실제사례 통해 데이터를 보석으로 활용하기 위한 세 가지 요소 3 상관

"데이터 간의 상관 관계를 이해하라"

데이터간 상관관계 이해…데이터를 유용하게 활용하는 상황 많아

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 세가지 요소, 즉 시간, 조감, 상관에 대하여 그 특징을 이해하여야 한다. 지난 호의 시간과 조감에 대한 설명에 이어서 이번 호에서는 상관 요소에 대하여 설명한다.

윤 태 성
KAIST 교수, 윤츠 사장
Yoon.taesung@kaist.ac.kr

필자는 KAIST 경영과학과와 기술경영전문대학원 겸직 교수이며 윤츠와 오픈놀리지(동경) 사장이다. 저서에는 [오픈놀리자-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.com 이며, 참고 사이트는 http://untz.co.kr.

1 상관 관계란

다른 사람 일에 잘 참견하거나 끼어들기 좋아하는 사람에게는 "너하고 무슨 상관인데?"라고 묻기도 한다. 그러면 돌아오는 것은 "친구니까" 하는 일반적인 대답이나, 경우에 따라서는 "사회 정의를 위해서"라는 전혀 기대하지 않았던 뜻밖의 대답이 돌아오기도 한다.

갱 영화를 보면 나를 중심으로 하여 "친구의 친구는 친구"라거나 "적의 적은 아군"이라는 표현도 많이 있다. 많은 사람이 섞여 있는 복잡한 사회 구조일수록 나를 중심으로 친구 아니면 적이라는 단순한 관계로 표현하여 사람을 구분하려고 하는 것이 사람의 본능인 것 같다. 이런 식이라면 처음 보는 사람이라도 친구인지 적인지 곧바로 구분이 될 것이다. 그런데 사람들이 가지고 있는 이런 특성은 데이터를 보석으로 활용하고 싶을 때에 매우 유용하게 활용된다. 왜냐하면 이 데이터와 저 데이터 사이에는 어떤 상관 관계가 있을까? 라는 의문을 가지고 데이터를 보면 마치 진흙 속에서 진주를 찾는 것 같은 발견을 하기 때문이다.

현실적으로 진흙과 같은 상황은 우리 주변에 일상적으로 존재한다. 만약 여러분이 이러한 진흙과 같은 상황에 부딪히면 어떻게 대응할 것인가. 대답은 간단하다. 우선 데이터간의 상관 관계를 이해하는 것이 전체적인 문제의 구조를 이해하고 현실적인 해결방법을 찾는데 매우 도움이 된다고 하는 것이다. 아래의 상황을 가지고 논의를 계속해 보자. 여기서는 상관을 관계 혹은 관련과

같은 의미로 사용한다.

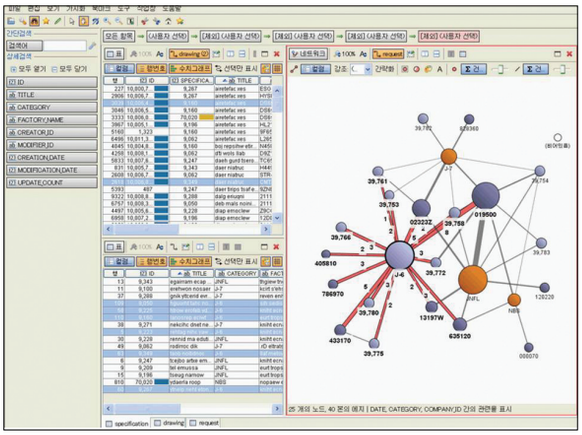
#상황1 (제조업의 도면 수정) 여기에 도면이 수백만 장 있다. 만약 여기 있는 이 도면을 수정한다면 이어서 어떤 도면을 수정하여야 할까?
#상황2 (자원의 국제 무역 현황) 자원 무역은 전 세계를 무대로 이루어진다. 특정한 자원의 주요 생산국과 주요 수입국간에는 어떤 특징이 있을까?
#상황3 (관련자 그룹 파악) 어떤 사건과 관련된 용의자가 있는데 아마 관련자가 있을 것으로 추측된다. 만약 이 용의자가 최근에 전화한 상대방 중에 관련자가 있다면 그건 누구일까?
#상황4 (소셜 네트워크 분석) 지금까지 우리 회사의 상품을 구입한 고객은 백만 명이다. 이번에 신상품을 개발하였는데 어느 고객에게 우선적으로 안내를 하면 좋을까?

