

미래 인재로의 도약, 스마트기술 기반 다지기	
06차시	빅데이터 분석 기법과 도구

## 1. 정형 데이터마이닝

### 1) 정형 데이터마이닝의 이해

#### 가. 데이터마이닝 개념

데이터마이닝(Data Mining)은 대용량의 데이터로부터 자동 또는 반자동적인 방법을 통하여 의미 있는 패턴, 규칙, 관계를 찾아내는 것이다. 데이터마이닝은 또한 많은 데이터베이스로부터 지금까지 잘 알려지지 않고 유용하며 활용이 가능한 정보를 추출하는 과정으로 정의가 되기도 한다. 기업이나 정부는 다양한 활동을 통해 대용량의 데이터를 축적해 왔다. 그러나 많은 양의 데이터들은 수치화가 되지 않았을 뿐만 아니라, 수치적 형태보다는 비수치적 형태로 저장되어 일반적인 통계 방법에 의해 분석과 활용이 될 수 없었다. 하지만 이러한 데이터에는 미처 발견하지 못한 패턴과 전략에 도움이 될 만한 정보들이 들어있을 수 있기 때문에 데이터를 정제하고 가공할 필요성이 생겨나게 된 것이다. 이러한 데이터를 분석하여 기업에 필요한 자산으로 만드는 디지털 기술이 바로 데이터마이닝이다.

#### 나. 데이터마이닝의 특징

기업은 업무의 효율적 수행을 위해 데이터베이스를 이용하고, 데이터베이스의 내용 및 결과를 단순히 활용하는 단계를 벗어나, 데이터 자체의 분석을 통해 패턴을 추출해내고 이 결과를 업무와 생산의 효율성 증대를 위해 이용하는 단계로 넘어가고 있다. 그러나 데이터가 방대해지고 기업의 업무가 복잡해지면서 데이터베이스를 관리하고 자료를 분석하는 전문가의 능력에 한계가 있고, 데이터에 내재된 유용한 지식 추출 작업을 사람이 손으로 직접 하는 것이 불가능하게 되었다.

데이터 마이닝은 이와 같은 문제를 해결하고 대량의 데이터에서 유용한 패턴과 지식을 추출하고자 하는 기법이다. 데이터마이닝은 사용자의 경험이나 편견을 배제하고 전적으로 데이터에 기반하여 지식과 패턴을 추출하기 때문에 영역 전문가가 간과해 버릴 수도 있는 지식과 패턴을 찾아낼 수 있다. 데이터마이닝의 활용 분야는 카드사의 사기 발견이나, 금융권의 대출 승인, 투자 분석, 기업의 마케팅 및 판매 데이터 분석, 생산 프로세스 분석, 기타 순수 과학 분야의 자료 분석 등 헤아릴 수 없이 많다. 정형 데이터마이닝 기법으로는 정형 데이터 분석을 다루는 연관관계분석, 군집분석, 의사결정나무, 인공신경망, 사례기반추론 등이 있다.

### 2) 연관관계분석

#### 가. 연관관계 분석의 개념

연관관계 분석은 상품 혹은 서비스 간의 관계를 살펴보고 이로부터 유용한 규칙을 찾아내고

자 할 때 이용될 수 있는 기법이다. 동시 구매될 가능성이 큰 상품들을 찾아내는 기법으로 장바구니 분석과 관련된 문제에 많이 적용되어 왔다. 측정의 기본은 얼마나 자주 구매되었는가 하는 빈도를 기본으로 연관 정도를 정량화하기 위해서는 지지도, 신뢰도, 향상도를 계산하여 기준으로 한다.

연관성 규칙의 기본적인 개념은 장바구니 품목들을 식별하는 것에서부터 시작되었다. 다시 말하면 사건들은 동시 다발적으로 발생하며, 이러한 사건들은 상호 영향을 주면서 결과를 나타나게 되는데 이와 같이 사건 또는 품목 간에 일어나는 연관성을 규명하려는 것이 연관성 규칙이다. 즉, 연관성 규칙이란 두 항목 간 그룹 사이에 강한 연관이 존재하는지에 대한 기술을 말한다.

### **나. 연관관계 분석의 특징**

연관성 규칙은 상품 또는 서비스 간의 관계를 살펴봄으로써 그들 간의 유용한 관계가 존재하는지 알아보고자 할 때 적합한 방법이라고 할 수 있다. 구체적인 행위를 언급하여 규칙을 도출하기 때문에 이해하기 쉽고 명쾌한 특성을 가지고 있으며, 실질적인 정보를 도출할 수 있는 장점을 가지고 있다. 이러한 이유로 연관성 규칙은 마케팅 문제 뿐만 아니라 광범위한 의사결정을 하는데 널리 사용되고 있다.

