□ 기술해설 □

국내 웹 정보 검색 기술의 동향

한국통신 권혜진·김영민·김형근·이상엽·정일형
조강래·신봉기*·송주원**·장희순

1. 서 론

월드와이드 웹(웹, WWW) 사용자가 폭발적으로 늘어남에 따라 대량의 정보가 개인, 기업 홍보 또는 상업적인 목적으로 생성되고 있다. 1997년 8월 시점을 기준으로 국내는 약 백만개 수준의 문서가, 세계적으로는 억단위 수준의 웹 페이지가 산재해 있다. 이처럼 웹 정보가 늘어나게 되면 하이퍼링크만을 이용해서 널리 분산된 정보를 모두 찾기는 불가능하다. 뿐만 아니라 웹의 정보가 하루가 다르게 변하고 매일 엄청난 양의 정보가 추가되는 상황에서 사용자는 어쩔 수 없이 정보 검색을 위한 가이드를 절실히 요구하게 된다. 여기서 사용자의 요구는 단순히 주소 기반의 탐색이 아니라 개념 또는 내용 위주의 검색을 의미한다. 그리고 원하는 정보가 지리적으로 얼마나 떨어져 있든 얼마나 널리 흩어져 있든 상관없이 내가 원하는 정보를 한눈에 찾아보기를 원한다. 이와 같은 요구는 1995년을 기점으로 갑자기 나타났다. 그리고 그 욕구는 곧 충족될 수 있었다.

웹에서 사용자들의 정보욕구를 충족시켜줄 수 있었던 것은 바로 정보 검색 엔진이다. 현재 전세계적으로 가동중인 수백개의 웹 정보 검색 엔진에 적용된 기술은 새로이 출현한 기술이 아니라 기존의 정보 검색의 기술이 분산 하이퍼미디어라는 특성에 맞게 응용되어 탄생된 기술이다[1~3]. 이 관점에 따라서 본고에서는 국내의 웹 정보 검색 엔진을 중심으로 한 기술의 발전 동향에 대해서 소개한다. 특히 그 중에서도 언어 처리 기술, 키워드 기반의 검색 기술, 로봇 에이전트 기술 등을 개괄적으로 설명하기로 한다.

본고의 구성은 다음과 같다. 먼저 제2절에서는 현재 국내에서 서비스중인 대표적인 웹 정보검색 엔진을 열거하여 비교 설명하고, 그 다음 제3절에서는 검색 서비스를 구성함에 있어서 웹이라는 특수한 정보원을 위해 고려해야 할 기술적 특징을 설명한다. 그리고 마지막 4절에서 결론을 맺는다.

2. 웹 정보 검색 시스템

현재 인터넷에서 서비스중인 웹 검색 엔진의 종류는 이미 수백개가 넘는다. 이들을 크게 키워드 검색형, 주제별 분류형, 메타 검색형의 세 가지로 나눌 수 있다. 본고에서는 국내의 키워드 검색형 엔진으로 한정하여 대표적인 한글 지원 검색 엔진을 살펴보기로 한다(표 1 참조).

1. 정보탐정(http : //www.idetect.com/)

한국통신 멀티미디어연구소에서 개발된 키워드 검색 엔진으로 국내의 웹 문서와, 유즈넷 뉴스를 대상으로 키워드 검색과 분류 서비스를 제공하며 신문기사 메타 검색도 제공한다. 키워드 검색에서는 일반적인 부울 연산(AND, OR, NOT), 구절 탐색이 가능하며, 기타 절단, 위치, 깊이 연산, 그리고 영역(domain)별, 필드별 제한 검색도 가능하다. 다른 엔진에 비해 상대적으로 검색 속도가 빠르고 영어, 일어 등 외국어 문서도 쉽게 수용할 수 있는 색인 구조를 사용하였다. 유즈넷 뉴스에 대해서도 필드

