특정단어의 우선순위지정을 시도하였다. 단어, 구, 큰따옴표, 대문자, 악센트, "*" 등을 사용하였고 특수키워드인 host, link, url, text 등을 지정하여 출력량을 줄이기를 시도하였다. 출력옵션의 지정은 "detailed form"을 사용하여 출력된 정보가 가능한 한 상세하도록 시도하였다. 본 연구에서는 탐색대상을 모두 Web으로 한정하여 실험을 수행하였다. 탐색식의 작성은 각 검색엔진의 특성에 비추어 가장 좋은 검색결과를 낼 수 있도록 적절한 탐색문과 기능을 이용하여 작성되었다.

질문 1 Alta Vista : volunteerism and society

Yahoo : Society and Culture 선택 : society +volunteerism

질문 2 Alta Vista : telecommunication and (resource* or policy)

Yahoo : Science : Engineering : Electrical Engineering : Telecommunications 선택 : telecommunication +policy

질문 3 Alta Vista : weather and forecast and annual and US

Yahoo : weather +forecast +1996

질문 4 Alta Vista : statistics and host:tw and government

Yahoo : taiwan +statistics +trade

질문 5 Alta Vista : "London University" and "Institute of Education"

Yahoo : Education: Universities and Colleges 선택 : university of london +institute of education

질문 6 Alta Vista : NAFTA

Yahoo : NAFTA

적합성이란 탐색결과가 이용자의 정보요구와 얼마나 관련이 있는가를 의미한다. 대부분의 Web의 검색엔진에서 보여주는 탐색결과는 수백, 수천건이 되므로 효과적인 적합성 판정을 위해서는 일정한 기준에 의해 순위를 부여하여 이용자에게 출력해 주는 것이 필요하다. 대부분의 검색엔진에서는 탐색어의 출현빈도나 출현위치, 텍스트의 길이 등을 이용해 탐색결과에 순위를 매겨 출력하고 있다. Alta Vista의 경우, Advanced query에서는 순위부여를 위해 가중치를 주고 있는데, 해당용어가 탐색식에 나타난 경우에는 그 용어가 출현한 문헌이 높은 순위를 가지고 출력되고, 가중치를 준 용어가 탐색식에 나타나지 않은 경우에는 이 용어를 포함하지 않은 문헌은 제거해 버리고 나머지 문헌을 순위에 따라 출력한다. 그러나 가중치를 주지 않으면 순위없이 무작위로 추출한다. Yahoo에서 적합성 순위를 결정하는 요인으로는 일치하는 키워드 수가 많을수록, 탐색어가 url이나 제목에 출현하는 경우, 그리고 주제카테고리가 Yahoo 주제트리의 윗쪽에 위치할수록 해당 Web 문헌은 높은 순위로 검색된다. 탐색대상을 Yahoo의 Web사이트로 선택하여 탐색을 수행하면 우선 해당 탐색문에 적합한 주제카테고리의 리스트를 제시한 다음 각 사이트들을 주제 카테고리별로 모아 적합성 순위별로 출력한다. 여러 검색엔진 중 출력양식이 가장 간단하며 Web사이트 검색결과의 경우 검색된 사이트의 주제 카테고리와 표제, 그리고 짧은 기술부로 구성된다.

이처럼 각 검색엔진이 나름대로 순위, 가중치 등을 줌으로써 탐색결과에 순위를 부여하고 있다. 그럼에도 불구하고 이들 순위에 따라 출력된 탐색결과의 내용만을 가지고 적합성의 판정을 한다는 것은 그 정보를 요구한 요구자에 의해 적합성 판정이 행해져야 한다는 것을 감안할 때 적절하지 않은 것으로 생각된다. 이 연구에서 적합성의 판정은 그 질문을 제기한 이용자가 출력물에 연결된 링크를 따라 해당되는 사이트로 가서 문헌자체를 살펴봄으로써 판단하였다. 적합성 평가의 대상은 각 검색엔진이 검색해낸 문헌가운데 적합성 순위가 상위인 20개의 문헌에 제한하여 판정하였다. 이미 다른 연구에서도 지적된 바와 같이 순위가 20위 이하로 내려가면 질문과 관련있는 적합문헌의 수가 급격히 감소하는 것으로 나타나고 있다 (Tillman, 1995). 적합성 판정은 "적합함"과 "비적합함"의 이분법에 의해 행해졌으며, 동일한 내용의 문헌이 같은 url이나 domain명에 속할 때에는 하나의 문헌으로 간주하였다.