## **3) 군집 분석**

### **가. 군집분석의 개념**

군집 분석(Cluster Analytics)은 전체 데이터를 군집을 통해 잘 구분하는 것으로 다양한 특징을 가진 관찰 대 상으로부터 동일 집단으로 분류하는데 사용한다. 이는 유사한 특성을 가진 개체를 합쳐가면서 최종적으로 유사 특성의 군집을 찾아내는 분류 방법으로 구분하려고 하는 각 군집에 대한 아무런 사전 지식이 없는 상태에서 분류하는 것이므로 무감독 학습(Unsupervised Learning)에 해당한다. 즉, 개체들에 대한 사전 지식 없이 유사도에 근거하여 군집들을 구분한다. 이러한 군집 분석의 종류는 대상을 어떻게 분석할지에 따라 다음과 같이 계층적 군집 분석과 비계층적 군집 분석으로 구분할 수 있다.

### **나. 계층적 군집 분석**

계층적 군집 분석은 개별대상 간의 거리의 의하여 가장 가까이에 있는 대상들로부터 시작하여 결합해 감으로써 나무 모양의 계층 구조를 형성해가는 방법이다. 계층적 군집 분석은 군집 간의 거리와 유사성을 정하는 방법에 따라 단일연결방식, 완전연결방식, 집단간 평균연결방식, 집단내 평균연결방식 등으로 구분할 수 있다.

계층적 군집 분석은 덴드로그램(Dendrogram)을 그려줌으로써 군집이 형성되는 과정을 정확히 파악할 수 있으나 데이터의 크기가 크면 분석하기가 어렵다는 단점을 갖는다. 또한 한 개체가 일단 특정 군집에 소속되면 다른 군집으로 이동될 수 없으며, 예외 값(outlier)이 제거되지 않고 반드시 어느 군집에 속하게 된다는 한계점을 갖는다.

### **다. 비계층적 군집 분석**

군집 분석에서 개체의 수가 많은 경우에는 개체들간의 유사성을 구하는 것이 번거롭고 어려운 일이다. 비계층적 군집 분석은 계층적 군집 분석과 달리 군집의 수를 한 개씩 감소시키는 것이 아니라 사전에 정해진 군집의 숫자에 따라 대상들이 군집들에 할당하는 방법이다. 즉, 구하고자 하는 군집의 수를 정한 상태에서 설정된 군집의 중심에 가장 가까운 개체를 하나씩 포함해 가는 방식으로 군집을 형성해간다. 이 방법은 많은 데이터를 빠르고 쉽게 분류할 수 있으나 군집의 수를 미리 정해 주어야 하고, 군집을 형성하기 위한 초기 값에 따라 군집 결과가 달라지는 단점이 있다.

## 4) 의사결정나무

### 가. 의사결정나무의 개념

의사결정나무(Decision Tree)는 데이터마이닝의 주요 기법 중 하나로서 분류 및 예측에 주로 사용이 되는 기법이다. 경영, 경제에 관련된 다양한 분야의 예측에 이용이 되고 있는 이 기법은 사용이 비교적 용이하고 그 결과를 이해하기가 수월하다는 장점을 가지고 있다. 데이터를 분석하여 나온 결과물이 의사결정나무라는 그래프 형식으로 표현이 되기도 하며, 또한 규칙 셋이라는 형식으로도 표현이 되기도 한다. 이러한 그래프와 규칙이라는 다양한 표현 형식은 다양한 다른 기법과 융합적 사용이 용이해 다양한 예측에 사용이 될 수 있다.

## 5) 인공신경망

### 가. 인공신경망의 개념

인공신경망은 생물학적 뇌의 작동 원리를 그대로 모방하는 방법으로, 데이터 안의 독특한 패턴이나 구조를 인지하는데 필요한 모델을 구축하는 기법이다. 인공신경망은 간단한 계산 능력을 가진 처리 단위, 뉴런(neuron) 또는 노드(node)들이 서로 복잡하게 연결된 컴퓨터 시스템으로서 외부에서 주어진 입력에 대하여 반응을 할 수 있다.

### 나. 인공신경망의 특징

인공신경망은 복잡하고 비선형적이며 관계성을 갖는 다변량을 분석할 수 있다. 인공신경망 기법은 회귀분석과 같은 선형 기법과 비교하여 비선형 기법으로서의 예측력이 뛰어나며, 자료에 대한 통계적 분석 없이 결정을 수행할 수 있다. 인공신경망은 통계적 기본 가정이 적고 유연하여 다양하게 활용이 된다. 특히 데이터 사이즈가 작은 경우, 불완전 데이터, 노이즈 데이터가 많은 경우 인공신경망 모델의 성능이 일반적으로 다른 기법과 비교해서 우수하다고 평가된다.