*정 회 원
**종신회원

표 1 국내의 키워드 검색 엔진의 비교

	정보탐정	심마니	애니서치	웹글라이더	코시크	까치네	와카노
제작사	한국통신	한글과 컴퓨터	삼성물산	삼성 SDS	충남대학교	대구대학교	계명대학교
검색서비스 분류	키워드 검색 주제분류 신문메타검색	키워드검색 주제분류 신문메타검색	키워드검색 주제분류 신문메타검색 헤와메타엔진	키워드검색 주제분류 허의데타검색	키워드검색 주제분류	키워드검색 주제분류	키워드검색
DB 크기	350,000	380,000	40여만	60여만	?	50여만	?
자료종류	웹페이지/유즈넷뉴스	웹페이지/일간신문	웹페이지/일간신문	웹페이지	웹피이지	웹피이지	웹피이지
언어	한글/영어/일어(예정)	한글/영어	한글/영어	한글/영어	한글/영어	한글/영어	한글/영어
검색연산							
부울연산	AND/OR/NOT	AND/OR/NOT	AND/OR/NOT	AND/OR/NOT	AND/OR	AND/OR	연산지원 없음
구문검색	○	○	×	○	×	×	×
검색제한 방 법	위치 깊이 질단 도메인	절단(부분) 도메인	절단(부분)	위치 깊이 짙단 도메인	없음	도메인	없음
결과내 재검색	○	○	×	○	×	×	×
필드별 제한검색	○	×	×	○	×	○	×
시소러스 사용	○	○	×	×	×	○	×
등의어/유사어	○	○	×	×	×	×	×
맞춤법	×	○	×	×	×	×	×
복합명사	×	○	×	○	×	×	×
기타서비스	광고	광고	광고	광고	×	통계서비스	×

*본 표는 검색엔진의 사용 결과와 도움말에 의한 것임

별 연산이 가능하여 원하는 문서를 단시간에 찾아준다. 찾은 결과안에서 재검색할 수도 있다. 주제별 분류의 데이타베이스 구축에 일반 사용자가 직접 참여할 수 있도록 되어있으나, 아직 데이타와 주석이 부족하고 분류 정보를 대상으로 한 검색 기능이 없다.

2. 미스다찾니(http://www.mocahnni.com/)

대표적인 국내의 메타 검색 엔진으로 한글과 영문을 모두 처리한다. 웹 검색은 정보탐정, 유니파인더 등의 엔진에 의뢰하며, 신문기사 검색은 중앙일보, 조선일보 등 국내 6개 신문사의 검색 엔진에 의뢰한다. 국내외의 주요 검색 엔진으로 연결하므로 검색 범위가 상대적으로 넓지만 다른 메타 검색 엔진과 마찬가지로 각 엔진의 특징을 제대로 살리지 못한다는 단점이 있다. 또한 처리시간 설정으로 검색 시간을 제한할 수는 있지만 의뢰한 각 엔진에서 결과가 나온 후에야 사용자에게 그 결과를 알려줄 수 있기 때문에 느린 속도는 어쩔 수 없다.

3. 심마니(http : //simmany.hnc.net/)

한글과 컴퓨터사에서 만든 검색 엔진으로 현재 국내 다른 엔진들에 비해 인지도가 높다. 키워드 검색에는 부울 연산, 괄호 연산, 구절 검색 등이 가능하다. 입력어에 대해 영어로 유의어 확장, 발음 확장이 가능하다. 또한 동적 색인, 신조어 인식이 가능하다고 하며, 검색 영역(domain) 또는 배제 영역을 설정할 수 있다. 국내 여섯개 신문의 기사를 사용자가 원하는 형태로 제공하며, 검색 기간 설정, 출력 방법 선택, 캐쉬 선택이 가능하다. 그러나 다른 엔진에 비해 상세 검색 기능이 부족하다.

4. 까치네(http : //www.kachi.com/)

대구대학교에서 만든 검색엔진으로 국내 초창기 검색 엔진중의 하나이다. 키워드 검색과 분류서비스를 제공한다. 주제별 분류에서는 정보탐정과 마찬가지로 사용자가 웹사이트를 입력, 삭제할 수 있는 열린 서비스를 하고 있다. 또한 특정 URL을 링크하는 사이트들을 역으로 찾아낼 수 있으며 원하는 사이트의 웹지도를 그려주는 서비스도 제공하는 등, 검색 이외의 여러가지 서비스를 구비하고 있다. 하지만 자연어 처리와 부가적인 검색 기능이 적고, 정

보도 웹으로 제한되어 있다.

