데이터베이스의 구조와 정보검색

유경희
(사)한국복지정보통신협의회
부회장

한사람이 관리할수 있는 정보의 양은 500-600건에 불과하다고 한다. 이정도 양의 정보를 관리를 잘하는 사람은 다년간의 경험에 의한 것이 분명하다. 어느 누구가 200-300개 정도의 전화번호를 기억할 수 있겠는가? 옛날에 전화 교환원이 2천여개의 전화번호를 외우고 있다는 말은 들었어도 그것도 프로이니까 그렇지 보통사람이 어떻게 2천여개의 전화번호를 외우고 있을까? 없다고 단정해도 좋다. 정보의 양이 1만건이하이면 사람을 여러사람을 동원해도 손과 머리로서 관리를 해낼 수가 있다.

그러나 문제는 1만건이상인 경우에는 어떻게 하겠는가 이다. 대체로 1만건이하 이면 분류법이니 색인방법이니 하는것들은 아무렇게나 해도 특별한 무리는 없다. 문제는 1만건이 훨신 넘는 정보량인 경우에는 반드시 데이터베이스화하여 관리하지 않으면 안된다.

지금 말하는 건이란 가령 주민등록정보인 경우 한사람의 정보를 한건이라고 할 수 있을 것이다. 회사의 인사기록카드도 한사람당 한 건이다. 예를 사람에 관한 것만 들었지만 데이터베이스화 해야할 가지수는 부지기수라 하겠다. 특허정보도 있고 문헌정보도 있고 어떤 경우에는 숫자만 나열되는 통계정보도 있다. 한때는 데이터베이스를 구분하기를 문자 데이터베이스, 수치 데이터베이스, 문자·수치혼합 데이터베이스 등으로 나눈 적도 있다. 오늘날에 와서는 별의미가 없어지고 있지만...

정보의 종류를 보면 제1차정보와 제2차정보로 구분하는 경우도 있었다. 제1차정보란 가령 "절대온도란 몇도인가?"라는 질문에 "섭씨영하 273도"라는 답변이 나오는 데이터베이스를 수치 데이터베이스일 수도 있고, 제1차정보일수도 있다.

그러나 "절대온도에 관하여 알아보려면 무슨 책을 봐야합니까?"라는 질문에 답할 수 있는 데이터베이스에 의하면 "××백과사전 ○○페이지 상단을 보시요."라는 대답이 나오는 데이터베이스라면 문자 데이터베이스일수도 있고 제2차정보일수 있다.

이것마저 급격히 달라지는 데이터베이스 환경때문에 시각장애자를 위한 점자도 나오고 소리, 정지화상 및 영상에 이르기까지 데이터베이스화 하는 세상이 되어버렸다.

멀티미디어 데이터베이스 이전의 것을 일단 고전적인 데이터베이스라고 해두자. 고전적인 데이터베이스에서는 정보량은

"레코드(Record)"의 수로 따진다. 우리말로는 그래도 건수라는 말이 적합할듯 하지만 요즘 건수란 말이 다르게 쓰이는것 같아서 레코드라고 쓰기로 한다.

보통 작문에서 다루는 단락(Paragraph), 문장(Sentence), 구(Phrase), 단어(Word)로 구분하듯이 데이터베이스의 한 레코드는 영역(Field)와 데이터 요소(Data Element)로 구분한다.

어떤 레코드는 몇 10개의 영역(필드)을 가지기도 한다. 문헌 데이터베이스인 경우 저자명 영역(필드)이 있다. 이 영역에 들어가는 것은 모두가 사람의 이름이다. 초록이라는 영역이 있다고 하면 초록에 쓰여져 있는 것은 대체로 문장이다. 영역의 아래에는 데이터 요소가 있다. 문장으로 치면 단어에 해당하는 것이다.

그러나 대체로 데이터 요소에 자연어를 그대로 못넣을 것은 아니지만 그렇게 하면 레코드의 길이가 너무나 길고 처리시간이 느려지기 때문에 부호화(코드화)하는 경우가 있다. 가령 국가명을 부호화 할때는 숫자로 나타내면 3자리면 충분하고 알파벳으로 표현하면 2자로 표기한 부호(가령 한국인 경우 KO), 3자로 표기한 부호(KOR) 등으로 통일 사용하는 경우가 많다. 이는 컴퓨터의 기억용량의 절약을 위한 목적도 있지만 처리시간의 절약이 더 큰 목적이다.

지금까지 데이터 요소를 부호화하려는 노력이 엄청나게 기울여져 왔다. 심지어는 주민등록번호까지 만들어 졌을 정도이다. 기억장치의 경비가 큰 문제가 되지 않게되고 또는 컴퓨터의 처리속도가 문제가 되지 않으만큼 발전해나가는 오늘에 와서는 굳이 부호화할 필요가 무엇이 있는가!

자연어를 그대로 사용한다면 입력 처리시간이 훨씬더 절약되는것이 아닌가라는 반론도 만만치가 않다. 이러한 논의는 어디까지나 고전적인 데이터베이스에 한한 것이다.

이제는 환경에 변혁이 일어나고 있다. 전화 700번 서비스와 같은 소리의 데이터베이스도 실용화 되었고, 700-1000번의 팩스 데이터베이스(화상)도 실용화가 되었다. 또한 영상을 담은 데이터베이스도 VOD(Video on demand)란 형식으로 변형된 데이터베이스도 조만간에 등장하게 된다.

이러한 데이터베이스가 발전하면 어쩌면 2차정보란 개념이 파괴되어 버리는게 아닌가 할만큼 데이터베이스의 환경이 급변하고 있다.