## 6) 사례기반추론

### 가. 사례기반추론의 개념

사례기반추론이란 과거에 있었던 사례들의 결과를 바탕으로 새로운 사례의 결과를 예측하는 기법이다. 과거에 발생한 문제는 미래에 다시 비슷한 형태의 문제로 발생할 가능성이 높고 새로운 문제를 해결할 수 있는 정답이 과거의 문제를 해결했던 정답과 유사할 것이라는

가정이다. 사례기반추론은 과거 사례들을 저장해 둔 사례기반으로부터 해결하고자 하는 새로운 사례와 가장 유사한 사례를 검색한 후, 유사 사례의 해결책을 바탕으로 당면한 문제의 해결책을 제안하는 과정으로 진행된다. 이때 제안된 해결책은 필요에 따라 적절히 수정된 후에 주어진 문제를 풀기 위해 재사용되며 이렇게 해결된 새로운 사례는 추후 다른 문제 해결에도 도움이 될 수 있도록 새로운 사례로 사례기반에 저장된다.

사례기반추론을 이용하기 위해서는 일반적으로 과거의 사례와 사례들 사이의 유사 정도를 측정하기 위한 유사도 척도가 준비되어야 한다. 유사도 측정 도구는 여러 가지 방법이 제안되고 있지만 일반적으로 근접이웃방법론이 가장 많이 이용되고 있다. 근접이웃방법론을 적용하려면 속성의 값들 간 유사성 정도를 측정할 수 있는 속성 유사성 함수를 정의하여야 하고, 이를 이용 하여 과거 사례의 속성 값들과 해결하고자 하는 문제의 속성 값들에 대한 유사성 정도를 측정하고, 이를 속성의 중요도에 따라 가중 합계를 계산하여 사례들 사이의 유사도를 측정하게 된다.

## **2. 비정형 데이터마이닝**

### **1) 비정형 데이터마이닝의 이해**

빅데이터 환경에서는 거의 80% 이상이 비정형 데이터이므로 빅데이터에서의 데이터마이닝은 비정형 데이터마이닝에 초점이 맞추어져 있다. 일반적으로 데이터마이닝은 통계 기반의 데이터 분석 도구를 사용하거나 OLAP(OnLine Analytical Processing) 분석을 통해 데이터를 다양한 각도의 관점으로 조명하여 의미 있는 것으로 해석하며, 데이터 사이의 숨겨진 관계와 패턴, 경향 등을 추출하는 것을 말한다. 결국 비정형 데이터 마이닝은 비정형 데이터에 대한 정제 과정을 통해 정형 데이터로 변환하고 난 다음에 분류, 군집화, 회귀분석, 요약, 이상 감지 등에 적용하여 의미 있는 정보를 발굴해낸다는 것을 의미한다.

비정형 데이터마이닝은 보통 탐색, 이해, 분석의 과정을 거친다. 탐색 과정에서는 질의, 집합 연산, 재귀 및 팽창 등의 작업을 수행한다. 이해 과정에서는 통계, 분배, 특징 선택, 군집화, 분류 편집, 시각화 등의 작업을 수행한다. 그리고 분석 과정에서는 경향, 상관관계, 분류 등의 작업을 수행한다. 정제된 데이터베이스를 기반으로 일정한 기준이 적용된 상식적인 범위에서 부분적인 데이터를 다루는 정형 데이터마이닝의 한계를 뛰어넘는 대표적인 비정형 데이터마이닝 기법으로는 텍스트마이닝, 웹마이닝, 오피니언마이닝, 소셜 네트워크 분석 등이 있다.

### **2) 텍스트마이닝**

#### **가. 텍스트마이닝의 이해**

인터넷 자료, 이메일, 여러 분야의 논문, 신문 또는 잡지의 기사, 여론조사 보고서 등 우리의 실생활에서 만들어지는 대부분의 자료는 텍스트 형태를 띤다. 텍스트마이닝(Text Mining)은 이러한 비정형 텍스트 데이터들을 자연어 처리(natural language processing) 방식을 이

용하여 정보를 추출하거나, 연계성을 파악하거나, 분류 혹은 군집화, 요약 등 빅데이터에 숨겨진 의미를 발견하는 기법을 말한다.

텍스트마이닝 기술을 사용하여 먼저 인간 중심의 비정형 데이터에서 자연어 처리 기술을 적용하여 추출한 텍스트에서 의미 있는 숫자나 단어 인덱스를 추출하고, 텍스트에 포함된 정보를 통계 및 규칙적인 기계학습과 같은 다양한 데이터마이닝 알고리즘에 의해 접근할 수 있도록 만들어 의미 있는 정보를 추출한다. 그리고 텍스트 정보에 의해 문서에 포함된 단어를 요약하거나 텍스트 정보 안에 포함된 단어를 기준으로 문서를 요약하기 위해 텍스트 정보를 추출할 수 있으므로 문서 등에 있는 단어나 단어의 군집을 분석할 수도 있고, 여러 문서를 분석하여 문서들 사이의 유사성이나 관련성을 파악할 수도 있다. 따라서 텍스트마이닝은 웹마이닝, 오피니언마이닝, 소셜 네트워크 분석 등과 같은 다른 비정형 데이터마이닝의 근간이 되는 기법이라고 할 수 있다.

