실제사례 통해 데이터를 보석으로 활용하기 위한 세 가지 요소 1 조감

"데이터 전체상을 조감하라"

가시화 · 조감 통해 사람 지식을 최대한 활성화시켜야

데이터는 모든 업무의 출발점이며, 과정이며, 종착점이다. 우리 주변에는 방대한 양의 데이터가 있으나 이러한 데이터의 본질을 제대로 이해하고 업무에 효과적으로 사용하는가에 대해서는 의문이 드는 경우가 많이 있다. 살아있는 데이터를 보석으로 만들어서 업무에 활용하기 위해서는 데이터를 그림으로 표현하는 게 효과적이다. 왜냐하면 그림을 보면 데이터가 말하려고 하는 내용이 들리기 때문이다. 그래서 전문가일수록, 어렵고 복잡한 개념일수록 그림으로 그리면서 설명하게 된다. 적절하게 데이터를 활용하기 위해서는 그림을 그리고 이해하기 위한 세 가지 요소, 즉 시간, 조감, 상관에 대하여 그 특징을 이해해야 한다. 지난 호에 이어 이번 호에는 조감에 관해 설명한다. 설명은 실제로 업무에 적용되었던 사례를 중심으로 요점을 간단히 요약하여 설명한다.

윤 태 성
KAIST 교수 · 윤츠 사장
yoon.taesung@kaist.ac.kr

필자는 KAIST 경영과학과와 기술경영전문대학원 겸직 교수이며 윤츠와 오픈놀리지(동경) 사장이다. 저서에는 [오픈놀리지-지식은 어떻게 비즈니스가 되는가(21세기북스)], [상대를 합리적으로 설득하는 막강 데이터력(매일경제신문사)], [테크놀리지 로드맵-기술지식의 조감과 분석에 의한 신산업 창조(일본어, 공저)] 등 다수가 있다. 이 원고에 관한 저자에의 연락은 untz.book@gmail.com이며, 참고 사이트는 http://untz.co.kr.

글 · 연 · 재 · 순 · 서

1 시간→데이터의 시간에 따른 변화를 이해하라(10월호)
2 조감→데이터 전체상을 조감하라
3 상관→데이터간의 상관관계를 파악하라(12월호)

1 조감이란

어떤 도시를 처음 가보면 방향 감각을 잃어버리고 헤매기 쉽다. 동서남북도 모르고 도시의 중심부도 모르니 그저 발길 닫는 데로 갈 뿐이다. 그러다 보면 묵고 있는 숙소 앞의 길을 중심으로 해서, 오늘은 이쪽으로 가보고 내일은 저쪽으로 가보고 한다. 만약 특별히 가본 곳도 없고 그 도시에 머무르는 시간도 짧은 경우라면 그 도시의 전체적인 구조를 마지막까지도 모르고 떠나게 된다.

이럴 때 도움이 되는 것이 관광버스다. 세계적으로 유명한 도시에는 관광객을 대상으로 해서 주요 루트를 따라서 관광버스를 운행하는 프로그램이 있다. 특히 2층 버스를 운행하는 곳도 많이 있는데, 이럴 때 2층 버스를 타고 우선 도시의 간선도로를 돌아보면 그 도시를 이해하는데 많은 도움이 된다. 2층 버스에서는 시선의 높이가 자신의 평소 시선보다 높이 위치하게 되니까 거리를 이해하기가 한결 쉬워지기 때문이다. 2층 버스를 타고 도시를 한 바퀴 돌은 다음에는 직접 골목마다 걸어 다닌다. 웬만한 규모의 도시라면 이 정도 하면 그 도시 전체의 구조를 매우 상세하게 알게 된다.

2층 버스를 타고 보아도 높은 시선으로 볼 수 있기 때문에 거리를 알기가 쉬운데, 만약 비행기를 타고서 내려다본다면 도시의 구조를 더욱 쉽게 한눈에 알 수 있을 것이다. 비행기를 타고 도시를 내려다보면 높이 나는 새가 멀리 본다는 속담을 실감하게 된다.

