메타 인덱스 카탈로그와 에이전트를 이용한

전문정보 검색 시스템 개발 방안

김화수/ 칼스텍

한국데이터베이스진흥센터 산하의 한국정보검색위원회에서 위원간의 연구의욕 고취와 새로운 검색 및 데이터베이스 관련 기술 보급을 위해 매월 연구발표회를 개최하고 있다. 본 코너는 매월 발표된 주제논문을 게재함으로써 정보검색과 관련된 정보를 제공하기 위해 마련된 것이다.〈편집자〉

도입

1969년 미 방위성(Department of Defense)의 연구 프로젝트로 시작한 인터넷이 이제 수백만대의 호스트 컴퓨터들이 연결된 거대한 네트웍으로 자리잡고 있다.[1] 특히 최근 2~3년간 인터넷의 성장은 세계적으로 매년 2배에 가까운 성장률을, 그리고 도메인 수는 3~4배의 성장률을 보이고 있다.[2]

이러한 추이는 학술 도메인(edu, ac)에 비해 상업용 도메인(com, co)에서 두드러지게 나타나고 있는데 이는 인터넷을 기업 경영 환경과 마케팅, 그리고 정보 기술에 접목시켜 그 활용도를 극대화하려는 노력의 산물로 파악할 수 있다. 새로운 정보기술로서 인터넷의 최종사용자 활용은 크게 두 가지 측면으로 볼 수 있다.

그 첫 번째는 인터넷을 새로운 통신 수단으로 바라보는 것과 두 번째는 인터넷을 정보자원(Information Resources)으로 보는 시각이다.[3]

통신수단으로서의 인터넷에 대한 시각은 인터넷의 역사가 시작하는 순간부터이지만 정보자원 관점에서 인터넷을 바라보는 것은 검색엔진들이 탄생한 시점인 1994년 상반기부터라 할 수 있다. 인터넷의 역사에 비해 매우 짧다.

그러나 현재 점차 상용화되고 있는 인터넷은 온라인 상용 데이터베이스들의 대안으로서 그 중요성이 커지고 있으며 또한 최근 1~2년간에 걸쳐 수백여개의 검색엔진과 수만여종의 전문 데이터베이스들이 출현함으로서 중요성을 뒷받침하고 있다.

인터넷의 규모는 어느 누구도 정확한 수치를 파악할 수 없지만 1998년 7월 현재 약 260만(2,594,622)[4]의 웹 사이트와 약 3억2천만개[5]의 웹 페이지가 있는 것으로 알려져 있다. 이 가운데 사이트의 경우 약 70만개가 Yahoo!에 등록되어 있으며 웹 페이지의 경우 검색엔진 AltaVista는 98년 7월 현재 약 2억(197,912,367)개의 웹 페이지를 데이터베이스로 구축해 둔 상태이다.

■ 전문정보 검색 시스템 개요 및 필요성

'에이전트를 이용한 전문정보 검색 시스템'(이하 전문정보 검색 시스템)의 개발 필요성은 현재 개발된 대표적인 두 가지 형태 검색엔진인 Yahoo!류와 AltaVista류의 기존 검색엔진이 갖는 검색 기능 한계에서 출발한다.

특정 분야의 전문 정보를 검색하고자 하는 최종 사용자들이 원하는 검색 서비스는 Yahoo!처럼 단순히 사이트 디렉토리를 통해 적당한 웹 사이트를 찾거나, AltaVista를 이용하여 두 세 개의 키워드 입력으로 수만 건이 쏟아져 나올 때 어느 URL로 이동해야 할지 판단하기 어려운, 그런 검색엔진이 아니다.

예를 들어 최종 사용자들이 자동차 분야 정보를 찾고자 할 경우 해당 분야의 웹 사이트들만 대상으로 검색하여 관련성 높은 URL로 이동하거나, 해당 분야의 산업정보를 제공해주는 데이터베이스로 직접 이동할 수 있기를 원하고 있다. 즉, 인터넷의 등장으로 점차 최종 사용자 검색(End User Searching) 환경이 확대되는 시점에서 최종 사용자 정보 만족도 제고는 본 논고에서 제시하는 전문 검색 시스템이 지향하는 목적이라 할 수 있다.

전문 정보 검색 시스템의 정의와 핵심 키워드는 다음과 같다.

전문정보 검색 시스템은 산재되어 있는 인터넷 정보 자원(information resources)을 사용자의 다양한 정보 요구(information requirement) 관점에서 접근, 호스트에 대한 정보를 사용자 시각에서 설계한 메타 인덱스 카탈로그와 각 호스트의 하위 URL에 대한 정보를 에이전트에 의해 수집한 전문(Full-text) 검색엔진이 결합한 검색 시스템을 일컫는다.