5. 와카노(http : //www.keimyung.ac.kr/wakano)

계명대의 박민우가 개발한 검색 엔진으로 실시간 색인을 지원하는 것이 특징이다[4]. 와카노 로봇은 영역을 구분하여서 문서를 수집하며, 불용어 처리, 형태소 분석, 사전 검색을 통해서 명사를 추출하며, 검색되지 않는 신조어는 사전에 등록할 수 있다. 아직은 색인 구조의 구현에 치우쳐 있어 데이타가 적고, 지원하는 연산자는 AND, OR 뿐이다.

6. 애니서치(http : //www.anysearch.com/)

삼성전자에서 개발되었던 엔진인 마당발을 토대로 만든 애니서치는 분류 서비스, 국내외 웹페이지 검색(해외 웹페이지는 전적으로 메타 검색을 이용한다) 서비스를 제공하며, 일간 신문도 검색해준다. '대한민국→Korea | 코리아'와 같이 간단한 대용어 확장 색인 방법을 이용하였다. 단점으로는 속도가 느린 편이고, 검색 연산도 부분적으로만 제공된다.

7. 웹글라이더(http://www.infoglider.com/)

삼성데이타시스템 제공 국내 웹사이트 검색 서비스이다. 한국과학기술원(KAIST)에서 개발한 저장 엔진 COSMOS/IR-S[5]과 까치네의 로봇 에이전트를 엮어 엔진을 구성하였다. 형태소 분석, 한글, 한자, 영어 처리, 그리고 복합명사를 분석한다. 일반적인 부울 연산 외에 위치, 깊이, 절단, 영역 제한 등의 연산도 가능하다. 검색어 분포를 기반으로한 랭킹 방법을 쓰며, 한글 키워드 추출, 별도의 한자어 색인, 영어 어근 처리 등의 기법을 적용하였다.

8. 코씨크(http : //kor-seek.chungnam.ac.kr/cgi-bin/korea)

충남대학교에서 만든 엔진으로 국내에서 역사가 가장 길다. 키워드 검색은 최대 6개의 단어까지 가능하다. 기본 검색을 선택하면 가장 상위의 문서들을 결과로 볼 수 있으며 관리자의 간략한 주석이 제공된다. 그러나 양적으로는 결과가 다소 부족하므로 다시 확장 검색을 할 수 있도록 해준다. 확장 검색은 결과를 세부적인 하위 문서까지 검색하기 때문에, 보다 나은 결과를 얻을 수 있다. 한글과 영문 분류 서비스를 동시에 제공하고 있으나, 영어는 전문성이 떨어진다. 링크에 대한 주석과 연산자 기능이 없고 정확한 정보를 찾기 위한 복합 검색 기능이 없는 것도 단점이다.

3. 웹 정보검색 기술 개요

3.1 언어 분석 기술

웹 공간의 정보는 멀티미디어 데이타로 되어 있지만 정보 검색의 초점은 주로 문자로 표현된 언어 정보에 맞춰져 있다. 일반적으로 정보 검색은 사용자의 질의를 분석하여 사용자가 원하는 정보를 찾아주는 것을 말한다. 사용자 질의 속에 담긴 의도를 알아내기 위하여 정보검색 시스템은 질의의 구성과 핵심어의 의미 등을 분석할 수 있어야 하고, 검색 대상이 되는 정보가 무엇에 관한 것이고 사용자의 의도와 얼마나 일치하는지 그 유사성을 판단할 수 있어야 한다. 하지만 투자에 비해 효과가 미미할 뿐만 아니라 시간 제약이 있기 때문에 분석 수준이 높지 않으며, 특히 대량의 정보, 다수의 사용자, 느린 반응 속도 등의 이유로 인터넷 정보 검색 서비스에 따라서는 언어적 분석을 하지 않고 전적으로 단순한 패턴 매칭에 승부를 걸기도 한다.