옛날 사람들이 알고 싶어 한 게, 날아가는 새가 높은 곳에서 내려다보는 풍경은 어떨까 하는 것, 즉 조감이다. 조금이라도 높은 위치에서 내려다보는 것이 전체적인 풍경과 구조를 이해하는데 결정적인 도움이 되기 때문이다. 데이터를 보석으로 활용하기 위해서는 데이터를 조감할 수 있어야 하는데 그 이유는 너무나 명백하다. 데이터의 전체적인 풍경과 구조를 이해하고 있으면 데이터의 본질을 쉽게 이해할 수 있기 때문이다.

2 조감의 방법

조감에는 실체에 의한 조감과 모델에 의한 조감이 있다. 풍경을 예로 들면, 실체에 의한 조감은 비행기나 인공위성에서 내려다보는 풍경이고 모델에 의한 조감은 지구의를 통하여 보는 풍경이라고 할 수 있다. 실체에 의한 조감은 비행기의 고도에 따라서 풍경이 변하는데 마치 카메라의 줌 렌즈를 당겼다 밀었다 하면서 제한된 앵글에 담을 피사체의 범위를 넓혔다가 좁혔다가 하는 작업과 유사하다. 모델에 의한 조감은 실체를 재현하기 위한 목적과 관점에 따라서 변하는데 마치 도시를 행정구역별로 구분하여 보거나 지하철 노선 별로 보는 것과 같다.

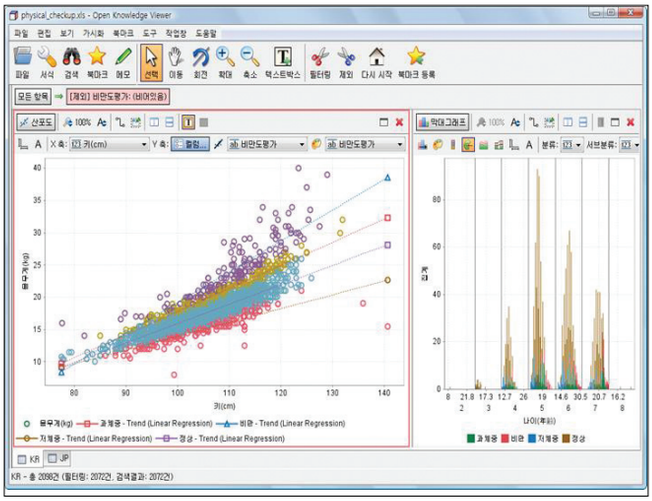
실체에 의한 조감

데이터를 조감하기 위해서는 실체에 의한 조감과 모델에 의한 조감이 모두 사용된다. 데이터를 실체에 의해서 조감하는 방법은 예를 들어 모든 데이터를 실제로 나타내보는 것이다.

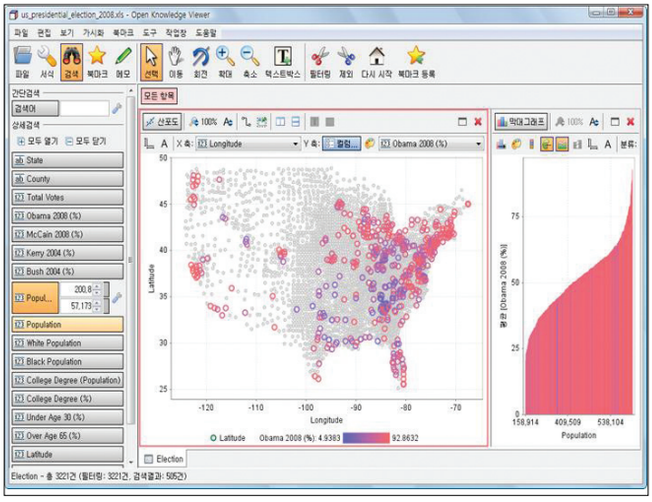
〈그림 1〉에서는 2천명 이상의 어린이를 대상으로 실시한 건강 진단 데이터를 나타내고 있다. 데이터의 항목에는 신장, 체중, 비만도 평가 등이 있다. 이런 데이터가 2천명 이상을 대상으로 수집되어 있으면 우선 데이터 전체의 특징을 조감하는 것이 데이터의 본질을 이해하는데 도움이 된다.

