정보검색엔진 – 국내기술동향과 발전방향

황 규 영(E-mail:kywhang@cs.kaist.ac.kr)
KAIST 첨단정보기술연구센터 소장/전산학과 교수

Ⅰ. 서 론

최근, WWW(World Wide Web)로 대표되는 인터넷은 정보홍수라는 말을 실감케 할 정도로 빠르게 성장하고 있다. 2000년 1월 기준으로 전세계에 존재하는 웹 페이지의 수는 10억 개 이상이[1] 존재한다고 한다. WWW을 이용하는 사용자 측면에서 이 숫자는 여러 가지 의미를 가진다. 긍정적인 의미로는, 필요로 하는 모든 자료를 WWW을 통해 찾을 수 있다는 가능성을 제시하는 반면, 부정적인 의미로는, 필요로 하는 자료를 찾기 위해 얼마나 많은 시간을 소비해야 하는가라는 문제점을 내포한다. 방대한 WWW이 가지는 이러한 문제점을 극복하기 위해, 필요한 정보를 빠르게 찾아주는 정보검색엔진이 널리 사용되고 있다.

정보검색엔진은 방대한 자료들로부터 사용자가 원하는 자료를 빠르고 정확하게 찾아주는 시스템으로 WWW 검색과 전자도서관, 전자상거래, 전자문서관리와 같은 정보산업 분야에서 널리 사용되는 핵심 시스템이다. 대표적인 정보검색엔진으로는, 국외의 경우, 1994년에 서비스를 시작한 야후!(Yahoo!)를 비롯하여 알타비스타(Altavista), 익사이트(Excite), 라이코스(Lycos)등 수많은 정보검색서비스와 데이터웨어(Dataware), 펄크럼(Fulcrum), BRS, 엑스칼리버(Excalibur), 서치97(Search97), 잉크토미(Inktomi)와 같은 정보검색엔진 패키지들이 있으며, 국내의 경우, 1996년 서비스를 시작한 심마니를 비롯하여 네이버, 한미르, 엠파스등의 정보검색서비스와 오디세우스, 헤이다, 크라스탈, 나모두레박과 같은 정보검색엔진 패키지들이 있다.

WWW의 성장과 정보산업의 발전과 더불어 정보검색엔진에 대한 수요가 급속히 커지면서 그 종류가 다양해지는 이 시점에서, 정보검색엔진을 효과적으로 활용하기 위해서는 정보검색엔진에 대한 보다 정확한 이해가 필요하다. 이를 위해 정보검색엔진을 이루는 구성요소와 핵심 기술을 살펴보고 국내 정보검색엔진 기술의 현황을 살펴 그 이해를 돕고자 한다. 또한 정보검색엔진이 가지는 문제점을 지적하여 앞으로 정보검색엔진의 발전 방향에 대해 생각해보도록 한다.

Ⅱ. 정보검색엔진의 구성 요소 및 기술

정보검색엔진은 문서로부터, 검색대상이 되는 단어

를 추출하고, 이를 색인으로써 구축한 후, 사용자가 원하는 문서를 찾고자 할 때, 색인으로부터 해당 문서를 찾아 결과로 보여주는 시스템이다. 정보검색엔진은 이와 같은 과정들을 처리하기 위해 크게 형태소 분석기, 색인 구축기, 질의 처리기의 세 가지 요소로 구성된다. 각 구성요소와 구성요소에 사용되는 기술에 대해 살펴보자.

■ 형태소 분석기

형태소 분석기는 검색하고자 하는 문서로부터 검색 대상이 되는 단어와 이와 연관된 메타 정보를 추출하는 기능을 수행한다. 정보검색의 결과는 형태소 분석기에서 추출되는 단어의 질에 따라 그 질이 좌우된다. 예를 들어, 검색에 사용되지 않는 단어를 추출하는 경우에는 시스템 자원을 낭비하여 성능을 떨어뜨리는 결과를 초래하며, 필요한 단어를 추출하지 못하는 경우에는, 원하는 자료를 정확히 찾지 못하는 결과를 초래한다.

형태소 분석기에는 정확한 단어의 추출을 위해, 문장 구조를 파악하여 검색대상이 되는 단어, 특히 명사를 뽑아내는 기술이 필요하며 복합명사와 새로 만들어진 명사를 구분할 수 있는 기술이 필요하다. 형태소 분석 단계는 정보검색시스템 구축 작업에서 가장 시간을 많이 차지하는 단계이므로, 빠른 형태소 분석 기능 역시 중요한 기술 중의 하나라 할 수 있다.