#### 나. 텍스트마이닝 과정

텍스트마이닝 과정은 여러 종류의 텍스트 데이터로부터 지식을 발견하는 과정이다. 지식 발견이라는 측면에서 텍스트마이닝의 목적은 비정형 데이터나 정형 데이터, 반정형 데이터를 처리하여 의사결정을 위해 필요한 고차원적이고 의미 있는 정보나 지식을 추출하는 것이다. 따라서 텍스트마이닝 과정의 입력 데이터로는 처리 과정을 위해 텍스트 문서로부터 수집되거나 저장되거나 만들어진 비정형 데이터나 정형 데이터, 반정형 데이터가 해당된다. 그리고 텍스트마이닝 과정의 출력물로는 의사결정을 위해 사용될 수 있는 텍스트의 패턴이나 관계와 같은 특별한 의미의 지식이 해당된다.

결국 거시적 측면에서 텍스트마이닝 처리 과정은 입력, 처리, 출력이라는 정보 처리의 기본 처리 과정을 따른다고 할 수 있다. 입력과 출력을 제외하고 처리 과정에만 국한하여 미시적으로 살펴보면 텍스트마이닝 과정은 준비 단계, 전처리 단계, 지식 추출 단계를 밟는다.

준비 단계는 입력되는 여러 가지 텍스트 문서의 데이터들을 문제 범위에 적절한 것으로 확립하는 것이다. 일부 텍스트 분석에서는 진보된 통계 방법을 적용하기도 하지만 대부분은 정보검색이나 텍스트 식별을 말하며, 웹상에서나 파일 시스템, 데이터베이스, 내용 관리 시스템 등에서 문제 범위에 맞는 일련의 텍스트들을 수집하거나 식별하는 것이다. 이렇게 수집된 텍스트들은 텍스트 파일과 같은 컴퓨터 처리에 적합하게 통일된 형태로 디지털화되고 조직화된다.

전처리 단계는 준비 단계에서 문제 범위에 맞게 조직화된 텍스트들을 정형화된 표현 양식으로 만드는 것이다. 용어와 텍스트 문서의 행렬이 너무 크게 되면 처리하기 힘들어지므로 빈도가 지나치게 적은 것과 전문가 입장에서 문제 영역에서 멀다고 생각되는 것을 제거하고 특이 값 분해를 통해 행렬의 전반적인 의미 구조가 나타나도록 하여 다루기 쉬운 크기로 줄인다. 즉 전처리 단계의 목적은 준비 단계의 문제 범위에 조직화된 데이터를 의미 구조를 갖으면서도 다루기 쉬운 정형화된 데이터로 변환시키는 것이다.

지식 추출 단계는 문제에 맞게 변환된 정형 데이터에서 의미 있는 패턴이나 관계와 같은 지

식을 발견하는 것이다. 여기에는 분류, 클러스터링, 개념 및 개체 추출, 세분화된 분류 체계의 생산, 심리 분석, 문서 요약, 개체 관계 모델링 등이 있다. 여기서 텍스트 분류는 분류 체계를 가지고 텍스트 내용을 보고 주제에 따라 분류하는 방법이다. 텍스트 클러스터링(clustering)은 분류 체계를 모르는 상태에서 성격이 비슷한 것끼리 같은 군집으로 묶어주는 방법이다.

### 3) 웹마이닝

웹마이닝(Web Mining)은 데이터마이닝 기술의 응용 분야로서 인터넷을 통해 웹 서비스를 이용하면서 웹에서 패턴을 발견하는 것을 말한다. 웹마이닝은 전통적인 데이터마이닝의 분석 방법론을 사용하면서도 웹 데이터의 속성이 반정형 이거나 비정형이고, 링크(link) 구조를 가지고 있기 때문에 전통적인 데이터마이닝 기술에 추가적인 분석 기법이 필요하다. 웹마이닝은 분석 대상에 따라 웹 사용 마이닝, 웹 구조 마이닝, 웹 콘텐츠 마이닝(Web Content Mining) 등으로 구분할 수 있다.

#### 가. 웹 사용 마이닝

이는 웹상에서 사용자가 찾고자 했던 것을 기록하고 있는 웹서버 로그(Web Server Log)에서 유용한 정보를 추출하는 과정을 말한다. 웹 사용 마이닝은 웹 기반 애플리케이션이 필요로 하는 것을 이해하고 서비스해 주기 위해 웹에서 흥미 있는 사용 패턴을 발견하는 데이터마이닝 기술의 응용으로 이것을 이용하여 웹 사용자가 웹사이트에서 사용한 데이터를 통해 나타난 행위에 따라 그들의 특성과 성향을 뽑아낸다.