3. 4 연구의 제한점

1) 연구대상을 Alta Vista와 Yahoo의 두 검색엔진에게만 국한시킴으로써 다른 검색엔진에서도 출력될 수 있는 적합문헌의 양을 제한하였고 최적의 검색효과를 가지지 못하였다. 따라서 이 연구의 결과를 모든 주제별 디렉토리와 키워드 검색엔진에 일반화시킬 수는 없다.
2) 본 연구의 대상 데이터베이스를 Web으로만 제한시킴으로써 Web 이외에 ftp, gopher, e-mail에서의 검색결과는 포함시키지 않았다. 또한 Alta Vista는 Web 이외에 usenet news group의 기사전문을 대상으로 데이터베이스를 구축하고 있으며, Yahoo도 usenet news group, e-mail의 정보를 제공하고 있으나 여기서는 일관된 비교를 위하여 Web의 문헌만을 연구의 대상으로 하였다.
3) 하나의 질문에 대한 적합성의 평가는 검색엔진이 검색해낸 문헌가운데서 적합성 순위가 상위인 20개의 문헌들에만 제한하여 20위 이하의 출력물에서도 적합한 문헌이 출력될 가능성을 배제하였다. 따라서 실제 탐색과정에서 이루어질 최대의 효과를 거두지 못하고 있다.
4) Yahoo가 적합한 문헌을 검색해내지 못할 때는 자동적으로 Alta Vista의 문헌을 검색해 주는데, 자연상태에서 두 검색엔진의 결과를 비교하기 위하여 이러한 경우에 Alta Vista가 검색해낸 문헌을 Yahoo가 검색해낸 내용으로 포함시켰다.
5) 이 연구는 조사하고자 하는 현상에 대한 이해도를 높이고 개념을 정립함으로써 향후 연구에 대한 제시를 위한 탐색적 연구(Exploratory Study)의 형태를 취하고 있다. 다시 말해서, 연구문제에 대한 통찰력과 추후 연구의 가설정립에 도움을 주지만 기초연구에서 목표하고 있는 가설의 검증을 통한 모집단에의 일반화를 추구하고 있는 것은 아니다. 좀 더 정교한 기초연구가 이 연구의 결과를 토대로 이루어질 수 있을 것이다.

4. 탐색결과 분석

4. 1 검색된 문헌의 양

두 검색엔진에 의해 검색된 문헌의 양은 〈표 1〉에서 보여진 바와 같이 각 질문마다 다른 양상을 보여주고 있으나 대체로 Alta Vista에 의해 검색된 문헌의 양이 Yahoo에 의해 검색된 문헌의 양보다 월등히 많은 것을 알 수 있다. 질문 1, 2, 5, 6의 경우, Alta Vista는 각각 700, 10,000, 100, 10,000건의 문헌을 검색해 내었으나 Yahoo는 평균 30건의 문헌을 검색해 내었다. 이것은 Alta Vista가 웹문헌을 색인대상으로 하는데 비해 Yahoo는 웹카테고리

〈표 1〉 검색된 문헌의 양

	질문 1	질문 2	질문 3	질문 4	질문 5	질문 6	평 균
Alta Vista	700	10,000	600	60	100	10,000	***
Yahoo	10	24	2,500	14,700	12	75	***

