웹 정보자원의 색인과 초록 요소에 관한 연구

A Study on the Elements of Indexing and Abstracting on the World Wide Web

최 재 황(Jae-Hwang Choi)*

목 차

1. 서 론
2. 웹 정보자원의 수집, 색인, 검색
   2. 1 로봇 에이전트
   2. 2 색인 및 저장
   2. 3 검색
3. 웹 검색엔진의 색인과 초록 요소
   3. 1 주요 웹 검색엔진
   3. 2 웹 검색엔진과 색인 요소
   3. 3 웹 검색엔진과 초록 요소
   3. 4 웹 검색엔진과 메타 요소
4. 결 론

초 록

인쇄물의 색인과 초록은 통제어나 시소러스를 이용하여 합의된 방식으로 작성되지만, 웹 정보자원의 색인과 초록은 인간의 작업을 거치지 않고 자동으로 작성된다. 웹 환경에서의 색인과 초록은 인쇄물의 색인·초록과 비교하여 어떠한 과정을 거치며, 어떠한 요소들을 포함하는가에 대한 연구는 매우 의미 있는 일이라고 본다. 본 연구에서는 웹 정보자원의 수집, 색인, 저장, 검색의 과정을 살펴보았고, 6개의 웹 검색엔진을 통하여 색인과 초록작성에 영향을 미칠 수 있는 17개의 색인 요소와 11개의 초록 요소, 그리고 2개의 메타 요소를 조사하였다. 전반적인 웹 정보자원의 색인과 초록에 대한 경향과 전망에 대해서도 살펴보았다.

ABSTRACT

Although traditional printed materials are indexed and abstracted by human beings with tools like thesaurus or controlled vocabulary, information resources on World Wide Web(WWW) are automatically indexed and abstracted without human beings efforts. It would be useful to investigate what major differences are in processes and in elements of indexing and abstracting between traditional printed materials and information resources on WWW. After discussing how WWW search engines work, six major WWW search engines were chosen for this study. Then, 17 indexing elements, 11 abstracting elements, and 2 meta elements were examined. Overall trends and issues for the future development of indexing and abstracting on WWW are also discussed.

* 성균관대학교 및 경기대학교 문헌정보학과 강사
접수일자 1999년 2월 9일

1. 서 론

정보기술의 발전과 함께 정보의 배포와 이용 방법도 변화하고 있다. 정보의 배포와 이용에 있어서 전통적인 방법은 저자가 그의 발견이나 발상을 적어 여러 가지 통신수단을 이용하여 동료에게 알리고, 그들의 의견을 들어 본래의 아이디어를 수정하고 출판의 과정을 거쳐 문헌화하면, 이용자는 일차 혹은 이차 정보원을 통해 그의 메시지를 읽게 된다. 이러한 출판을 통한 전통적 방법이 지금은 저자가 그의 생각을 워드프로세서로 정리하고, 전자우편을 이용하여 동료에게 그것을 알려 그들의 비평을 들어 수정한 다음, 디지털형태로 인터넷에 띄우고, 이용자가 웹 검색엔진을 이용하여 접근하는 방식으로 바뀌어 가고 있는 것이다.

인터넷은 가장 빨리 성장하는 커뮤니케이션과 출판의 매개체이다. 1998년 1월 현재 세계에는 약 3천 만개의 호스트가 인터넷상에 접속되어 있으며 이는 1997년 1월의 1천 6백 만개에 비해 1년 사이에 그 수가 거의 배로 증가 하였음을 말해준다(Network Wizards, 1998). 인터넷을 통한 정보의 배포와 이용은 기존 인쇄물에 대하여 몇 가지 특징을 가진다. 인터넷에 의한 정보의 배포와 이용은 즉각적으로 이루어 질 수 있고, 적은 비용으로, 많은 사람이 정보를 공유할 수 있으며, 배포되는 자료를 저자가 완전히 통제할 수 있다는 점이다(Missingham, 1996). 이와 같은 특징으로 인터넷에는 수천 혹은 수억 개에 달하는 문서나 파일이 존재하게 되었고, 수시로 바뀌고, 사라지고, 이동하게 되었다. 이들 내용의 통제는 어렵게 되었고 일관성을 기대하기도 힘들게 되었다.

색인과 초록은 지식기록의 내용과 물리적 위치에 대한 체계적인 안내이며 대용물이라 할 수 있다. 웹 환경에서 색인과 초록은 인쇄물의 색인과 초록과 어떠한 차이가 있으며, 어떠한 과정을 거치며, 어떠한 요소들을 포함하는지에 대한 조사는 의미 있는 일이라고 본다. 구체적으로 본 연구의 목적은 웹 정보자원의 수집, 색인, 저장, 검색의 과정을 살펴보고, 주요 웹 검색엔진의 색인과 초록작성에 영향을 미치는 요소들을 메타 요소와 함께 살펴보려는 것이다.

2. 웹 정보자원의 수집, 색인, 검색

웹 정보자원의 색인과 초록은 인쇄물의 색인과 초록과는 다른 특징이 있다. 기존 인쇄물의 색인과 초록은 인간의 작업을 통하여 자료가 선정되고, 합의된 방법으로 색인과 초록작성의 과정을 거쳐 인쇄물, CD-ROM, online형태로 저장되고, 이용자는 이들을 다시 인쇄물, CD-ROM, online형태로 자료를 탐색하여 이용하게 된다. 그러나 웹 정보자원의 색인과 초록은 로봇 에이전트라는 프로그램에 의해 자료가 주기적으로 수집되고, 자동색인 과정을 거쳐 웹 데이터베이스에 저장되고, 이용자는 검색엔진에 키워드를 이용하여 관련자료에 접근하게 된다. 본 장에서는 웹 정보자원의 색인과 초록의 과정을 로봇 에이전트, 색인과 저장, 검색의 세 단계로 나누어 살펴본다.

