데이터베이스와 시소러스

(Database & Thesaurus)

김 동훈
中央日報社 데이터뱅크局 常務理事
Kim, Dong-Hoon
Executive Director, The Joong and Daily News

1. 정보기술혁명과 정보환경의 변화

21세기 情報化社會는 無人運轉, 홈뱅킹, 音聲認識 컴퓨터, 補聽器 형태의 自動通譯 시스템 등과 같이 기계와 인간의 교감으로 능력을 극대화하는 시대가 된다. 腦波 檢出器와 推論컴퓨터 기술을 접목해 인간의 생각까지도 기계가 대신하는 사회는 공상과학소설이 아니라 불과 5년 앞으로 성큼 다가선 2000年代에 펼쳐질 가정과 사회의 실제 상황이다.

한편 컴퓨터와 마이크로일렉트로닉스 그리고 情報通信 技術 革命의 대조류는 컴퓨터가 창조하는 人工의 세계에 사람이 직접 들어가 보고(視覺) 듣고(聽覺) 만지는(觸覺) 등 실제와 같은 경험을 할 수 있는 「假想現實」 기술을 이미 일부 실현하였다. 또한 출판, 신문 등의 印刷媒體와 영화 등의 映像媒體 그리고 音聲媒體를 하나로 통합하는 미디어 統合을 가속화 하여 국내에서도 멀티미디어 電子新聞 時代의 서막을 열었다.

최근 각국은 이상과 같은 情報化 시대를 겨냥한 투자에 박차를 가하고 있다. 기존 개념의 사회간접자본 투자가 情報高速道路網 등과 같은 情報化 時代를 대비한 투자로 바뀐 것이다.

가까운 일본 해도 "情報 인프라의 구축에 일본은 미국보다 10년 이상 뒤졌다. 미국경제의 부활에는 80년대의 情報化 관련 투자가 밑거름이 되고 있다. 일본이 이제라도 情報化 투자를 서두르지 않으면 美-日 역전의 구도는 정착되고 만다. 情報化社會로 이행할 수 있는 기반을 서둘러 정비해야 한다"는 위기의식에서 光파일網의 부설 등 情報 인프라의 구축에 정부와 민간이 힘을 쏟고 있다고 한다. 사실 情報通信 분야에서 미국과 일본의 격차는 매우 크다. 퍼스컴 보급률, CATV 보급률, 商用데이터베이스 數, 휴대용 전화기 보

급률 등 모든 분야에서 뒤져있다.

이같은 상황은 물론 남의 나라들만의 얘기는 아니다. 우리도 情報化 時代에 대비해 여러가지 준비를 해야 한다. 大容量 통신이 가능한 光케이블 포설과 행정전산망을 이용한 각종 공공서비스 제공을 촉진함은 물론 서류없는 사무실, 사회를 만드는 작업도 서둘러야 한다. 모든 의사전달과 공문서를 電子化 하는 한편, 고부가가치의 다양한 데이터베이스가 개발되면 보다 신속 정확한 情報 전달과 보다 많은 사람들이 情報를 공유할 수 있는 情報化는 자연히 이루어진다고 볼 수 있다.

이처럼 다가올 2000년대의 세계는 情報와 그 情報를 활용할 줄 아는 知識, 그리고 창의력이 국제경쟁을 관가름하고 전략적 자원이 될 情報化社會, 혹은 知識社會가 될 것이다. 이에 따라 情報는 知的 通貨로서의 힘을 더욱 강력하게 발휘하여 정치, 경제, 사회 그리고 모든 과학적 諸要因은 情報를 중심으로 움직이게 될 것이다.

本稿에서는 정보 인프라의 중점 부문인 데이터베이스에 대해 살펴보고, 이를 보다 효과적으로 구축, 이용할 수 있는 방안에 대해 언급하기로 한다.

2. 데이터베이스의 개념과 데이터베이스산업 동향

(1) 데이터베이스의 개념

多種多樣한 情報가 범람하는 현대 사회에서는 情報의 입수시기와 입수방법이 매우 중요하며, 情報化社會가 고도화할수록 유통되는 情報量도 많아지기 때문에 필요한 情報를 필요할 때, 필요한 만큼 입수하는 일이 매우 어려워 진다.

이와 같은 문제의 해결책으로 각광받고 있는 것이 데이터베이스라고 하는 情報處理 개념이다. 데이터베이스는 많은 情報 가운데서 체계적으로 情報를 수집, 선택, 가공해 새로운 부가가치를 가진 情報를 재창출해서 제공할 수 있는 기능을 가지고 있다. 도서, 신문, 통신, TV방송과 같은 종래의 미디어가 一方向性인데 비해 데이터베이스는 雙方向性이며 데이터를 컴퓨터 시스템에 축적하여 이용자들에게 도움이 될 수 있도록 일정한 테마로 조직화된 情報의 집대성인 셈이다.