〈그림 1〉의 왼쪽에 있는 산포도는 모든 데이터를 플롯팅한 것인데, X축은 신장을 나타내고 Y축은 체중을 나타낸다. 즉 모든 데이터를 신장과 체중을 기준으로 해당하는 곳에 단순히 플롯팅했기 때문에 하나하나의 데이터를 알기 위해서라기

■그림 1 신체검사 결과를 산포도와(좌측) 각 연령별로 분포(우측)를 나타낸다



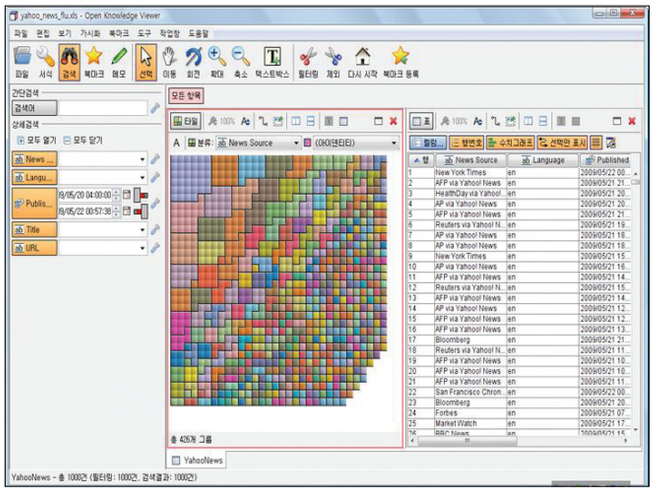
■그림 2 미국 대통령 선거결과 인구밀도가 1평방 마일 당 200인 이상인 카운티의 득표결과



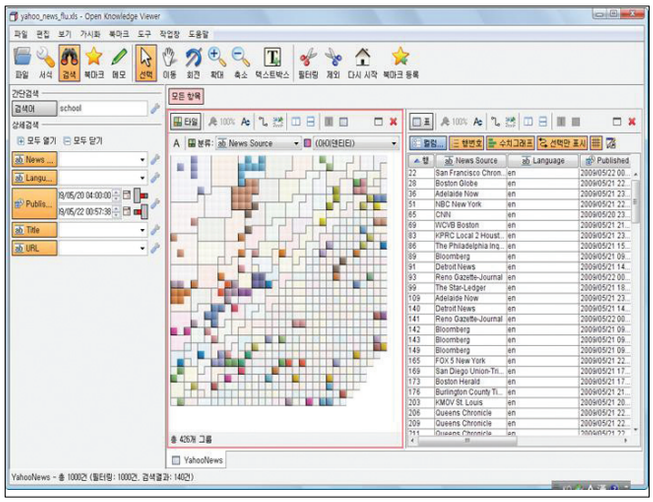
보다, 데이터 전체의 분포가 어떻게 되어 있는지를 알기에 편리하다. 이 데이터에 비만도 측정 결과에 따라서 색깔을 칠하여 구분하였다. 이 그림을 보면 신장이 커질수록 체중이 늘어나긴 하지만 신장이 커질수록 체중의 분포는 넓어지고 있는 풍경을 쉽게 알 수 있다. 〈그림 1〉의 오른쪽에 있는 막대 그래프는 데이터를 각 연령별로 나누어서 분포를 나타낸 것이다. 전체의 데이터를 몇 개의 그룹으로 나누어서 조감함으로써 데이터의 특징을 보다 자세하게 이해할 수가 있다.

데이터를 조감하는 작업에는 데이터와 지리 정보를 동시에

■ 그림 3 키워드를 "flu"로 하여 검색한 결과



■그림 4 그림 3에 대해 "school"을 추가로 검색어로 사용한 결과



나타냄으로써 데이터의 본질을 쉽게 이해하게 되는 경우가 있다. 〈그림 2〉에서는 2008년에 실시된 미국의 대통령 선거 결과 데이터를 나타내고 있다. 왼쪽 그림은 인구 밀도가 1평방 마일당 200인 이상인 카운티만의 득표 결과를 색깔을 칠해서 보여주고 있는데 빨간색 원은 오바마 후보가 승리한 카운티를 나타낸다. 이 그림에서는 오바마 후보의 득표 수를 상세한 숫자로 보지 않고 어느 지역에서 승리하였는지 그 풍경을 조감하고 있는 상태이다. 오른쪽의 막대 그래프는 X축의 오른쪽으로 갈수록 인구 밀도가 높아지면서 오바마 후보가 상대적으로 승리한 것을 보여준다. 이 두 가지 그림을 동시에 조감함으로써 오바마 후보가 승리한 지역에 대한 특징을 쉽게 이해할 수 있다.

