특집2

데이터베이스의 실용 색인에 대한 소고

김경주
중앙일보사 데이터뱅크국
조사정보팀

한국데이터베이스진흥센터의 조사에 의하면 국내에서 제작, 유통되는 데이터베이스 수는 95년을 기준으로 1,061개이고, 이중 53.3%가 검색방식으로 "명령어" 방식이나 "명령어+메뉴" 방식을 제공하는 것으로 나타났다. 이중 색인에 의한 키워드검색을 제공하는 데이터베이스의 규모는 밝혀지지 않았지만 상당수로 추정된다. 더구나 정보량이 많아질수록 키워드에 의한 정보검색은 중요한 수단이 될 것이므로 실용적인 시스템에 대한 다각적인 고려가 필요하다. 여기서는 색인에 대한 비용개념의 허실과 현실적으로 적용가능한 실용색인의 방향에 대해 언급하고자 한다.

비용 대비 효과의 개념

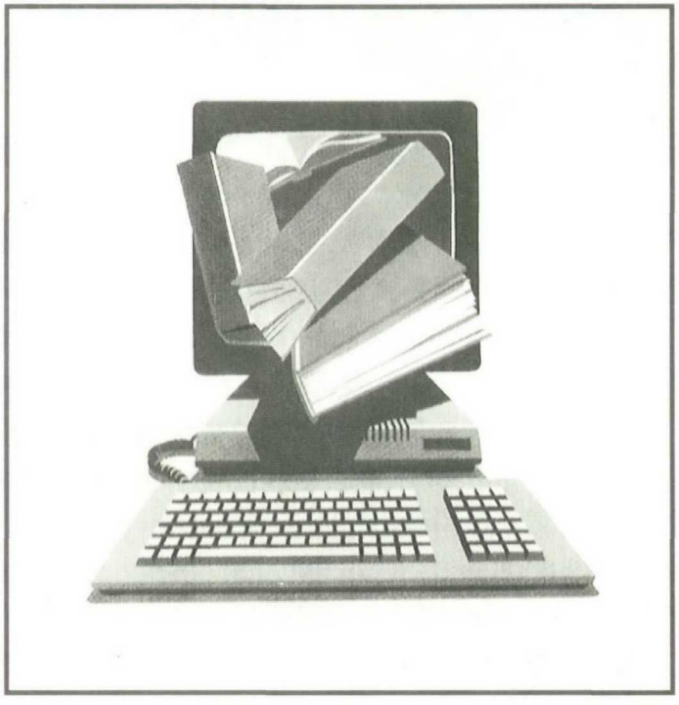
비용 대비 효과는 어떤 부문에서나 투자를 결정하는 핵심적인 요인이다. 데이터베이스 구축과 운영에 있어서도 비용 대비 효과는 중요한 개념이다. 전통적인 색인은 사람이 문서의 내용을 파악하고 주제를 추출하는 주제분석 과정을 거쳐 적합한 색인어를 부여하는 것이다. 그러나 사람에 의한 수작업 색인은 비용이 많이 들고 질적인 면에서도 수준이 다양하다. 또한 폭증하는 정보에 대응하기 위해 계속적으로 전문인력을 투입하는 것은 바람직한 해결책이라고 할 수 없다. 이처럼 경제적인 문제뿐만 아니라 전문가의 확보도 용이하지 않기 때문에 자동색인에 대한 요구가 발생한 것은 당연한 귀결이다.

1950년대 후반부터 연구되기 시작한 자동색인에 의한 free keyword 추출방식은 별도의 인력을 투입하지 않고도 대량의 문서를 처리할 수 있다는 점에서 상당한 매력으로 받아들여져 왔다. 그러나 시스템에 따라 다소의 차이는 있지만 전통적인 자동색인 시스템의 효율이 70%의 재현율(recall rate)과 10%의 정확율(precision rate)에 머무르고 있다는 사실은 시사하는 바가 매우 크다.

정확율이 10%라는 것은 검색된 정보 가운데 10%만이 적합하고 나머지 90%는 부적합한 정보임을 의미한다. 이처럼 정확율이 낮은 이유는 자동색인 시스템이 수작업 색인자 처럼 문서의 내용을 파악한 다음 주제분석을 통해 적합한 색인어를 부여하지 못한다는데 있다.