정보 검색에서 언어처리를 하는 중요한 이유는 정보를 추출할 때 내용을 고려한 검색을 하기 위한 것이다. 즉 사용자 질의와 패턴 매치되는 문서뿐만 아니라 관련 내용의 문서도 일부 찾아주기 위한 것이다. 예를 들어 사용자가 /주식시세/라고 질의하더라도/주식동향/만 포함된 문서도 검색할 수 있으면 바람직할 것이다. 이것이 가능하려면

/주식시세/→/주식/+/시세/
/주식동향/→/주식/+/동향/

으로 분석할 수 있어야 하고 또 /시세/와 /동향/간의 관계를 알고 있어야 한다. 이와 같은 언어처리를 위해서는 계산 알고리즘과 사전, 기타 여러 가지 지식이 필요하다. 한편 언어처리는 검색 엔진의 연산자 및 그 기능에 직접적인 영향을 주기도 한다. 일례로 두 개의 단어가 이어서 나오는 문서를 찾을 때(예 : '가상기업') 복합명사를 분석하지 않는 엔진이라면

(위치 연산자는) 어절을 기준으로 처리한 것과 같은 잘못된 결과를 초래할 수도 있다(예 : '가상현실의 헤드세트제조기업인 A사').

언어처리는 문장이나 구절 형태로 주어진 텍스트로부터 가능한 많은 정보를 추출하는 것을 목표로 하는데, 그 형태는 단어나 구절의 처리에서부터 대용어처리에 이르기까지 그 수준이 다양하다. 그러나 최근의 웹 정보 검색 엔진에서는 언어처리를 위한 알고리즘 개발보다는 사전 및 지식정보에 의존하는 경향이 강세를 이루고 있다. 동의어와 유의어, 발음확장, 맞춤법, 영어 및 한자어 확장 등이 그러한 예이다. 동의어, 유의어, 발음, 언어확장 등으로 불리는 기능은 일명 전거어 사전에 관련된 단어를 저장해놓고 활용하는 방법이다. 예를들어 사전에 /한국통신/과 관련된 단어들로 /한국전기통신공사/, /Korea Telecom/, /KT/, /韓國通信/, /Corea Telecom/ 등을 관련있는 단어로 기록해 놓고 적절히 활용한다. 또한 시소러스를 사용하여 주어진 단어를 개념적으로 확대 또는 축소하여 관련 단어를 추가하기도 하는데 /가축병원/의 /가축/을 /동물/로 확대하여 /동물병원/을 관련 정보로 생산하는 방법을 말한다. 이들 방법이 구축된 사전이나 지식기반에 의존하는 것이라고 한다면 형태소 분석, 구문 분석, 의미 분석 등의 언어분석은 상대적으로 알고리즘에 대한 의존도가 높다. 지금까지 여러가지 알고리즘이 개발되었지만 그 완성도가 다소 떨어지고 부하도 상당하기 때문에 현재 몇가지의 단순 기능 정도만이 사용될 뿐이다. 그리고 이 분야에는 현재 복합명사, 단음절 명사 처리, 용언 활용, 미등록어 처리 기법 실현 등이 이슈로 되어 있다.

한편 최근에는 사용자의 편의를 더하기 위해서 정보 검색 결과에 요약 정보를 자연어로 보여주는 요약기능, 또 문서를 분류체계에 따라 분류하는 자동 분류 기능 등이 추가되고 있는데 이들 모두 언어분석 기술의 고도화를 요구하고 있다. 요약기능은 현재 사용자 질의로부터 주요한 단어를 추출하고 이를 포함한 문장들로 이루어진 요약 정보를 제공하는 수준에 와있다. 그리고 문서 분류는 아직 전적으로 수동 분류에 의존하고 있는데 앞으로 재구성까지 가능한 자동 요약과 자동 분류로 기술이 발전할 것으로 기대된다. 다른 한편으로 자연어를 통한 대화형 검색 방법이 오랫동안 연구되어 왔는데 현재 제한 영역의 응용 시스템에서는 가능하다. 하지만 분야 제한이 없는 웹에 적용하기 위해서는 대용량의 사전, 학습 기능, 담화분석과 같은 고도의 언어 분석 기술이 필요하다. 이를 위하여 보다 체계적인 알고리즘과 사전, 지식베이스를 지속적으로 개발해야 하며 아울러 사용자 모델 등의 연구나 인지과학, 심리학 등 제반 분야에서 실용적인 결과를 내놓아야 할 것이다. 실제로 웹 정보검색 서비스의 최근 동향은 사용자 중심의 주문형 검색, 정보여과 등을 제공하는 방향으로 나아가고 있다. 또 하나의 빠뜨릴 수 없는 것은, 예컨대 한글로 질의하지만 각종 외국어 문서도 검색할 수 있는 다국어 처리 및 검색 기술이 있다. 머지않아 그 비중이 매우 커질 것으로 예상된다.

