□ 기술해설 □

인터넷 정보검색 서비스 동향

한국통신 신봉기*·김영환**

1. 서 론

월드와이드웹(WWW)이 폭발적으로 팽창함에 따라 각자가 소화해야 할 정보는 그만큼 늘어나면서 오늘을 사는 정보인의 부담은 점점 커지고 있다. 그 부담을 해소해줄 수 있는 정보검색 서비스는 실제로 인터넷과 웹이 갑자기 폭발하던 그 시점부터 있었다. 바로 정보검색 엔진이다. 그런데 실제로는 인터넷 정보검색 엔진이 폭약의 뇌관 역할을 했다고 하는 편이 더 정확할 것이다.

인터넷의 성장은 이제 본격적 궤도에 오르고 이제 단순한 엔터테인먼트 공간이 아닌 사업, 교육, 생활 등 다양한 모습의 공간으로 성장하고 있다. 인터넷은 수많은 벤쳐 기업의 사업 터전이 되었고, 치열한 경쟁의 열기가 전세계적으로 퍼져있다. 이에 따라 정보 검색엔진도 각종 부가의 기능과 정보를 더하여 특징있는 모습의 상용 정보 서비스로 자라나고 있다. 인터넷 정보 검색 서비스는 독자적인 콘텐트를 파는 것도 아니고 소프트웨어를 파는 것도 아니며 정보 검색 서비스 자체를 파는 것도 아닌 독특한 모양을 하고 있다. 본고에서는 이와 같은 정보 검색 서비스의 현황과 앞으로의 발전 방향을 살펴보기로 한다.

21세기에 들어서면 지식 산업이 국가 경쟁력을 좌우하며 선, 후진국을 가름할 것이다. 지식 산업은 어느날 갑자기 성장하는 것은 아니다. 사회 구성원에 의한 지식과 정보의 창출, 가공, 표현, 활용 및 응용 등과 같은 능력이 발전하면서 성장하는 것이다. 이러한 능력은 하루 아침에 이루어지는 것은 아니다. 인간과 기술을 포함한 사회 전반적인 정보화가 고르게 이루어져야 가능한 것이다. 이런 맥락에서 본고에서는 인터넷 정보검색 서비스의 관점에서 본 정보 서비스(기술)의 현황과 요구사항/문제점, 그리고 앞으로 정보검색 서비스 발전 방향에 대하여 간단히 논의하고자 한다.

*정 회 원
**종신회원

2. 검색 서비스 현황

아무리 인터넷 도사라고 하더라도 URL이나 전자우편 주소를 모르면 철자를 모르고 사전 찾는 것과 같다. 따라서 인터넷에 접속한다고 하면 내가 원하는 정보의 위치를 알아야 하고, 그러기 위해서는 누구나 한번씩 검색 서비스 페이지를 방문하게 된다. 검색 서비스 페이지를 거쳐서 인터넷 정보 공간의 어딘가에 있을 웹페이지를 검색하는 수밖에는 별 도리가 없기 때문이다. 최근 이와 같이 여러 사용자가 거쳐가는 개념의 서비스를 포털(portal) 서비스라고 한다. 지금까지 웹페이지를 개설한 여러 대형 콘텐트 제공자들은 가능한 많은 사용자들을 끌어들이기에 혈안이 되어있다. 그리고 그 현상은 인터넷 정보 검색 엔진 홈페이지부터 시작되었다.

지금까지 내노라하는 검색 엔진은 자사의 대용량(색인 DB 크기), 고속처리(질의어 처리 속도) 기술을 강조하는 방향으로 발전해왔다. 그 대표적인 예가 미국 Digital사의 초고속 검색 엔진AltaVista이다. 그 외에도 Infoseek, Lycos, HotBot, Excite 등이 있으며 국내에는

정보탐정, 심마니, 네이버, 까치네, 다찾니 등이 있다[2]. 이들에게 공통된 것이 있다면 검색 엔진으로 키워드 검색만 하는 것이 아니라 분류 정보, 유즈넷 뉴스, 신문 기사 메타 검색, 기타 각종 관심거리를 제공해주는 온라인 PC 통신 서비스 모습으로 발전해 간다는 점이다. 야후도 예외는 아니다. 현재 인터넷 공간에는 검색 서비스 상호간에 포털 서비스로 살아남기 위한 치열한 경쟁이 진행되고 있다.