2. 1 로봇 에이전트

사회에는 많은 종류의 에이전트가 있다. 입찰에 대신 참가하여 일감을 가져다주는 에이전트, 주식시장에서 적당한 주식을 사고 팔아주는 에이전트, 심부름을 대신해주는 에이전트 등 사람의 일을 대신해주는 에이전트는 많다. 많은 종류의 에이전트 중에서 웹에 흩어져 있는 웹 정보자원을 자동으로 찾아주는 에이전트가 로봇 에이전트(robot agent)다. 에이전트의 사전적 의미는 대행자 또는 대리자이며 컴퓨터 분야에서는 '작업을 대행해 주는 프로그램'이라는 의미로 사용된다.

일명 '거미'(spiders), '떠돌이'(wanderers), '벌레'(worms), '개미'(ants)라고 불리는 로봇 에이전트는 웹 상의 수많은 사이트를 돌아다니며 자료를 모아오는 프로그램이다. 로봇 에이전트들은 웹 페이지에 나타난 하이퍼텍스트(hypertext)링크들을 쫓아다니는 방식으로 새로운 페이지에 접근한다. 즉, URL(Uniform Resource Locator)로부터 문서를 다운로드 받고, 다운로드 받은 문서 속에 포함된 URL들을 추출해 내고, 추출된 URL을 가지고 또 다른 문서들을 다운로드받아 문서를 수집하게 된다. 로봇 에이전트마다 웹 정보자원을 수집하는 방법이 다르기 때문에 저장되는 정보도 각기 다르다.

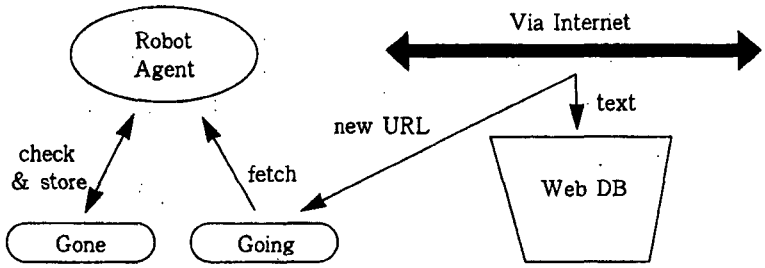
2. 2 색인 및 저장

인쇄물, CD-ROM, online상에 저장된 대부분의 색인정보는 통제어휘나 시소러스를 이용하여 색인되며, 합의된 방식으로 선정된 주제, 저자, 표제, 기관 등이 색인의 주 요소를 이룬다. 이러한 색인은 보통 대학, 학회, 연구소, 정부 등 권위 있는 기관에서 발행한 자료를 대상으로 작성되며 영구적인 성격을 가진다. 그러나 웹 데이터베이스에 저장되는 색인 정보의 대부분은 인간의 합의와는 관계없이 로봇 에이전트라는 프로그램에 의해 자동으로 수집된 HTTP(Hypertext Transfer Protocol), FTP(File Transfer Protocol), gopher, 유즈넷 뉴스 그룹(Usenet News Group), 이미지, 비디오, 영상 자료 등을 대상으로 하며 일시적인 성격을 가진다.

로봇 에이전트들은 인터넷을 돌아다니다가 방문한 웹 페이지의 정보를 자신이 가지고 있지 않으면 그 정보를 수집한다. 물론 방문한 웹 페이지의 정보를 이미 가지고 있으나 그것이 갱신된 상태이면 갱신된 내용만 등록시키고 나머지는 무시하고 지나간다. 로봇 에이전트를 이용하여 웹 정보자원을 수집하므로 색인되는 문서의 양이 많다는 것도 하나의 특징이다. 이는 검색결과를 많게 하여 이용자의 선택범위를 넓혀 주는 이점이 되지만 너무 많은 검색결과로 이용자의 판단을 흐리게 하여 정확률(precision ratio)을 감소하게 만드는 단점이 되기도 한다. 로봇 에이전트의 기본적인 자료수집과 저장 알고리즘은 <그림 1>과 같으며, 그 순서와 내용은 <표 1>과 같다(김성훈, 1996).

2. 3 검색

검색엔진의 검색 프로그램은 이용자들의 질의에 대해 구축되어 있는 색인을 바탕으로 검색결과를 보여준다. 즉, 이용자가 질의를 하면 그때마다 인터넷을 검색하는 것이 아니고, 구축된 색인정보로부터 검색을 하는 것이다. 검



〈그림 1〉 로봇 에이전트의 자료수집 및 저장과정

〈표 1〉 로봇 에이전트의 자료수집 및 저장과정

순서	내 용
①	로봇 에이전트가 방문하여야 할 URL을 Going(앞으로 방문해야할 URL의 저장 시스템)에 저장한다.
②	Going에서 URL을 하나 꺼내온다.
③	Gone(이미 방문한 URL의 저장 시스템)을 통해 다녀온 URL인지를 확인한다.
④	다녀온 URL이면 ②로 복귀한다.
⑤	다녀오지 않은 URL이면 이 URL을 Gone에 저장한다.
⑥	URL을 방문하여 새로운 링크가 있으면 Going에 추가 저장한다.
⑦	⑥에서 얻어진 본문을 Web DB에 색인하고 저장한다.
⑧	Going에 아무것도 없을 때까지 ②부터 반복 수행한다.

색엔진마다 구축된 색인 정보의 차이와 검색 알고리즘의 차이로 동일한 질의에 대하여 검색엔진들은 서로 다른 결과를 보여준다.

검색에서 중요한 문제는 검색결과에 순위를 매기는 일이다. 검색결과로 나오는 문서는 다 읽어볼 수 없을 정도로 많기 때문이다. 따라서 검색엔진의 성능평가 척도는 검색결과의 앞부분에 이용자의 의도와 일치하는 문서가 얼마나 많은가에 초점이 맞추어진다. 검색결과의 출력 순위는 검색엔진마다 조금씩 다르지만, 보통 검색어의 문서 내 위치(location)와 빈도수(frequency)에 근거한 점수제를 도입하고 있다(Sullivan, 1998a).

