인터넷 탐색엔진의 분류체계에 관한 연구

A study on the classification scheme of the Internet search engine

김영보, 성균관대학교 문헌정보학과

Kim Young-Bo
Dept. of Library & Information Sci., Sung Kyun Kwan Univ.

인터넷 도구 중의 하나인 탐색엔진은 월드 와이드 웹의 보편화와 함께 중요한 매개체로 자리잡고 있다. 탐색엔진은 서비스 제공형태에 의해 크게 분류체계 제공형과 주제어 검색 제공형으로 나뉘어 지는데, 분류체계 제공형 엔진에 대한 연구는 그 이용빈도에 비해 부족한 편이다. 따라서, 인터넷 이용자의 탐색노력을 줄이는데 보다 유용한 분류체계 제공형 엔진에 대한 연구가 필요하다. 본 연구에서는 분류체계 제공에 중점을 두고 있는 국내외의 대표적인 탐색엔진 6종과 문헌 분류이론인 KDC와 DDC를 선정하여 그 분류체계를 비교·분석하여 적합한 형태의 탐색엔진 분류체계의 모형을 구축하고자 한다.

I. 서 론

1.1 연구의 필요성

인터넷 도구 중의 하나인 탐색엔진(Search Engine)은 월드 와이드 웹의 성장과 함께 중요한 매개체로 자리잡고 있다.

탐색엔진에 대한 지금까지의 연구는 분류이론의 도입에 의한 체계적 분류테이블의 구축에 대한 연구는 상대적으로 적은 편인데, 인터넷 상에 흩어져 있는 수많은 자원을 체계적으로 분류하여 접근하도록 하는 것은 이용자의 탐색노력을 줄이는데 보다 유용한 방법이라고 여겨진다.

1.2 연구의 목적

본 연구의 목적은 탐색엔진에 적용될 수 있는 분류체계를 구축하는데 있다. 이런 목적을 위하여 본 연구에서 구체적으로 밝히고자 하는 것은 다음과 같다.

첫째, 기존의 인터넷 탐색엔진들이 제공하는 분류체계들은 분류의 원칙이 결여되어 있어, 주제의 망라성, 세분성, 계층성 및 항목간 균형성에서 적절하지 않음을 밝힌다.

둘째, 인터넷 탐색엔진들이 제공하는 분류체계와 DDC, KDC의 분류이론을 비교·분석하여 적합한 형태의 탐색엔진 분류체계의 노형을 구축한다.

II. 이론적 배경

2.1 인터넷의 성장과 탐색엔진

인터넷 호스트와 도메인 수는 급격히 증가하고 있다. 국내의 경우, 1993년 12월과 1997년 4월 사이에 각각 1135%, 8851% 증가하였고,[1]

1) 한국인터넷정보센터, 1997. "국내 인터넷 호스트, 도메인 수" http://www.krnic.net/

외국의 경우, 1993년 1월과 1997년 1월 사이에 각각 1229%, 3942% 증가하였다.[2] 이와 함께 호스트에 연결된 서버 역시 기하급수적으로 증가하고 있다.

최종이용자들의 정보탐색을 돕기 위한 도구의 하나인 탐색엔진은 형태에 의해 ①분류체계 제공형 엔진, ②주제어 검색 제공형 엔진, ③분류체계와 주제어 검색 제공의 통합형 엔진, ④메타정보 엔진으로 구분할 수 있다. 분류체계 제공형 엔진이란 인터넷상의 서버들의 정보를 수집한 뒤 계층 구조로 제공하는 형태의 엔진이고, 주제어 검색 제공형 엔진이란 이용자가 부여한 주제어를 중심으로 데이터베이스내의 해당자료를 탐색하는 형태의 엔진이다. 통합형 엔진이란, 두가지 서비스를 동시에 제공하는 형태를 말하며, 메타정보 엔진은 탐색용어를 일단 입력하고 여러 개의 탐색 엔진에 동시에 적용하여 질의하는 방식을 취하는 형태를 말한다. 오늘날의 추세는 대부분의 탐색엔진들이 통합형 엔진으로 발전하고 있으며, 메타정보 엔진의 경우는 인터넷 서비스 제공자들의 사이트에서 제공되는 것이 보통이다.

탐색엔진은 월드 와이드 웹에서 빈번한 접속 회수를 보이고 있는데, 외국의 경우 20여 개가, 국내의 경우 12개가 존재한다. 탐색엔진의 이용도가 높아지는 것은 그만큼 인터넷 상에서 정보를 찾는 과정에 체계적인 도움이 필요함을 역설하는 것이라 하겠다.

