WWW 탐색도구의 검색성능에 관한 실험적 연구

An Experimental Study on Retrieval Performance of WWW Search Tools

이은주, 연세대학교 대학원 문헌정보학과
정영미, 연세대학교 문헌정보학과
Lee Eun-Joo, Chung Young-Mee
(Dept. of Library and Information Science, Yonsei Univ.)

본 연구는 WWW 탐색도구의 검색성능을 평가하고, 또 검색성능에 영향을 미치는 요인들을 밝혀내기 위하여 수행되었다. 탐색도구의 데이터베이스 구축방식과 색인기법, 탐색기법, 이용자 인터페이스에 관련된 현황분석을 토대로 탐색도구의 검색성능에 영향을 미치는 요인들을 알아보기 위하여 검색실험을 수행하였다. 실험결과의 분석은 각 탐색도구의 검색효율과 검색결과의 중복도 및 유사도, 검색결과의 순위 및 적합성 순위부여 알고리즘, 웹 문서의 수집기법, 탐색도구의 최신성을 기준으로 이루어졌다.

1.서론

World Wide Web(WWW 또는 Web ; 웹)이 빠르게 대중성을 확보함으로써, 웹 상에는 다양한 유형의 정보가 폭발적으로 증가하게 되었다. 결과적으로 이용자들이 원하는 정보를 찾는 일이 극히 어렵게 되었으며 많은 수의 WWW 탐색도구가 등장하였음에도 불구하고 각기 제공하는 특성이나 검색성능에 있어 상당한 차이를 보이고 있다.

본 연구에서는 여러 가지 WWW 탐색도구의 색인 데이터베이스 특성과 색인기법 및 검색기법, 적합성 순위부여 알고리즘, 탐색 인터페이스와 관련된 제반 특성을 살펴보고, 그러한 현황 분석을 토대로 하여 각 탐색도구의 검색성능을 알아보기 위한 실험을 수행하였다. 각 탐색도구의 검색성능이 어느 정도이고, 또 검색성능에 영향을 미치는 요인은 무엇인가를 파악하여 탐색결과의 질을 향상시키고 검색성능을 강화할 수 있는 여러 가지 방법들을 모색해 보고자 한다.

2. 이론적 배경

WWW 탐색도구는 크게 2가지 유형으로 나누어 볼 수 있다.

우선, 목록(catalog), 디렉토리(directory), 색인(index) 등으로 불리는 디렉토리 유형에 속하는 탐색도구는 미리 정해진 주제에 따라 이용자가 브라우징을 할 수 있도록 웹 문서를 일정 체계에 따라 계층적으로 조직한 것으로, 색인전문가가 다른 이용자들에게 적합할 것이라고 판단하는 웹 정보들을 포함하고 있는 데이터베이스이다. Yahoo!나 Galaxy, WWW Virtual Library, Magellan 등이 이에 속하는데, 다양한 주제분야를 포괄하면서 가치 있는 사이

트를 선택하는 것이 목적이므로 웹에 익숙하지 않은 초심자들이 탐색을 시작하기에 좋은 출발점이 되거나, 정부단체나 기업, 연구센터. 혹은 기타 잘 알려진 항목들을 찾는 데 특히 유용하다.

또 다른 유형의 하나인 탐색엔진은 일반적으로 자동화된 로봇(robot) 프로그램에 의해서 만들어지고, 이용자가 자신의 탐색문을 스스로 구성하도록 하는 서비스를 가리킨다. 웹 로봇은 웹의 하이퍼텍스트 구조를 따라 다니면서 문서를 검색하고, 참조되어 있는 모든 문서들을 반복적으로 검색해 내는 프로그램으로서, AltaVista, Excite, Infoseek, Lycos, OpenText, HotBot, WebCrawler 등이 이러한 로봇프로그램을 사용하는 대표적인 탐색엔진이다.

3. WWW 탐색도구의 현황 비교·분석

본 연구에서는 현재 널리 사용되고 있는 WWW 탐색도구의 현황을 데이터베이스의 구축, 색인기법, 탐색기법, 이용자 인터페이스의 4가지 측면에서 비교·분석하였다.

분석대상 탐색도구는 1) AltaVista, 2) Excite, 3) Infoseek, 4) Lycos, 5) Magellan, 6) OpenText, 7) Yahoo! 7가지 등이다. 1997년 3월 현재 각각의 탐색도구가 가지는 특성들을 <표 1>부터 <표 4>에서 제시하였다.