3. 웹 검색엔진의 색인과 초록 요소

여기서는 본 연구에 필요한 주요 웹 검색엔진을 선정하고, 선정된 웹 검색엔진에서 색인과 초록의 요소에는 어떤 것들이 있는지 살펴본다. 웹 정보자원의 색인과 초록작성에서 메타 태그의 역할도 함께 살펴본다.

3. 1 주요 웹 검색엔진

1998년 NEC 연구소(NEC Research Institute)의 한 연구에 의하면 세계에는 약 3억 2천 만개의 색인 가능한 웹 페이지가 존재하며(Lawrence and Giles, 1998a), 이들에 대한 최

신 자료를 광범위하게 담고 있는 주요 검색엔진으로 AltaVista, Hotbot, Northern Light, Excite, Infoseek, Lycos가 있다고 보고하고 있다(Lawrence and Giles, 1998b). 한편, Sullivan(1998b)은 1998년 12월 현재 각 검색엔진의 색인된 웹 페이지 수를 조사한바 있다(표 2의 색인량 참조). Sullivan(1998b)이 밝힌 상위 6개의 검색엔진도 NEC 연구소가 조사한 주요 검색엔진과 일치하고 있다.

〈표 2〉에 나타난 바와 같이 모든 검색엔진이 모든 웹 페이지를 색인하는 것은 아니다. NEC 연구소에서 추정한 3억 2천만(320 백만) 개의 색인 가능한 웹 페이지를 기준으로 살펴보면 상위 6개 검색엔진의 색인비율은 모두 50%를 넘지 못함을 알 수 있다.

색인된 웹 페이지가 많다고 하여 주요 검색엔진이라고 말할 수는 없다. 어느 검색엔진이 많은 양의 자료를 색인하였다면 재현률(recall ratio)은 향상될 지 몰라도 정확률(precision ratio)은 하락할 수 있기 때문이다. 그러나 본 연구의 목적을 이루기 위하여 주요 검색엔진을 선정할 필요가 있으며, 이를 위해 편의상 NEC 연구소(Lawrence and Giles, 1998b)와 Sullivan(1998b)의 연구를 참고하였다. 모두 여섯 개의 검색엔진을 선정하였으며 선정된 검색엔진의 개요와 URL은 〈표 3〉과 같다.

〈표 2〉 주요 검색엔진의 색인량과 색인비율

	AltaVista	Hotbot	NLight	Excite	Infoseek	Lycos
색인량(백만)	140	110	80	55	30	30
색인비율(%)*	44	34	25	17	9	9

*색인비율=(색인량×100)÷320

〈표 3〉 선정된 검색엔진의 개요와 URL

AltaVista	개요	Digital Equipment Co.에서 개발하여 1995년 12월부터 서비스
	URL	http://www.altavista.com/
Hotbot	개요	HotWired와 Inktomi가 개발하여 1996년 5월부터 서비스
	URL	http://www.hotbot.com/
Northern Light	개요	Northern Light Technology에 의해 1997년 8월부터 서비스
	URL	http://www.northernlight.com/
Excite	개요	스탠포드 대학원생 6명이 ArchitextSoftware Co.(후에 Excite Co.로 변경)을 설립하여 1995년 말부터 서비스
	URL	http://www.excite.com/
Infoseek	개요	Steven Kirsch가 Infoseek Co.를 설립하여 1994년 초부터 서비스
	URL	http://infoseek.go.com/
Lycos	개요	카네기 멜론대학의 Dr. Michael L. Mauldin이 개발하여 1994년 말부터 서비스
	URL	http://www.lycos.com/

본 연구에서 분석되는 데이터는 1998년 말까지의 상태를 나타낸다. 이들 데이터는 본 연구자가 직접 분석하여 얻은 것 외에 각 검색엔진의 홈페이지, 뉴스레터, 관련 연구논문 등에서 수집된 것도 포함된다. 데이터 분석에서 각 보고서의 수치와 내용이 상반되는 경우가 있어 어려움을 겪었다. 검색엔진의 홈페이지 내용과 연구논문 내용이 맞지 않는 경우는 검색엔진의 홈페이지 내용을 채택하였고, 연구논문 간에 상반되는 경우는 공백으로 남겼다. 얻을 수 없는 데이터도 역시 공백으로 남겼다. 본 연구의 목적은 각 검색엔진을 분석하여 성능을 평가하기 위한 것이 아니고, 주요 웹 검색엔진을 통하여 웹 정보자원의 색인과 초록 요소들을 살펴보는데 있으며, 데이터의 정확성은 위에 언급한 정보원에 의존하였다.

3. 2 웹 검색엔진과 색인 요소

웹 정보자원의 색인과 관련하여 모두 17개의 색인 요소를 조사하였으며 이들 결과의 요약은 〈표 4〉에 표시된 바와 같다.

3. 2. 1 색인 대상

어떠한 자료들이 색인되는가하는 것이 색인 대상이다. 기본적으로 조사된 모든 검색엔진이 웹 페이지에 대하여 색인하고 있다 (Birmingham,1997; Northwestern University Library 1999; Westera, 1996) (색인 대상의 세부사항은 표 4참조). 조사된 검색엔진 중 가장 최근에 서비스를 시작한 Northern Light는 EBSCO, UMI, H.W.Wilson과 같은 전문 상업 데이터베이스까지 색인 하고 있다. Northern ght의 특이한 점은 도서관의 사서들을 직접 용하여 인터넷으로 제공되는 자료들은 무료 , 상업 데이터베이스를 활용하는 자료는 유 로 운영되고 있다는 것이다(Notess, 1998).

3. 2. 2 색인 필드

색인 필드는 색인 대상(3.2.1 참조) 자료의 느 부분이 색인되는가를 나타낸다. 검색엔 의 색인 필드 조사를 위하여 전문(full- xt), URL, 표제(title), 주석(comment) 태그 의 문장이 조사되었다. AltaVista, Excite, foseek은 전문(full-text)에 대하여 색인하고 지만 Lycos는 전문을 색인하지 않는다)ong and Su, 1997). 조사된 검색엔진 모두 RL에 대하여 색인하고 있으며, Excite만이 제에 대하여 색인하지 않는다. 주석 태그 안 문장은 Hotbot만이 색인하고 있다.