2.2 선행연구의 개관

2.2.1 국내의 연구

정영미와 김성은은 탐색엔진의 성능을 평가하였다.[3] 저자는 알타비스타(Alta Vista), 익사이트(Excite), 핫봇(HotBot), 인포식(Infoseek), 라이코스(Lycos), 마젤란(Magellan), 오픈 텍스트 인덱스(Open Text Index), 웹크롤러(WebCrawler), 야후(Yahoo!)를 분석대상으로 선정하여, 각 탐색엔진의 색인 및 탐색 기능과 검색된 문서의 순위부여 방법을 비교한 후, 탐색실험을 통해 검색효율, 중복탐색의 정확도, 탐색결과의 유사도 등을 측정하였다. 연구의 결과에서 대부분의 탐색엔진이 질문의 성격과 작성된 탐색문에 따라 탐색결과에 있어 차이가 있는 것으로 나타났고, 재현율이 전반적으로 낮은 수치를 기록하였다. 또한 각 탐색도구간 유사도가 매우 낮았으며 탐색질문에 따라서도 탐색도구간의 분포가 다른 것으로 나타났다.

이명희는 주제별 디렉토리와 키워드 검색엔진의 검색효율에 관한 연구를 수행하였다.[4] 저자는 분류체계 제공형 엔진인 YAHOO!와 주제어 검색 제공형 엔진인 ALTA VISTA가 대학도서관 이용자들에게 의해 제기된 탐색질문에 대해 얼마나 적합한 문헌을 탐색해 내는지 알아보기 위하여 탐색적 연구를 시도하였다. 저자에 의한 탐색결과는 검색된 문헌의 양, 검색된 적합문헌의 양, 재현율, 정확률의 측정기준에 의해 평가되었는데, ALTA VISTA는 특정적이고 기술적인 용어의 탐색에 적합한 반면 YAHOO!는 평이하고 일반적인 용어의 탐색에 적합한 것으로 드러났다.

2.2.2 외국의 연구

스베노니우스(E. Svenonius)는 온라인 검색에 있어서의 분류체계의 이용에 관해 연구하였다.[5] 저자는 분류체계를 이용하는 것이 적합율과 재현율을 높이고 이용자의 시간을 절약하는데 유용하며, 특히 통계자료와 같은 비 서지데이터베이스의 설계와 관련 색인, 유사 초록, 자연어 용어의 배치시에 중요한 방법이라고 보았다. 이는 분류체계가 용어간 상관관계를 계층적으로 보여주며 의미론적 브라우징에 도움이 되기 때문이며 따라서, 분류체계의 이론과 실제 적용이 필요하다고 보았다.

비자인-고츠(D. Vizine-Goetz)는 1983년 스베노니우스의 연구를 이어받아 인터넷 자원에

---

2) "Internet Domain Survey."
http://www.nw.com/zone/WWW/report.html

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

4) 이명희, 1997. "네트웍 데이터베이스에서의 주제별 디렉토리와 키워드 검색엔진의 검색효율에 관한 탐색적 연구." 한국문헌정보학회지 31:2, 177-197.

5) E. Svenonius, 1983. "Use of classification in online retrieval." Library Resources & Technical Services, 27:1, 76-80.

대한 도서관 분류표의 적용을 연구하였다.[6] 저자는 DDC와 LC를 선정하여 탐색엔진 YAHOO!의 분류체계중 1-10, 35-45 범주와 비교·분석하였는데, 각 항목들의 용어와 하위 구성에 포함된 자원의 수를 조사하여 항목간 균형성과 적절성을 밝혀냈다.

달버그(I. Dahlberg)는 네트워크 환경에서의 분류이론 적용에 관한 연구를 수행했다.[7] 저자는 LC와 DDC를 선정하여 새로운 분류체계로서 적합한지를 분석하였는데, 분류이론의 적용은 각 주제의 분석과 추적 및 계층의 구분에 유용하며 또한 기존에 분류되어 있는 자료의 네트워크 상 재조직에도 필요하다고 보았다.

윌리스(J. Wallis)와 버든(P. Burden)은 월드와이드 웹 자원에 대한 분류체계 기반 탐색을 분석하였다.[8] DDC를 응용한 WWlib과 같은 분류체계기반 탐색엔진이 특히 텍스트기반 탐색에서 우수한 성능을 보였는데, 이를 통해 Automated Classification Engine(ACE) 시스템을 제안하였다. 저자들은 분류체계 기반 탐색엔진설계가 많은 탐색엔진에 적용되어야 한다고 보았다.

III. 연구의 방법

3.1 연구 대상 및 표본의 선정

3.1.1 인터넷 탐색엔진 분류체계 분석을 위한 대상의 선정

본 연구에서는 분석의 대상으로 분류체계 제공에 중점을 두고 있는 국내외의 대표적인 엔진 6종을 선정하여 분석한다.

- 심마니 (http://simmany.hnc.net)
- 애니서치 (http://www.anysearch.com)
- 정보탐정 (http://idetect.kotel.co.kr)
- 줌 (http://zoom.cyso.net)
- 집! (http://www.zip.org)
- Excite (http://www.excite.com)

3.1.2 분류이론 분석을 위한 대상의 선정

본 연구에서는 문헌 분류이론으로서 KDC 4판과 DDC 21판을 선정하여 분석한다.

3.2 연구의 방법

본 연구에서는 연구방법으로서 문헌연구를 채택한다.

IV. 기대되는 결과

본 연구를 위해 탐색엔진 심마니와 Excite의 분류체계 중 컴퓨터와 인터넷 분야의 매칭 테이블을 작성했다.