<표 1> 각 탐색도구의 색인 데이터베이스 특성

	Alta Vista	Excite	Info seek	Lycos	Magellan	Open Text	Yahoo!
로봇이용	○	○	○	○	○	○	○
색인전문가	×	×	×	×	○	×	○
이용자등록	○	○	○	○	○	○	○
이용자삭제				○			○
수집원칙							
· 서버단위	○	○				○	○
· 파일단위	○		○	○	○	○	

<표 1> 각 탐색도구의 색인 데이터베이스 특성 (계속)

	Alta Vista	Excite	Info seek	Lycos	Magellan	Open Text	Yahoo!
수집원칙							
· 서버단위	○	○				○	○
· 파일단위	○		○	○	○	○	
선정기준	모든 자료	대중성		모든 자료	유용성		질통제
자원유형							
· 웹	○	○	○	○	○	○	○
· 유즈넷	○	○	○		○		○
· 고퍼			○	○	○	○	
· ftp			○	○	○	○	
갱신빈도	매일	1-2주	매일	2-4주	매주	매일	이용자의 지적시
규모(백만)	31	50	50	66/34			

<표 2> 각 탐색도구의 색인대상 필드

	Alta Vista	Excite	Info seek	Lycos	Magellan	Open Text	Yahoo!
전문	○	○	○			○	
서명				○	○		○
표목				○	○		
URL				○	○		
요약/초록				○	○		○
메타태그 지원	○	×	○			×	
날짜	○						

<표 3> 각 탐색도구의 순위부여 기준

	Alta Vista	Excite	Info seek	Lycos	Magellan	Open Text	Yahoo!
하이퍼링크 구조		○		○			
탐색어의 출현빈도	○	○	○	○	○	○	
탐색어의 출현위치	○	○	○	○	○	○	○
탐색어간 인접성	○			○			
매치되는 탐색어의 수	○		○	○	○		○

<표 4> 각 탐색도구가 제공하는 탐색기능

	Alta Vista	Excite	Infoseek	Lycos	Magellan	Open Text	Yahoo!
매치유형							
· 완전일치	○	○				○	○
· 최적일치	○	○	○		○		
· 혼합	○	○		○			
탐색문유형							
· 자연언어		○	○				
· 단어리스트	○	○	○	○	○		○
· 불논리질문	○	○		○		○	
불논리연산			×		×		
· AND	○	○		○		○	○
· OR	○	○		○		○	○
· NOT	○	○				○	
· 혼합	○					○	
· nesting	○	○					
+/-	○	○	○	○	○	×	○
인접탐색	○					○	
구절탐색	○	○	○	×	×	○	○
대소문자인식	○	○	○	×	×	×	
절단탐색					×	×	
· 자동		○	○	○			
· 수동	○			○			○
제한탐색							
· 특정필드별	○		○			○	○
· 자료유형별	○			○			
· 데이터베이스 유형별		○	○		○		○
· 날짜별	○						○

4. 검색실험

(1) 검색성능에 관한 실험

탐색도구의 성능평가를 위한 검색실험에 앞서, 각 탐색도구의 데이터베이스 규모를 분석하기 위한 실험을 수행하였다. 이를 통해 데이터베이스의 규모별, 색인방식별로 각기 차이를 보이거나 유사한 탐색도구들을 실험대상으로 선정하였다.

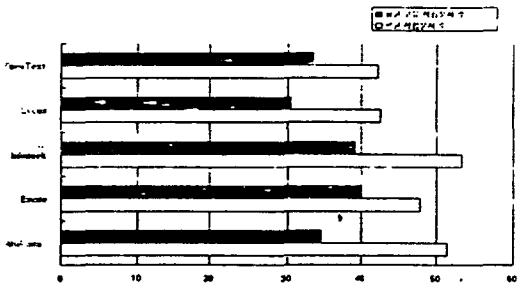
실험대상으로 삼은 탐색도구는 AltaVista, Infoseek, Excite, Lycos, OpenText였고, 8개의 실험질문에 따라 검색을 수행하였다.

본 연구에서는 탐색대상을 모두 웹 문서로 한정하여 실험을 수행하였고, 각 탐색문에 대한 검색결과 중 100개의 검색 결과에 대하여 적합성을 판정하였다. 탐색결과에 출력된 간단한 요약만으로는 적합성을 판정하기에 부적절한 경우가 많기 때문에 연결된 링크를 따라서 실제 웹 문서에 접근, 내용을 살펴본 후 적합성을 판정하였다. 정확히 일치하는 URL을 가지고 있는 웹 문서들을 중복으로 계산하였고, 미러사이트, 즉 상이한 URL을 가지고 있는 동일한 웹 문서는 중복으로 간주하지 않았다.

(2) 검색결과 및 평가

Infoseek가 평균적으로 질문당 53개의 적합문서를 검색하였고, 고유한 적합문서는 Excite가 질문당 평균 40개로 가장 많았다. Excite의 경우 고유한 적합문서가 검색된 적합문서의 84%에 달하는 것으로 나타났다. Lycos와 OpenText의 평균 적합문서 수는 나머지 탐색도구에 비해 다소 적은 것으로 나타났다.

<그림 1>은 각 탐색도구가 검색해 낸 평균 적합문서 수와 평균 고유한 적합문서 수를 비교해 놓은 것이다.