3. 2. 3 하루 색인가능한 웹 페이지의 수

하루 색인가능한 웹 페이지의 수는 검색엔진 하루에 최대로 얼마나 많은 웹 페이지를 색 할 수 있는가를 보여준다. 검색엔진의 하루 인 능력이 높으면 높을수록 특정 정보와 최 의 정보를 수록할 가능성이 높다고 할 수 있 (Sullivan, 1998c). AltaVista, Hotbot, Lycos는 루에 최대 약 천만 개까지의 웹 페이지를 색 할 능력이 있는 것으로 조사되었다.

3. 2. 4 보유자료의 갱신주기

웹 상의 자료는 끊임없이 변화한다. 어떤 자료는 하루 혹은 이틀 밖에 지나지 않은 것이지만, 어떤 자료는 몇 달 혹은 몇 년이 지난 것일 수 있다. 자료의 나이가 검색엔진마다 다

〈표 4〉 웹 검색엔진과 색인요소

항목	색인요소		AltaVista	Hotbot	NLight	Excite	Infoseek	Lycos
①	색인 대상		웹 페이지, 유즈넷 뉴스그룹	웹 페이지, 유즈넷 뉴스그룹	웹 페이지, 주요 뉴스 정보원, 미정부 정보 DB 대부분, EBSCO, UMI	웹 페이지, 2주간의 뉴스그룹, 유즈넷 광고, 웹 사이트 비평	웹 페이지,유즈넷 뉴스그룹,웹 사이트 비평, 주요 뉴스 정보원, E-mail 주소, 회사양력, FAQs	웹 페이지, FTP와 Gopher 링크, A2Z, 웹 사이트 비평
②	색인필드	전문(full-text)	○	--	--	○	○	×
		URL	○	○	○	○	○	○
		표제(title)	○	--	--	×	○	○
		주석(comment)	×	○	×	×	×	×
③	하루 색인가능한 웹 페이지의 수 (백만)		10	10까지	3	3	--	6-10
④	보유자료의 갱신 주기		하루-6주	하루-2주	2-4주	하루-6주	하루-2달	--
⑤	등록 대기 시간		1-2일	2주내	2-4주	2주내	2일내	2-3주
⑥	연결된 자료의 등록 대기 시간		하루-한달	2주	2-4주	6주	1달-2달	2주-3주
⑦	연결된 자료의 추적 차수		제한없음	제한없음	제한없음	제한없음	제한	제한
⑧	프레임 링크와 이미지 맵	프레임 링크	○	×	○	×	×	×
		이미지 맵	○	×	○	×	○	×
⑨	암호 처리 능력		×	×	○	○	○	○
⑩	웹 문서의 링크 인기도		×	○	×	○	×	○
⑪	특정 웹사이트의 갱신 주기		○	○	×	×	○	×
⑫	로봇 배제의 표준 준수		○	○	○	○	○	○
⑬	메타 로봇의 인지		○	○	○	○	○	○
⑭	URL 등록상태 파악		○	○	×	×	○	○
⑮	불용어 제어		○	○	×	○	×	○
⑯	스테밍		×	×	○	×	○	○
⑰	대소문자의 구별		○	△	△	×	○	×

(○=긍정, ×=부정, △=부분적 긍정, --=조사되지 못함)

른 이유는 여러 가지이다. 등록 대기 시간(3.2.5 참조)과 연결된 자료의 등록 대기 시간(3.2.6 참조)에 따라 자료의 최신성이 달라 질 수 있으며, 검색엔진이 얼마나 인기 사이트의 자료를 자주 수집하느냐(3.2.10 참조)에 따라 달라질 수도 있다. 보유자료의 갱신주기는 최신성을 유지하기 위해 얼마나 자주 자료를 갱신하느냐에 대한 척도로써 보유자료의 나이를 나타낸다. 보유자료의 갱신주기는 이용자의 검색시간을 줄여준다는 면과 검색엔진이 최신의 정보를 반영한다는 면에서 중요한 요소이다. 조사된 검색엔진들의 보유자료 갱신주기는 빠른 경우는 하루였고 오래 걸린 경우는 2주에서 2달까지였다.

3. 2. 5 등록 대기 시간

등록 대기 시간은 어떤 자료가 웹 상에 떠워져서 색인되기(검색엔진의 보유자료가 되기)까지의 대기 시간을 나타낸다. AltaVista가 기대할 수 있는 등록 대기 시간은 하루에서 이틀이었고, Northern Light의 경우는 빠르면 2주, 늦으면 4주까지였다.

3. 2. 6 연결된 자료의 등록 대기 시간

색인(등록)된 자료는 보통 여러 개의 관련 자료를 연결(link)하고 있다. 연결된 자료들은 검색엔진에 따라 즉시 색인되기도 하지만, 뒤로 미루어지는 경우도 있다. 이는 검색엔진에 따라 발견된 자료는 즉시 색인하고, 나머지 연결자료의 색인은 이후 작업으로 미루기 때문이다. AltaVista의 경우 연결된 자료의 등록 대기 시간은 빠르면 하루가 걸리고, Infoseek의 경우 늦으면 2개월까지 걸린다.

3. 2. 7 연결된 자료의 추적 차수

연결된 자료의 추적 차수(depth)는 연결된 자료들을 얼마나 깊이 따라가서 색인 하는가를 나타낸다. 검색엔진에 따라 추적 차수에 제한을 두지 않고 끝까지 관련자료를 추적하여 색인 하기도 하고 추적 차수에 제한을 두어, 자료의 중요도에 따라 차수를 정하여 색인하기도 한다. 추적 차수에 제한을 두지 않더라도 프레임이나 이미지 맵(3.2.8 참조)과 같은 물리적인 방해물에 의해 색인되지 못하는 경우도 있다. AltaVista, Hotbot, Northern Light, Excite는 추적 차수에 제한을 두지 않고, Infoseek, Lycos는 추적 차수에 제한을 두고 있다.