3.2 색인 및 검색 기술

인터넷 정보 검색에는 전통적인 검색에서는 고려되지 않았던 몇가지 중요한 특징이 있다. 우선 웹문서 내용의 변화가 매우 빈번하다. 또한 표준화된 용어를 기대하는 것도 불가능하고 분야도 제한이 없다. 흔히 정보검색기를 설계할 때 생기는 미등록어 문제, 띄어쓰기 문제, 신조어 문제 등도 심각하며, 외국어 수용문제까지 고려할 필요가 있다.

이와 같은 문제는 고도의 언어 분석이 필요한 '지능형 또는 지식기반' 정보검색 기법들을 사용하기 어렵다는 것을 뜻한다. 실제로 우리의 관찰에 의하면 "한국어처리"를 아주 잘 한다고 주장하는 검색 엔진조차도 점차로 복잡한 언어처리 기능을 일부 포기하고 마치 언어처리를 하지 않은 검색 엔진처럼 동작하도록 바뀌어 가고 있다. 검색어를 선정하는 수준이 거의 N-gram 수준으로 떨어진 것이다. 이런 현상은 결코 기술적인 후퇴를 나타내는 것이 아니라 현실적인 선택이라고 해야 할 것이다. 제어가 불가능한 인터넷 환경에서 극단적으로 제어된 환경에서만 사용하는 기술을 쓸 수는 없기 때문이다. 한편 검색의 범위를 해외로 확장하려면 무수한 외국어를 처리해야 하고 그 범위

를 국내로 한정한다고 해도 상당수의 문서가 영어 전용 또는 한영 혼용으로 되어 있으므로 최소한 영어까지는 수용해야 하는 다국어 처리 문제가 발생한다. 물론 문서가 위치하는 도메인 이름과 문서의 인코딩 방식에 따라서 언어를 판별할 수 있지만[6] 그 문서를 검색 가능하게 만드는 것은 전혀 다른 차원의 문제이다.

인터넷 정보 검색의 경우 문서의 양이 일반적인 정보 검색보다 훨씬 많다는 것도 하나의 특징이다. 세계 최대의 알타비스타는 수천만 문서를 색인한다고 하며, 국내 엔진들의 경우 현재 수십만건의 문서를 색인하고 있는데 조만간 백만건 이상의 문서를 색인하게 될 것이다. 그리고 도서관과 같은 기존의 검색 환경에서는 검색 빈도가 매우 낮기 때문에 검색 속도가 문제되지 않지만, 인터넷 검색의 경우 수많은 사용자가 동시다발적으로 검색엔진을 호출하기 때문에 처리 속도가 매우 중요한 성능 평가 기준이 된다. 이 때문에 검색 서비스중에 새로운 문서를 추가하거나 삭제할 수 있는 방법보다는 이미 색인된 정보에 대해서만 고속 검색을 하고, 색인은 별도의 색인 엔진에서 일괄처리 방식이 일반적으로 사용되고 있다. 일반적으로 어떤 식으로든 검색의 결과를 보장한다는 것은 매우 어렵고 또 사용자들도 반드시 보장받기를 원하지는 않으므로 색인된 정보와 실제 문서 사이에 약간의 불일치는 어쩔 수 없는 것으로 보는 것이다. 물론 문서의 수가 늘어남에 따라 색인된 정보와 실제 문서간의 괴리가 커지고 있지만 아직 검색 속도에 대한 요구가 너무 크기 때문에 문서들의 변화를 즉각 반영해주지 못하고 있는 것이다.

인터넷 정보 검색에서 한층 더 중요한 문제는 검색 결과에 순위를 매기는 일이다. 왜냐하면 검색 결과로 나오는 문서조차도 도저히 다 읽어볼 수 없을 정도로 많기 때문이다. 따라서 검색 엔진의 성능 평가 척도는 검색 결과의 앞부분에 사용자의 의도와 일치하는 문서의 수 등에 초점이 맞춰져야 한다. 정보 검색 관련 분야의 용어로 말하자면 '재현률' 보다는 '정확성'이 훨씬 더 중요한 요소인 것이다.