엔진간의 경쟁은 각 엔진의 정보검색 서비스가 상업화 하면서 본격화되기 시작했다. 각 서비스가 벌이는 경쟁 시장은 그러나 만만치 않다. 인터넷의 특성상 정보검색 서비스가 다수 공존한다기 보다는 대부분 자연도태하거나 합병하여 궁극적으로는 5개 이하 소수개 만이 살아남을 것으로 짐작된다. 정보검색 서비스의 수입원을 살펴보자. 현재 인터넷 공간의 주된 수입은 웹페이지 광고에서 나오고 있다. 광고 수익을 올리는 유일한 방법은 현재로선 가능한 많은 사용자가 접속하여 광고를 많이 보게 하는 것인데, 이를 위해서 모두 인기있는 포털(portal) 서비스를 지향하고 있다.

검색 엔진을 판매하는 것도 점보 검색 서비스 판매 방법 중의 하나라고 볼 수 있다. 이것은 단순히 SW로 개발된 검색 엔진을 판매하는 것이다. 국내에서는 아직 시장이 소규모이지만 곧 크게 성장할 것으로 예상된다. 한편 엔진 판매와는 약간 다르게 전자화된 정보 DB를 다운로드받아 검색 엔진을 구축하고 검색 서비스를 대행해 주는 방법도 있다. 이 경우 시스템 유지보수와 업그레이드도 대행하기 때문에 고객의 관점에서는 매우 편리한 형태이며, 따라서 잠재적 성장 가능성이 크다고 판단된다.

3. 검색 서비스의 문제점

90년대 중반 인터넷과 월드와이드 웹이 갑자기 팽창하고 급부상하게 된 데는 두 가지 요소가 있었다. 쓰기 쉽고 편리한 브라우저가 그 첫째이다. 웹브라우저는 복잡한 분산 하이퍼텍스트를 마우스 클릭 한가지만으로 쓸 수 있게 해준 소프트웨어이다. 웹 사용자가 폭발적으로 늘어나게 된 두번째 이유는 인터넷/WWW 공간 어딘가에 숨겨져 있는 정보를 순식간에 찾아주는 검색 엔진이 있었기 때문이다.

검색 엔진은 과연 수천만개가 넘는 웹 문서를 뒤져 원하는 것을 즉시 찾아주는 경이로운 소프트웨어이다. 특히 초기에는 그랬다. 하지만 이제는 인터넷 정보가 너무 많아졌다. 또 하루가 다르게 변하고 주체하기 힘들 정도로 늘어간다. 현재 순수한 국내 웹문서는 200만쪽, 전세계 웹문서는 대략 3억쪽 정도로 추정된다.

표 1 'latex software' 질의 결과

검색기	OR 질의	AND 질의	구절 검색	첫 10문서에 포함 여부
Alta Vista	1,781,529	52,019	416	
Excite	134,669	29,287	29,287	AND, OR, 구절
HotBot	3,696,449	61,830	17,630	AND, OR, 구절
InfoSeek Guide	3,111,835	427	100	AND질의
Lycos	29,881	26	NA	
OpenText	481,846	2,541	6	구절 질의
WebCrawler	158,751	864	6	구절 질의
WWW Worm	4,999	2	NA	AND, OR, 구절
Galaxy	6,351	20	NA	
Magellan	17,658	17,658	NA	AND, OR, 구절
Yahoo	373 범주 18,344 사이트	1 범주 3 사이트	NA 101	
MetaCrawler	29	32	34	AND질의
정보탐정	28,701	105	1	AND, 구절
* 네이버	61,112	359	0	

* 마지막 행은 LaTeX SW 다운로드할 수 있는 페이지 유무 여부를 나타냄[3]