文獻情報 데이터베이스란 이용자가 필요로 하는 자료를 찾을 수 있도록 안내하는 데이터베이스를 말하며 팩트데이터베이스(Fact Database)란 검색자가 요구하는 情報를 직접 터미널에서 볼 수 있도록 해주는 것으로 情報는 문자, 수치, 화상, 그래픽, 음성 등 다양한 형태로 출력되는 장점을 지니고 있다.

데이터베이스는 또 구축 목적에 따라 社內用 데이터베이스와 商業用 데이터베이스로 구분되기도 한다.

얼마전까지는 情報를 축적하는 기능만이 데이터베이스라고 정의하고 있었는데 최근에는 네트워크 속에서 유통하며, 계속해서 update되는 즉시성 情報와 같은 非축적 情報도 데이터베이스의 개념에 포함시키는 경향이 있다. 또한 기존의 단순한 1,2차 情報의 가공, 제작뿐만 아니라 보다 부가가치가 높은 3차 情報, 즉 豫測情報, 分析情報 등에 대한 관심과 투자가 늘어가고 있다.

(2) 해외 데이터베이스산업 동향

1960년대 소련의 스푸트닉 쇼크에 따라 미국 정부의 전폭적인 지원으로 육성되기 시작한 MEDLINE, ERIC, AGRIS 등과 같은 공공부문의 데이터베이스 개발과 함께 록히드社가 세계 최초의 商用데이터베이스인 DIALOG를 개발한 이래 미국은 전세계의 데이터베이스산업을 석권하고 있다.(DIALOG는 87년 신문그룹인 나이트리더社가 3억7천만 달러에 인수했음)

미국 情報시장의 규모를 알아보기 위해 데이터

베이스 수를 예로 들면, 1979년부터 89년까지 10년동안 529개에서 5,043개로 약 10배 가까이 증가하였고 IP는 6배로 늘어났으며 情報통신 사업자는 3배로 증가하였다.

초기에는 과학기술정보가 주류였지만 최근에는 경제, 비즈니스 정보가 주류가 되었고 리얼타임 정보가 현저하게 증가하였다. 또한 과거의 전통적인 文獻情報 데이터베이스와는 달리 데이터베이스化한 정보를 기초로 쇼핑, 發注, 豫約 등을 리얼타임으로 행하는 Transaction Service도 등장하였다. 더구나 PC통신이 급속하게 발전해 PC통신으로 데이터베이스를 이용하는 경우가 늘어나고 있다. 이같은 변화의 배경에는 컴퓨터의 하드웨어, 스포트웨어, 통신기술의 눈부신 발전이 있었음은 말할 것도 없다.

이처럼 데이터베이스의 내용, 질, 서비스형태의 변화로 데이터베이스 서비스의 개념도 확대되었고 미국에서는 "電子情報 서비스"라는 호칭도 일반화 되어가고 있다. 이같은 개념 속에는 데이터베이스 계열과 Transaction계열 등이 모두 포함된다.

실제로 이들 사이에는 경계선이 없어지고 있으며 전통적인 정보검색형 데이터베이스 서비스에 株價의 온라인 서비스가 부가되기도 하고 반대로 去來(또는 시세)의 리얼타임 서비스에 축적정보가 게재되는 경우도 늘고 있다. 이같은 전자정보서비스의 개념으로 미국 정보시장의 규모를 보면, 지난 89년에는 79억 1천 500만 달러의 거대한 시장이 되었다. 10년 전인 79년에 12억 7천 500만달러였던 것에 비하면 실로 45배나 증가한 숫자인 것이다.

미국 商務省에서 1989년 발간한 "Industrial Outlook"에 의하면 온라인 데이터베이스의 시장규모는 88년에 이미 37억달러에 달했다. 최근 수년간 평균 성장율은 약 20%라는 높은 비율을 보이고 있으며 90년대 들어서도 고수준의 성장을 계속할 것으로 전망하고 있다.

온라인 데이터베이스 서비스의 내용을 보면 시장규모가 가장 큰 것은 역시 기업과 개인의 신용정보 서비스이다. 이는 온라인화 이전부터 기업경영상 이용빈도가 높았던 분야로서 온라인화 이후에도 여전히 이용자가 많다. 다음으로 시장규모가 큰 서비스는 金融市況 情報서비스인데 성장율은 약간 둔화되는 경향을 보이고 있다. 근래에 눈에 띄게 발전하는 것이 신문, 잡지 기사 등의 全文 出力이 가능한 비즈니스情報 분야이다. 온라인 데이터베이스 가운데는 經濟데이터로 대표되는 數値데이터나 文獻檢索情報(2차情報)의 이용은 성장의 고비를 넘어 평균 이하의 성장율만을 기록하고 있다. 그 대신으로 신장하고 있는 것이 비지니스情報로 대표되는 오리지널의 1次情報이다.

