실무에 적용 가능한 Big Data 분석 개론

빅데이터 분석 방법론

한국기술교육대학교
온라인평생교육원

▣ 빅데이터 분석 방법론의 개요

1. 빅데이터 분석

빅데이터 분석 목적과 역할

빅데이터 분석

대량의 데이터로부터 숨겨진 패턴과 알려지지 않은 정보를 찾아내기 위한 과정

빅데이터 분석 목적

데이터 사이언티스트들에 의해 분석된 정보를 토대로 각 분야의 의사결정을 수행

- 의사결정 시 최선의 대안을 선택할 수 있도록 근거를 제시하는 중요한 역할을 함
- 불확실성이 높고 의사결정이 초래하는 파급효과가 큰 의사결정일수록 실제 데이터 분석을 바탕으로 의사결정을 해야 함
- 많은 기업에서 빅데이터를 활용하여 주요 의사결정을 내리고 있음
- 효과적인 빅데이터 분석을 위해서 일반적으로 빅데이터 분석 플랫폼을 구축함

빅데이터 분석 목표와 기술

더 짧은 시간 안에 보다 많은 정보를 빅데이터로부터 추출하는 것

빅데이터 분석

- 데이터 마이닝
  - ➢ 대용량의 데이터베이스에 저장된 데이터에 숨겨진 중요한 정보와 지식을 추출하는 기술
- 예측 분석
  - ➢ 현황 정보 대신 예측 정보를 제공할 수 있는 분석

빅데이터 분석 관련 기술

- NoSQL
- 데이터베이스
- 하둡과 맵리듀스 등

▣ 빅데이터 분석 방법론의 개요

1. 빅데이터 분석

빅데이터 분석 진행

데이터를 보다 효율적으로 정확하게 분석하고 비즈니스 등의 영역에 적용하려는 노력이 꾸준히 진행되고 있음

분석	새로운 개념이 아니며 이미 오래 전부터 여러 영역에서 효과적으로 활용해온 기술임

분석 단계(마케팅 조사)

연구 목적	연구 설계	표본 설계
시장 조사인지 고객의 요구사항 파악인지를 고려함	목적에 맞게 어떻게 조사를 하고 어떤 데이터를 확보하고 어떻게 분석할지를 고려함	조사 데이터 수집 방법과 관련하여 데이터 샘플을 어떻게 취할 것인지를 고려함

자료 수집 → 자료 분석 → 결과 제시

▣ 빅데이터 분석 방법론의 개요

2. 비즈니스에서의 분석 수행 과정

분석 수행 단계(Forrester)

1 문제인식

- 문제가 무엇인지, 왜 이 문제를 해결해야 하는지, 문제 해결을 통해 무엇을 달성할 것인지를 명확히 하는 단계

2 관련 연구 조사 단계

- 문제와 직간접적으로 관련된 지식을 각종 문헌(예 잡지, 책, 보고서, 논문 등)을 조사하면 문제를 더욱 명확히 할 수 있을 뿐 만 아니라 문제와 관련된 주요 요소(변수)들을 파악할 수 있는 단계