와 웹사이트를 대상으로 색인을 하기 때문인 것으로 생각된다. 그러나 질문 3과 4의 경우는 Yahoo에서 한 건의 문헌도 검색해내지 못했기 때문에 자동적으로 Alta Vista로 연결되어 검색된 경우를 보여준다. 이들의 경우, 결과적으로 두 질문 모두 Alta Vista에 의해 검색된 결과를 보여주지만 검색과정에 있어서 Alta Vista 자체에 의해 검색된 경우와 Yahoo의 검색식을 그대로 사용하면서 Alta Vista에 의해 검색된 결과는 상당히 다른 것을 알 수 있다. 질문 3의 경우, 처음부터 Alta Vista를 통한 경우는 600건의 문헌이 검색되었으나 Yahoo를 경유한 Alta Vista의 경우는 2,500건의 문헌이 검색되었다. 전자의 경우는 3개의 AND 연산자로 연결된 검색이었기 때문에 적은 양의 문헌이 검색되었으나 Yahoo를 경유한 경우에는 2개의 AND 연산자를 사용했기 때문에 전자보다 많은 양의 문헌이 검색된 것으로 여겨진다. 또한 질문 4의 경우, 처음부터 Alta Vista로 검색한 경우는 60건의 문헌이, Yahoo를 경유해서 Alta Vista로 검색한 경우는 14,700건의 문헌이 검출되었다. 이 경우는 두 탐색식 모두 AND 연산자를 두번씩 사용한 경우이지만 Alta Vista만을 사용한 경우에는 "host:tw"를 지정해 줌으로써 도메인명이 Taiwan인 것만으로 검색의 범위를 제한했기 때문에 Yahoo를 경유한 경우보다 훨씬 적은 양의 문헌이 검색된 것으로 보인다. 특히 질문 3과 4를 위해서 Yahoo는 전혀 문헌을 검출해내지 못했는데 그 이유는 두가지로 요약된다. 첫째, Yahoo에서 키워드 검색을 진행하면 문헌의 내용을 검색대상으로 하지 않고 웹 카테고리나 웹 사이트를 대상으로 검색하기 때문에 Alta Vista보다 훨씬 적은 양의 문헌이 검색되는 것으로 보인다. 둘째, Yahoo에서 AND 연산자를 3개 이상 사용하면서 검색했기 때문에 검색된 결과가 없다는 메시지가 나타난 것으로 보인다.

4. 2 검색된 적합문헌의 양

각 검색엔진이 검색해낸 문헌의 양 중에서 출력순위 상위 20위까지를 판단해 보았을 때, 각 질문당 검색된 적합문헌의 양의 평균은 Alta Vista에서 3.8건이고 Yahoo에서는 5.5건이었다. 이 결과는 수백만개의 항목탐색이 가능하여 많은 양의 문헌이 검출되는 Alta Vista에서 소량의 문헌이 검출되는 Yahoo보다 더욱 많은 적합문헌이 검색되리라는 예상을 벗어난 결과이다. 전문데이터베이스인 Alta Vista에서 검색된 적합문헌의 양은 초록데

〈표 2〉 검색된 적합문헌의 양

	질문 1	질문 2	질문 3	질문 4	질문 5	질문 6	평 균
Alta Vista	5	4	0	6	2	6	3.8
Yahoo	5	4	4	7	0	13	5.5

〈표 3〉 재현률

	질문 1	질문 2	질문 3	질문 4	질문 5	질문 6	평 균
Alta Vista	0.50	0.50	0	0.46	1	0.32	0.46
Yahoo	0.50	0.50	1	0.53	0	0.68	0.54

〈표 4〉 정확률

	질문 1	질문 2	질문 3	질문 4	질문 5	질문 6	평 균
Alta Vista	0.25	0.20	0	0.30	0.10	0.30	0.19
Yahoo	0.50	0.20	0.20	0.35	0	0.65	0.31

이터베이스인 Yahoo에서 검색된 적합문헌의 양보다 많을 것이라는 예측은 거부되었다. 물론 질문 3과 4를 위해 Yahoo에서 검색된 문헌은 사실상 Alta Vista에서 검색된 것으로서 이들을 제외하면 두 검색엔진이 거의 유사한 정도의 적합문헌을 검색해 내었으나 이 연구에서는 자연상태에서의 연구결과를 원했기 때문에 Yahoo를 사용해서 검색할 때 매치되는 문헌이 없으면 자동적으로 Alta Vista로 연결시켜 검출해내는 결과를 Yahoo의 검색결과로 인정하였다.