모델에 의한 조감

데이터를 모델에 의해서 조감하는 방법 중에는 예를 들어 회귀분석에 의한 방법이 있다. 즉 방대한 양의 데이터에 대하여 전체적인 구조를 설명하기 위해서 회귀식을 사용하는 방법이다. 모델에 의한 조감이 가지는 장점 중에는 데이터 전체에 대한 추가, 삭제, 변경과 같은 처리를 쉽게 함으로서 특정의 목적과 관점에 집중할 수 있다고 하는 점이 있다. 즉 데이터의 전체상에 대하여 부분상의 확대와 축소, 부분상간의 연산, 부분상의 집합 등을 자유롭게 실행할 수 있다.

〈그림 1〉의 왼쪽 그림은 4가지로 구분한 비만도 평가 결과에 대하여 각각 회귀직선을 표시하고 있는데, 이것은 각 그룹별로 모델이 다르게 생성되는 것을 나타낸다. 〈그림 1〉에서 데이터 전체에 대한 회귀직선은 각 비만도 별로 작성된 회귀직선과는 다른 모델이 되는데, 회귀직선은 소프트웨어에서 범위를 설정하면 자동으로 생성해 주니까 사용자는 데이터를 조감하는 목적과 관점에 따라서 전체에 대한 회귀직선과 그룹별로 생성된 회귀직선을 구분하여 사용하여야 한다.

검색결과의 조감

데이터를 조감하는 게 전체적인 풍경과 구조를 이해하는데 도움이 되는 것은 확실하지만, 아직도 전혀 조감이 되지 않는 분야가 있다. 바로 검색결과의 이용이다. 특히 인터넷 검색을 하면 검색결과가 수 천 건 이상이 되는 경우가 빈번하며 많을 경우에는 몇 억 건이 검색결과로서 제시되는데, 사실 몇 십 건 이상의 결과가 제시된다면 그 양은 사람에게는 별로 의미가 없다. 어차피 많은 양의 결과를 전부 다 볼 수 있는 시간도 없으며, 혹시 본다 하더라도 모든 내용을 다 이해할 수 있는 능력도 없기 때문이다. 그래서 대부분의 경우에, 검색결과는 제시된 내용 중에서 몇 건만을 살펴보고 나머지 결과는 모두 다 버리게 된다. 열심히 검색해서 찾아온 결과 중에는 보석과 같은 내용도 있을지 모르며, 쓰레기와 같은 내용이 있을지도 모른다. 어느 쪽이던지 전혀 살펴보지도 않은 채 검색결과를 거의 모두 버리는 것은 정보 자원을 제대로 활용하지 못하고 있는 현실적인 문제이다.

신종 플루에 대하여 인터넷 검색하는 경우를 생각해 보자. 검색결과는 수 천 건이나 그 이상이 되는데 한 페이지당 수십 건씩 해서 수 십 혹은 수 백 페이지가 제시된다. 이렇게 제시되는 방대한 양의 검색결과는 전혀 이용되지 않은 채 그대로 사라지는 경우가 대부분이다. 만약 검색결과를 조감할 수 있다면 검색결과의 이용도는 급격히 증가할 것이다.

검색결과를 조감할 수 있는 방법은 결과를 가시화하는 것이다. 〈그림 3〉은 인터넷 검색결과를 가시화하여 조감하고 있는 것을 나타내는데, "flu"를 키워드로 사용하여 모 검색 사이트의 뉴스를 검색한 결과 중에서 다운로드가 가능한 1,000건을 나타내고 있다.

오른쪽에 있는 테이블은 검색결과를 표 형식으로 나열하여 정리한 결과인데 이것만 가지고는 전체적으로 어떤 풍경을 하고 있는지 알 수가 없다. 그래서 가시화한 것이 왼쪽 그림인데, 검색결과 1,000건을 타일 모델로 가시화하여 표현한 결과이다.