3 모형화(변수 선정) 단계

- 모형은 문제(연구 대상)를 의도적으로 단순화한 것을 말하며, 모형화는 문제와 본질적으로 관련된 변수만을 추려서 재구성하는 단계

4 자료 수집(변수 측정) 단계

- 인식된 문제는 모형화를 통하여 주요 변수로 재구성되고 측정이라는 과정을 거치면서 자료가 되는 단계

1차 자료	조사자가 관찰, 설문조사, 실험을 통하여 직접 자료를 수집하는 것
2차 자료	다른 사람에 의해 이미 수집, 정리되어 있는 자료

5 자료 분석 단계

- 나열된 숫자에서 변수 간의 규칙적인 패턴, 즉 변수간의 관련성을 파악

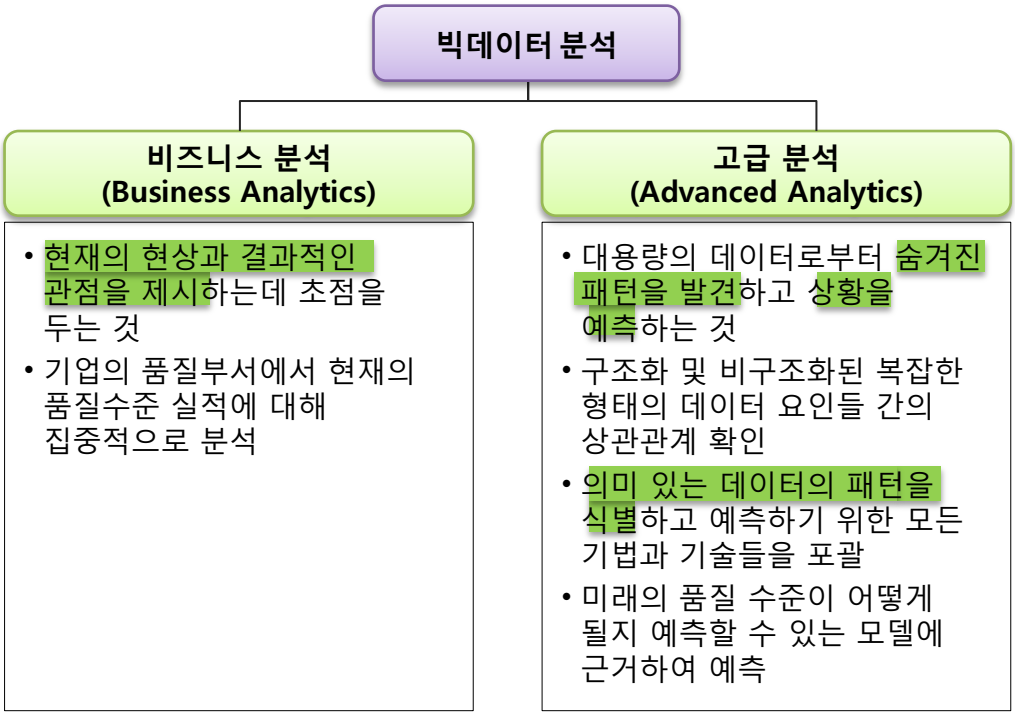
6 결과 제시 단계

- 자료 분석 결과가 의미하는 바를 해석하여 의사결정자에게 구체적인 조언을 하는 단계

▣ 빅데이터 분석 방법론의 개요

2. 비즈니스에서의 분석 수행 과정

빅데이터 분석 분류



3. 빅데이터 분석 도구

R 프로그래밍 언어

R 프로그래밍 언어 개요

오픈소스 프로젝트로 통계 계산 및 시각화를 위한 언어 및 개발 환경을 제공함

- 기본적인 통계 기법부터 모델링, 최신 데이터 마이닝 기법까지 구현이 가능함
- 통계적 컴퓨팅 언어로 다양한 통계 분석에 용이함

현재 R 프로그래밍 언어를 이용하여 다양한 빅데이터 분석 및 예측 분석 등을 포함한 고급 분석 기술들이 연구 및 개발되고 있음

R 프로그래밍 언어 장점

- 사용자가 제작한 패키지를 추가하여 기능을 확장 가능
  - ➢ 핵심적인 패키지는 R 프로그래밍 언어와 함께 설치됨
  - ➢ CRAN(the Comprehensive R Archive Network)을 통해 700개 이상의 다양한 기능을 가지는 패키지를 내려 받을 수 있음
- 그래픽 기능
  - ➢ 수학 기호를 포함할 수 있는 출판물 수준의 그래프를 제공함

▣ 빅데이터 분석 방법론의 개요

3. 빅데이터 분석 도구

빅쿼리(BigQuery)

빅쿼리(BigQuery) 개요

- 구글의 대용량 데이터를 처리할 수 있도록 개발된 쌍방향 서비스
- 사용자 혹은 개발자 등은 SQL과 같은 익숙한 쿼리문 등을 이용해 인사이트를 전달할 수 있음
  - ➢ 일반적으로 SQL문이라고도 불리는 쿼리문이 작성됨

쿼리문	SQL
• 데이터베이스에 저장된 값을 불러내기 위함 • 절차적 언어로 작성된 프로그램 문장	• Structured Query Language 의 약자 • 구조화된 절차적인 데이터베이스 언어

빅쿼리(BigQuery)를 이용하는 방법

먼저 이용자가 데이터 세트를 구글 시스템에 업로드 함 → **빅쿼리 API를 이용하여 이에 대한 쿼리를 던지는 방식으로 이용함**

빅쿼리(BigQuery)를 출시한 목적

구글이 자체 데이터센터가 없는 기업도
쉽게 데이터를 분석할 수 있는 환경을 만들어주기 위해 출시함

- 웹 광고나 실시간 관리 시스템, 온라인 게임의 데이터 현황을 쉽게 관리할 수 있음
- 예 제약회사
  - ➢ 전세계 판매량과 광고 데이터를 바탕으로 일일 마케팅 최적화 전략을 세울 수 있게 됨
  - ➢ 사용자 클릭을 바탕으로 제품 권고 사항을 만드는 일도 쉬워짐