Keywords:LINK, GATEWAY, TARGET, COMBINATION, RESOURCE, CATALOG, AGENT, SUBJECT, INDUSTRY, META INDEX

1) Richard Selfzer, Eric J. Ray, and Deborah S. Ray, AltaVista Search Revolution: How to find anything on the Internet, (Berkeley: Osborne McGraw-Hill, 1997)
2) Network Wizards, "Internet Domain Survey", http://www.nw.com/zone/WWW/report.html (1997)
3) John F. Lescher, Online Market Research: cost-effective searching of le internet and online databases, (Berkeley:Addison-Wesley Publishing Company, 1997), pp.43-44, 167-177
4) http://www.netcraft.com/Survey/
5) NEC Research Institute, 1998. 04. 06, http://www.nua.ie/surveys/index.cgi?service=view_survey&survey_number=678&rel=no

■ 정보검색 시스템의 분류와 일반 검색엔진(인터넷)의 한계

일반적으로 정보검색 시스템은 i) DIALOG, KINITI, IR 등과 같은 專門 온라인 시스템, ii) CompuServe, 천리안, 하이텔 등과 같은 일반 온라인 시스템, iii) 인터넷 시스템, 그리고 iv) CD-ROM 오프라인 시스템으로 구분한다.[6]

이 네 가지 분류 가운데 온라인 검색 시스템은 CD-ROM(Stand-alone Database)을 제외한 나머지 세 가지를 들 수 있다. 그러나 위 분류와 달리 정보 데이터베이스의 제공형태로 분류할 경우 정보검색 시스템을 ON-LINE, CD-ROM, INTERNET의 세 가지로 분류하기도 한다.[7]

이러한 분류는 김현희의 분류에서 전문 온라인 시스템과 일반 온라인 시스템을 ON-LINE으로 통합하여 보는 시각이다. 이러한 온라인 검색 시스템은 1995년 현재 전세계에 총 8,522개가 서비스 중인 것으로 파악되고 있으며 이 가운데 북미지역에서 개발된 것이 69%에 달하는 5,865개로 파악되고 있다.[8]

본 논고에서 범주로 하는 전문정보 검색 시스템은 온라인 데이터베이스 시스템 가운데 인터넷 시스템만을 대상으로 하며 여기에는 위 분류에 의하면 검색엔진과 인터넷 전문 데이터베이스를 포함할 수 있다.[9] 검색엔진과 인터넷 전문 데이터베이스의 특징은 다음과 같다.[10]

그러면 이상에서 설명한 내용을 좀 더 뒷받침하기 위해 외국의 정보 전문가들은 인터넷의 검색도구들을 어떻게 분류하고 있는지 과거의 시각과 최근의 시각을 함께 보고자 한다. 국내에서는 과거 '인터넷 시스템'으로 포괄적인 시각으로 보았지만 이미 미국을 비롯한 선진국에서는 인터넷 시스템을 보다 상세하고 분류하고 있다.

본 논고에서는 아래 분류에 기초하여 전문 정보 검색시스템의 개발 아이디어를 도출했다. 먼저, 과거의 시각은 AIIP[11] 회장을 역임했던 Reva Basch의 저서 'Secrets of the Super Net Searchers'[12]에서 분류한 것으로 일반적인 검색도구 분류법을 그대로 적용하고 있지만 'Subject Hubs' 개념에 유의할 필요가 있으며, 최근의 시각에서는 본 자료에서 일컫고 있는 '전문 검색엔진' 개념이 그대로 소개되고 있음을 알 수 있다.

〈표 1〉 일반 검색엔진, 인터넷 전문 DB 비교

구분 / 항목	일반검색엔진	인터넷 전문DB
인터넷내 전체 DB 개수	1000	10,000 이상
레코드 규모	10,000,000 이상	10,000 ~ 1,000,000
데이터 정형성	비정형적 (반정형적)	정형적
검색 기능	다양	단순
로봇이용	이용	이용하지 않음
상대적 검색 난이도	어려움	쉬움
소유 기관	기업	정부, 교육기관, 온라인 서비스
스폰서	광고	공공기관, 시협서비스
색인	프로그램 (Robot agent)	사람에 위한 수작업
검색공간	인터넷 내부 문서 검색	인터넷 외부 문서 연동
Gateway 속도	90Mbps 이상	90Mbps 이하
요금	무료	유료 / 무료
예	AltaVista, Infoseek, Lycos	EDGAR, QPAT, MicroPatent

과거의 시각 (1996)

by Reva Basch

A search engine (1) is not always the most effective approach to a research project. Web spiders and bots aren t as smart or discriminating as human-mediated aids. .... The less you know about a subject at the outset, the more sense it makes to start with already-filtered source, a catalog, guide, or virtual library (2) like the Argus Clearinghouse, where subject experts have already done a preliminary evaluation and identified some good places to begin.