제조업의 도면 수정

제조업은 설계 도면과 기술 자료가 가지고 있는 정보를 바탕으로 제조 프로세스를 통하여 재료를 제품으로 바꾸는 업종이다. 설계 도면은 한번 작성하면 그대로 사용되는 것이 아니라 일반적으로는 여러 번의 수정을 거치면서 갱신되어 간다. 도면을 수정하기 위한 이유나 시기가 타당하다면 수정 그 자체가 문제가 되는 것은 아니다. 현실적인 문제는, 관련되는 도면이 방대하게 있을 경우에 특정한 도면이 수정되면 그 영향을 받아서 연속적으로 수정되어야 하는 도면이나 기술 자료를 쉽게 파악하기가 어렵다는 점이다. 특히 대형 구조물이나 복잡한 인공물의 경우에는 관련된 도면과 기술 자료가 수백만 건으로 방대하기 때문에 수정의 영향 범위를 파악하고 연속적으로 수정해나가야 하는 도면을 금방 알기가 어렵다. 이런 상황에서는 도면간의 상관 관계를 중심으로 데이터를 가시화 하면 수정의 영향을 파악하기 쉬워진다.

〈그림 1〉은 일본의 모 제조기업에서 데이터간의 상관 관계를 이용하여 원자력 발전소의 도면과 기술 자

■ **그림 1 도면을 수정할 경우의 영향 범위**



료를 관리하고 있는 사례이다. 왼쪽 위에 있는 테이블에서 진하게 나타나 있는 부분은 특정 도면이 수정될 경우에 일차적으로 영향을 받는 도면의 리스트를 나타낸다. 여기에 있는 도면에 의해서 새롭게 영향을 받는 도면은 왼쪽 아래에 있는 테이블에 진하게 나타난 부분이다. 즉 특정한 도면이 수정될 경우에 일차적으로 영향을 받는 도면이 위에, 이차적으로 영향을 받는 도면이 아래에 나타나 있다. 오른쪽에 있는 그림은 도면이 수정될 경우에 일차적, 이차적으로 영향을 받는 도면을 네트워크 그림으로 가시화한 결과이다.

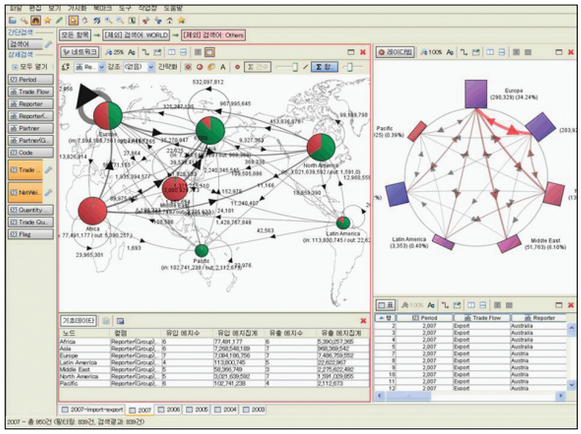
자원의 국제 무역 현황

지금은 전 세계적으로 자원 확보 전쟁이 일어나고 있다. 특히 희소 금속 등 매장량이 매우 적으면서 산업에 절대적으로 필요한 자원의 경우에는 수출국과 수입국 사이에 특징이 있다. 무역 현황을 가시화하여 지도 위에 표시하면 이러한 특징을 쉽게 이해할 수 있다.

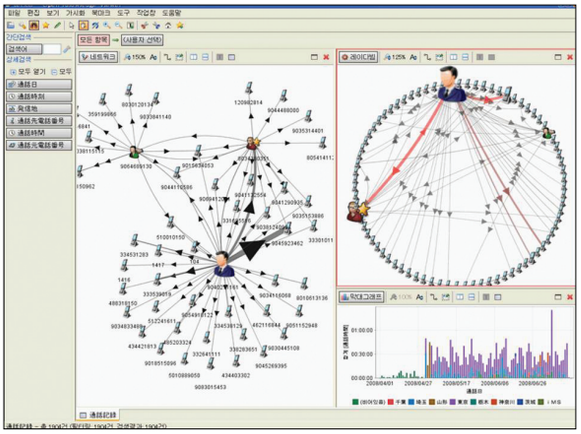
〈그림 2〉는 백금에 대한 국제 무역 현황을 가시화한 결과이다. 백금은 공업용 촉매로 사용되거나 보석으로 사용되거나 하는데 전 세계적으로 매장량이 적은 희소금속이다. 무역 현황을 나타내는 네크워크의 노드는 크기와 색깔이 다르게 나타나 있다. 노드의 크기는 무역 금액의 크기에 비례한다. 노드의 색깔은 수입과 수출을 구분하고 있는데 빨간색은 수출을, 초록색은 수입을 나타낸다.