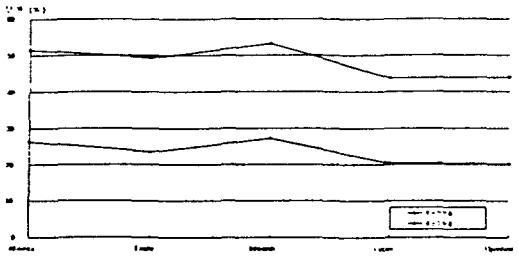
<그림 1> 각 탐색도구의 평균 적합문서 수와 평균 고유 적합문서 수

(3) 검색효율 측정

검색효율을 측정하기 위하여 각 탐색도구별로 재현율과 정확률을 산출하여 비교하였다.

Infoseek와 AltaVista의 재현율이 평균 27.21%와 26.25%로 다른 탐색도구에 비해 상대적으로 높은 것으로 나타났지만, 그리 큰 편차를 보이지는 않았다. 정확률의 경우 각 탐색

도구의 정확률 평균은 48.4%이다. 평균을 넘는 정확률 값은 갖는 것은 Infoseek와 AltaVista로 평균정확률이 각각 53.32%와 51.54%이다. 이는 재현율과 마찬가지로 나머지 세 개의 탐색도구보다 상대적으로 높은 성능을 보였다. <그림 2>는 각 탐색도구별 평균재현률과 평균정확률을 나타내주는 것이다.



<그림 2> 탐색도구별 평균재현율과 평균정확률

(4) 탐색도구간의 유사도 측정

각 탐색도구의 검색결과가 어느 정도로 유사한지 혹은 상이한지를 알아보기 위해 검색결과의 유사도를 측정하였다.

유사계수 공식 가운데 다이스계수를 이용하여 유사도를 측정하였으며, 먼저 탐색도구간에 동일하게 검색한 파일을 대상으로 유사도를 계산하였다. 탐색도구들간에 유사도는 대부분 0.1을 넘지 않았으며, 평균적으로 약 0.07의 유사도를 보였을 뿐이었다. 서버를 기준으로 완화시켜 검색결과의 유사도를 살펴보았을 때는, Infoseek와 Lycos간의 유사도가 0.4 정도를 보였고, 그 다음으로 유사도가 높은 경우는 AltaVista와 Infoseek의 탐색결과로 약 0.28의 유사도를 보였다.

<표 5> 각 탐색도구간 평균 유사도 행렬

	AltaVista	Excite	Infoseek	Lycos	OpenText
AltaVista	1	.0481	.1029	.0879	.0909
Excite	.1670	1	.0672	.0584	.0332
Infoseek	.2790	.2278	1	.0970	.0396
Lycos	.2632	.2087	.3911	1	.0406
OpenText	.1941	.1301	.1064	.1055	1

(5) 탐색도구의 최신성

검색결과 제시된 링크를 따라 실제 웹 문서를 검색할 때, 지정된 URL에 실제 문서가 없는 경우가 발생한다. 평균적으로 약 9개의 링크가 사용불능링크였다. Infoseek가 질문당 평균 12개의 사용불능 링크를 검색결과로 제시하였다. 이러한 사용불능 링크는 검색결과의 정확률에 저하를 가져오게 되며, 탐색도구에 대한 신뢰를 떨어뜨리는 원인 가운데 하나로 볼 수 있었다.

5. 결론 및 제언

본 연구의 실험결과, 각 탐색도구들이 검색해 낸 적합문서는 총 적합문서의 20% 정도에 지나지 않았고, 검색된 적합문서의 비율은 평균 약 48%의 결과를 보였다. 재현율과 정확률의 검색효율을 측정한 결과, 탐색도구 가운데 Infoseek와 AltaVista가 각기 평균재현율 27.21%, 26.25%와 평균정확률 53.32%, 51.54%로 탐색도구의 평균적인 검색성능보다 우수한 성능을 보여주었으며, 나머지 3개의 탐색도구에 비해 상대적으로 높은 검색효율을 보여주었다.

기존의 WWW 탐색도구가 이상적인 시스템이 되기 위해서는 결국 다음과 같은 특성과 기능을 지녀야 할 것으로 보여진다. 일정한 선정기준, 데이터베이스의 규모 확장, 완벽한 갱신이 가능한 로봇프로그램 구현, 언어에 대한 이해를 기반으로 한 색인, 정교한 적합성 순위부여 알고리즘, 중복 URL에 대한 관리와 언어에 따른 제한탐색을 할 수 있도록 하는 기능을 제공해야 하며, 다양한 유형의 자료들이 웹에 추가됨에 따라서 좀 더 효율적인 분류체계 및 여과장치가 적용되어서, 수집된 자료를 그 유형이나 성격에 따라 자동으로 분류하고 검색할 수 있도록 해주어야 할 것이다.