4. 3 재현율

Alta Vista로부터 검색된 문헌의 평균재현율은 〈표 3〉에서 보는 바와 같이 0.46이었고 Yahoo로부터 검색된 문헌의 평균재현율은 0.54로서 두 결과가 유사했으나 Yahoo의 평균재현율이 약간(0.08) 높은 것으로 나타났다. 이것은 일반적으로 키워드형 검색엔진이 전문데이터베이스로서 수백만개의 항목탐색이 가능한 반면에 주제별 디렉토리 검색엔진은 초록데이터베이스로서 웹사이트를 담고 있는 소분류체계를 색인하기 때문에 적은 양의 문헌이 검출된다는 일반적인 연구결과와 다른 모습을 보여주고 있다. 물론 Yahoo 대신에 Alta Vista의 결과를 보여주는 질문 3과 4의 결과를 제외한다면 평균재현율은 Alta Vista와 Yahoo에서 각각 0.58:0.42로서 Alta Vista에서의 비율이 높은 것으로 나타나고 있다.

4. 4 정확률

〈표 4〉에서 보여주는 것과 같이 Alta Vista의 평균정확률은 0.19이고 Yahoo의 평균정확률은 0.31로서 전문데이터베이스인 Alta Vista보다 초록데이터베이스인 Yahoo에서 평균정확률이 더욱 높을 것이라는 예측은 적중되었다. 이 연구의 결과를 다른 연구결과와 비교해 보면, Alta Vista의 평균정확률이 0.1이라고 보고한 정영미와 김성은(1997)의 결과와 비슷한 모습을 보여준다. 그러나 Chu와 Rosenthal(1996)의 결과에서 Alta Vista의 정확률이 0.78인 것과 비교하면 매우 다른 모습을 보여준다. 그 이유는 Chu와 Rosenthal의 연구에서는 대개 2개 이상의 AND 연산자를 사용하거나 단일 단어, 구탐색에 치중하였으므로 높은 정확률을 가진 적합논문을 많이 검출했으나 본 연구에서는 평균 3개 이상의 AND 연산자를 사용한 복잡한 질문이었기 때문에 정확률이 낮은 것으로 생각된다. 선행연구에서 밝혀진 바와 같이 복잡성의 정도와 검색된 적합문헌의 양 또는 정확률은 서로 반비례 관계에 있다(이명희, 1994). 복잡성은 탐색전략과 관계된 탐색개념의 수에 근거하여 질문을 등급화한 것으로 표시되고, 탐색개념은 탐색자에 의해 파악된 개념을 나타내는데 이를 위해 탐색파셋의 수는 불리안 연산자 AND로 구분된다. 탐색파셋이 많으면 많을수록 적합문헌의 양은 적어지고 정확률은 낮은 것으로 알려져 있다. 이 연구결과가 Chu와 Rosenthal의 연구결과보다 정확률이 낮은 것은 질문의 유형에 따른 복잡성의 정도에 기인하는 것으로 보인다.

4. 5 각 질문에 대한 개별 분석

질문 1과 2는 비교적 평이한 일반적인 질문으로서 두 검색엔진으로부터의 결과, 재현율은 공히 0.5로서 어느 정도 높았으나 정확률은 그리 높지 않은 것으로 나타났다. 하지만 질문 1의 경우 Yahoo의 정확률은 0.5로서 상대적으로 높은 것으로 보인다. 이로 미루어 보아 Yahoo는 일반적이고 평이한 주제에 적합한 검색엔진인 것으로 생각된다. 특히 Yahoo의 검색시에는 처음부터 키워드 검색을 시작할 것이 아니라 주제별 카테고리를 몇 단계 클릭해서 검색의 범위를 좁힌 다음 거기서 "options"를 선택한 후 거기에 나타난 검색어 입력상자를 이용해서 해당 분류안에 있는 정보만을 대상으로 몇개의 키워드를 조합하여 검색을 진행하면 효과적으로 정보를 찾을 수 있다. 이때 특히 한두개만의 키워드를 "+"부호를 사용하여 검색하는 것이 좋다.