3. 2. 8 프레임 링크와 이미지 맵

프레임(frame)을 이용하여 작성된 웹 문서나 이미지 맵(image map)은 보통 색인과 탐색을 어렵게 만든다. 검색엔진이 프레임으로 연결된 자료나 이미지 맵들을 색인할 수 있는가하는 문제는 매우 중요하다. 이들을 색인 할 수 없다면 많은 자료를 잃게 되기 때문이다. AltaVista와 Northern Light는 프레임을 이용해 작성된 웹 문서들을 색인할 수 있고, 이미지 맵에 대해서는 AltaVista, Northern Light, Infoseek이 색인 가능하였다.

3. 2. 9 암호 처리 능력

로봇 에이전트는 ID와 암호가 필요한 서버는 접속하지 못하는 것으로 생각되지만, 특정서버의 ID와 암호를 미리 준비하여 서버로부터 요구가 있을 때 제시하면 불가능하지도 않다. Northern Light, Excite, Infoseek, Lycos는 ID와 암호를 요구하는 웹 정보자원에 접근할 수 있다.

3. 2. 10 웹 문서의 링크 인기도

특정 웹 문서가 다른 웹 문서에 의해 얼마나 자주 인용(연결)되고 있는지를 측정할 수 있다. 웹 문서의 링크 인기도를 색인 여부의 결정수단으로 이용하고 있는 검색엔진으로 Hotbot, Excite, Lycos가 있다.

3. 2. 11 특정 웹 사이트의 갱신 주기

특정 웹 사이트의 갱신 주기를 검색엔진이 파악할 수 있다면 이는 로봇 에이전트의 방문 주기를 설정하는 자료로 쓰일 수 있다. 웹 사이트가 자주 바뀐다면 방문 주기도 이에 따라 빈번하게 해야 할 것이고, 그렇지 않다면 방문 주기를 길게 정할 수 있다. AltaVista, Hotbot, Infoseek은 자료의 갱신 주기를 파악할 수 있다.

3. 2. 12 로봇 배제의 표준 준수

정보 제공자의 입장에서 보면 로봇 에이전트는 달갑지 않은 손님이다. 예를 들어, 이용자에게 서비스를 제공해야할 웹 서버가 로봇 에이전트들의 계속되는 서비스 요청 때문에 과도한 부하가 생겨 주어진 임무를 수행하지 못하게 되는 경우가 생길 수 있기 때문이다. 이에 로봇 에이전트 개발자들은 '로봇 배제의 표준'(A Standard for Robots Exclusion)을 만들어 각 웹 서버 관리자가 robots.txt 라는 로봇의 출입을 통제하는 파일을 둘 수 있도록 하고 있다. 로봇 배제의 표준은 다른 로봇의 접근으로부터 검색엔진을 보호하기 위한 수단으로 쓰인다. 로봇 배제의 표준은 어떤 표준화 단체에서 표준으로 만든 것은 아니지만, 이 표준을 따르는 것이 트래픽을 감소시키는 데 도움을 준다고 하여 따르기를 바라는 것이다. 조사된 모든 검색엔진이 로봇 배제의 표준을 준수하고 있다.

3. 2. 13 메타 로봇의 인지

메타 로봇(meta robot)은 특정 웹 페이지가 색인되는 것을 막기 위하여 쓰이는 특별한 메타 태그다. 메타 로봇은 robot.txt 파일을 만들 수 없는 웹 페이지에 유용한 도구이다. 로봇 에이전트로부터 웹 페이지가 색인되는 것을 막기 위해서는 〈META NAME="ROBOTS" CONTENT="NOINDEX"〉의 문장을 웹 페이지의 헤더 태그(header tag) 사이에 첨가하면 된다(Koster, 1998?). 조사된 모든 검색엔진이 메타 로봇 태그를 인지할 수 있었다.

3. 2. 14 URL 등록상태 파악

검색엔진 중에는 특정 웹 문서가 색인되었는지를 알려주기도 한다. 특정 웹 문서가 색인되었음을 확인할 수 있다는 것은 특수 키워드로 해당 자료를 검색할 수 있음을 의미하기도 한다. 예를 들어, AltaVista와 Infoseek의 경우 url:asis.org처럼 URL을 키워드 박스에 집어넣으면 URL의 등록상태를 파악할 수 있다. Hotbot과 Lycos도 방법은 다르지만 URL의 등록상태를 파악할 수 있다.

3. 2. 15 불용어 제어

불용어(stopwords)란 색인할 때와 검색할 때 검색엔진이 무시해 버리는 단어들을 의미한다. 이들 단어들(주로 전치사, 관사, 접속사)은 색인될 때 저장공간을 줄이고, 검색될 때

검색의 속도를 높이기 위해 제어(무시)되게 된다. 검색어에 불용어를 꼭 사용해야 할 경우는 이중 따옴표(" ")를 사용하면 된다. Northern Light와 Infoseek은 색인과 검색시 불용어를 제어하지 않는다.

3. 2. 16 스테밍

검색엔진에 따라 키워드의 어간(변하지 않는 부분의 형태소)에 근거하여 단어의 변형(어미)을 찾는 경우가 있다. 예를 들어, 검색엔진에 따라 질의어 'swim'이라는 단어로 'swims' 'swimming'도 함께 검색할 수 있는 경우이다. 반대로 'swimming'이라는 단어로 'swim'이라는 어간도 함께 검색할 수 있는 능력, 즉 탐색시 단어의 어간을 포함시키는 능력도 아울러 포함한다. 여섯 개의 검색엔진 중 Northern Light, Infoseek, Lycos가 스테밍(stemming)이 가능하다.