.....

a third approach to the research process, one that takes advantage of the ease of personal publication on the Web, and that offers what futurist Paul Saffo predicted would become a highly valued Net-based commodity:point of view. Michel Bauwens refers to such resources as centers of excellence. Gail Clement talks about subject hubs. (3) . Whatever you call them, they represent a individual s or institution s informed opinion of the best and most useful Internet resources in a particular field.

최근의 시각 (1998. 6)

by Richard Wiggins

System Architect, NEM Online

and Judith A. Matthews

Physics-Astronomy Librarian, Michigan State University

(1998. 6)

Common themes emerged from the various speakers presentations. First, ...Second,....Third, we can expect to see catalogs(e.g. Yahoo!) and Full-text indexes (e.g. Infoseek, Lycos, AltaVista) evolve to become more like each other. Indeed, several speakers predicted that future search engines will combine the best features of the hand-crafted catalog with the specificity and currency of full-text search indexes.

6) 김현희, 배금표, 안태경, 정보검색론, (서울:오름, 1997), p. 66

7) 홍정화, "Web을 통한 정보 DB 검색 시스템 구축 방안", '96 데이터베이스 심포지움 및 학술대회, (1996), pp.83-84

8)이의택, 유성준, "디지털 영상 정보의 시장 전망", Multimedia Database on the Internet, 1997, pp.141-143

9) · 김현희 외, op. cit., pp. 185-186,

· 넥서스컨설팅(주), 교육부, op. cit., pp.273-287;

· 김준오, "Web상에서의 정보검색기법에 관한 고찰", '97 International Conference MULTIMEDIA DATABASES on INTERNET, 1997, pp.270-273

10) 金和秀, "인터넷 정보검색의 마지막 노하우", 바다출판사, p.23, 1997

11) http://www.aiip.org

12) "Secrets of the Super Net Searchers", Reva Basch, 1996

Basch는 지난 1996년 자신의 저서에서 'Subject Hubs'라는 개념을 도입하였는데 이것이 특정 주제나 산업에 특화된 검색서비스를 일컫는 것으로 당시 디렉토리 형태로 구성된(에이전트가 탑재되지 않은) 소스맵[13]의 원형이라 할 수 있다.

그러나 최근에는 이러한 형태에서 보다 발전된 에이전트 기술과 'hand-crafted catalog'가 혼합, 연결(combined)되고 특정 주제 및 산업분야에 대한 정보 서비스를 제공하는 검색엔진이 부상할 것이라는 주장이 제기되고 있다. 본 논고는 여기서 설명하는 개념과 상통한다. 그리고 이러한 전문 주제 분야 서비스는 기존의 검색엔진들도 최근 들어 적용하고 있는 방식이다.

Infoseek의 Channel 서비스, Yahoo!의 http://movie.yahoo.com 서비스를 예로 들 수 있다. 이러한 차이점에 따른 전문정보 검색 시스템의 장점(검색의 적합도 측면)을 기존 검색도구의 검색 한계와 비교하여 설명하면 다음과 같다.

전문 정보 검색 시스템

■ 개요

소스맵이라는 개념을 끌어 온 것은 이 개념이 특정 주제 분야의 정보를 제공하는 Site Catalog를 구성하는 것을 일컫기 때문이다 따라서 '에이전트를 이용한 전문 검색 서비스'란 결국 사람의 노하우와 지식이 부여된 Site Catalog와 Robot Agent의 결합 형태로 볼 수 있다.

그러나 이 정도의 해석에서 머물 경우 단순히 Yahoo!형과 AltaVista형이 결합된 형태로만 이해할 수 있는데, 그렇지 않다. 만약 단순 결합 형태라면 두 서비스는 별개로 운용될 것이다.

전문 정보 검색 시스템은 특정 산업 및 주제 분야의 일정 웹 사이트를 주제 전문가와 정보검색 전문가가 주제별로 분류하고 해당 분류에 구성된 여러 사이트 집단을 사용자가 나름대로 구성(My Catalog)하여 특정 주제만을 대상으로 사이트 검색(Yahoo!처럼)과 웹 페이지 검색(AltaVista)을 모두 적용할 수 있도록 서비스할 수 있다.

또한 이 검색 시스템은 에이전트의 URL 색인이 Catalog에 기초하여 출발하게 되고 이러한 Catalog는 최종적으로 모든 사용자들이 개인화된(Personalized) 서비스를 통해 소스맵이 나타내는 특정 주제 내에서 하위 주제들의 조합을 선정, 이렇게 구성된 하위 Catalog만을 대상으로 검색할 수 있으므로 이는 최근의 푸시 기술을 접목할 수 있다.