일반적으로 정보검색 엔진은 문서의 종류가 특정 분야로 한정되어 있고 문서들의 길이나 내용상으로 일정 수준의 동질성이 있어서 검색 결과를 내용(정확히 말해서 문서내 키워드에 관한 통계적 정보)에 따라 우선 순위를 결정하였는데, 인터넷 환경에서는 문서의 구성이나 내용이 너무도 다양하여 내용만으로는 판단이 어려운 경우가 많다. 오히려 문서가 있는 위치 등 구조적인 특징에 따라 우선 순위를 결정하는 것이 종종 더 유리할 수도 있다. 예를 들면, URL만 보고 상위 디렉토리에 있는 문서를 하위 디렉토리에 있는 문서보다 앞에 놓고, 문서의 참조 횟수(즉 외부로 부터의 링크수)가 큰 문서를 우선으로 처리한다든지 하는 것이 더 효과적이다. 그래서 몇몇 검색 엔진에서는 선택적으로 URL상으로 상위 디렉토리에 속하는 문서로 제한하여 검색할 수 있다.

3.3 로봇 기술

몇년 전만 해도 정보 검색이란 어떤 식으로든 다량의 자료가 이미 한 곳에 집중되어 있는 상태에서 그 자료를 대상으로 검색하는 것이 일반적이었다. 그러나 범지구적 분산 환경을 특징으로 하는 인터넷에서는 광범위하게 분산된 자료를 검색해야만 한다. 그리하여 기존의 정보 검색기술을 이용하기 위해서 분산된 문서들을 한곳에 모을 필요가 생겼는데 그 일을 하는 소프트웨어를 흔히 '로봇' 혹은 '웹 스파이더'라고 부른다. 좀 더 정확하게는 '웹 문서 수집 로봇'라고 한다. 인터넷에서 정보 검색 서비스를 제공하려면 일단 웹 문서를 수집해야 하는데, 웹문서를 어떻게 수집하느냐에 따라 검색 결과도 크게 달라진다. 특히 사용자가 원하지도 않는 쓸모 없는 문서를 많이 모아 놓으면 검색 엔진 사용자의 부담을 가중시키기 때문에 문서량이 많다고 해서 반드시 좋은 것만은 아니다.

3.3.1 문서 수집 성능 향상

양질의 웹 페이지 검색 서비스를 제공하기 위해서는 일단 로봇의 문서 수집 속도가 충분히 빠를 필요가 있다. 왜냐하면 주기적으로 웹 페이지가 변경된 상황을 파악하여 사용자들에게 최신 정보만을 제공해야 하기 때문이다. 예를 들어 1주일의 주기로 한번씩 전체 문서의

현행화를 체크한다면, 초당 1개의 문서를 처리해도 하루에 8만 6천, 일주일 내내 작업해도 60만개의 문서밖에 처리할 수 없다.

그런데 1초에 문서하나를 처리하는 것이 쉬워보이겠지만 결코 그렇지 않다. 거기에는 두가지 문제가 있다. 우선 망의 지연을 극복해야 한다. 지리적으로 로봇으로부터 가까이 있는 문서라면 금방 다운로드 할 수 있겠지만 먼곳에 있는 문서를 1초 내에 다운로드하기를 기대하는 것은 현재로선 큰 무리다(또하나의 문제는 웹페이지의 중복성 체크를 위하여 URL DB를 검색해야 하며). 1초내에 그 문서 속에 있는 URL을 뽑아 내고 중복성을 체크하는 것까지 완료해야 한다. URL에는 서버 주소가 있어서 주소의 유일성을 확인하기 위해서는 망을 통해 네임서버를 통해야 한다. 평균적으로 문서 당 URL이 11.2개라는 측정치에 따르면 문서당 네임서버를 평균 11번 이상 호출해야 한다는 것이다. 현재의 망 속도하에서 단일 프로세스로는 도저히 불가능하다.

당연히 다중프로세스나 스레드 기법이 도입되어야 하며, 그렇게 될 경우에는 유일성 체크를 위해서 같은(URL) DB를 서로 다른 여러 프로세스나 스레드가 접근하는 것을 허용해야 하므로 전형적인 자원의 경쟁적 공유문제가 발생한다.

3.3.2 유일성 문제

겉보기에는 서로 다른 URL들이 실제로는 같은 대상을 지칭할 수 있는 이유는 인터넷 주소의 명명법과 웹 서버가 그것을 허용하기 때문이다. 예를 들어 다음과 같이 하나의 컴퓨터를 가리키는 여러개의 주소 스트링과 하나의 파일을 나타내는 여러 가지 방법이 가능하다.