3. 2. 17 대·소문자의 구별

검색엔진 중에는 문자의 입력시 대·소문자를 구별하는 것이 있다. 대·소문자를 구별하는 검색엔진에 있어서 'Library'의 검색결과는 'LIBRARY'나 'library'의 검색결과와 다를 수 있다(Sullivan, 1998d). AltaVista와 Infoseek의 경우 소문자로만 쓸 경우는 아무 문제가 없지만 대·소문자를 혼합해서 사용할 때는 주의해야 한다. Northern Light는 표제형과 혼합형의 문자들을 쓸 때, Hotbot은 혼합형의 문자들을 쓸 때 주의해야 하고, Excite와 Lycos는 대·소문자에 관계없이 자유로이 문자를 사용해도 된다. 이를 도표화하면 〈표 5〉와 같다. 여기서 "O"은 대·소문자를 구별하지 않는 경우를 나타내고, "X"는 대·소문자를 구별함을 나타낸다.

3. 3 웹 검색엔진과 초록 요소

웹 정보자원의 초록에서 '초록'(abstract)이라는 말은 흔히 '요약'(summary), '배열'(display), '기술'(description), 혹은 '추록'(extract)이라는 말과 함께 쓰이지만 본 연구에서는 '초록'이라는 말로 이들을 대신 한다. 모두 11개의 웹 정보자원의 초록요소를 조사하였으며 조사 결과의 요약은 〈표 6〉에 표시된 바와 같다.

3. 3. 1 표제 길이

표제 길이(title length)는 표제 태그로부터 나타내주는 문자수를 나타낸다. 조사된 검색엔진이 보여주는 표제의 길이는 60문자(Lycos)에서 115문자(Hotbot)까지 다양하다.

〈표 5〉 유형별 대·소문자의 구별

유형 \ 검색엔진	AltaVista	Hotbot	NLight	Excite	Infoseek	Lycos
소문자형(library school)	O	O	O	O	O	O
대문자형(LIBRARY SCHOOL)	X	O	O	O	X	O
표제형(Library School)	X	O	X	O	X	O
혼합형(LibrarySchool, NeXT)	X	X	X	O	X	O

〈표 6〉 웹 검색엔진과 초록 요소

항목	초록요소		AltaVista	Hotbot	NLight	Excite	Infoseek	Lycos
①	표제 길이(문자수)		78	115	80	70	70	60
②	표제가 없을 경우의 대체 표제		'No Title'	URL	URL	'Untitled'	페이지 첫 째 줄	페이지 첫 째 줄
③	초록 길이(문자수)		150	249	150-200	395	170-240	135-200
④	URL 표시		○	○	○	○	○	○
⑤	날짜 표시		○	○	○	×	○	×
⑥	웹 문서 크기		○	×	×	×	○	×
⑦	적합도		×	○	○	○	○	×
⑧	적합도에 영향을 미치는 요소	자체 평가	×	×	×	○	○	×
		메타 태그	×	○	×	×	○	×
⑨	유사 자료의 연결		×	×	×	○	○	○
⑩	검색 결과 수의 옵션		10	10, 25, 50, 100	25	10, 20, 30, 40, 50	10, 20, 25, 50	10, 20, 30, 40
⑪	출력 옵션		text-only	full description, brief description, URLs only	default (기본)	summaries, titles only, list by web site	summaries, titles only, sort by date, sort by score, group results, ungroup results	titles only, URLs only, selected web site

(○=긍정, ×=부정)

3. 3. 2 표제가 없을 경우의 대체 표제

표제 태그에 표제가 없을 경우 표제를 대신하여 나타내는 방법은 검색엔진에 따라 다양하다. AltaVista는 'No Title'로, Hotbot과 Northern Light는 해당 URL로, Excite는 'Untitled'로, Infoseek과 Lycos는 웹 페이지의 첫째 줄로 표제를 대신하고 있다.

3. 3. 3 초록 길이

일반적으로 초록은 웹 문서에서 발견된 문장의 처음부터 작성해 나가지만 웹 검색엔진이 메타 태그를 인지할 수 있는 경우는 디스크립션 메타 태그로부터 작성된다(3.4.2 참조). 초록의 길이는 135자(Lycos)에서 395자(Excite)까지 검색엔진에 따라 다양하다.

3. 3. 4 URL 표시

특정 웹 문서에 접근하고 문서의 소재를 지정하는 수단으로 URL이 사용된다. 조사된 모든 검색엔진이 URL을 초록에 포함하고 있다.

3. 3. 5 날짜 표시

검색엔진 중에는 웹 문서가 만들어지거나 수정된 날짜를 초록에 보여주기도 한다. 이것은 검색엔진의 보유 자료들이 얼마나 최신의 것인지, 또는 얼마나 오래된 것인지를 판단하

게 하는 근거가 되는 것이다. 그러므로 날짜 표시는 검색엔진의 명성과도 관련된 요소라 할 수 있다. 제시된 날짜들이 완벽하다고 할 수는 없다. 웹 문서가 만들어지거나 수정된 날짜를 웹 서버가 정확하게 보내주는 경우, 부정확한 날짜를 보내주는 경우, 날짜를 보내주지 않는 경우도 있기 때문이다. 날짜를 보내주지 않는 경우는 로봇 에이전트가 정보를 수집한 날짜를 표시하기도 한다. Excite와 Lycos를 제외한 검색엔진이 초록에 날짜를 포함하고 있다.

3. 3. 6 웹 문서 크기

검색된 웹 문서의 크기는 Kbyte를 단위로 이들의 크기를 나타낸다. AltaVista, Infoseek은 웹 문서크기를 초록에 포함하고 있다.

3. 3. 7 적합도

검색결과의 적합 정도를 보여주기 위하여 적합도(relevance score)를 초록에 포함하기도 한다. 적합도를 계산하는데 가장 큰 영향을 주는 것은 웹 페이지에서 키워드의 위치와 빈도수다. 정도의 차이는 있지만 모든 검색엔진은 이 두 요소를 적합도 측정에 포함하고 있으며, 여기에 다른 요소들도 영향을 미친다(3.3.8 참조). AltaVista와 Lycos를 제외한 검색엔진이 적합도를 초록에 포함하고 있다.