따라서 '에이전트를 이용한 전문정보 검색 시스템'은 Catalog 제작에 부여되는 필터링 기술, 소스맵 모델링, 푸시 기술, 그리고 에이전트의 탑재라는 이상적인 형태의 서비스로 구현될 수 있는 것이다.

이러한 '이상적'인 형태란 바로 사람의 지식과 자동화된 프로그램이 적절하게 혼합된 것을 의미하기도 한다. 최종 사용자의 검색 시스템 이용 단계를 살펴보면 다음과 같다. 일반 검색엔진의 이용과 유사하지만 'My Catalog' 설정 단계가 추가되어 있음에 유념할 필요가 있다.

■ Meta Index Catalog

메타 인덱스[14] 카탈로그란 에이전트에 의해 형성될 전문 인덱스(Full-text Index)의 상위 인덱스로 호스트에 대한 정보(테이블, Host Information Table)와 주제 분류에 대한 정보(테이블, Subject Index Table)를 저장하는 데이터베이스 테이블로 구성된다.

호스트 정보를 지니고 있는 DB는 에이전트에서 중요한 'starting point' 역할을 하며 자체적으로 보유하고 있는 호스트 정보는 단순 디렉토리 형태의 인덱스보다 한 단계 발전한 인덱스로 소스맵 구현에 있어 가장 중요한 역할을 수행한다.

주제 분류 데이터베이스는 사용자들이 분류항목을 필터링하여 개인화서비스(Personalized Service)가 가능하도록 지원한다. 이 Catalog를 생성하는 전체 과정은 사람의 지식과 검색 능력에 의존하므로 구성 호스트 선정 기준에 주관적인 판단이 개입될 위험이 많다.(그러나 이는 위험요소이면서 오히려 장점으로 작용한다. Yahoo!의 주제분류는 학술적 이론에 바탕하지 않고 있지만 인터넷 이용자들이 이 분류 체계를 수용하고 있다).

메타 인덱스의 테이블 구조와 규모를 예측해 본다. 우선 Catalog의 호스트 규모를 1천개로 설정했을 때 전체 URL 규모를 산정하면 다음과 같다.

The Number of Host	: 1000
The Average Pages of a Host	: 100
The Total Pages of Catalog(DB)	: 100,000

※ 사이트별 평균 페이지 수는 100~200 규모로 파악할 수 있으나 여기서는 100으로 산정

〈표 2〉 기존 검색 시스템의 한계와 전문정보 검색 시스템의 장점

검색도구 / 기준	Yahoo! 형태	AltaVista 형태	전문정보 검색 시스템
구체성 및 정밀성	검색하고자 하는 최종 URL에 대한 정보를 제공하지 않는다. 정보의 구체성 결여 (Because of site directory)	검색 결과의 정밀성이 떨어진다. (Because of Too many results)	검색결과의 정밀성과 구체화 정도가 높다.
전문성	모든 주제 분야를 다루므로 전문성이 떨어진다.	무작위 로보팅에 의한 결과물을 제시하므로 전문성이 떨어진다.	정보자원 설계 전문가에 의한 Catalog 생성과 이를 바탕으로 한 로보팅이 이루어지므로 상대적 전문성이 높다.
산업정보 검색	특정 산업에 대한 정보원을 검색하기에 부적합하다.	특정 산업에 대한 정보원을 검색하기에 부적합하다.	산업별, 주제별 검색이 가능하므로 가장 적합하다.
조직체에 대한 응용	특수한 산업, 주제를 다루어야 하는 조직체에 단순 디렉토리로 응용하기에는 역부족	URL 정보만 지니고 있으므로 사용이 불편. 오히려 조직 구성원이 많은 시간을 검색에 낭비	조직내에서 필요로 하는 정보자원만을 요구분석, 모델링하므로 구체적인 주제 분야를 설정할 수 있다. 업무와 관련된 검색에 있어 기존 검색엔진의 필요성이 줄어든다.
최종사용자 검색 지원	간접지원(Reference역할)	간접지원(Reference역할)	직간접 지원

기대 성과 및 효과

1. Information Hub 역할을 통한 트래픽 증가

향후 인터넷 컨텐트 비즈니스에서는 Information Hubs[15] 정보 연결통로)가 될 수 있는 사이트의 중요성이 점차 커진다. 이런 점에서 검색엔진의 잠재 가치를 높게 평가하는 것은 자연

13) http://www.searchplus.com/source_map/

14) '메타 인덱스'를 인터넷에서는 특정 주제 분야에 대한 링크를 다량으로 보유하고 있는 색인집을 일컫는다.
(예: Meta-Index of Nonprofits: http://www.philanthropy-journal.org/plhome/plmeta.htm) Catalog란 이러한 메타 인덱스를 기반으로 구성, 호스트에 대한 정보를 제공하는 데이터베이스를 말한다.