주소 : aistar.kotel.co.kr = idetect.com = www.idetect.com = 147.6.4.185

화일 : / = /index.cgi = /~idetect2/ = /~idetect2 / index.cgi = / ~ idetect2 / front.cgi

현재 4개의 주소명과 5개의 서로 다른 파일명이 있으므로 이들을 조합하면 총 20개의 서로 다른 스트링이 나온다. 이 20개의 스트링 모두가 합법적으로 쓰일 수 있으므로, 20번까지 중복해서 똑같은 문서를 다운로드 할 가능성이 있다. 이와 같은 낭비를 방지하기 위해서는 URL을 DB화하여 중복성을 체크해야만 한다. 한편 네임서버를 과도하게 호출하는 것을 막기위해서 호스트 주소에 대한 캐쉬도 만들 필요가 있으나 호스트 주소가 바뀔 경우 유일성을 체크하기가 더욱 어려워진다는 문제도 안고 있다.

3.3.3 현행화 문제

현행화도 결코 쉬운 문제는 아니다. 우선 망 상황이 매우 자주 변화하므로 많은 URL들이 금방 쓸모 없어진다. 네트웍에 문제가 있어서 현행화하는 시점에 그 URL을 접근할 수 없는 경우도 있고, 그 URL이 실제로 없어져 버렸을 경우도 있고, 혹은 다른 곳으로 이동된 경우도 있을 수 있고, 심지어는 그 URL에 쓰여 있는 호스트의 주소가 바뀌었을 수도 있다.

3.3.4 로봇 배제 표준

로봇 배제 표준은 실상 어떤 표준화 단체, 기관에서도 표준이라고 제정한 것은 아니지만, 그 표준을 따르는 것이 망의 트래픽을 감소시키는데 도움을 준다고 하여 도의적으로 따르기를 호소하는 것이다[7]. 로봇 배제 표준이라는 것이 프로토콜로 구현된 것이 아니기 때문에, 그것을 따르기 위해서는 역시 그것만을 위한 일종의 로컬캐쉬 같은 것을 둬야만 한다. 게다가 최근에는 각 웹 페이지별로 로봇의 행동을 지시할 수 있도록 하는 확장된 표준안이 제기되고 있어서 개발이 점점 복잡해지고 있다.

3.3.5 저작권 문제

웹은 현재 완전히 공개된 공간이므로 웹에 공개된 문서에 대한 저작권 문제가 전혀 정의되지 않은 상태이지만 앞으로 그 문제가 심각하게 제기될 소지가 남아 있다. 예를 들어 특정 신문사의 기사를 로봇을 통해 자동적으로 다운르드 받아서 이차적으로 다른 상업적인 목적에 이용한다면 어떻게 할 것인가? 아직까지 국내에서는 이 문제로 법적인 사건이 일어나지는 않았지만 로봇들이 활개를 치고 또한 상업적으로 이용될 경우는 크게 문제가 될 수 있을

것이다.

4. 결 론

인터넷상의 정보 검색에 대한 다양한 면모와 실제 응용 시스템들에 대해서 알아보았다. 90년대초 인터넷의 팽창이 시작되기 전에 자연언어 처리와 함께 대학 연구실 수준에서 시작되었던 정보검색 연구가 90년대 중반에 들어 인터넷의 팽창과 함께 그 필요성이 갑작스럽게 부각되고 곧 여러 검색 엔진이 속속 개발되었다. 현재는 문서의 양이 너무 많고 대규모로 적용할 수 있는 검색기술이 미흡하여 아직 낮은 수준의 키워드 검색에 머물러 있다. 문서량의 증가 속도가 매우 크기 때문에, 복잡한 계산이 요구되는 세련된 정보검색 기술들을 인터넷에 곧바로 적용하는 것은 쉽지가 않겠지만 이 분야의 연구개발 인력도 빠르게 증가하고 있는 추세이어서 조만간 정교한 검색 엔진이 나오리라 기대한다.

