인터넷 WWW검색엔진 비교 평가

신동한
(주)위스정보 대표이사

인터넷에 산재해 있는 각종정보들을 모아 주제별로 게재하며, 이번호에는 본격적으로 주제별 데이터베이스 검색에 들어가기에 앞서 개요를 살펴보았다. <편집자 주>

1. 들어가면서

인터넷은 정보의 바다라고 일컬어지는 것처럼 전세계에 걸쳐 다양한 정보가 유통되고 있는 정보의 보고이지만, 약간 거친 형태의 비정형적인 정보이다.

즉, DIALOG 등 정형화된 상용 데이터베이스에 익숙한 정보검색사들의 입장에서 비교해 볼 때 전반적 통제나 표준 형식, 주제어 색인이 잘 되어있지 않은 비정형적인 다양한 정보의 혼합 서비스이다.

특히 인터넷에서는 각각의 정보제공자들이 정보제공 형식에 제약을 받지 않고 독자적이고 다양한 방식으로 정보를 제공하기 때문에 혼동스러운 면이 있다. 또한 원하는 정보의 소재를 파악하기 어려운 면도 있다.

이러한 비정형성과 정보소재 파악의 어려움을 보완하고 이용자들의 정보검색 편의를 위해 제공하기 위해 Yahoo, Lycos, Webcrawler 등 각종 검색엔진 서비스들이 등장하고 있다.

인터넷은 전자메일, 뉴스그룹과 메일링 리스트, FTP, Gopher, WWW 과 Telnet 명령 사용으로 접근할 수 있는 게시판이나 서비스들 등으로 구성되어 있다고 볼 수 있다. 현재 이들 인터넷 자료의 위치를 찾아주는 검색엔진에는 뉴스그룹 자료를 검색해주는 Deja News, 메일링 리스트를 검색해주는 UK Mailbase, Gopher서비스를 검색해 주는 Veronica, FTP 자료를 찾아주는 Archie, WWW 문서를 검색해주는 Yahoo,

Lycos, WebCrawler, Infoseek, Alta Vista, Search Magellan, Savvy Search, Inktomi, Open Text 등이 있다.

한편 BUBL Subject Tree와 같이 인터넷 자료들을 주제별로 분류해 놓은 서비스들도 있다. WWW 검색엔진에는 뉴스그룹이나, FTP 자료도 함께 검색해 주는 것도 있고, 검색엔진과 주제별 분류를 함께 서비스 하는 것도 있다.

여기서는 주요 검색엔진의 사용법과 알아보고 그 특징들을 비교해 보겠다.

가. 메일링 리스트 검색엔진

UK Mailbase lists : 영국의 다양한 주제에 대한 메일링 리스트 검색을 제공한다.

(URL: http://gopher://mailbase.ac.uk/

나. 뉴스그룹 검색엔진

Deja News : 뉴스그룹 자료를 검색해 준다.

(URL: http://www.dejanews.com/forms/dnq.html)

다. 고퍼 검색엔진

베로니카 프로그램(Veronica program) : Gopher 서비스의 요약물(heading) 검색은 가능하나, 텍스트 검색은 안된다.

(URL:gopher://veronica.psi.net:2347/7/)

라. FTP 검색엔진

Archie : FTP file을 검색해 준다.

(URL: telnet://archie.sogang.ac.kr login에서 archie 입력)

마. WWW 검색엔진

웹(Web)은 고퍼와 같이 서버의 통일된 중앙 등록이 없기 때문에 웹 뿐아니라 다른 인터넷서비스까지 정교한 방법으로 주제에 접근할 수 있도록 하는 많은 검색엔진들이 고안되고 있다. 대체로 이러한 것들은 인터넷의 각 사이트들을 돌아다니면서 문서나 정보 집합 (URL, 문서타이틀, 본문의 키워드 등)을 자동적으로 검색하는 crawler나 worm 프로그램을 사용한다.

따라서 검색엔진을 분석함으로써 인터넷 정보에 접근하는 원칙 수립의 가이드로서 사용될 수 있다. 모든 검색 툴을 커버하는 것은 불가능하므로 좀더 체계적이고 많이 사용되고 있는 다음과 같은 4개의 검색엔진을 살펴보고자 하며 검색 기능과 함께 주제별 소장이 되어 있는 것도 2가지 살펴보고자 한다.

검색엔진(URL)	개발자 및 서비스 기관
world Wide Web Worm http://www.cs.colorado.edu/wwww/)	Oliver McBryan, University of Colorado
WebCrawle (http://webcrawler.com/)	Carnegie Mellon, 현재 Microsoft 부분 출자
Lycos (http://www.lycos.com/)	Brian Pinkerton, Washington state, 현재 America Online 소유
Galaxy (http://www.einet.net/)	Einet, Texas Internet services company
Yahoo (http://www.yahoo.com/)	Stadford University, 현재 독자 회사로 운영
InfoSeek (http://guide.infoseek.com/)	Steve Kirsch, InfoSeek Corporation
Open Text (http://www.opentext.com/omw/f-omw.html)	Tim Bray, Open Text Corporation
Alta Vista (http://www.altavista.digital.com/)	Digital Equipment Corporation

사용자들이 적합한 검색엔진을 선택하는 것을 결정할 수 있도록 비교하려는 데에 중점을 두고 데이터베이스의 내용, 검색 특징, 출력 결과물 등을 테이블 형태로 살펴보고자 한다.