15) 특정 주제 분야에 대한 각종 정보자원을 연결하고 있는 사이트 (예:The Airline Information Hub : http://airlineinfo. com). 앞서 제시한 Subject Hubs와 동일한 개념이다.

스런 귀결(Yahoo!의 자산 가치: 40억달러)이다.

웹 공간은 사이트 개발 1차 목적이 트래픽[16] 증가에 있으며 이를 실현하기 위해서는 규모(Size)보다 연결(linkage)이 중요하다. 특정 주제 분야 다수 사이트의 정보 연결망이 될 수 있는 웹 사이트를 '인포메이션 허브', 즉 정보 연결통로라 일컫는다.

예를 들어 법률분야 전문 검색엔진이 LawCrawler가 전세계 인터넷 사이트에서 어느 정도 연결하고 있는지를 파악해보면 다음과 같다. 세계적으로 유명한 법률 서비스 전문 업체인 West Group의 웹 사이트를 비교대상으로 선정했다. 기업규모로 볼 때 LawCrawler는 West Group에 비교할 수 없을 정도로 작지만 인터넷에서의 인기도[17]는 오히려 LawCrawler가 더 높다는 것을 알 수 있다.[18]

인터넷에서 특정 웹 페이지를 찾아가기 위해 '검색'(데이터베이스나 검색엔진)을 수행하는 것이 가장 일반적임은 많은 조사에서 판명되었다. 그만큼 검색 서비스에 대한 수요가 많다는 것을 증명한다. 이는 미 조지아공대의 GVU센터[19]에서 설문 조사한 결과에서 파악할 수 있다.

그리고 동일한 조사 자료를 보면 전문가일수록 점차 '검색'을 통해 인터넷 웹 페이지를 열람하는 빈도가 높다는 것을 알 수 있다. 따라서 향후 인터넷을 이용하는 사용자가 점차 확대되고, 이용 시간이 늘어날수록 검색 도구의 중요성이 커질 것으로 예상할 수 있다.

2. 웹 사이트 홍보 효과의 극대화

인터넷 이용자들의 인터넷 이용목적 중 가장 큰 부분이 정보검색이다[20]. 이 경우 특정 업종 영역을 차지하는 기업체가 관련 산업정보를 별도의 데이터 생성 노력없이 웹 자원으로 조합한 결과를 제공하는 것은 저비용으로 높은 홍보효과를 거둘 수 있는 좋은 방안이 된다.

즉, 이용자들이 얻고자 하는 정보를 미리 제공함으로써 자연스러운 접속을 유도할 수 있다. 인터넷의 이용목적이 정보검색, 또는 기업 정보를

〈표 3〉 LawCrawler와 West Group 비교

링크 검색	LawCrawler.com	WestGroup.com
AltaVista	4,232	2,680
Infoseek	2,038	2,409
평균	3,135	2,545

〈표 4〉 이용자 수준에 따른 '검색' 이용도 차이 조사

	Books	Directories	Friends	Other	Print	Search	Sigs	TV	Usenet	WWW Pages
Noyice	29.36	64.77	51.8	24.08	59.44	72.96	28.6	43.35	23.11	84.2
Inter	27.01	64.74	60.61	22.62	64.08	85.19	32.06	34.71	27.04	89.2
Expert	26.64	66.13	64.08	27.66	62.9	88.32	40.03	31.57	41.54	90.08

Source : GYU's Elghth WWW User Survey™ (Conducted October 1997)
(URL : Http : //WWW.gyu.gatech. edu/user_surveys/)
Copyright 1997 GTRC - ALL RIGHTS RESERVED
Contact : www-survey@cc.gatech.edu

〈표 5〉 인터넷 사용자들의 이용 목적 조사 (GVU 센터)