3. 3. 8 적합도에 영향를 미치는 요소

적합도의 측정방법은 검색엔진 마다 조금씩 다르다. 웹 페이지에서 키워드의 위치와 빈도수의 분석 외에 각 검색엔진은 다른 요소들을 웹 페이지의 적합도 측정에 포함시키고 있다. 그 예가 검색엔진의 '자체평가'와 '메타 태그'다. 자체평가는 검색엔진이 디렉토리 서비스를 함께 제공할 경우 디렉토리에서 평가(review)가 이루어졌는지를 나타낸다. 만약 평가가 이루어졌다면 그렇지 않은 웹 문서보다 더 높은 적합도를 얻게 된다. Excite와 Infoseek이 자체평가를 적합도에 포함하고 있다.

색인과 초록에 대한 메타 태그(3.4 참조)를 적합도 측정에 이용하기도 한다. 즉, 키워드와 디스크립션 메타 태그에서 검색어를 찾으면 본문(text)에서 찾은 것보다 점수가 높게된다. 메타 태그를 인지하는 4개의 검색엔진(AltaVista, Hotbot, Excite, Infoseek)중 Hotbot과 Infoseek이 메타 태그를 적합도 측정에 포함하고 있다.

3. 3. 9 유사 자료의 연결

각 검색결과에 유사자료(관련자료)가 연결되는 경우가 있다. Excite는 'Search for more documents like this one', Infoseek은 'Find similar pages', Lycos는 'similar pages'라는 링크로 이용자에게 유사한 자료를 제시한다.

3. 3. 10 검색 결과 수의 옵션

이용자가 한 번에 볼 수 있는 검색 결과의 수는 검색엔진에 따라 다르다. Northern Light(25개)를 제외한 5개 검색엔진의 기본(default) 결과표시 수는 10개이고, 옵션에 따라 최고 100개(Hotbot)까지 한 번에 볼 수 있다. AltaVista와 Northern Light는 각각 10개와 25개의 결과표시만을 볼 수 있으며 이용자가 이를 바꿀 수는 없다(AltaVista와 Northern Light는 기본 외에 다른 옵션을 제공하지 않는

다). 〈표 6〉에서 기본 결과표시 수는 밑줄로 표시하였다.

3. 3. 11 출력 옵션

검색 결과의 출력 옵션 요소는 검색엔진에 따라 다양하게 선택될 수 있다. Infoseek의 경우, 초록을 포함하는 'summaries', 한 줄의 표제에 의한 'titles only', 날짜와 적합도 순위에 의해 정돈되는 'sort by date'와 'sort by score', 그리고 관련된 검색결과를 묶어주는 'group results'와 그렇지 않은 'ungroup results' 등 많은 출력 옵션을 가지고 있다 (자세한 웹 검색엔진의 출력 옵션은 표 6 참조).

3. 4 웹 검색엔진과 메타 요소

정보를 웹 상에 비교적 쉽게 올려놓게 되고, 축적·검색되는 정보의 양이 기하급수적으로 증가됨에 따라 웹 이용자가 원하는 정보를 얻기는 더욱 어려워지게 되었고, 얻은 정보에 대하여도 만족하지 않게 되었다. 이에 웹 정보자원을 간단하게 기술하고 쉽게 접근 가능하며, 정확한 자원발견에 기여할 수 있는 새로운 관심분야가 나타났는데 이것이 메타데이터에 관한 연구이다. 메타데이터는 일반적으로 '데이터에 관한 구조화된 데이터'(structured data about data)로 정의된다(Iannella, 1998). 본 연구에서는 여러 메타 태그 중에서 웹 정보자원의 색인과 초록에 관련 있는 'Keywords'와 'Description' 메타 태그에 대해서만 살펴본다.

3. 4. 1 색인작성과 키워드 메타 태그

키워드 메타 태그의 역할은 사람이 직접 작성한 키워드로 검색엔진이 자동으로 생산해낼 색인을 일부 대신하여 만드는 것이다. 키워드 메타 태그를 인지할 수 있는 검색엔진은 색인할 때 키워드 메타 태그의 내용을 우선으로 색인하게 되며 문서의 나머지 부분들도 뒤따라 색인하게 된다. 즉, 사람이 작성한 색인은 검색엔진이 자동으로 작성한 색인보다 우선시되며 이에 따라 키워드 메타 태그는 웹 정보자원의 정확성을 높여주는데 기여한다. 예를 들어, 어떤 웹 문서가 resource discovery라는 단어들을 포함하고 있지 않지만, 주제가 resource discovery일 경우, resource discovery라는 검색어로 이 문서가 검색될 확률은 없다. 키워드 메타 태그가 이를 가능케 한다. 키워드 메타 태그의 작성 예는 〈표 7〉과 같고, 여기서 키워드는 〈resource discovery, Z39.50, X.500, urn, urc, metadata, information retrieval, WWW, internet〉이 된다. 세 개의 웹 검색엔진 (AltaVista, Hotbot, Infoseek)이 키워드 메타 태그를 인지할 수 있는 것으로 조사되었다.

3. 4. 2 초록작성과 디스크립션 메타 태그

디스크립션 메타 태그의 역할은 사람이 직접 작성한 초록이 검색엔진에서 자동으로 생산해낼 초록을 대신하여 만드는 것이다. 디스크립션 메타 태그를 인지할 수 있는 검색엔진은 디스크립션 메타 태그의 유무를 먼저 확인하고, 있으면 디스크립션 태그 안의 내용을 초록으로 출력하고, 없으면 본문의 첫 문장부터 초록을 작성하게 된다. 디스크립션 메타 태그의 작성 예는 〈표 7〉과 같고, 여기서 초록은 〈The Resource Discovery Unit researches emerging technologies for the seamless

〈표 7〉 웹 검색엔진과 메타 요소

항목	메타요소		AltaVista	Hotbot	NLight	Excite	Infoseek	Lycos
①	키워드	인지	○	○	×	×	○	×
		예	〈META NAME = "Keywords" CONTENT = "resource discovery, Z39.50, X.500, urn, urc, metadata, information retrieval, WWW, internet"〉					
②	디스크립션	인지	○	○	×	○	○	×
		예	〈META NAME = "Description" CONTENT = "The Resource Discovery Unit researches emerging technologies for the seamless discovery and retrieval of information and services on the internet and WWW"〉					

(○=긍정, ×=부정)

discovery and retrieval of information and services on the internet and WWW〉가 된다. 디스크립션 메타 태그를 인지하는 검색엔진은 키워드 메타 태그를 인지하는 검색엔진들(AltaVista, Hotbot, Infoseek)외에 Excite가 더 있다.