#### 나. 웹 구조 마이닝

이는 웹사이트의 노드와 연결 구조를 분석하기 위해 그래프(Graph) 이론을 사용하는 것을 말한다. 웹 구조 마이닝은 웹 구조 유형에 따라 웹에서 하이퍼링크로부터 패턴을 추출하는 것과 문서 구조를 분석하는 것으로 구분할 수 있다.

#### 다. 웹 콘텐츠 마이닝

웹 콘텐츠 마이닝은 웹페이지에서 유용한 데이터, 정보, 지식을 마이닝하고 추출하고 통합하는 것을 말한다.

### 4) 오피니언 마이닝

#### 가. 오피니언 마이닝의 이해

오피니언 마이닝(Opinion Mining)은 어떤 사안이나 인물, 이슈, 이벤트 등과 같은 원천 데이터에서 의견이나 평가, 태도, 감정 등과 같은 주관적인 정보를 식별하고 추출하는 것과 관련되므로 오피니언 분석, 평판 분석, 정서 분석이라고도 한다. 일반적으로 말해서 오피니언 분석은 어떤 화제나 문서의 전반적 문맥 특성과 관련된 작성자나 화자의 태도를 파악하는데 도움을 준다. 여기서 태도는 판단이나 평가, 효과적 상태나 의도된 감정적 의사소통 등에 대한 것일 수 있다. 오피니언 분석의 기본적인 작업은 문서, 문장, 특징, 관점 수준에서 표현된 견해가 긍정적인지, 부정적인지, 중립적인지, 진보적인지 주어진 텍스트의 특성을 분류하는

것이다.

오피니언 마이닝에서 주요 분석 대상은 포털 게시판, 블로그, 쇼핑몰과 같은 대규모의 웹 문서이기 때문에 자동화된 분석 방법을 사용하며 분석 내용이 주로 텍스트로 이루어져 있으므로 텍스트마이닝에서 활용하는 자연어 처리, 텍스트 분석, 컴퓨터 언어학 등의 기술을 사용한다. 오피니언 마이닝은 상품이나 서비스에 대한 시장 규모를 예측하거나 소비자의 반응 및 입소문을 분석하는데 활용되고 있는데 이를 위해서는 전문가들에 의해 선호도를 나타내는 표현이나 단어 등에 대한 자원을 축적해두는 것이 필요하다.

#### **나. 오피니언 마이닝의 활용**

오피니언 마이닝을 활용하여 온라인 쇼핑몰에서의 잠재 구매자의 상품평 검색 효율을 높이기 위해 상품평 데이터에 순위를 결정하는데 이용할 수도 있다. 영화 관람의 후기를 요약하고 긍정/부정을 평할 수도 있고, 법률 분야의 블로그를 대상으로 오피니언 마이닝을 이용해 고객의 반응이나 법률적 이슈에 대한 모니터링을 할 수도 있다. 또한 소셜 미디어에서 나타나는 오피니언들을 조기 감지하여 기업의 위기 상황을 인지하고 위기에 대응할 수 있는 위기관리 모델의 핵심 정보가 될 수도 있고 오피니언을 경제적 관점에서 정량화하여 금액으로 환산할 수도 있다.

### **5) 소셜 네트워크 분석**

#### **가. 소셜 네트워크 분석의 이해**

데이터마이닝을 통해 판매와 수익성을 개선할 수 있었던 회사들은 고객의 데이터와 온라인 행위를 포함하는 고객 프로필을 만들었고 최근에는 이러한 SNS 환경에 발맞추어 고객의 관계망을 형성함으로써 이를 통한 성향 분석 및 관계 분석을 통해 마케팅 전략을 수립하고자 하는 욕구가 증대되었다. 대표적인 SNS라 할 수 있는 메타의 경우 관심사와 취미, 위치 정보, 사회적 관계망을 기반으로 맞춤형 광고까지 하기에 이르렀다. 이와 관련하여 개인의 일상 정보가 연결된 사회적 관계망을 분석하는 것이 필요한데 그것이 바로 소셜 네트워크 애널리틱스(Social Network Analytics) 즉, 소셜 네트워크 분석이다.

소셜 네트워크 분석은 노드와 링크로 구성되는 네트워크 이론에 의해서 사회적 관계를 보여주는 것을 말한다. 이러한 네트워크는 사회적 관계도에서 노드(node)의 경우 점으로, 링크(link)의 경우 선으로 표현된다. 여기서 노드는 행위자를 의미하고 링크는 각 노드들의 관계에 해당된다. 관계는 우정, 연대감, 조직력, 성향 등을 나타낸다. 소셜 네트워크 연결 구조 및 연결강도 등을 바탕으로 노드의 복잡도를 측정하여, 소셜 네트워크 상에서 연결의 중심 역할을 하는 영향력이 있는 행위자를 파악한다. 이러한 영향력 있는 행위자를 파악하고 관리하는 것이 마케팅 관점에서 매우 중요하다.