인터넷 사용목적	All	Male	Female	USA	Europe	10-18	19-25	26-50	50+	Novice	Intermediate	Expert
통신	2371.00	1470.00	872.00	2004.00	616.00	152.00	417.00	1397.00	295.00	904.00	822.00	616.00
	32.84%	32.54%	33.41%	33.43%	30.11%	42.11%	32.94%	33.41%	27.26%	38.19%	30.29%	30.11%
교육	4357.00	2634.00	1670.00	3660.00	1266.00	263.00	843.00	2397.00	643.00	1355.00	1683.00	1266.00
	60.35%	58.31%	63.98%	61.05%	61.88%	72.85%	66.59%	57.33%	59.43%	57.25%	62.01%	61.88%
오락	4720.00	3032.00	1631.00	4066.00	1134.00	308.00	942.00	2636.00	629.00	1813.00	1716.00	1134.00
	65.38%	67.12%	62.49%	67.82%	55.43%	85.32%	74.41%	63.05%	58.13%	76.59%	63.23%	55.43%
기타	646.00	401.00	240.00	571.00	176.00	38.00	100.00	349.00	115.00	259.00	206.00	176.00
	8.95%	8.88%	9.20%	9.52%	8.60%	10.53%	7.90%	8.35%	10.63%	10.94%	7.59%	8.60%
개인정보욕구	5249.00	3295.00	1894.00	4456.00	1509.00	234.00	879.00	3074.00	828.00	1740.00	1940.00	1509.00
	72.71%	72.95%	72.57%	74.33%	73.75%	64.82%	69.43%	73.52%	76.52%	73.51%	71.48%	73.75%
쇼핑	2866.00	1923.00	900.00	2476.00	970.00	90.00	436.00	1801.00	414.00	797.00	1056.00	970.00
	39.70%	42.57%	34.48%	41.30%	47.41%	24.93%	34.44%	43.08%	38.26%	33.67%	38.91%	47.41%
시간 때우기	2903.00	1894.00	982.00	2480.00	815.00	249.00	698.00	1543.00	321.00	954.00	1107.00	815.00
	40.21%	41.93%	37.62%	41.37%	39.83%	68.98%	55.13%	36.91%	29.67%	40.30%	40.79%	39.83%
업무	3949.00	2585.00	1317.00	3130.00	1620.00	80.00	657.00	2594.00	465.00	651.00	1631.00	1620.00
	54.70%	57.23%	50.46%	52.21%	79.18%	22.16%	51.90%	62.04%	42.98%	27.50%	60.10%	79.18%

16) (Traffic, Transaction : 사이트 접속을. 일반적으로 Traffic은 bytes 전송량으로, Transaction은 접속 횟수로 계산한다)
17) 링크의 정도를 인터넷 이용자들의 인기도로 간주할 수 있다. 논문의 인용 횟수를 중요한 논문 평가 요소로 간주하는 것과 동일한 이치로 해석링크의 정도를 인터넷 이용자들의 인기도로 간주할 수 있다. 논문의 인용 횟수를 중요한 논문 평가 요소로 간주하는 것과 동일한 이치로 해석
18)WestGroup 사이트도 낮지 않은 수치가 나오는 것은 검색엔진은 보유하고 있지 않지만 Law Hubs 수준의 정보 제공이 이루어지고 있기 때문이다.
19) GVU센터에서는 1994년 1월부터 현재까지 1년에 2회 정기적(1995년부터 매월 4월과 10월)으로 인터넷 이용자 조사를 수행하고 있다.
인터넷 홈페이지 주소는, http://www.gvu.gatech.edu/user_surveys/
20) CyberAtlas(http://www.cyberatlas.com), I-PRO(http://www.ipro.com)

얻기 위한 것임은 한국도 예외가 아니어서 지난 1997년 7월 한겨레신문사와 LG애드가 조사한 바에 따르면 아래 그림과 같이 이용자의 37.6%가 인터넷을 기업정보를 검색하기 위해 사용하는 것으로 나타났으며 이러한 현상은 인터넷 전세계 이용자를 대상으로 조사한 설문에서도 그대로 드러난다.

아래 표는 미 조지아 공대에서 지난 1997년 10월 조사한 결과로 지역별, 남녀별, 그리고 연령별 구분없이 인터넷 이용 목적이 일(Work)과 개인적인 정보 욕구(Personal Info)라는 것에서 알 수 있다.

3. 해당 주제(산업) 분야에서 소스맵 보유 기관의 전문성 강조

인터넷 이용자들은 기업의 전문성을 판단하는 중요한 변수로 웹으로 접근 가능한 컨텐트에 둔다. 따라서 이들을 향한 전문 정보 서비스가 해당 기업의 전문적 이미지 고양에 많은 역할을 할 수 있다. EPRI의 EnergySearch는 미 버지니아공대 도서관에서 추천한 엔지니어링분야 검색 데이터베이스에 INSPEC, Chemical Abstract 등과 나란히 등록되어 있다.[21]

4. 높은 자산 가치를 지니는 검색 서비스

인터넷 웹 사이트의 자산가치 평가는 일반적으로 '방문자 수', 즉 사이트 인지도 및 인기도에 근거한다.[22] 미국의 인터넷 시장조사 전문 업체인 사이버아틀라스에서 RelevantKnowledge사의 조사를 인용하여 작성한 'Top 25 Web Properties'는 자산 가치가 높은 25개 사이트를 보여준다. 순위별로 각 사이트의 면면을 보면 다음과 같다. 25개 가운데 9개를 제외한 16개 사이트가 모두 다양한 분야의 정보검색 서비스를 전문적으로 제공하고 있음을 알 수 있다.