〈그림 2〉를 보면 아프리카의 경우에는 거의 다 빨간

■ 그림 2 자원의 국제 무역 현황



■ 그림 3 특정 인물간의 전화 통화 데이터



색인데, 이는 아프리카는 백금을 거의 수출만 하고 있는 현황을 나타낸다. 유럽은 수입액과 수출액이 거의 동일한데 무역이 지역 내에서 많이 일어나고 있는 점이 재미있다. 호주는 액수는 적지만 거의 다 수입에 의존하고 있는 현황도 파악이 된다.

오른쪽에 있는 레이더 빔 그림은 대륙 간의 수출 수입 관계를 나타내고 있는데 데이터간의 상관 관계를 나타낸다는 점에서는 왼쪽에 있는 네트워크 그림과 본질적으로 동일하다. 왼쪽과 오른쪽의 아래에 위치한 테이블에는 상세한 수치 데이터가 나타나 있는데, 만약 네크워크상의 노드나 엣지를 선택하면 그와 관련된 수치가 진하게 표현되므로 상세한 수치를 파악하기가 쉽다.

관련자 파악

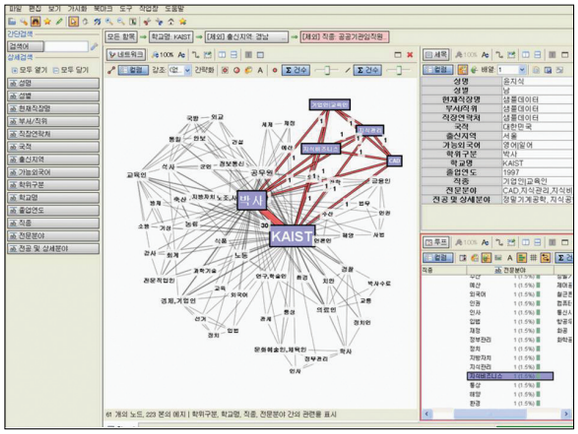
요즘은 휴대전화를 소지하는 것이 상식적인 일이 되었기 때문에 어떤 사건이 발생하면 수사 기관에서는 우선 특정한 인물이나 특정한 지역을 중심으로 전화 통화 내역을 분석하게 된다. 통화 내역 중에 의심이 가는 부분이 있으면 그 통화를 중심으로 보다 상세히 내역을 분석할 것이다. 여기서 현실적인 문제는, 방대한 양의 통화 기록 중에서 어떻게 의심이 가는 통화나 관련자를 파악하는가 하는 것이다.

전화 통화 분석이라는 상황에 대하여 통화 데이터를 가시화한 것이 〈그림 3〉이다. 왼쪽에 있는 네트워크 그림에서는 지정한 시간 중에 통화한 모든 내역을 가시화 한 결과이다. 네트워크는 유향 그래프로 표현되어 있는데 화살표의 방향은 전화를 걸은 방향을 나타낸다. 그림의 중심에 위치한 사람 모양의 아이콘은 통화 분석을 위한 특정 인물을 나타낸다. 이 그림을 보면 특정 인물 외에도 다른 두 사람이 전화 통화의 중심이 되어 있는 것을 알 수 있다. 물론 전화 통화의 중심이 되는 인물이 반드시 관련자라는 것은 아니지만 적어도 특정 인물을 중심으로 한 전화 통화의 특징은 쉽게 파악할 수 있다. 오른쪽 위에 있는 레이더 빔 그림은 어느 번호간에 전화 통화가 빈번하게 이루어 지는지를 나타내고 있다. 오른쪽 아래에 있는 막대 그래프는 모든 통화에 대하여 일자 별로 구분하여 나타낸 결과이다. 특정한 날짜에 통화량이 많거나 적거나 한 것을 알 수 있다. 여기에 나타낸 세가지 그림은 서로 동기화 되어있기 때문에 만약 막대 그래프 상에서 특정한 날짜를 선택하면 그 날 통화한 내용이 네트워크 그림과 레이더 빔 그림에 반영되어 색깔이 변화한다.