#### **나. 소셜 데이터마이닝의 활용**

주요 데이터 서비스 기업들이 사용자의 로그, 관심사, 정보를 분석하여 트렌드를 감지하고, 브랜드 모니터링, 감성 분석, 마케팅 등을 제공할 수 있는 기반 환경을 서비스하고 있다. 구글에서는 구글 트렌드를 통해 실시간 핫 이슈 검색 기능을 제공, 실시간 순위 및 순위 차트

제공, 카테고리별 이슈 분류 기능, 기간별 설정 및 검색 기능을 제공하고 있다. 국내에서도 소셜 미디어 분석에서 언어 분석 기술을 적용해 검색어에 대한 기간별 소셜 모니터링, 연관어 탐색, 감성 분석 서비스 등을 제공하고 있다. 다음소프트의 경우 소셜 매트릭스 서비스를 제공하고 있는데, 소셜 매트릭스는 소셜 미디어 정보를 모니터링, 주별 급증한 키워드 순위를 제공, 연관어 기반 탐색 건수 제공, 감성 기반 연관어를 제공하고 있다.

### 3. 데이터 시각화

#### 1) 데이터 시각화의 개념

빅데이터는 데이터의 풍부함을 드러내기 위한 새로운 방식으로 방대한 양의 데이터를 탐색하거나 이해할 때 가장 좋은 방법으로 시각화(Visualization)를 활용한다. 데이터 시각화는 데이터 분석 결과를 사용자가 쉽게 이해할 수 있도록 시각적 수단을 통해 제시하는 것으로 도표나 이미지, 단어 구름 등을 이용하여 한눈에 이해할 수 있도록 하는 것이다. 최근 빅데이터 시대에서 그 중요도가 높아지는 것이 바로 데이터 시각화이다. 방대한 데이터가 빠르게 증가하는 빅데이터에서 통찰력을 얻기 위해 분석도 중요하지만 이를 한눈에 알아볼 수 있는 인지성 또한 중요하다. 바로 이러한 역할을 하는 것이 '데이터 시각화(Data Visualization)'이다.

데이터의 시각화 작업은 오랜 시간 동안 단순한 수치의 그래프나 데이터의 패턴을 파악하는 방법으로 사용되어 왔다. 최근 빅데이터의 이슈(issues)가 두드러지면서 다른 학문과 융합하여 다양한 정보 전달이나 상황 분석을 위한 시각적 도구로 메시지 전달을 위한 시각적 표현으로 많이 사용되고 있다.

#### 2) 데이터 시각화의 특성

빅데이터의 트렌드가 가속됨에 따라 분석 기법을 사용하는 기업들이 늘어나고, 데이터의 단순 나열보다는 분석된 데이터를 표현해 주는 '데이터 시각화'와 같이 예전에는 요구되지 않았던 새로운 기술들과 전문가가 요구되고 있다.

데이터 시각화는 다양하고 방대한 데이터를 탐색하는 가운데 데이터의 특징을 쉽고 빠르게 알 수 있도록 도와준다. 또한, 데이터에 감춰진 의미(narrative)를 찾아내어 이를 논리성과 심미성의 균형을 이루며 보여주는 것 역시 데이터 시각화의 주된 특성이다. 시각화는 커뮤니케이션 측면에서 다음과 같은 특성을 지닌다.

첫째, 시각화는 인간의 정보 처리 능력을 확장시켜 정보를 직관적으로 이해할 수 있게 한다.  
둘째, 많은 데이터를 동시에 차별적으로 보여줄 수 있다.  
셋째, 시각화는 다른 방식으로는 어려운 지각적 추론(Perceptual Inference)을 가능하게 한다. 예를 들어 눈에 보이지 않는 구조나 원리를 다양한 다이어그램, 상징, 기호로 시각화할 때 이해하기 쉽다.



넷째, 시각화는 보는 이로 하여금 흥미를 유발하여 주목성이 높아지며 인간의 경험을 풍부하게 한다.

다섯째, 시각화를 통해 문자보다 친근하게 정보를 전달하며, 다양한 계층의 사람들에게 쉽게 다가갈 수 있다.

여섯째, 시각화는 데이터 간의 관계와 차이를 명확히 드러냄으로써 문자나 수치에서 발견하기 어려운 이야기를 창출할 수 있다. 즉 데이터 시각화는 데이터 이면의 의미를 만든다.

일곱째, 시각화를 통해 데이터를 입체적으로 만들 수도 있으며, 필요에 따라 거시적 혹은 미시적으로 표현이 가능하고 위계를 부여할 수도 있다.

### 3) 데이터 시각화의 원리

데이터 시각화의 원리는 다음과 같이 요약할 수 있다.