예를 들어 "승용차끼리 충돌해 불이 났다"는 신문기사를 색인하는 경우를 살펴보자. 자동색인은 십중팔구 "승용차" "충돌" "불"이라는 키워드를 선정할 것이다. 수작업 색인자는 아마도 "승용차" "충돌" "불" 외에 "교통사고"와 "화재"라는 키워드를 추가할 것이다. 이같은 색인결과가 이용자에게 가져다주는 결과는 충분히 짐작할 수 있을 것이다. 결국 이용자는 정교하지 못한 데이터베이스 속에서 자신이 원하는 정보를 찾아내기 위해 많은 시간과 지적노력과 통신비용 등을 지불해야 하는 결과를 가져온다. 색인에 소요되는 인력을 절감하고 시간을 단축시키려는 합리적인 비용개념이 자칫 데이터베이스의 품질 저하는 물론 이용자의 비용 과다 지출을 초래할 수 있는 것이다.

그러나 여기서 자동색인이 수작업색인에 비해 비효율적이라는 시대착오적인 얘기를 하려는 것은 결코 아니다.



색인은 반드시 자동화 되어야 한다.

색인은 데이터품질 평가의 요소 정보검색시스템의 효율을 결정하는 요인은 처리속도, 이용자 인터페이스 등 여러가지가 있다. 이중 처리속도는 지난해부터 일본 마쓰시다 전기산업이 1초간 3억 문자를 탐색하는 고속 검색시스템을 개발, 판매하기 시작한 사실에서도 짐작할 수 있듯, 이미 크게 문제가 되지 않는다고 본다. 이용자 인터페이스 역시 GUI등 새로운 기술이 보편화 되면서 많은 불편이 해소되었다. 시간이 갈수록 더욱 심각해지는 문제는 정확성이다. 즉 적합 정보는 반드시 나와야 하고, 부적합 정보는 나오지 말아야 하는 것이다.

이같은 문제는 정보의 보고라는 인터넷을 항해하다 보면 극명하게 드러난다. 알타비스타 같은 검색엔진은 여러 가지 검색엔진중에서도 우수하고 편리한 것으로 알려져 있다. WEB 환경에서 제공되는 만큼 조작이 간편하고 무엇보다도 놀라울 정도의 검색속도를 자랑한다. 웬만한 키워드의 조합이라면 특별히 기다린다는 느낌이 들기도 전에 수백건, 수천건의 검색결과를 가져다 준다. 문제는 수백건, 수천건이 검색된다는 사실이다. 열심히 범위를 좁혀나가다 보면 수십건정 도로 줄일 수 있지만 이 속에서 원하는 정보를 찾기란 쉬운 일이 아니다. 화려한 환경에서 검색된 정보중 상당부분은 불필요한 정보라는 얘기다.

인터넷뿐 아니라 국내에서 유통되는 대부분의 정보들이 "좀 빠져주었으면" 하는 정보는 너무 많고 "꼭 나와주었으면" 하는 정보는 도대체 어디에 들어있는지 알 수 없다는 것이 아쉬운 점이다. 물론 이같은 불편도 비교적 정보가 충분하게 수록된 데이터베이스의 경우이니 그나마 국내에서는 귀한 정보임에 틀림없다.

그러나 국내의 데이터베이스 시장도 급성장 추세이다. 머지않아 색인의 질과 검색의 효율로 데이터베이스의 품질이 평가되는 날이 올 것이다.

이상적인 색인시스템

이상적인 자동색인은 본문에 나와있는 키워드를 정확하게 찾아주는 real matching 뿐만 아니라 본문에 나와있지 않은 키워드나 관련단어까지 추출하는 semantic matiching의 수준도 뛰어나야 한다.