국내외 검색 엔진들은 모두 경쟁적으로 색인 DB를 키우고 있다. AltaVista는 약 1억 4천만, 국내 정보를 위주로 한 정보탐정은 약 130만쪽의 웹문서를 색인하고 있다. 이렇게 많은 데서 검색하면 그 결과는 어떨까? 검색 엔진이 서로 다르지만 기본적으로 질의 단어와 문서 사이의 문자 패턴 정합에 기초로 하고 있다. 따라서 같은 문자열이 있기만 하면 단어의 의미와 문서의 내용에 상관 없이 선택된다. 그 결과 굉장히 많은 문서가 출력되며, 출력된 문서 중에서 필요한 것을 찾기가 점점 어려워진다(표 1 참조).

참고로 웹 페이지는 극히 불안정하고 불완전한 정보이다. URL이 바뀌는 것은 다반사이고 다른 호스트로 옮겨 가거나 아예 사라지는 경우도 많으며 내용도 수시로 바뀐다. 또한 웹페이지는 엄밀히 말해서 개인적인 기록일 뿐이며 검증될 수 있는 정보로서의 가치는 크게 떨어진다. 따라서 색인 정보의 currency나 completeness를 보장하기는 어렵다.

그럼에도 불구하고 인터넷 검색 엔진 사용자들의 요구는 까다로와지기만 한다. 정보검색 작업이 점점 익숙해지고 생활의 일부분이 되어감에 따라서 사람들은 (A) 내가 원하는 정보만을 (B) 빠뜨리지 않고 (C) 편리하게 (D) 빨리 찾아주기를 기대한다. 지금까지는 D에 많은 노력을 기울여 왔으며 상대적으로 그리 어렵지 않게 고속화될 것으로 예상된다. C는 사용자의 편리를 위한 자연어 질의어 처리, 다양한 입력 방식 등에 관한 인터페이스 이슈로서 여기서는 논외로 한다. B는 전통적인 정보 검색 분야의 성능 척도인 recall 지수를 높이는 것을 말한다. B를 '빠뜨리지 않고 모두'라고 한다면 질의에 따라 출력이 지나치게 많겠지만 precision 지수까지 높이라는 것이다. Recall과 precision 지수는 현실적으로 상충하는 목표인데 두가지를 동시에 달성하기란 매우 어렵다. 마지막으로 A는 불특정 다수의 사용자 개개인을 식별하고 그의 취향에 맞는 서비스를 해주는 것으로서 사용자 모델을 포함한 순수한 인공지능/에이전트 이슈이다. 앞으로 정보검색 서비스는 바로 A와 같은 요구를 만족시켜주는 쪽으로 발전해야 할 것이다.

지금까지는 정보검색 서비스에 대한 일차적 요구에 관한 것이다. 여기서 한걸음 더 나아가면 널려있는 대량의 정보를 적절히 가공하여 압축된 형태의 축약 또는 요점만 받기를 원할 것이다. 최소한 야후의 정보보다는 더 정제되었고 가능하면 요약 보고서와 같은 책자에 버금가는 것을. 지금까지는 제공 정보 DB의 크기와 고속 검색으로 엔진이 평가되곤 했지만 이제는 정보가 많다고 능사가 아니다. 효율도 중요하지만 이제부터는 효과, 즉 정확성을 높이는 것이 주요한 이슈로 되어있다. 21C에는 지식 산업의 수준이 곧 국가 경쟁력의 척도가 될 것이다. 그러기 위해서는 일차적으로 정보의 창출 능력이 필요로 한다. 그 다음에는 정보의 가공 기술이 필요하다. 거기에는 단순하며 품질이 낮은 정보 검색만으로는 한계가 있고 고도의 언어처리 기술과 지식/정보 mining과 같은 기술을 도입해야만 한다.