<표 1-1>에서 보는 바와 같이 두 탐색엔진의 분류 체계는 많은 차이점을 가지고 있었다. 우선 대분류 항목과 보다 세분화된 항목의 수에 있어 현저한 차이가 있었다. 동시에 항목에 수록하고 있는 주제어의 선택에 있어서도 차이가 있었으며, 심마니의 대분류 항목에 수록된 단어가 Excite의 하위 항목에서 나타나기도 했고, 그 반대의 경우도 있었다. 이런 결과는 두 탐색엔진의 분류 체계 구축에 있어서 단어의 선택, 중요도의 부여, 용어집의 활용, 컴퓨터와 인터넷 분야에 대한 이해도의 차이에서 기인한 것이라 하겠다.

연구를 보다 확장해서 6개의 탐색엔진을 비교할 경우 이런 차이는 보다 커질 것으로 기대된다. 동시에 KDC, DDC의 컴퓨터 분야 분류체계와 비교를 할 경우 양자간의 유사성은 보다 낮게 나타날 것으로 기대된다.

---

6) D. Vizine-Goetz, 1996. "Using library classification schemes for Internet resources (Position Paper)." Proceedings of the OCLC Internet Cataloging Colloquium, San Antonio, Texas, January 19, 1996. Dublin, Ohio: OCLC.
http://www.oclc.org/oclc/man/colloq/v-g.htm

7) I. Dhalberg, 1995. "The Future of Classification in Libraries and Networks, a Theoretical Point of Veiw," Cataloging & Classification Quarterly, 21:2, 23-35.

8) J. Wallace and P. Burden, 1995. "Toward a classification-based approach to resource discovery on the Web." Wolverhampton: University of Wolverhampton, School of Computing and Information Technology.
http://www.scit.wlv.ac.uk/wwlib/position.html

<표 1.1> 심마니와 Excite의 항목비교

심마니		Excite	
BBS	Text BBS,Web BBS,인트라넷		
검색시스템 (검색엔진)	DB서비스,Web검색,회사(업체)	Starting Point	
단체,기관	동호회,사설학원,연구소	Education	ComputerScience,ComputerTraining,InternetTraining
		Computer Science	Algorithms,AI,Conferences,DistributedComputing,ECAD,Education,HCI,Organizations,Parallel&Supercomputing,Technical Reports
보안		Security	Anti-Virus,Cryptography,DigitalCash,Encryption,Firewalls,Hardware,SecureHTTP,Software
신문		News&Magazines	CD-ROM,Games,GraphicDesign,Internet,Multimedia,PCs,Telecommunications
웹매거진			
인터넷교실	HTML,WWW,자바(JAVA),정보검색	Internet	AccessProviders,Advertising,Beginner'sGuides,Chat,Consultants,Directories&Searching,DomainRegistration,E-mail,Events,Intranet,Magazines,MailingLists,Organizations,SiteAdministration,Software,Usenet,UserGroups,WebCulture,WebPageDesign
서비스 (전문업체)	Web서버구축,웹호스팅,인터넷접속서비스,정보서비스,홈페이지제작		
출판류	SW,인터넷,종합	Desktop Publishing	ClOipArt,DTPPrograms,Fonts,PostScript,Services,TeX,WordProcessing
프로그래밍		Programming	Basic,C/C++,Companies,Delphi,Fortran,Java,JavaScript,Miscellaneous,ParallelLanguages,Pascal,Perl,PostScript,TeX,VisualBasic,VRML
회사 (업체)	HW,SW,교육,기타,멀티미디어,유통, 컴퓨터그래픽,통신,통합솔루션	Hardware	CDRecorders&Players,Companies,Components,DesktopComputers,DigitalCameras,InputDevices,LaptopComputers,Miscellaneous,Modems,Monitors,NC,Palmtops&PDAs,Parallel&Supercomputers,Printers,Scanners,Security,StorageDevices,Workstations
		Multimedia	Authoring,CD-Rom,Education,Events,Magazines,Organization,Sound,Video,VirtualReality
		Shareware	Archives,Games,Internet,Miscellaneous,ProgrammingLanguages,Windows
		Software	Archives,Companies,DBs,DTP,Educational,Financial,Games,Graphics,Internet,Multimedia,Networking,Platforms,Reviews,Scientific,ScreenSavers,Security,Utilities,VirtualReality,WordProcessing
		Networking	Consultants,Hardware,Intranets,LAN's&WAN's,MobileComputing,Software,Standards
		Telecommunication	Education,ISDN,Magazines,MobileComputing,Software,Telephony,Wireless
메킨토시 (MAC)		Operating Systems	Companies,DOS,Linux,MacOS,MSWindows,Miscellaneous,Netware,Nextstep,OS/2,Unix,VMS,XWindowsSystem
CAD		Graphics	Archives,ClipArt,DesignFirms,Magazines,Organizations,Software
자료실	이미지		
카페			
행사,대회			
기타		Help	BuyersGuides,FAQ's,ProducReviews,TechnicalSupport,UserGroups
		Miscellaneous	