예를 들어 "성수대교가 무너졌다"라는 신문기사라면 "성수대교"라는 real matching 외에 "붕괴사고"나 "다리" "교각" 같은 의미적인 연관이 있는 semantic matiching 키워드도 추출되어야 한다. "불이 났다"는 기사는 "화재"라는 키워드로 찾을 수 있어야 하고 "불을 질렀다"는 기사는

"방화"라는 키워드로 찾을 수 있어야 한다. 또한 "노트북"의 판매동향을 다루는 신문기사는 PC나 컴퓨터라는 검색어를 입력한 경우에도 찾을 수 있어야 한다.

이처럼 유의어, 관련어등을 망라한 검색결과를 제공하는 것은 이용자의 정보검색 효율을 상당히 개선함으로써 데이터베이스의 부가가치에 기여하는 바가 크다.그러나 국내에서 이같은 수준의 자동색인을 기대하기는 어렵다.

국내 상용 색인시스템의 한계

국내에서는 상용화된 색인시스템이 아직 다양하지 못하다. 상용 시스템의 시장규모는 구매자가 시스템 환경, 가격조건, 성능 등을 비교해 보고 구미에 맞는 시스템을 도입할 수 있을 정도가 되지 않았다고 본다. 이 때문에 몇몇 기관이나 업체들은 자체적으로 개발한 시스템을 사용하기도 한다.

국내에서 상용화된 대부분의 자동색인 시스템은 단순한 형태소분석과 불용어제거에 의한 키워드 추출기법을 채용하고 있다. 엄격히 말하면 조사나 형용사, 부사같은 불용어를 제외한 거의 모든 명사구를 키워드로 추출하는 방식으로 이해할 수 있을 것 같다. 이들 시스템의 경우 재현율은 크게 문제되지 않는다. 그러나 정확율에 대해서는 아직 충분한 연구나 배려가 이루어지지 않고 있다.

물론 업체측의 고충도 이해된다. 국내 데이터베이스 시장규모가 아직 크지 않은데다 일반 도서에서 논문, 기술문서, 잡지기사, 신문기사 등 적용해야 하는 문서의 범위도 다양해 범용시스템을 구성한다는 것이 여간 쉬운 문제가 아닐 것이다. 결국 이러한 문제는 시스템을 도입하는 기관에 따라 별도의 어플리케이션을 적용할 수 밖에 없는데 이 경우 시스템을 판매하는 쪽이나 도입하는 쪽의 공통된 요구사항은 최소비용의 원칙이므로, 다양한 기능의 어플리케이션을 개발하는 것도 쉽지 않을 것이다.

따라서 십중팔구 색인의 결과는 정확한 k.e.y.w.o.r.d.가 아닌 free term일 수 밖에 없고, 한글정보처리의 곤란한 문제인 띄어쓰기나 동음이의어 같은 문제는 항상 과제로 남아 있다. 색인의 결과는 자연어(natural language)가 아닌 자유어(free term)임에 유의해야 한다. 간혹 비용절감을 위한 단순 free term 색인을 natural language 처리와 혼돈하는 예가 있는데 free term은 앞서도 언급했듯 문서내에 표현된 용어와의 real matching 수준이다.

결국 적은 비용으로 최소한의 기능을 탑재한 시스템을 도입하여, 자동색인한 free keyword 방식의 정보를 생산하므로써 이용자에게 garbage를 감수하도록 하는 것이 불가피한 것처럼 보인다. 그리고 이같은 상황과 수준은 데이터베이스를 제작하는 어느 곳이나 마찬가지이므로 일단 크게 문제되지 않는 것처럼 보일 수도 있다.

그러나 앞에서도 언급했듯 국내의 정보시장도 급성장 추세다. 머지않아 색인과 검색의 수준이 데이터베이스의 품질을 평가하는 중요한 요소가 될 것이다.

현실적인 대안 - 기계와 사람의 공동 색인

유통되는 색인시스템의 up-grade는 단시일내에 해결될 수 없는 문제일 수 있다. 더구나 우리가 바라는 semantic matiching 수준이 뛰어난 자동색인 시스템은 요원하다. 그러나 매일매일 데이터를 생산하고 서비스해야하는 데이터베이스 제작자들에게 있어 색인은 "좋은 시스템이 나올 때까지" 라고 미루어둘 수 있는 문제가 아니다. 이같은 현실은 두가지 대안을 가능하게 한다. 하나는 정보량이 많지 않은 경우 어차피 정보량이

많지 않으니 garbage를 감수하는 것이다. 두번째는 정보량이 많고 정교한 검색결과를 희망하는 경우 자동색인과 수작업 색인을 병행하는 것이 불가피 하다.