4. 결 론

사서들은 인쇄형태의 자료들을 수집, 조직, 저장, 검색, 배포, 이용하는데 지금까지 중요한 역할을 수행해 왔지만, 현존하는 웹 정보자원의 색인과 초록 도구들은 사서들의 관점에서 만들어 졌다고 말할 수 없다. 사람이 아닌 로봇 에이전트에 의해 정보가 수집되고, 수집된 정보에 대해 통제되지 않은 언어로 접근하는 것은 지금까지 도서관 및 정보센터에서 행해온 업무와는 본질적으로 다른 것이기 때문이다. 웹 데이터베이스가 어떠한 모습으로, 어떻게 작동하는지에 대해 사서 및 정보관리자들은 주의 깊게 살펴보아야 한다. 웹 데이터베이스에서 사서가 기여할 부분이 분명히 있기 때문이다.

로봇 에이전트가 시소러스나 통제어휘에 부합되는 색인을 얻기 위해 웹 문서들을 정밀히 분석하고, 웹 문서에서 키워드가 집중되어 있는 곳으로부터 초록이 만들어지는 때가 올 것이다. 이때에는 사서와 정보관리자가 로봇 에이전트에 의해 수집된 웹 문서에 대해 자료의 선정기준은 만족스러운지, 통제어휘가 적절히 적용되었는지, 색인과 초록이 적절히 작성되었는지 등에 대해 검토하게 될 것이다. 부분적으로 일부 검색엔진은 메타 태그를 이용하여 인간의 노력을 이미 포함하고 있지만 사서와 정보관리자들이 본격적으로 이러한 일을 담당하게 될 시기는 분명히 올 것이라고 본다. 이와 관련하여 본 연구에서는 문헌정보학 측면에서 웹 정보자원의 색인과 초록에 영향을 미칠 수 있는 17개의 색인 요소와 11개의 초록 요소를 2개의 메타 요소와 함께 살펴보았다.

웹 환경에서 사서나 정보관리자의 역할 못지 않게 중요한 것이 메타데이터의 역할이다. 사서들이 수 백년동안 작성해온 도서와 잡지에 대한 목록도 메타데이터의 일종이라고 말할 때 메타데이터는 새로운 것이 아니라고 말할지 모르지만 웹 환경에서의 메타데이터는 검색의 관점에서 정보의 검색효율을 향상시키

는데 보다 많은 초점이 맞추어진다. 앞으로 몇 년간 메타데이터에 관한 연구는 서로 다른 메타 데이터 세트를 결합하고, 메타데이터를 생산하고 관리하는 도구를 개발하며, 메타데이터의 표준을 확장하는데 노력이 집중될 것이다.

참 고 문 헌

김성훈(1996). "까치네 로봇의 원리와 서비스 형태," 웹 코리아 웹 워크샵 자료모음. 제4회 WWW 워크샵 자료. WWW-KR. pp. 219-230.

Birmingham, Judy(1997). "Major Internet Search Engines," [웹 문서]. http://www.stark.k12.oh.us/Docs/search/info.html.

Dong, Xiaoying and Louise T. Su(1997). "Search Engines on the World Wide Web and Information Retrieval from the Internet: A Review and Evaluation," Online & CDROM Review. 21(2): 67-81.

Iannella, Renato(1998). "Mostly Metadata: A Bit Smarter Technology," [웹 문서]. http://www.dstc.edu.au/RDU/reports/VALA1998.

Koster, Martijn(1998?). "The Web Robots FAQ...," [웹 문서]. http://info.webcrawler.com/mak/projects/robots/faq.html.

Lawrence, Steve and C. Lee Giles(1998a). "How big is the Web? How much of the Web do the search engines index? How up to date are the search engines?" [웹 문서]. http://www.neci.nj.nec.com/homepages/lawrence/websize.html.

Lawrence, Steve and C. Lee Giles(1998b). "Tips for Searching the Web," [웹 문서]. http://www.neci.nj.nec.com/homepages/lawrence/searchtips.html.

Missingham, Roxanne(1996). "Indexing the Internet: pinning jelly to the wall?" LASIE. 27(3): 32-42. [웹 문서]. http://www.zeta.org.au/~aussi/resources/conferencepapers/MissinghamR.htm.

Network Wizards(1998). "Internet Domain Survey, January 1998," [웹 문서]. http://www.nw.com/zone/WWW/report.html.

Northwestern University Library(1999). "Beyond Northwestern Internet Resources: Evaluation of Selected Internet Search Tools," [웹 문서]. http://www.library.nwu.edu/resources/internet/search/evaluate.html.

Notess, Greg R(1998). "Northern Light: New Search Engine for the Web and Full-Text Articles," Database. 21(1): 32-37. [웹 문서].

http://www.onlineinc.com/database/awards/award2.html.

Sullivan, Danny(1998a). "How Search Engines Rank Web Pages," [웹 문서]. http://www.searchenginewatch.com/webmasters/rank.html.

Sullivan, Danny(1998b). "Search Engine Sizes," [웹 문서]. http://www.searchenginewatch.com/reports/sizes.html.

Sullivan, Danny(1998c). "Search Engine Features," [웹 문서]. http://www.searchenginewatch.com/webmasters/features.html.

Sullivan, Danny(1998d). "Search Engines and Capitalization," [웹 문서]. http://www.searchenginewatch.com/webmasters/capitalization.html.

Westera, Gillian(1996). "Search Engine Comparison Page: Background Information," [웹 문서]. http://www.curtin.edu:80/curtin/library/staffpages/gwpersonal/senginestudy/sengines.htm.