좀 더 기대하자면 정보를 검색하는 것 뿐만이 아니라, 대량의 자료를 분석하고 해석하여 좀더 함축적이고 추상화된 정보로 만들어 내는 것까지 생각해 볼 수 있다. 예컨대 현재 기술로는 미약한 감이 있지만, 앞으로는 지능적인 언어처리 능력과 대량의 자료를 소화할 수 있고 고도의 계산능력을 갖춘 에이전트 등이 실험적인 수준에서 시도할 만한 방향으로 생각된다.

참고문헌

[1] M. Agosti and A. Smeaton(eds), *Information retrieval and hypertext*, 279p, Boston : Kluwer academic publishers, 1996.

[2] T. Berners-Lee and R. Fielding and H. Frystyk, Hypertest Transfer Protocol-HTTP/1.0, Internet Draft draft-ietf-http-v10-spec-04.html, HTTP Working Group, Work in progress, 1995.

[3] T. Berners-Lee and L. Masinter and M. McCahill, Uniform Resource Locator (URL). RFC 1783, Network Working Group, 1994.

[4] 박민우, 실시간 인덱싱 서비스, http : // www.keimyung.ac.kr/ wakano/ rt-index. html, c. 1997. 8.

[5] 한국과학기술원 전산학과 황규영 교수 연구실, http : //dblab.kaist.ac.kr/, c. 1997. 8.

[6] G. Kikui, "Identifying the coding system and Languages of On-line Documents on the Internet," *Proc. of the 16th Int. Conf. on Computational Linguistics*, pp. 652-657, Copenhagen, Denmark, Aug. 1996.

[7] M. Koster, "World Wide Web Robots, Wanderers, and Spiders," found in http : / / info.webcrawler.com / mak / projects/robots/robots.html, c. 1997. 8.

권 혜 진

1991~1995 서강대학교 전자계산학과 졸업
1995~1997 포항공과대학교 전자계산학과 졸업
1997~현재 한국통신 소프트웨어연구소 멀티미디어 연구실 연구원
관심분야 : 한국어 정보처리, 의미분석, 정보검색

김 영 민

1995 연세대학교 전산학과(학사)
1997 한국과학기술원 전산학과(석사)
1997~현재 한국통신 근무(멀티미디어연구소 정보검색연구팀)
관심분야 : 정보검색, 지능형에이전트, Multimodal Interface, HCI.

김 형 근

1993 한국과학기술대학 전산학과 졸업
1995 한국과학기술원 전산학과 졸업
1997~현재 한국통신 멀티미디어 연구소 근무
관심분야 : 자연언어 처리, 정보 검색, 인공지능

이 상 엽

1988 고려대학교 전산과학과(학사)
1991 한국과학기술원 전산학과(석사)
1991~현재 한국통신 근무
관심분야 : 지능형 정보검색, 자연어 처리

정 일 형

1991 동국대학교 전산학과(학사)
1993 포항공과대학교 전자계산학과(석사)
1993~현재 한국통신 근무
관심분야 : 자연어 처리, 지능형 정보검색

조 강 래

1994 경북대학교 컴퓨터공학과 졸업(공학사)
1996 경북대학교 컴퓨터공학과 대학원 졸업(공학석사)
1996~현재 한국통신 멀티미디어 연구소 전임연구원
관심분야 : 에이전트, 정보 검색

신 봉 기

1985 서울대학교 공과대학 자원공학과 졸업
1987 한국과학기술원 전산학과(석사)
1987~현재 한국통신 근무
1995 한국과학기술원 전산학과(박사)
관심분야 : 패턴인식 및 모델링, 지능형 에이전트, 인공지능

송 주 원

1981 경북대학교 전자공학과 컴퓨터전공(학사)
1983 한국과학기술원 전산학과(석사)
1983~현재 한국통신 근무(현재 멀티미디어 연구소 선임연구원)
1992~현재 한국과학기술원 전산학과 박사과정
관심분야 : 데이타베이스 시스템, 공간 데이터베이스, 지리정보 시스템, 멀티미디어 데이타베이스, 다차원 동적 회일임

장 희 순

1977 한국항공대학 전자공학과 졸업(학사)
1980~현재 한국통신 근무, 현재 멀티미디어 연구소 휴먼인터페이스 연구실장
1989 연세대학교 산업대학원 전자계산전공(석사)
관심분야 : 통신망 관리, 교환기 질좋 운용, 멀티미디어, 휴먼인터페이스, 정보검색