▣ 빅데이터 분석 방법론의 개요

3. 빅데이터 분석 도구

프레스토(Presto)

페이스북에서 개발한 빅데이터 분석 도구	하둡을 위한 SQL 처리 엔진
• 페이스북이 300페타바이트에 달하는 엄청난 내부 데이터를 분석하기 위해 만듦	• 데이터 분석가가 기존의 SQL 언어로 대용량의 데이터를 대화형 분석을 수행할 수 있도록 해줌

기존에 많이 쓰는 하이브/맵리듀스 보다 CPU 효율성과 대기 시간이 10배 빠름

실무에 적용 가능한 Big Data 분석 개론

빅데이터 분석 방법론

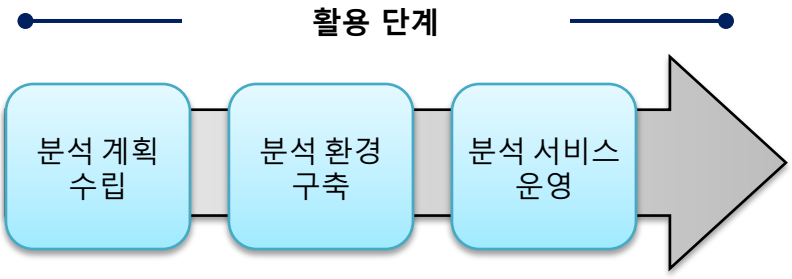
한국기술교육대학교
온라인평생교육원

▣ 플랫폼을 활용한 빅데이터 분석방법론

1. 빅데이터 분석 플랫폼 활용

수집 및 저장 된 데이터를 분석하여
서비스를 개발하고 통찰력(Insight)을 끌어내는 작업을 수행함

통찰력	주어진 데이터 분석을 통해 전체적인 상황을 한번에 파악할 수 있는 능력



▣ 플랫폼을 활용한 빅데이터 분석방법론

1. 빅데이터 분석 플랫폼 활용

분석 계획 수립

1. 빅데이터 분석 전 분석결과를 통해 해결하고자 하는 문제를 명확히 정의함

2. 분석절차, 기법 등을 포함한 세부 시나리오를 마련해야 함

3. 분석에 필요한 인프라 구축 조건 등 분석환경을 조사하여 자체 구축 및 외부 인프라 활용여부를 결정해야 함

자체 구축	• 빅데이터 분석과 활용을 위해 분석 시스템과 운영환경을 기관 내에 구축하는 방식 • 내부 데이터의 관리 정책과 보안문제로 외부 서비스를 활용하기 어려운 경우나 분석 요구사항을 외부 서비스 기관에서 지원하지 못하는 경우에 대한 적절한 대응책이 필요함
외부 활용	• 외부 분석업체의 분석 서비스를 활용하는 방식 • 외부 분석 시스템의 기능과 분석 품질이 활용 목표 수준에 부합할 경우에 대한 대응책이 필요함

4. 세부 추진 계획을 수립해야 함

- 분석 목적
- 분석 방법론
- 분석 시나리오 작성
- 분석 인프라 구축 방식 및 운영예산

▣ 플랫폼을 활용한 빅데이터 분석방법론

1. 빅데이터 분석 플랫폼 활용

분석 계획 수립

하드웨어와 소프트웨어 구축

하드웨어

빅데이터 수용 용량 및 분석작업에 대한 부하 등을 감안하여 하드웨어 인프라 구축

분석에 필요한 수집, 관리, 분석, 이용자 환경 등 관련한 소프트웨어 구축

소프트웨어

분석 용도에 따른 데이터 분석 기법

1 통계적 분석

➢ 전통적인 분석 방법으로 주로 수치형 데이터에 대하여 확률을 기반으로 어떤 현상의 추정, 예측을 검정하는 기법

기술통계량 | 상관분석 | 회귀분석 | 분산분석 | 주성분분석

- 대표적으로 평균(산술평균, 중앙값, 최빈값), 분산, 표준편차 등을 구하는 것
- 전체 데이터 그룹이 주로 어디에 위치하고 있으며 이를 중심으로 얼마나 산포를 가지는지를 확인 가능함