첫째, “각각의 음식은 고유한 맛과 향이 있다.” 하나의 시각화는 그 데이터 셋에 표현하는 유일한 특성들만을 표현해야만 한다. 일반적인 시각화 도구들을 이용해서 표현하는 것은 제대로 된 시각화가 어렵다. 그 데이터만을 위한 고유의 시각화를 만들어 내야만 한다.

둘째, “확실한 일품요리를 차려라. 데이터 시각화는 당신이 좋아하는 뷔페가 아니다.” 덜 세세한 설명이 더 많은 정보를 전달한다. 너무 많은 정보는 청중을 혼란스럽게 하고, 정작 중요한 것을 전달하지 못한다. 가능한 한 소중한 정보만으로 최소화해야만 한다.

셋째, “손님이 원하는 식사를 제공해라.” 청중이 누구인가? 시각화에 접근하는 이들의 최종 목적은 무엇인가? 그들은 무엇을 얻으려고 하는가? 모바일 디바이스와 데스크탑을 통한 이용자의 목적은 다르다.

## 4. 빅데이터 분석 도구

### 1) 엑셀

엑셀은 마이크로소프트사에서 개발한 윈도우 환경의 스프레드시트 프로그램이다. 이 프로그램은 사용자에게 그래픽 환경을 제공하는데 스프레드시트 기능을 비롯해 매크로, 그래픽, 데이터베이스 기능과 차트 작성 등 문서 작성에 필요한 기능을 제공한다. 마이크로소프트사는 1985년에 엑셀 초기 버전을 개발한 뒤 꾸준한 개발을 통해 현재 엑셀 2022년 6월 기준으로 엑셀 2022(MS 오피스 2022 버전)을 제공하고 있다. 수식 작성과 함수 생성 및 계산이 편리하여 전 세계적으로 많은 사용자들이 사용 중인 프로그램이다.

엑셀의 장점은 다른 분석 툴에 비해 사용이 비교적 쉽다는 것이다. 다른 분석 툴이 데이터를 입력하기 위해 다양한 명령어를 사용해야 하는 것과 달리 엑셀에서는 복잡한 명령어 없이 사용자가 직접 해당 셀에 원하는 데이터를 입력할 수 있고, 기존에 존재하는 데이터를 불러와 수정, 사용하는 방법이 있어 사용자의 상황에 따라 선택적 사용이 가능하다. 뿐만 아니라 데이터 핸들링이 어렵고 명령어를 직접 암기하여 입력해야 하는 다른 분석 툴과 달리 엑셀은 ‘데이터’ 리본 메뉴에서 제공되는 다양한 방법을 마우스 클릭을 통해 사용할 수 있다. 단순 평균 비교부터 회귀분석과 시계열 분석과 같은 고급 통계 분석 또한 데이터 → 분

석 → 데이터 분석 클릭을 통해 손쉽게 할 수가 있다.

## 2) SPSS

SPSS는 Statistical Package for Social Science의 약자로 사회과학의 자료 분석을 위해서 고안된 프로그램이다. 이는 광범위한 데이터의 핸들링이 가능하고 다양한 통계 분석이 가능하며 널리 사용되고 있는 통계 분석 전용 프로그램이다. SPSS는 1969년 사회과학 분야의 데이터 분석을 위해 시카고 대학의 전미여론조사센터(National Opinion Research Center)에서 컴퓨터 프로그램의 모음집으로 출발하게 되었다. 그러다가 2009년 IBM사에 인수되면서 정식 명칭이 IBM SPSS Statistics로 변화하였고, 2022년에 나온 IBM SPSS Statistics 28 버전이 가장 최신 버전으로 판매되고 있다.

IBM SPSS Statistics(SPSS)는 비즈니스 사용자나 분석가 또는 통계 프로그래머에게 적합한 프로그램으로 만들어졌지만, 엑셀과 유사하고, 사용이 간편하여 비전문가도 단기간에 사용법을 습득할 수 있다는 장점이 있다. 특히 사용자의 니즈에 맞춰 사용자가 속한 기관에 따라 교육기관용, 의학연구기관용, 공공 기관용, 병원용, 그리고 일반기관용 등으로 분류된 프로그램을 제공한다.

## 3) SAS

SAS는 Statistical Analysis System의 약자로 1966년에 노스캐롤라이나 주립대학에서 고안해낸 프로그램이다. 현재는 SAS라는 회사가 설립되어 패키지를 판매중이고, 현재 SAS 9.4 버전까지 출시되었다. SAS는 상당히 고가인 제품으로 라이선스 없이는 사용이 불가능하고, 일정 기간이 지난 후에는 라이선스 갱신이 필요하다. 그러나 고가로 판매 및 서비스되는 프로그램인 만큼 현재 공인되어 있는 거의 모든 통계 분석을 포괄하여 수행할 수 있고 매우 정밀한 결과를 제공한다는 장점이 있다. 뿐만 아니라 보고서 작성과 그래프도 가능하여 통계를 전문적으로 사용하는 전문가의 경우 다른 통계 프로그램보다 SAS 사용을 선호한다.