질문 3의 경우, Alta Vista에서는 비록 600건의 문헌을 검출했음에도 불구하고 상위 20건의 문헌에 대해 적합성 판정을 했을 때 하나의 문헌도 적합한 것으로 나타나지 않았다. 이것은 키워드 검색엔진이 수백만개의 항목색인의 탐색에도 불구하고 부적합문헌의 검출에 의한 탐색의 실패를 초래한다는 것을 보여주는 단적인 예로서 정보량의 홍수에도 불구하고 막상 원하는 정보의 탐색실패를 초래하는 현대

상용 데이터베이스의 전형적인 문제점을 그대로 보여 준다. 네트웍 데이터베이스에서 키워드 검색엔진의 좀 더 정교한 검색 테크닉이 요구된다 하겠다. Yahoo를 사용해서 질문 3과 4에 대해 검색했을 때에도 위와 유사한 결과가 나타났는데, 앞에서 지적한 바처럼 Yahoo를 사용해서 검색할 때에는 먼저 주제 카테고리로 검색범위를 좁힌 다음 한 두개의 키워드를 사용하여 검색하는 것이 바람직한 것으로 보인다. 한편 질문 4의 경우, Alta Vista를 사용했을 때 "host:tw"의 도메인 네임을 지정함으로써 특정한 국가의 정보를 제한하는 기능을 사용하였다. 이 방법은 정확률을 별로 낮추지 않으면서도 검색문헌의 수를 줄임으로써 이용자 노력을 최소화시킬 수 있는 좋은 방법으로 보인다. 이 이외에도 가능하다면 title, url, text, link 등의 옵션 기능을 사용함으로써 검색결과를 줄이는 것이 키워드 검색엔진인 Alta Vista의 사용시에 좋은 방법인 것으로 보인다.

질문 5의 경우는 특정대학의 특정기관에 관한 정보를 찾는 경우인데, Alta Vista는 상당한 규모의 적합문헌을 검색해 내었으나 Yahoo는 전혀 검색하지 못했다. 그러므로 희귀한 단어나 자주 사용되지 않는 특정적인 단어 또는 기술적인 용어는 Yahoo보다 Alta Vista를 사용해서 검색하는 것이 나은 것 같다. 그러나 일반적이고 보편적인 용어는 Yahoo를 사용하는 것이 효과적인 것처럼 보인다.

질문 6의 경우는 비교적 특정적인 단일의 단어를 사용한 경우로서 두 검색엔진의 결과는 모두 양호한 것으로 보인다. 그 이유는 NAFTA(North American Free Trade Agreement)가 비록 특정적인 단어이기는 하지만 대상범위가 상당히 포괄적이기 때문인 것으로 보인다.

5. 결론 및 제언

본 연구는 주제별 디렉토리 검색엔진인 Yahoo와 키워드 검색엔진인 Alta Vista가 대학도서관 이용자들에 의해 주어진 6건의 탐색질문에 대해 얼마나 적합한 문헌을 검색해 내는지를 알아보기 위하여 시도되었다.

탐색결과 밝혀진 사실은 다음과 같다. 첫째, 대체로 Alta Vista에 의해 검색된 문헌의 양이 Yahoo에 의해 검색된 문헌의 양보다 월등히 많은 것을 알 수 있는데, 그 이유는 Alta Vista가 웹문헌을 대상으로 색인작업을 진행하는데 비해 Yahoo는 웹 카테고리나 웹 사이트를 대상으로 색인하기 때문인 것으로 생각된다.

둘째, 각 검색엔진이 검색해낸 문헌의 양 중에서 출력순위 상위 20위까지를 판단해 보았을 때 각 질문당 검색된 적합문헌 양의 평균은 Alta Vista에서 3.8건이고 Yahoo에서는 5.5건이었다.

셋째, Alta Vista의 재현율은 0.46이었고 Yahoo의 재현율은 0.54로서 두 검색엔진으로부터의 결과는 유사했는데 이것은 전문데이터베이스인 키워드형 검색엔진이 초록데이터베이스인 주제디렉토리보다 많

은 양의 적합문헌을 검출한다는 일반적인 예측과는 다른 결과이다.

넷째, Alta Vista의 정확률은 0.19이고 Yahoo의 정확률은 0.31로서 전문데이터베이스인 Alta Vista보다 초록데이터베이스인 Yahoo에서 정확률이 높을 것이란 예측은 적중되었다.