그러나 고전적인 것도 확실히 알아두어야 한다.

데이터베이스에서 정보를 찾을때에는 대체로 두가지 방식이 있다. 그 첫째는 메뉴(Menu) 선택방식이고, 둘째는 키워드(Keyword)방식이다. 키워드방식 에서도 단순 키워드검색방식과 부울논리를 적용한 키워드 검색방식이 있다.

오늘날 우리주변에서 보는 각종 데이터베이스는 거의 모두가 메뉴 선택방식이다. 가령, 식사를 하고자 할때 양식, 한식, 중국식, 일본식중에서 먹고싶은 것을 선택하라고 하면 번호로서 선택한다. 만약 중국식을 선택한 경우에는 다시금 밥, 면, 요리 등에서 골라라고 나온다. 여기서 면을 선택한다.

그러면 짜장면, 우동, 짬뽕 등의 메뉴가 나온다. 여기서 먹고싶은 것을 선택하면 된다. 여기서 선택해야할 기회가 네번이나 주어진다.

이와 마찬가지로 하이텔, 천리안, 유니텔, 나우콤 등은 거의 모두가 메뉴 선택방식으로 데이터베이스를 만들고 있으며, 그 밖에도 만들고 있는 수많은 데이터베이스도 이러한 방식으로 만들고 있다. 그러나 유의해야 할 것은 미국의 Dialog, Orbit, BRS 등과 같은 문헌 데이터 위주의 데이터베이스에서는 모두가 키워드 검색방식으로 찾도록 되어있다.

키워드 검색이란 복잡한 데

이터베이스의 구조에 익숙하지 않는 사람들이 직접 찾고자 하는 정보를 "용어"로서 찾도록 꾸민 데이터베이스이다.

대체로 메뉴 선택방식은 데이터를 제작하기는 용이하지만, 검색의 정확도나 효율면에서는 크게 뒤지게 되며, 필요없는 선택기회가 너무 자주 부여되어 이용자가 금방 식상하게 되는 경우가 많으며, 정확한 정보를 찾지못한채로 사장되어버리는 정보가 많다.

이것은 아직도 데이터량이 약 1만건 미만에서는 그러한 불편을 별로 느끼지 않지만 1만건 이상의 대량 데이터베이스인 경우에는 이러한 단점이 자주 나타난다. 키워드로 검색하도록 꾸민 데이터베이스는 아무래도 제작경비가 더 많이 들게 마련이다.

그러나 정보검색의 정확도에서 뛰어난 효과를 나타낸다. 사람들이 정보를 검색할때 정확한 키워드를 모를 경우가 많다. 정확한 키워드를 알고있는 경우에는 정작 정보검색의 필요성마저 없어지는 경우도 많다. 또한 사람들이 정보를 검색하려고 할때에는 키워드가 2개 이상인 경우가 허다하다. 이러한 개념들을 적절히 조합해서 검색을 하도록 만든 경우가 많다.

키워드 개념의 조합을 위하여서는 부울논리(Boolean Logic)를 사용한다. 부울논리란 대체로 논리적(AND), 논리화(OR) 및 논리부정(NOT)등 세가지로 표현한다. 키워드 A와 키워드 B와의 사이에 논리적(AND)을 적용시키면 그 뜻은 키워드 A와 키워드 B를 동시에 함유하고 있는 정보를 찾게 된다.

한편, 키워드 A와 키워드 B와의 사이에 논리화(OR)를 적용시키면 그 뜻은 키워드 A와 키워드 B중 어느 하나라도 함유한 정보는 모두 찾아낸다. 끝으로 키워드 A와 키워드 B와의 사이에 논리부정(NOT)을 적용시키면 이것은 키워드 A를 포함한 정보에서 키워드 B를 포함한 정보를 제외한다는 뜻이 된다.

시내를 걸어가다가 몇년전에 TV에서 얼핏본듯한 서양사람을 봤다. 무심코 지나면 그만이지만 이경우를 정보검색한다고 생각하여 보자.

먼저 몇년전인가를 생각하여 본다. 3-4년전이라고 한다면 우선 검색범위를 1990-1995년으로 잡게 된다. 이 년도범위를 하나의 키워드(A)라고 생각하자. 두번째 키워드를 생각한다. TV에서 얼핏볼때 서울에서 개최된 무슨 국제회의인듯 했다는 것이 상기된다. 그러면 키워드는 서울(B)과 국제회의 (C)가 된다.

가령, 한국내의 어느신문이 전량 데이터베이스가 되어 있다고 한다면, 여기서 검색공식(Formula)을 구성하여 본다. 즉 A and B and C라고 정리할수 있다. 이 말은 1990년과 1995년 사이에 서울에서 개최된 국제회의를 찾게 된다.

만약에 이것이 수천건 검색된다면 이 검색공식은 실패한 것이다.

그러나 30-40건밖에 검색되지 않는다면 여기서 충분히 조금전에 본 서양사람이 누구인가를 찾아낼 수가 있다. 키워드 가운데는 자연어에 의한 키워드와 통제어에 의한 키워드가 있다. 문자로 표현된 자연어를 그대로 컴퓨터로 자동색인하는 경우에는 자연어로써 검색하면 좋으나 색인자와 검색자의 표현방식의 차이로서 적합한 정보가 누락될 가능성이 높다.

그러나 색인비용이 훨씬 싸게 먹힌다. 한편, 통제어로서 검색하는 경우에는 적합한 정보를 정확히 찾아낼 수가 있지만, 용어를 통제하는데 많은 경비(인건비)가 소요된다. DC