2. 데이터베이스의 내용

어떤 데이터베이스에 접근하려면 무엇이 포함

된 것을 명확이 알아야 할 필요가 있으며 아래 표는 색인된 소스의 유형들을 요약해 놓았다.

웹 문서들을 서지적인 용어로 타이틀 (또는 머릿글)과 Uniform Resource Locator 또는 URL(예를 들면, 다음과 같은 형태의 인터넷 주소 http://www.wis.co.kr), 소재지를 명시하는 정보원 등으로 비교할 수 있다.

기 능 \ 검색엔진	Worm	Webcrawler	Lycos	Harvest	Galaxy	Yahoo
telnet					*	
gopher			*		*	
ftp			*			
www	*	*	*	*	*	*
타이틀/머릿글	*	*	*	*	*	*
요약 및 발췌				*	*	
전문(全文)	*				*	
크기		1,000,000 docs	4.2m urls	42,000 objects		
사용자 url등록	*		*			*

Lycos는 2개의 데이터베이스를 제공한다. Main Database는 4백만개 이상의 아이템을 가지고 있으며, 좀 더 작은 것은 약 5천만 아이템을 가지고 있는데 이런 차이의 구분은 명확하지 않다. 인용된 크기는 검색 툴마다 표현되는 것이 다르다. 예를 들면 WebCrawler의 문서 경우에는 URL을 더 많이 포함하게 될 것이다. Lycos 경우는 모든 유형의 WWW 화일을 색인하지 않는다. 즉 이미지나 비디오 압축화일은 색인되지 않으며 각 레코드는 한 파일의 처음 20행만 포함하고 있다. Galaxy와 Yahoo는 링크된 소장 주제내의 자료에서만 색인되며, Galaxy에 포함된 고퍼자료들도 등록된 모든 고퍼들이 아니라 Gopher Jewels의 소장 주제에 포함된 것 뿐이며 텔넷 서비스도 Hytelnet database의 색인들이다. 또한 어떤 시스템 들은 데이터베이스의 향상을 위해 새로운 URL을 사용자들이 등록할 수 있는 것도 있다.

3. 검색 특징

가능한 선택사양과 검색할 수 있는 특징을 요약하면 다음과 같다

부울연산자(Boolean Operators)

기 능 \ 검색엔진	Worm	Webcrawler	Lycos	Harvest	Galaxy	Yahoo
and	*	*	*	*	*	*
or	*	*	*	*	*	*
not			*			

절단(Truncation)

기 능 \ 검색엔진	Worm	Webcrawler	Lycos	Harvest	Galaxy	Yahoo
자동절단			*	*	*	
우측절단				*		
좌측절단	*			*		*
중간절단				*		
근접연산자				*		
소스유형명시					*	

검색필드

기 능 \ 검색엔진	Worm	Webcrawler	Lycos	Harvest	Galaxy	Yahoo
url	*	*	*	*	*	*
머릿글	*	*	*	*	*	*
디스크립트			*	*	*	*
주제어	*		*	*		
전문(全文)				*		
인용 url	*			*	*	
필드명시	*			*	*	*
검색결과수의 명시	*	*	*	*	*	*
대소문자 구분				*		*
불용어(Stopwords)		*			*	

4. 검색 결과 출력

검색에 의한 정보는 검색엔진과 사용자가 출

력 수를 통제 가능한 방법에 따라 다양하다

기능 \ 검색엔진	Worm	Webcrawler	Lycos	Harvest	Galaxy	Yahoo
출력수 랭크		*	*	*	*	
url	*	*	*	*		
인용 url		*				
머릿글	*	*	*	*	*	*
전문(全文)			*			
주제어			*			
문맥				*		
문서크기			*	*	*	
핫 링크	*	*	*	*	*	*
내용명시			*	*		
최대 힛트수 명시		*	*		*	*

5. 마치면서...

인테넷에서의 효율적인 정보검색은 검색엔진의 수행능력도 중요하지만, 제공하는 정보의 레코드 구조와 이용자의 검색 기술도 중요하다. 따라서 각각의 인터넷 정보검색 엔진들의 특성을 잘 이해하고 이용 목적에 맞는 적절한 검색엔진의 선택이 필요하다.

검색 힛트수가 많아 보다 적합도가 높은 정밀한 검색을 원할 경우 Yahoo나 Galaxy를 이용하는 것이 좋고 찾기 어려운 정보는 자료건수를 많이 보유한 Lycos나 Web Crawler를 이용하는 것이 좋다. 또한 각각의 검색엔진에서 제공하는 부울연산자 (AND, OR, NOT) 기능을 활용하면 원하는 정보를 정확하게 찾아내는데 도움이 된다.

검색엔진에 의한 자연어 검색 뿐만 아니라 주제별로 메뉴형태로 분류되어 있는 BUBL 주제구조도 (BULB Subject Tree)와 같은 일람목록을 사용하는 것도 유용한 방법이다. 최근에는 Alta Vista 등 한글자료도 검색할 수 있는 검색엔진이 등장하고 있다.

본 내용은 인테넷 월드와이드웹 http://www.wis.co.kr/에서 최신 정보를 볼 수 있다. 다음호부터는 인터넷의 주제별 검색에 들어가 보기로 하겠다. DC