■ 색인 구축기

색인 구축기는 형태소 분석기로부터 뽑혀진 단어와 메타 정보를 검색에 알맞은 구조로 재구성하는 단계로, 어떤 색인 구조를 사용하는가에 따라, 구축 속도, 검색 속도, 검색 기능 등의 특성이 결정되기 때문에 검색시스템에서 가장 핵심이 되는 부분이라 할 수 있다. 즉, 색인 구조가 만들어지면 검색 엔진을 완성하였다고 말할 수 있다. 때문에 색인의 구조는 잘 공개하지 않는 것이 일반적인 추세이다.

색인 구조에는 일반적으로 인버티드 파일(Inverted File) 기법에 기반으로 한 여러 가지 변형을 사용하는 것으로 알려져 있다. 인버티드 파일은 색인 단어를 키(key)로 하여 이와 연관된 문서들의 식별자 집합을 빠르게 찾는 구조이다. 인버티드 파일은 그 구조가 매우 복잡하여 구축 시 많은 시간을 요하므로 대용량 데이터 구축을 위한 벌크로딩(bulkloading) 등의 기술을 사용하여 빠르게 처리해야 한다. 또한, 새로운 내용을 기존 색인에 추가하는 등적 수정기능

은 복잡한 색인 구조를 실시간으로 고쳐야 하기 때문에 높은 난이도를 가지는 기술이다. 따라서 이 문제들을 어떻게 잘 해결하느냐가 검색엔진의 유용성을 좌우한다고 할 수 있다.

■ 질의 처리기

질의어로는 일반적으로 불리언 연산을 기반으로 한 키워드 나열 방법을 사용하나, 좀더 사용자와 친숙한 형태의 질의어로 자연어 질의 처리 방식이 시도되고 있다. 질의처리기는 찾을 문서를 정확하게 서술할 수 있는 질의어를 제공해야 하며, 이를 빠르게 처리해 사용자가 이해하기 쉬운 형태로 제시할 수 있어야 한다. 이를 위해 색인구조와 밀 결합된 질의처리기술과 질의 결과 관리 및 표현 기술이 필요하다. 최근 정보검색분야에서는 단순한 키워드 검색이 아닌, 다양한 데이터베이스 속성 조건과 복합된 정보검색이 사용되고 있기 때문에, 데이터베이스와의 밀 결합 기술이 요구되고 있다.

Ⅲ. 국내 정보검색엔진

국내 정보검색엔진은 1996년 심마니를 시작으로 다양한 서비스와 제품들이 나와 있다. 특히 1999년을 기점으로 정보검색엔진 시장은 급성장을 하여 1998년에는 40억원 정도 하는 시장 규모가 1999년에는 150억원 대로 늘어났고, 올해에는 400억원 규모로 급성장[2] 할 것으로 예상되어 각종 제품들이 있는지를 살펴보아 국내 정보검색엔진의 현황을 파악해 보고자 한다. 참고로 여기서 다루고 있는 정보검색엔진 들은 시간 및 지면 관계상 국내에 있는 모든 검색엔진을 모두 소개한 것이 아님을 미리 밝혀둔다.

■ 심마니

1996년 초에 처음 검색서비스를 시작한 이후, 국내에서 가장 오래된 정보검색 서비스로 누적방문사용자 횟수로는 국내 최다를 기록하고 있다. 기본적으로 키워드 기반 불리언 검색을 지원하며, 한글의 특징을 고려한 부분적인 자연어 검색도 가능하다. 특징적으로 동의어, 등음어에 의한 확장 검색을 지원하는데, 이 검색은 주어진 키워드와 같은 뜻을 가지거나 같은 발음을 가진 단어를 포함하는 문서를 찾는 기능으로 다양한 문서를 찾아 준다.