Yahoo는 일반적이고 추상적인 주제에 적합한 검색엔진인 것으로 나타났는데, 특히 Yahoo의 검색시에는 처음부터 키워드 탐색을 시작할 것이 아니라 주제별 카테고리를 몇 단계 클릭해서 검색의범위를 좁힌 다음에 "options"를 선택한 후 나타난 입력상자를 이용해서 해당 분류안에 있는 정보만을 대상으로 키워드를 조합하여 "+"부호를 사용하여 검색하는 것이 효과적이다. Alta Vista를 사용해서 검색했을 때에는 수백만개의 항목색인에 의해 많은 문헌이 검출되었음에도 불구하고 부적합문헌의 검출에 의한 탐색실패를 초래하는 전형적인 키워드 검색의 문제점을 보여주고 있다. 키워드 검색시의 중요한 사항은 어떻게 검색결과를 줄이는가 하는 점이다. Alta Vista에서는 키워드를 2-3개 정도 사용하는 것이 좋은데, 2개 정도의 용어는 일반적인 용어로, 하나의 용어는 전문적인 용어를 사용하여 AND 연산자로 묶는 것이 효과적이다. 구검색시에는 " "를, 불리안 연산자를 비롯한 고유한 연산자를 많이 사용하는 것이 좋으며 특히 url, host, link, text 등의 고유한 옵션기능을 사용함으로써 검색범위를 줄이는 것이 효과적이다. 희귀한 단어나 자주 사용되지 않는 특정적인 단어, 전문적인 용어는 Yahoo보다 Alta Vista가 더욱 효과적인 것 같다. 그러나 질문 타입의 특성과 검색엔진의 사용은 좀 더 지속적이고 광범위한 연구가 요청되는 분야이다.

실험결과에서 보여주는 바와 같이, 두 검색엔진은 각각 색인 구축방법, 중점적인 서비스 및 탐색기법 등이 다르기 때문에 동일한 정보요구에 대해 서로 다르게 반응하는 것을 알 수 있다. 어떠한 하나의 검색엔진도 완벽한 정보서비스를 제공하지 못하기 때문에 다양한 여러 검색엔진을 함께 사용함으로써 포괄적인 탐색이 이루어진다고 볼 수 있다. 현재 인터넷상에서 300여개 이상의 검색엔진이 활용되고 있는데, 이용자 입장에서 이들의 성능과 서비스에 대해 충분히 숙지하고 검색에 임하는 경우는 드물다고 봐야 할 것이다. 따라서 이들 검색엔진의 탐색목적, 색인 구축범위, 서비스의 내용 등에 관한 좀 더 체계적인 연구가 수행되어야 할 것이다. 특히 정보량의 홍수에도 불구하고 탐색실패라는 상반되는 결과는 정보검색 현장에서 기존의 검색엔진의 다각적인 평가를 통하여 효과적으로 탐색할 수 있는 차세대적인 검색엔진의 설계가 절실히 요구된다 하겠다.

참고문헌

남영준. 1987. 전문데이터베이스의 탐색 효율성 분석. 석사학위논문. 중앙대학교.

이명희. 1994. "수자원문헌의 주제탐색과 인용탐색의 검색효율 비교연구", 한국문헌정보학회지 26 : 213-233.

정영미, 김성은. 1997. "WWW 탐색도구의 색인 및 탐색기능 평가에 관한 연구", 한국문헌정보학회지 31 (1) : 153-184.

Chu, H. and M. Rosenthal. 1996. " Search Engines for the World Wide Web : A Comparative Study and Evaluation Methodology." Presented at the 96' ASIS Conference.

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

Janes, J. W. and L. B. Rosenfeld. 1996. "Networked Information Retrieval and Organization : Issues and Questions." JASIS. 47(9) : 711-715.

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

Leighton, H. V. 1995. "Performance of Four World Wide Web (WWW) Index Services : InfoSeek, Lycos, WebCrawler, and WWW Worm", http://www.winona.msus.edu/services-f/library-f/webind.htm.

Leonard, A. J. 1996. "Where to Find Anything on the Net", http://www.cnet.com/Content/Reviews/Search

Pollack, A. and A. Hockley. 1997. "What's Wrong with Internet Searching", D-Lib Magazine. March, 1997. http://www.dlib.org/dlib/march97/bt/03pollock.htm/

Taubes, G. 1995. "Indexing the Internet", Science. 269. 1354-6.

Tenopir, C. 1985. "Full Text Databases Retrieval Performances." Online Review. 9(2) : 149-164.

Tillman, O. N. ed. 1995. Internet Tools for the Profession : A Guide for Special Librarians. Washington DC : SLA.

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