4. 서비스의 발전

정보검색 결과는 질의를 던진 본인의 주관적 판단 외에는 의미가 없다. 그런데 유감스럽게도 빈틈없이 완전한 질의를 던지는 사람은 아무도 없다. 앞뒤의 문맥이나 사회적 관습, 개인적 편향 등에 따라 대부분 생략하고 당장 머리에 떠오르는 두어가지 단어만 입력하는 것이 보통이다. 따라서 불특정 다수 사용자들을 모두 만족시켜 주는 것은 거의 불가능하다. 기술적으로는 어떨까? 약간 과대 평가를 하자면, 패턴 매치에 의한 오늘날의 정보검색기의 성능은 이제 거의 이론적 한계에 도달하였다고 생각된다. 이를 간파한 사람들은 벌써부터 에이전트 개념으로 포장된 인공지능 기술을 응용하여 정보 부하를 해소하려는 방향으로 나아가고 있다. 본 절에서는 바람직한 정보검색 서비스의 발전 방향에 대하여 고찰해보기로 한다.

• 정보의 가공 서비스

대량의 정보를 그대로 출력하는 것이 아니라 언어처리 기술을 최대한 활용하여 원문을 축약 또는 요약한다. 현재의 언어처리 기술은 형태소 분석을 상당한 수준인 대략 98%내외의 성능까지 해결한 단계에 있다. 유감스럽게도 이

정도 성능으로는 아무 것도 하지 못한다. 예를 들어 30개의 형태소로 구성된 문장을 생각해 보자. 그 결과로 구문 분석을 한다면 제아무리 잘해도 0.98^{30}~0.5, 즉 두 문장 중 하나는 틀린다는 이야기이다. 이 정도면 문단 전체로 본 의미 분석 결과는 거의 0에 가까워진다. 그러나 최근 상당한 성능의 문서 축약 기술이 개발되고 있는 것으로 볼 때, 머지 않은 장래에 효과적인 문서 요약 기술이 등장할 것으로 예상된다.

• 정보 mining/여과

최근 이슈로 떠오르고 있는 지식/정보 mining 기술을 활용한 지능형 정보 검색 서비스나, 이보다 간단한 형태로서 주문한 정보만을 솎아 제공하는 정보 여과, 맞춤 등의 서비스가 있다. 후자의 경우 각 개인의 관심사를 학습하는 기계 학습 기술이 지원되어야 한다.

• 외국어 처리

웹 공간에는 다양한 언어로 된 다양한 내용의 문서가 섞여 있다. 현재의 언어간 번역 기술로(특정 언어의) 내용에 제한이 없는 불특정 문서를 만족스러울 정도로 번역하기는 어렵다[1]. 하지만 현재의 기술로도 어느 정도의 내용은 전달할 수 있을 정도는 가능하다. 외국어 번역 서비스의 가장 바람직한 형태는 최근 정보탐정(http://www.idetect.com/)에서 제공하는 일본 웹페이지 번역 서비스이다. 정보탐정 일본웹여행 페이지를 통해 들어가면 선택한 웹페이지를 정보탐정 서버에서 한글로 번역하여 출력해주는 서비스로서 마치 국내 한글 웹페이지를 보는 듯한 착각을 일으키는 서비스이다.

이와는 약간 다른 것으로는 AltaVista에서 제공하는 사용자 선택형 번역 서비스이다. 마지막으로 다소 번거롭지만 클라이언트 PC에 있는 일한 번역 소프트웨어로 사용자가 직접 번역하여 보는 형태도 가능하다.

• 멀티미디어 정보

앞의 서비스와는 다르고 외국의 일부 검색 서비스에서는 이미 시행중인 것이긴 하지만 멀티미디어 정보가 산재하는 국내의 멀티미디어 정보를 검색하는 것을 말한다. 웹 공간의 멀티미디어 정보로는 영상과 그림이 대부분이며 음성과 음악 정보가 나머지의 대부분을 차지한다.