> "국내 정보검색엔진은 1996년 심마니를 시작으로 다양한 서비스와 제품들이 나와 있다. 특히 1999년을 기점으로 정보검색엔진 시장은 급성장을 하여 1998년에는 40억원 정도 하는 시장 규모가 1999년에는 150억원대로 늘어났다. 올해에는 400억원 규모로 급성장할 것으로 예상되어…"

■ 네이버

1997년 웹글라이더로 시작하여 이후, 네이버라는 이름으로 서비스되고 있는 정보검색 서비스로 국내에서 가장 많은 웹 문서를 가장 짧은 주기로 갱신하는 특징을 가진다. 검색 시스템의 저장시스템으로 오디세우스/COSMOS를 사용하여 빠른 검색 및 안정적인 성능을 제공한다. 검색은 기본적으로 키워드 기반 불리언 검색을 지원하며, 다양한 결단연산과 단어의 위치정보를 기반으로 한 위치검색이 가능하여 좀더 정확한 결과를 찾아낼 수 있다.

■ 한미르

한국통신에서 정보탐정이라는 이름으로 시작하여 한미르라는 이름으로 서비스되고 정보검색 서비스로 일본어 검색기능과 전화번호부 검색기능을 제공한다. 검색은 기본적으로 키워드 기반 불리언 검색을 지원하며, 연속 단어 검색과 대 소문자 검색을 지원하는 특징을 가지고 있다.

■ 엠파스

1999년 지식발견연구소에서 AidSearch 검색엔진을 기반으로 작성된 정보검색 서비스이다. 검색은 키워드 기반 불리언 검색과 자연어 검색 기능을 제공한다. 불리언 검색에서는 결단 연산과 위치 검색기능을

제공하며 자연어 검색에서는 "세종대왕과 이순신 장군의 공통점은 무엇입니까" 와 같은 질의문을 통해 관련이 있는 문서를 찾아주는 특징을 가지고 있다.

■ 와카노

1999년 라스21에서 개발한 메타검색서비스이다. 메타검색서비스란 사용자가 준 검색 질의문을 타 검색서비스의 정보검색서비스를 사용하여 검색을 한 후, 그 결과를 재구성하여 보여주는 서비스로 기존 검색서비스들에 대한 동일한 사용 방법의 제공과 부가서비스를 지원할 수 있는 특징을 가진다. 와카노에서 제공하는 가장 특징적인 부가서비스로는 각 검색엔진으로부터 얻어진 결과를 바로 나열하는 방식이 아닌, 관련 있는 내용을 실시간으로 분류하는 온라인 클러스터링 기술이다. 이 기능은 사용자들이 수십만 건 이상의 검색 결과를 효과적으로 파악하는데 도움이 된다.

■ 오디세우스

1995년 한국과학기술원 데이터베이스 및 멀티미디어 연구실에서 개발된 객체지향 멀티미디어 데이터베이스 관리시스템을 기반으로, 정보검색기능을 추가하여 1997년에 개발된 세계 최초의 밀 결합 정보검색용 데이터베이스관리 시스템이다.

2000년에는 대용량 데이터베이스 지원기능을 추가하여 버전 3.0이 발표되었다.

오디세우스는 데이터베이스 관리시스템의 특징인 과손회복, 동시성제어, SQL 표준 질의어, 대용량 데이터베이스 지원, 최적화된 벌크로딩(bulkloading) 기능, 동적 수정기능과 이를 바탕으로 일반 데이터베이스 속성과 밀 결합 정보검색기능을 제공하여 성능과 기능면에서 뛰어난 시스템이다. 검색은 불리언 검색을 기본으로 결단연산과 위치 검색, 가중치 연산 등을 지원한다.

■ 레이더

한국정보공학에서 개발한 정보검색엔진으로 1992년 처음 개발된 이후, 1999에 버전 3.0이 발표되었다. XML, HTML, HWP, DOC, PDF등의 다양한 포맷의 문서를 자동 색인하며, 이종 데이터베이스 시스템과의 연동 및 다중 시소러스 및 전문 시소러스 기능을 제공하여 인터넷뿐만 아니라 인트라넷 및 지식관리 시스템에 적합한 기능들을 제공한다. 검색은 불리언 검색과 자연어 검색을 지원하며 결단연산과 인접어 검색을 지원한다.

■ KRISTAL

연구개발정보센터(KORDIC)에서 개발한 정보검색엔진으로 1991년 텔넷(telnet) 기반의 검색 서비스를 제공하는 KRISTAL-I이 처음 개발된 이후, 1996년 정보검색 엔진인 KRISTAL-II가 개발되었다. SGML 문서 처리 기능과 이종 데이터베이스 시스템과의 연동, 그리고 Z39.50 프로토콜(protocol)을 지원하는 것이 특징이다. 검색은 불리언 검색을 지원하며 결단연산과 근접연산을 제공한다.