자동색인과 병용하는 수작업색인은 일반적으로 자동색인 결과를 숙련된 색인자가 훑어보고 불필요한 키워드 삭제와 함께 기계가 처리하지 못한 필요 키워드를 추가하는 방식으로 이루어진다. 따라서 이같은 후통제색인은 자동색인의 재현율을 비롯한 여러가지 효과와 함께 정확율을 상당히 개선한다. 그러나 색인의 품질을 유지하기 위해서는 수작업 색인을 담당하는 색인자의 전문성이 유지되지 않으면 안된다. 색인자는 그 분야에 대한 지식이 풍부하고, 색인과 검색시스템을 충분히 이해하고 있어야 하며 보편성과 객관성, 일관성을 유지하고 색인작업 매뉴얼등에 의한 규칙을 준수해야 한다.

일본 과학기술연구소(JICST)의 경우도 자동색인과 수작업 색인을 병용하는 예인데, 특히 특히 전문가에 의한 색인으로 데이터베이스에 대한 철저한 품질관리를 유지하는 것이 인상적이다.

의미처리의 시도 - 시소러스 기반 색인

수동색인과 병용하는 자동색인은 유통되는 형태소 분석 수준의 시스템이라 해도 크게 문제 되지 않는다. 그러나 이용자 측면의 수작업 색인의 효율과 이용자의 정보이용 비용절감을 생각한다면 몇가지 시스템의 색인효율 개선을 위한 제안을 하고 싶다. 하나는 최근 주류를 이루는 형태소 처리같은 단순한 언어학적 접근 외에 용어의 출현위치나 출현빈도등 전통적인 확률적, 통계적 기법을 병용하므로써 대상 키워드가 얼마나 중요한 의미로 사용되었는지를 계산하는 방법을 병용하는 것이 하나다. 이같은 방법으로도 어느정보의 garbage는 축소가 가능하다. 또 하나는 완전한 구문분석이나 대규모 지식베이스에 기초한 의미분석 까지는 어렵다 해도 동의어사전, 시소러스 같은 용어정보를 활용하는 등의 초보적인 의미처리에 대한 시도는 꽤 생산성이 있는 노력이라고 생각한다.

일본 NTT의 경우 시소러스상에서의 어구의 상하관계, 문장중에서의 특정표현, 색인 전문가들이 이용하는 규칙, 문장중 출현위치, 출현빈도수 등을 이용하여 색인어로서의 자격을 판정하는 색인시스템을 개발함으로써 정확율을 상당히 개선한 것으로 알려져 있다. 동의어사전은 일관성 있는 키워드 선정과 함께 이용자의 검색효율에 기여하는 바가 크다. 시소러스는 정보검색에 사용되는 용어들의 집합이나 사전 정도로 이해하면 될 것같다. 이 용어들은 정보를 색인하거나 검색할 때 사용되는데, 단순히 용어를 모아 놓은 것이 아니라 각 용어들이 논리적인 구조를 갖는다. 유사개념, 상위개념, 하위개념, 관련개념 등이 그것으로, 이같은 어의정보는 색인 및 검색시스템에서 초보적인 의미처리를 가능하게 한다. 시소러스의 용어간 계층관계를 이용하여 보다 일반적인 탐색어를 선택하거나 특정한 탐색어를 선택함으로써 검색되는 문헌의 수를 조절하거나, 관련된 다른 용어를 탐색어로 사용하여 보다 광범위한 탐색을 할 수 있게 하는 것이 그 예이다.

이상에서 언급한 것처럼 자동색인과 함께 수작업 색인을 병용하거나 동의어사전, 시소러스 같은 어의정보를 유지하는 것은 적지 않은 비용이 드는 일일 수도 있다. 그러나 색인에 소요되는 단순한 비용개념을 벗어난다면 데이터베이스의 품질과 서비스 측면에서 가장 현실적이고 실용적인 대안이 아닐까 생각한다. DC