데이터 1,000건을 타일모델로 나타내면 타일 1,000개로 구성된 그림이 된다. 조그마한 타일 하나하나는 각각 검색결과 한 건과 대응하고 있다. 타일 모델을 사용하여 데이터를 가시화 하는 경우에, 어떤 속성을 기준으로 가시화 하는가에 따라서 타일의 모양이 변하게 된다. 〈그림 3〉에서는 뉴스 소스를 기준으로 가시화하고 있는데, 이 그림을 보면 뉴스를 많이 발신하고 있는 곳은 상위 15사 정도라는 것을 알 수 있다. 만약 뉴스 내용과 세계 지도를 함께 가시화하게 되면 〈그림 2〉와 같이 지도상에 뉴스 발신 건수나 발신 시각 등을 함께 나타낼 수 있게 된다. 이런 그림을 보면 전 세계적으로 신종 플루에 관한 뉴스가 어느 지역에서 언제 어느 정도 발신되고 있는지를 간단히 이해할 수 있게 된다.

인터넷 검색을 하다가 결과로서 제시되는 건수가 너무 많으면 연속적으로 키워드를 입력해서 검색결과를 줄이는 경우가 일반적이다. 〈그림 4〉는 〈그림 3〉의 결과를 대상으로 키워드 "school"를 사용하여 더욱 상세하게 검색한 결과를 나타내고 있는데 타일의 색깔이 진하게 칠해져 있는 게 검색된 결과이다.

이 그림을 보면 flu에 관한 뉴스 발신 건수가 적은 뉴스 소스 중에도 school이라는 내용을 포함하고 있는 뉴스가 골고루 분포되어 있는 풍경을 볼 수 있다. 이러한 그림을 보면, 뉴스는 대형 신문사나 통신사가 발신하는 내용만 볼게 아니라 발신 건수가 적은 소규모 신문사나 전문적인 조직이 발신하는 내용도 주의 깊게 관찰할 필요성을 느끼게 된다. 그러나 인터넷 검색결과를 한 건 한 건 확인하는 것은 현실적으로 어려운 작업이다. 이럴 때에 검색결과를 가시화하여 전체적인 풍경을 조감하면 검색결과 전체를 효율적으로 이용할 수 있게 된다.

검색결과의 조감은 회사 내의 ERP 시스템의 운영을 통한 자료의 검색과 이용에도 영향을 끼치는 문제이다. 만약 특정 부품의 재고 데이터가 ERP 시스템에 저장되어 있다고 가정해 보자.

그러나 처음부터 이 데이터의 어디까지를 조감할 수 있는지를 각 사원 별로 혹은 각 업무별로 미리 정의하기는 어렵다. 왜냐하면 어떤 사원의 어떤 업무에는 조감이지만 다른 사원의 다른 업무에는 상세한 작업이 될 수도 있기 때문이다. 그래서 회사의 각 업무와 관련한 데이터를 조감 한다는 게 구체적으로 데이터를 어떻게 보는 것인지를 절대적인 기준으로 정의하기는 어렵다. 마치 새가 어느 정도의 높이에서 내려다보는 게 조감인지를 정하는 기준은 없는 것과 마찬가지이다.

이럴 때에는 우선 이용하려고 하는 모든 데이터를 산포도로 플롯팅해 본다. 일반적으로 데이터를 구성하는 항목은 여러 개가 있으므로 산포도의 X축과 Y축에 사용하는 항목을 바꾸어 가면서 데이터 전체가 어떤 풍경이 되는지를 살펴본다. 만약 〈그림 1〉과 같은 신체검사 데이터라면 체중과 신장을 기준으로 플롯팅해 보고, 비만도 평가 결과와 측정 날짜를 기준으로 플롯팅 해보고, 성별과 비만도 평가 결과를 플롯팅해 본다. 그 다음에는 중요한 항목을 기준으로 조금 상세하게 구분하여 전체 데이터를 나타내 본다. 예를 들어 연령별고 구분하거나, 지역별로 구분하거나, 성별로 구분한다.

이러한 작업을 계속 진행하다가 어느 순간에 데이터 전체의 풍경과 구조를 바라보는 목적과 관점을 만족시켜주는 그림을 발견하게 된다면, 이 그림이 나타내는 것을 조감이라고 정의한다. 데이터를 가시화하고 조감하는 목적은 데이터를 보석으로 이용하기 위해서이다. 이를 위해서는 가시화와 조감을 통하여 사람의 지식을 최대한으로 활성화시키는 것이 데이터의 본질을 이해하기 위한 가장 기본적이고 중요한 작업이다.