5. Target 광고 게재

Traffic 증가와 함께 실현 가능한 것이 광고 유치이다. 일반적인 광고 단가를 낮게 산정하여 $20 CPM으로 볼 경우 하루 5만회의 Transaction으로 (이론적으로 볼 때) $1,000 광고 유치가 가능해 진다.

인터넷의 광고 시장 규모는 매년 2.5배 규모로 성장하고 있으며[23] (Forrester Research, Jupiter Communications, Internet Advertising Bureau), 1997년 3/4분기까지 9개월간 인터넷 광고 매출 규모는 5억7천1백만 달러이며 이는 1996년도 전체의 2억6천7백만 달러보다 2배 이상 높은 수치이다.

인터넷의 배너 광고는 일반적으로 대부분 검색엔진이 ROI(Run of Impression, 또는 General Impression)형태와 Target Banner 두 가지 종류로 서비스중인데 이 가운데 후자의 경우 높은 CPM(Cost Per Thousand Impression) 단가를 유지하고 있으며 본 검색 시스템은 Target 광고에 적합하다.

아래 표는 검색엔진 Lycos에서 적용하는 광고 단가로 General Impression과 Target Impression의 광고 단가 차이를 비교하고 있다. 작게는 평균 33%, 크게는 60% 이상 높은 단가로 산정되어 있음을 알 수 있다.

7. 일반 검색엔진의 무작위 Roboting이 지니는 단점 극복

사이트별 검색엔진 등록현황 비교를 통한 전문 검색 시스템의 차별화를 꾀할 수 있다. 아래 표는 개발 주제 분야를 에너지/전력 산업으로 설정한 것으로 가정하고, 비교적 인해도 높은 '에너지' 분야 사이트를 일반 검색엔진이 어느 정도 검색할 수 있는지를 보여준다.

실제 보유하고 있는 페이지[24] 에 비해 검색엔진에 등록된 URL[25] 이 현저히 낮음을 알 수 있

〈표 6〉 Top 25 Web Properties 선정 사이트

순위	서비스(사이트)명	검색 서비스	분야
1	Yahoo!/Four11*	☑	일반 검색엔진, 전문(인물)
2	Netscape*	·	
3	Excite/Webcrawler*	☑	일반 검색엔진
4	Microsoft*	·	
5	AOL.com	·	
6	Lycos/Tripod*	☑	일반 검색엔진
7	Geocities	·	
8	MSN/Hotmail*	·	
9	Infoseek	☑	일반 검색엔진, 전문(뉴스)
10	CNET*	☑	전문(뉴스, 소프트웨어 등)
11	Walt Disney/ABC/ESPN*	·	
12	Alta Vista*	☑	일반 검색엔진
13	ZDNet*	☑	전문(컴퓨터산업, 기업)
14	WhoWhere/Angelfire*	☑	전문(인물)
15	CNN*	☑	전문(뉴스)
16	RealNetworks*	·	
17	Wired	☑	전문(뉴스)
18	Amazon	☑	전문(도서)
19	Looksmart	☑	일반 검색엔진
20	Mirabilis*	·	
21	Mapquest*	☑	전문(지리)
22	Pathfinder	☑	전문(뉴스)
23	GTE	☑	전문(인물, 기업)
24	The Weather Channel	☑	전문(날씨)
25	Compuserve	·	

〈표 7〉 일반 배너 광고와 Target 광고의 단가 비교

Package	Impression 횟수	General Impression(Month)	Target Impression(Month)
1	1,000,000	$20,000 ($20 CPM)	$32,000 ($32 CPM)
2	750,000	$16,500 ($22 CPM)	$25,000 ($33 CPM)
3	500,000	12,000 ($24 CPM)	$18,000 ($36 CPM)
4	200,000	$6,000 ($30 CPM)	$8,000 ($40 CPM)

21) http://www.lib.vt.edu/Subjects/engr/dbheadin.htm
22)http://www.cyberatlas.com/market/usage/index.html, http://www.relevantknowledge.com/
23) http://www.iab.net/advertise/adsource.html
24) 실제 보유 페이지는 LinkBot을 이용하여 확인한 수치이다.
25) AltaVista의 경우 'host:' 제한검색, Infoseek의 경우 'site:' 제한 검색을 수행했다.