SAS의 사용은 크게 두 가지 단계를 거쳐 이루어진다. 데이터 입력 및 편집을 위한 데이터 스텝(DATA STEP)과 본격적인 데이터 분석이 이루어지는 프록 스텝(PROC STEP)이다. DATA STEP에서는 데이터의 입력, 데이터의 오류 판단 및 수정, 데이터의 샘플링 및 병합 등이 가능하다. PROC STEP에서는 DATA STEP에서 가져온 데이터를 출력, 정렬, 요약할 수 있고, 더 나아가 여러 분석 기법을 이용해 통계 분석을 수행할 수 있다.

## 4) R

R은 데이터 분석을 위한 통계 분석 기법과 알고리즘, 시각화 기능을 지원하는 오픈 소프트웨어 환경으로서 R project 홈페이지에서 무료로 제공하기 때문에 누구든지 사용할 수가 있다. 또한 R은 오픈 소스(open source) 프로그램 형태로 제공되기 때문에 사용자들이 직접 분석 패키지를 만들어 업로드 할 수 있고, 반대로 타인이 업로드한 분석 프로그램들을 다운

받아서 사용할 수도 있다. R은 무료이지만 강력한 분석 기능 및 뛰어난 확장성을 가지고 있다.

R은 최신 통계 분석 및 마이닝 기능을 가진 패키지 및 샘플이 지속적으로 업데이트 되고 있고 전 세계적인 커뮤니티 생태계를 형성하고 있어 다른 통계 분석 툴에 비해 최신 통계 기법이 적용되는 속도가 빠르다. 또한 통계 패키지 중 유일하게 저널이 발행되어 그 사용법을 익히기도 편리하다. R을 가지고 통계 및 데이터 분석을 하는 가장 일반적인 도구가 알 스튜디오(RStudio)이다. RStudio는 운영 체제 환경(Linux, Mac, Windows 등)에 상관없이 R 기반으로 데이터 분석을 할 때 가장 보편적으로 사용되고 있는 툴(tool)이다.

## 5) 기타 최신의 분석 도구들

이 외에도 최근 다양한 빅데이터 분석 도구들이 출현되면서 관련 시장을 달구고 있다.

첫째, 웨카(Weka)는 자바(JAVA)로 개발된 오픈소스 데이터마이닝 프로그램으로 머신러닝을 수행할 수 있는 프로그램이다. Weka를 설치하기 위해서는 먼저 자바가 설치가 되어 있어야 한다. Weka는 다른 분석 프로그램과는 다르게, 클릭 몇 번으로 다양한 분석을 수행하게 해 준다는 장점을 가지고 있다. 일반적으로 R과 파이썬(Python)은 최소한의 코딩(프로그래밍)을 할 줄 알아야 하지만, Weka는 (코딩을 알면 좋지만) 코딩에 대한 사전 지식이 부족해도 간단한 UI로 분석 수행을 할 수 있도록 지원한다.

둘째, 래피드마이너(RapidMiner)는 데이터 과학 분야에서 사용하는 컴퓨터 소프트웨어 프로그램으로 2006년에 개발되었다. 데이터 전처리, 머신러닝, 딥러닝, 텍스트마이닝, 예측 분석 등에서 쉽게 활용할 수 있다. 래피드마이너는 예측 분석을 위한 모델 개발부터 관리까지 완전 GUI 기반으로 작업을 할 수 있는 통합 플랫폼으로 다양한 스크립트 언어를 지원하고 있으며, 분석을 위한 모든 과정을 단일 환경(Single Suite)에서 관리하기 때문에 업무의 생산성 및 효율성을 증가시킬 수 있다.

셋째, 최근 들어 큰 주목을 받고 있는 파이썬(Python)은 귀도 반로섬(Guido van Rossum)이 개발한 언어로 초보자부터 전문가까지 다양한 사용자층을 보유하고 있다. 동적 타이핑(dynamic typing) 범용 프로그래밍 언어로 펄 및 루비(Ruby)와 자주 비교된다. 파이썬은 다양한 플랫폼에서 쓸 수 있고, 라이브러리(모듈)가 풍부하여 국내·외를 막론하고 대학, 교육기관, 연구기관, 산업계 등에서 점차 이용이 증가하고 있는 추세이다. 또한, 파이썬은 순수한 프로그램 언어 기능 외에도 다른 언어로 쓰인 모듈들을 연결하는 풀 언어(glue language)로 자주 이용이 된다.

파이썬은 많은 상용의 응용 프로그램에서 스크립트 언어로 채용되어 활용되고 있다. 최근에 데이터 분석에서는 아나콘다(Anaconda)를 많이 사용한다. 아나콘다를 설치하면, 파이썬, 주피터 노트북 등을 활용하여 데이터 분석을 보다 쉽게 할 수 있도록 해준다.