■ 나모두레빅

나모인터랙티브에서 개발한 단일 홈페이지 전용 검색엔진이다. 동의어 및 등음어 검색을 지원하며 사용자가 입력한 오타를 자동으로 수정하는 등의 사용자를 배려한 기능들이 많이 있다. 정보검색엔진에서 색인 파일은 원문보다 큰 것이 일반적인데, 나모두레빅은 색인파일이 원문 보다 작은 것이 특징이다. 검색은 불리언 검색을 지원한다.

Ⅳ. 검색엔진의 발전방향

정보검색엔진은 WWW를 효과적으로 활용하게 해주는 필수적인 도구임에 틀림없다. 그러나, 인터넷의 폭발적인 성장에 따라, 정보검색엔진이 얼마나 그 역할을 다 해 줄 수 있는지는 한번 고려 해 볼 필요가 있다.

정보검색엔진은 그 정의에 의하면 방대한 자료들로부터 사용자가 원하는 자료를 빠르고 정확하게 찾아주는 시스템이다. 그러므로, 정보검색시스템은 증가하는 자료의 양에 따라 보다 방대한 자료를 검색할 수 있어야 한다. 또한, 검색을 할 자료의 양이 늘어남에 관계없이 빠른 검색 속도를 제공할 수 있어야 하

며, 검색되는 자료의 양이 증가함에 비례하여 증가되는 검색 결과를 효과적으로 처리할 수 있어야 한다. 즉 정보검색시스템은 대용량 데이터 색인 문제(즉 벌크로딩), 대용량 질의 처리 문제, 그리고 대용량 질의 결과 처리 문제들을 가지고 있으며 이 문제들을 해결하는 것이 앞으로의 발전 방향이라 할 수 있다.

정보검색엔진이 가지는 이들 문제들을 해결하는 방법에는 여러 가지가 있을 수 있다. 대용량 데이터 색인 문제와 대용량 질의 처리 문제를 해결하기 위해서는 보다 효율적인 검색 구조와 저장 시스템의 개발이 필요하며 이에 따른 효과적인 질의 처리 방법이 개발되어야 한다. 또한 필요에 따라 정보검색엔진의 처리 용량을 늘릴 수 있는 방법이 제공되어야 한다. 대용량 질의 결과 처리 문제는 시스템의 처리 용량 및 처리 속도의 증가라는 해결 방법과 병행하여 질의 결과를 쉽게 분석할 수 있는 기능이 제공되어야 한다. 일반적으로 정보검색에 의해 구해진 결과는 그 수가 많기 때문에 분석하기가 쉽지는 않다. 이 문제를 해결하기 위해 질의 결과의 수를 줄이는 방법과 질의 결과를 재구성하는 방법이 개발되고 있다. 질의 결과의 수를 줄여주는 방법으로는 확장 불리언 질의 방법이 있으며, 질의 결과를 재구성해주는 방법으로는 랭킹 기법과 주제별 문서 분류 기법 등이 있다.

V. 결 론

정보검색엔진은 수많은 정보로부터 필요한 정보를 빠르게 찾아주는 시스템으로, 현재 직면하고 있는 정보화 사회에서 정보를 효과적으로 찾기 위해 없어서는 안될 필수 도구이다. 국내 정보검색엔진 기술은 선진국에 비하면 그 시작이 다소 늦어졌지만, 기술적인 측면에서 선진국과 대등하거나 앞서고 있는 부분이 많으며, 앞으로 기하급수적으로 늘어가는 정보의 양을 잘 처리할 수 있는 대용량 색인 구축 기술, 대용량 질의 처리 기술, 대용량 질의 결과 처리기술을 개발, 보유할 수 있다면 세계적인 정보검색엔진을 만드는 것이 불가능하지 않으리라 생각한다.

■ 주

1) Inktomi "Web Surpasses One Billion Documents" Inktomi News and Events Jan. 2000

2) 전자신문, "해즈니스 솔루션, 웹사이트 제작 솔루션", 검색엔진, 메일 및 메시징 솔루션, 전자신문, 2000년 3월