차세대 검색기술 웹 패러다임 바꾼다

김형자 | 과학칼럼니스트

1989년 3월 13일, 스위스의 유럽입자물리학연구소(CERN)가 세계 도처에 흩어진 정보를 컴퓨터로 손쉽게 찾아볼 수 있는 월드와이드웹(WWW)을 만든 날이다. 인터넷의 탄생을 알린 순간이다. 초창기에는 과학자들의 연구 성과를 논의하기 위한 공간이었지만, 1993년 CERN이 웹 특허 로열티를 포기하면서 세계를 하나로 묶는 정보의 보고(寶庫)로 거듭나게 됐다.

문장의 의미 파악해 즉답 내놓는 시맨틱 검색

검색의 시작은 알타비스타, 라이코스 등 미국 포털로 거슬러 올라간다. 처음에는 원하는 정보가 어디에 있는 지 알려주는 단순한 정보 제공 수준이었다면 지금은 생활과 밀접하고 복잡한 정보까지 알려주는 '정보 공유 플랫폼'의 개념으로 발전했다. 방대하게 쌓인 포털 사이트의 검색 데이터베이스(DB)는 원하는 정보를 손쉽게 찾을 수 있도록 해준다. 반면 정답과 무관하거나 연관 단어들이 들어간 문서들까지 검색 되어 스크롤바를 오르내리며 일일이 내용을 확인해야 하는 불편함도 따른다.

가령 검색창에 '김혜수 출연작'을 입력하면 '김혜수', '출연작'이란 단어가 들어가는 웹페이지를 우후죽순으로 보여준다. 또 '대한민국의 수도'를 입력하면 물론 결과에 나온 문서에서 대한민국의 수도가 서울임을 알 수 있지만, '대한민국의 수도를 방문해 보세요.' 라든지 '대한민국의 수도는 유명한 곳이다.' 와 같이 정답과 무관한 문장의 문서도 있어 서울이라는 정답을 찾으려면 문서 내용을 일일이 확인해야 하는 경우도 생긴다. 이는 대부분의 검색엔진이 단순히 '김혜수', '출연작', '대한민국', '수도' 등 입력한 단어가 들어 있는 문장만 찾아내기 때문에 생기는 문제이다.

이와 같은 단점을 해결하고, 누리꾼의 입맛에 딱 맞는 결과를 내놓기 위해 만들어진 검색기술이 시맨틱(의미 기반) 검색이다. 지금 정보기술(IT) 분야는 시맨틱과 멀티미디어 검색이라는 차세대 인터넷 검색기술 개발 경쟁으로 뜨겁다. 시맨틱 검색은 단어나 문장의 뜻을 논리적으로 추론해 상황에 맞는 결과를 찾는 의미 기반 검색이다.

가령 '올해 초복이 언제인가요?'를 입력하면 지금처럼 단순히 '올해', '초복', '언제'란 단어가 들어가는 웹페이지를 우후죽순으로 보여주는 게 아니라, 수많은 검색 데이터를 해석해 '7월 14일'이라는 검색 결

과를 가장 위에 보여준다. 또 '김혜수 출연작'을 입력하면 김혜수가 영화배우이며 출연작은 곧 영화 제목임을 검색엔진 스스로 인식해 '타짜', '모던보이', '바람피기 좋은 날'이라는 결과를 가장 먼저 보여준다. 즉답형 서비스인 셈이다. 네이트가 운영하는 '검색실험실(lab.nate.com)'에서 이러한 사실을 실제 확인할 수 있다.

차세대 검색엔진의 핵심은 이처럼 단어와 단어 사이의 연관성을 컴퓨터가 알게 하는 것이다. 사람의 머릿속의 언어에 대한 이해, 즉 문장 안에서 나타내는 의미를 컴퓨터 스스로 파악해 그 의미와 직접 관련이 있는 결과를 내놓는다는 얘기다.

이미지와 음성까지 활용하는 멀티미디어 검색

이런 방식의 검색기술은 평면적인 이미지뿐만 아니라 비디오, 오디오 파일 등 멀티미디어 검색으로도 확장되고 있다. 예를 들어 '노란 손수건'의 이미지를 찾고 싶다고 하자. 이때 대부분의 사람들은 이미지 검색창에 '노란 손수건'이라고 입력한 후 그 결과로 나온 많은 이미지 중에서 적당한 걸 골라 내려받는 게 보통이다. 그런데 이미지 검색도 키워드 창에 단어를 입력했을 때와 마찬가지로, 노란 손수건의 이미지뿐만 아니라 '노란 손수건'이라는 소설책, '노란 손수건'이라는 영화 장면, '노란 손수건이 걸려 있는 빨랫줄' 등의 사진들도 결과로 나온다. 심지어 손수건과 전혀 무관한 음란물이 나오기도 한다.

이것은 네티즌이 인터넷에 노란 손수건 사진을 올리면서 이미지에 대한 설명, 즉 그냥 '노란 손수건'이라든가 '슬픔과 기쁨을 전하는 노란 손수건', '감동의 영화 노란 손수건' 등의 내용을 함께 적어 올리기 때문에 나오는 결과이다. 이런 설명 자료를 이미지의 '메타데이터'라고 한다. 메타데이터(metadata)는 '데이터를 위한 데이터' 또는 '데이터를 설명하기 위한 데이터'이다. 메타데이터를 통해 원 자료(raw data)의 여러 특징들을 기술함으로써 그 자료를 이용하는 데 도움을 주고자 함이다. 하지만 편리함을 제공하기 위해 붙여진 사진의 설명 때문에 오히려 이미지 그 자체보다는 이미지에 연결된 문서를 검색하는 경우가 허다하다. 이런 단점을 보완해 내놓은 기술이 멀티미디어 검색이다.

멀티미디어 검색은 색깔이나 모양 등 이미지 자체의 정보를 뽑아내 검색에 이용하는 기술이다. 예를 들어 노란 손수건 사진을 찾을 때 '노란'과 '손수건'이란 단어와 함께 이미지 파일에서 노란색 정보를 골라내 이 3가지를 모두 만족하는 이미지를 검색 결과로 내놓는 방식이다. 멀티미디어 콘텐츠를 음성에 의해 직접 검색하는 기술도 개발됐다. 동영상과 같은 멀티미디어 안에서 이용자가 필요로 하는 정보를 찾으려고 할 때, 멀티미디어 콘텐츠에 포함되어 있는 음성 내용을 직접 검색하여 정보를 찾게 하는 것이 음성 검색기술이다.

IT 전문가들은 머지않아 컴퓨터가 사람을 대신해 정보를 읽고 이해하고 직접 가공도 해 새로운 정보를 만들어내는 '지능형 웹'까지 선보일 수 있을 것으로 전망한다. 내 상황에 꼭 필요한 개인 맞춤형 정보가 여기저기서 제공될 날을 기대해 보자. TTA