• 인터페이스

인터넷 검색 엔진 사용자의 상당수는 쓰기 쉬운 것을 선호한다. 엔진 개발자를 위시한 전문가를 제외하면 대부분의 기능은 거의 쓰이지 않고 있다. 실제로 사용자들은 오묘한 기능을 배우고 생각해 가며 질의어를 만들기 보다는 쓰기와 찾기 쉬운 것, 기억하기 쉬운 것을 선호한다. 따라서 인상적이거나 눈에 띄고 부담없으면 금상첨화. 앞으로 인터페이스의 발전은 VR이나 애니메이션 기술을 첨가한 실감/감각 인터페이스, 음성과 필기 등의 도입이 기대해 본다.

• 파생 공간의 정보검색

지금까지 정보검색은 WWW를 포함한 인터넷 공간에 대하여 이루어져 왔다. 앞으로는 그 속에 전자 상거래의 쇼핑몰 또는 전자 장터, 인트라넷 공간, 또는 가상환경 등의 파생 공간이 창출되어가고 있다. 이들 공간이 활성화되면 정보가 창출되고 그 양이 크게 늘어나면 그 파생 공간의 정보검색 서비스의 요구가 커질 것이다.

5. 결 론

정보검색 기술은 이제 충분히 무르익었고, 기초 기술 자체의 성능은 이제 이론적 한계에 근접하고 있다. 그렇다고 해서 발전이 완전히 멈추는 것은 결코 아니다. 이제부터는 인공지능 등 기존의 다른 기술을 통합, 응용하여 새로운 발전을 모색해야 한다. 그리고 자연 언어 처리 기술 등을 최대한 활용하여 한 차원 높은 정보 가공 서비스를 개발해 나가야 한다.

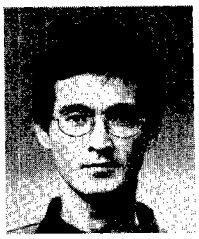
본고에서는 이와 같은 관점에서 현재 인터넷 서비스의 현황과 최근에 출현한 여러 가지 이슈를 살펴보았다. 그 다음 본고에서 주장하는 신규 서비스의 개발 방향에 대하여 소개하였다. 인터넷 정보 서비스 사업의 판도가 향후 어떻게 바뀌어 갈 지 알 수는 없지만 본고에서 지적한 방향으로 기술 개발과 사업을 적극적으로 추진하는 서비스만이 살아남을 수 있을 것이다.

참고문헌

[1] 김태석, "일한 기계 번역 시스템의 연구 및 개발", 정보과학회지 제15권 제10호, p. 9~15, 1997. 10.

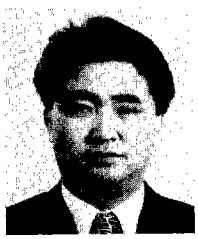
[2] 권혜진 외, "국내 웹 정보 검색 기술의 동향", 정보과학회지 제15권 제10호, p. 16~23, 1997. 10.

[3] V. Gudivada, V. Raghavan, W. Grosky, and R. Kasanagottu, "Information retrieval on the World Wide Web," IEEE Internet Computing, pp. 58~68, Sept-Oct. 1997.



신 봉 기

1985 서울대학교 공과대학 자원공학과 졸업
1987 한국과학기술원 전산학과 졸업(공학 석사)
1987~현재 한국통신 근무
1995 한국과학기술원 전산학과 졸업(공학 박사)
관심분야 : 패턴인식 및 모델링, 지능형 에이전트, 인공지능
E-mail:bkshin@multi.kotel.co.kr



김 영 환

1981 경북대학교 전자공학과 학사
1983 한국과학기술원 전산학과 석사, 한국통신 전임연구원 입사
1986~1990 한국과학기술원 전산학과 박사과정 교육 파견, 공학박사
1990~현재 한국통신 전략정보시스템연구실장, 인공지능연구실장, 초고속통신서비스연구실장, 멀티미디어개발팀장 등 역임, 한국통신멀티미디어연구소, 인터넷서비스개발팀장(책임연구원)
관심분야 : 정보탐정, 인트라넷 그룹웨어, 인터넷폰게이트웨이 시스템, 인터넷보안도구, 기업인터넷서비스 등
E-mail:ywkim@kt.co.kr