그밖의 情報 선진국인 영국, 프랑스, 독일, 일본 등에서도 데이터베이스産業이 꾸준한 발전을 보여왔다. 유럽의 경우 데이터베이스 서비스 시장 규모가 91년 33.5억 달러였다.

(3) 국내 데이터베이스산업 동향

지난 '88년 시작된 국내의 데이타베이스 서비스 시장규모는 1992년 약 739억원이었다. 전체 산업에서 차지하는 점유율은 극히 적지만 전년대비 30.2%의 급성장세를 보이고 있다.

PC통신을 통해 이용할 수 있는 데이터베이스 건수를 보면 데이콤의 천리안이 94년 3월 국내에서 처음으로 1,000개를 돌파했다. 지난 88년 5월 생활백과정보(연세대학교), 관광명소(한국관광공사) 등의 생활정보를 제공하면서 시작된 천리안이 데이터베이스 상용화를 시작한지 만 5년 9개월만에 1,014개의 데이터베이스를 확보하게 된 것이다. 물론 이는 90년도 미국의 5,500개와 일본의 2,400개에 비하염 걸음마 단계이다. 그러나 타산업의 신장세에 비하면 괄목할 만한 수치라 할 수 있겠다.

3. 시소러스의 개념과 동향

시소러스란 정보의 축적과 검색시 索引 작성자와 검색자가 사용하는 용어를 표준화된 어휘로 통일한 어휘집으로 용어간의 개념 관계를 동의어, 계층관계, 관련성 등의 측면으로 결합하여 체계적으로 배열해 놓은 용어 통제표를 의미한다.

情報檢索을 위한 시소러스는 索引 작업시에는 적절한 索引語의 선택과 索引語의 統制를 위해 필요하며, 검색시에는 적절한 탐색용어의 선택을 지원한다. 그밖에 시소러스는 개념간의 계층, 관련 관계를 이용한 탐색어의 확장과 축소를 통해 검색효율을 조절하는 데도 사용된다. 즉 특정 개념과 관련된 보다 제한적인 검색과 보다 포괄적인 검색이 가능하다.

현재 전세계적으로 다양한 주제영역에 걸쳐 약 500여종의 시소러스가 개발되어 정보검색 분야에서 활용되고 있다. 그러나 우리나라의 경우 외국의 시소러스를 한글로 번역한 교육관계 시소러스 1종, 과학기술용어 시소러스 1종 등이 나와 있을 뿐, 자체적으로 개발된 우리말 시소러스는 전무한 실정이었다. 그러다가 최근 중앙일보사에서 신문기사에서 다루는 모든 주제영역을 망라하는 종합 시소러스를 자체 개발함으로써 국내에서도 시소러스를 이용한 자동검색 시스템의 막을 열었다.

한편 최근에는 자연어 방식의 한계에 직면한 다수의 정부 또는 민간기관에서 시소러스개발을 추진중이거나 기 제작된 시소러스의 채용을 모색중인 것으로 알려진다.(表 1, 2 「국내외 시소러스의 사례」 참고)

表 1. 세계적으로 실용화 단계에 있는 시소러스

시소러스	1차개발	구 축 기 관	주 제 분 야
MeSH	1960	NATIONAL LIBRARY OF MEDICINE	의학
ERIC	1966	EDUCATIONAL RESOURCES INFORMATION OENTER	교육
INSPEC	1973	INFORMATION SERVICE FOR PHYSICS ELECTRONICS & COMPUTERING	물리, 전자공학, 컴퓨터
ROOT	1981	BSI(英國工業規格協會)	공업규격, 표준
日經시소러스	1988	日本經濟新聞社	신문기사용 종합
中日시소러스	1974	中日新聞社	신문기사용 종합
JICST시소러스	1971	日本科學技術情報센터	과학기술 중심 종합

表 2. 국내의 시소러스

시 소 러 스	1차개발	구 축 기 관	주 제 분 야
농업경제문헌검색어집	1985	한국농촌개발연구원	농업
KEDI교육시소러스	1981	한국교육개발원	교육
과학기술용어시소러스	1992	시스템공학연구소	과학기술
신문기사종합시소러스	1993	한국인로연구원	기사용 종합
JOINS시소러스	1991	중앙일보사	기사용 종합