소셜 네트워크 분석

고객의 정보를 별도로 수집하여 관리하는 회사도 있고 고객에게서 받은 명함을 공동으로 관리하는 회사도 있는데, 이는 고객의 정보를 회사 내에서 공유하여 고객관리를 효율적으로 하기 위함이다. 교환한 명함에 나타나 있는 내용만으로도 고객에 대한 많은 데이터를 알 수 있는데, 여기에 더해서 비즈니스를 통해서 알게

■ 그림 4 고객 데이터를 이용한 소셜 네트워크



된 데이터까지 포함한다면 고객 데이터의 축적은 쉽게 할 수 있다. 현실적인 문제는 이러한 데이터를 어떻게 활용할 것인가이다. 〈그림 4〉는 가공의 고객 데이터를 가시화한 결과이다. 왼쪽에 있는 네트워크 그림은 특정한 인물이나 특정한 조직 혹은 특정한 직업을 중심으로 모든 고객 간의 관계를 나타내고 있다. 왼쪽 위에 있는 테이블에는 특정 인물에 대한 상세한 데이터를 나타내고 있다. 왼쪽 아래에는 고객 전체에 대하여 통계치를 나타내고 있다. 왼쪽에 있는 네트워크상에서 특정한 인물을 선택하면, 네트워크상에는 관련이 깊은 사람을 빨간 색의 엣지로 나타내며, 오른쪽 위에는 특정한 인물 본인의 데이터가 나타나는 방식이다.

만약 회사에서 개발한 신상품을 우선적으로 안내를 하고 싶은 고객을 찾으려면 우선 신상품과 관련된 인물을 찾는다. 찾는 방법은 예를 들면 그림4의 왼쪽 상단에 위치한 검색 지원 창에 관련된 용어를 입력하여 적절한 사람을 발견한다. 그 후에 이 사람을 중심으로 소셜 네트워크를 만들어서 여러가지 관점에서 네트워크를 분석한다. 소셜 네트워크 분석은 사람과 사람의 관계를 중심으로 새로운 관계를 발견해 나가거나 만들어 나가기 위한 방법이다.

2 상관 관계 사례들

앞의 상황 이외에도 데이터간의 상관 관계를 이해함으로써 데이터를 유용하게 활용하는 상황은 현실적으로 매우 많다.

예를 들면, 편의점의 상품 배치가 그러하다. 편의점에서는 상점에 들어오는 손님의 움직임을 파악하여 상품의 진열 위치를 결정한다. 그래서 특히 체인화되어 있는 편의점에 가보면 상품이 진열된 위치가 거의 표준화 되어 있다. 만약 처음 가보는 지역의 처음 보는 편의점이라 하여도 과자를 사거나 캔 커피를 찾는데 별로 망설임이 없게 된다. 이는 손님이 구매하는 상품 데이터간의 상관 관계를 파악하여 상관 관계가 깊은 상품은 가급적 가까운 곳에 진열하기 때문이다.

이와 유사한 상황으로 인터넷 상점이 있다. 인터넷 상점에서는 손님이 접속하여 내용을 열람하는 특징을 파악하여 홈페이지 상에서 상품의 전시 위치를 정하게 된다. 회원 등록한 손님이 접속해 오면 그 손님의 개인 데이터를 바탕으로 지금까지의 구매 이력을 파악한 후에 추천 상품을 제시하기도 한다. 그런데 회원 등록하지 않은 손님이 접속하는 경우에는 그 손님의 데이터가 없기 때문에 과거의 구매 데이터를 분석할 수 없다. 이런 경우에는 그 손님이 내용을 열람하는 특징을 바탕으로 상품을 추천하게 된다. 만약 가죽 장갑을 열람하였다면 이어서 머플러나 내복을 추천하는 식인데, 특정한 상품을 구매한 손님은 통계적으로 이러이러한 상품을 구매하는 경우가 많다고 하는 확률론적인 상관 관계를 바탕으로 추천 상품을 정하는 방식이 일반적이다. 그러나 처음 접속하는 손님이 어떤 내용을 어떤 순서로 열람하는지 네트워크 그림으로 그려서 살펴보면 통계분석으로는 보이지 않던 특징을 발견할 가능성이 있다.

이러닝에서의 맞춤형 학습 지원도 상관 관계가 중요하다. 문제 은행이 있어서 학생 한 명마다 실력에 맞게 문제를 내고, 대답하는 결과를 보면서 개별적으로 지도할 수 있도록 자동화된 이러닝 시스템을 개발한다고 하자. 어떤 문제를 풀거나 혹은 못 풀은 학생이 있다면 이어서 어떤 문제를 내면 좋을까 정하기 위해서는 문제간의 상관 관계가 기준이 된다.