다. 전문 정보 검색 시스템의 메타 인덱스 카탈로그가 일반 검색엔진에 비해 장점 요소로 작용할 수 있는 중요한 요인이다.

4. 유사 서비스 사례 및 확장 가능성

EnergySearch[26], LawCrawler[27]

EPRI(Electric Power Research Institute)에서 제공하는 서비스로 검색 방식은 일반 검색엔진(Infoseek)과 동일하다. 차이점은 에너지 관련 사이트 500개에서 수집한 18만개 URL을 검색할 수 있다는 것인데, 앞서 설명한 Catalog를 이용하지 않는 형태이다(97년 11월 서비스 시작).

에이전트를 이용한 검색 서비스와 법률 관련 디렉토리 주제 서비스를 동시에 제공한다. 본 시스템과 가장 유사한 서비스이지만 Catalog와 에이전트과 연계(combined)되지 않은 형태이다. 즉, 두 서비스가 별도로 제공되고 있다. 최근에는 'LawCrawler International' 이라는 이름으로 각 국가의 법률 전문 검색 서비스로 성장하고 있다.

한편, 확장 가능성은 크게 시스템 자체의 확장 가능성과 제안 서비스의 응용 확장 가능성으로 나누어 볼 수 있다. 결론적으로 두 가지 모두 높은 확장 가능성을 제공한다. 전문정보 검색 시스템의 가장 큰 특징은 무엇보다도 전 산업 및 모든 주제에 대한 확장성이 용이하다는데 있다.

시스템의 중요한 핵심 요소 중 하나인 메타 인덱스 카탈로그가 주제 테이블과 사이트 정보 테이블로 크게 나뉘는데 이 가운데 전자와 후자는 별도로 구축되지만 쉽게 연동될 수 있도록 상호간 결합력(Cohesion)이 낮다.

따라서 주제 테이블이 새로운 주제로 대체되거나, 사이트 정보 테이블이 새로운 필드, 또는 필드 값을 추가하더라도 변경이 용이하며 쉽게 응용할 수 있다. 이러한 점은 최근 데이터베이스 개발 방법론상의 특징이기도 하다. 두 가지 확장 수준과 항목에 따른 확장 용이성을 설명하면 다음과 같다.

〈표 8〉 에너지 분야에 대한 검색엔진 비교

번호	사이트 명	사이트 주소	확인 검색엔진*	등록 페이지수	실제 보유 페이지
1	Energy Intelligence Group	http://www.energyintel.com/	AltaVista	1	123
2	Alternative Energy Engineering	http://www.alt-energy.com/	AltaVista	2	64
3	Clean Air Engineering, Inc.	http://www.cleanair.com	AltaVista	9	327
4	ALSTOM Energy	http://www.energy.alstom.com/	AltaVista	1	95
5	Nuclear Electric Ltd.	http://www.nuclear-electric.co.uk/	Infoseek	44	446
6	Energy & Business Newsletters	http://www.mhenergy.com/main.html	Infoseek	3	147
7	Glossary of Energy Terms	http://www.energy.ca.gov/glossary	Infoseek	5	13
8	California Energy Collection	http://www.energy.ca.gov	Infoseek	925	4,000(about)
9	Energy Info	http://www.energyinfo.co.uk/	Infoseek	10	28
10	Nuclear Energy Institute	http://www.nei.org/	Infoseek	5	310
총계 (8번 제외)				80	1553
비율 (검색엔진 등록 수 : 1.00)				1.00	19.41

〈표 9〉 확장 수준과 항목에 따른 확장 용이성 비교

확장 항목 / 확장 수준	주제(또는 산업) 확장(또는 변경)	사이트 정보 확장(또는 변경)
서비스 자체 확장	주제 분야가 세분화되거나 새로운 상위 주제가 생성되더라도 이미 제공중인 서비스에 영향을 미치지 않고 확장 가능	사이트 정보에 포함되는 필드 및 필드값이 추가되더라도 주제 테이블 및 에이전트로 수집한 전문 인덱스에 영향을 미치지 않음
서비스 응용 확장	새로운 주제/산업 분야(예를 들어 자동차, 철강, 환경, 법률, 국제기구 등)에 대해 응용, 개발할 경우 주제 테이블구조만 수정하면 되므로 프로그램개발비가 초기 개발보다 훨씬 적게 소요된다. 그러나 사이트 정보 수집에는 초기에 투입된 직접 인건비만큼 소요된다. 이러한 직접 인건비 투입은 본 검색 시스템이 부가가치를 갖는 중요한 요소이다.	

26) http://www.energysearch.com
27) http://www.lawcrawler.com