▣ 플랫폼을 활용한 빅데이터 분석방법론

1. 빅데이터 분석 플랫폼 활용

분석 용도에 따른 데이터 분석 기법

평균

- 데이터 집합의 중심적인 경향을 표현하는 값
- 전체 데이터의 합을 전체 데이터 개수로 나누어 산출

분산

- 평균을 중심으로 각각의 데이터 들의 편차를 구하고 편차의 제곱을 모두 더한 후 전체 데이터 개수에서 하나를 뺀 값으로 나눈 값

표준편차

- 분산 값의 제곱근으로 산포를 의미

1 통계적 분석

➢ 전통적인 분석 방법으로 주로 수치형 데이터에 대하여 확률을 기반으로 어떤 현상의 추정, 예측을 검정하는 기법

기술통계량 | **상관분석** | 회귀분석 | 분산분석 | 주성분분석

- 두 변수간에 어떤 선형적 관계를 갖고 있는지를 분석하는 방법
  - ✓ 하나의 변수가 증가할 때 비례 또는 반비례적으로 다른 한 변수가 증가 또는 감소하는 정도를 규명
- 분석 시 서로 관계를 가지는 변수들을 찾아 낼 수 있음

▣ 플랫폼을 활용한 빅데이터 분석방법론

1. 빅데이터 분석 플랫폼 활용

분석 용도에 따른 데이터 분석 기법

1 통계적 분석

➢ 전통적인 분석 방법으로 주로 수치형 데이터에 대하여 확률을 기반으로 어떤 현상의 추정, 예측을 검정하는 기법

기술통계량 | 상관분석 | **회귀분석** | 분산분석 | 주성분분석

- 연속형 변수들에 대해 독립변수와 종속변수 사이의 상관관계에 따른 수학적 관계식을 구하여 어떤 독립변수가 주어졌을 때 이에 따른 종속변수를 예측하는 방법
  - ✓ 종속변수 값을 예측할 수 있는 수학적 모델식을 구성
  - ✓ 특정한 독립변수의 값을 가지는 경우
  - ✓ 종속변수의 값을 예측 가능함

독립변수

- **종속변수에 영향을 주는 요인을 가지는 변수**

종속변수

- **독립변수의 값에 의해 종속적으로 영향을 받는 변수**

연속형 변수

- **독립변수와 종속변수가 일반적으로 연속형의 값을 가지는 경우**

▣ 플랫폼을 활용한 빅데이터 분석방법론

1. 빅데이터 분석 플랫폼 활용

분석 용도에 따른 데이터 분석 기법

1 통계적 분석

➢ 전통적인 분석 방법으로 주로 수치형 데이터에 대하여 확률을 기반으로 어떤 현상의 추정, 예측을 검정하는 기법

기술통계량 | 상관분석 | 회귀분석 | **분산분석** | 주성분분석

- 3개 이상의 집단에 있어서 평균치 차이가 존재하는지를 검증함
- F분포를 이용하여 가설검정을 하는 방법
- F분포
  - ✓ 두 개 이상 다수의 집단을 비교하고자 할 때 집단 내의 분산, 총평균과 각 집단의 평균의 차이에 의해 생긴 집단 간 분산의 비교를 통해 만들어짐
- 다수의 집단에 있어서 평균치가 차이가 있는지 유의성을 판정할 수 있음

기술통계량 | 상관분석 | 회귀분석 | 분산분석 | **주성분분석**

- 다양한 변수들에 대해 분석하는 다변량 분석으로 많은 변수들로부터 몇 개의 주성분들을 추출하는 방법
- 많은 변수들을 관리할 수 있는 관리의 로드가 줄어들 수 있음

▣ 플랫폼을 활용한 빅데이터 분석방법론

1. 빅데이터 분석 플랫폼 활용

분석 용도에 따른 데이터 분석 기법

2 데이터 마이닝

- 대용량 데이터로부터 패턴인식, 인공지능 기법 등을 이용하여 숨겨져 있는 데이터간의 상호 관련성 및 유용한 정보를 추출하는 기술
- 기존 데이터베이스에 마이닝 기술을 적용하여 이들 데이터 간에 숨은 의미 있는 관계성을 다양한 방법으로 발견한 후 이를 현실에 효과적으로 적용하는 방법론으로 사용됨

3 텍스트 마이닝

- 텍스트 기반의 데이터로부터 새로운 정보를 발견할 수 있도록 정보 검색, 추출, 체계화, 분석을 모두 포함하는 Text-processing 기술 및 처리 과정
- 텍스트 내에 존재하는 단어의 등장횟수 등을 평가하여 문서간의 유사성을 수치화 하는 텍스트 데이터를 분석하는 방법
- 유사 문서 분류 및 문서 내 정보 추출과 같은 결과를 산출이 가능함

4 소셜 네트워크 분석

- 대용량 소셜 미디어를 언어분석 기반 정보 추출로 탐지함
- 시간의 경과에 따라 유통되는 이슈의 전체 과정을 모니터링하고 향후 추이를 분석함
- 소셜 네트워크 연결 구조 및 연결 강도 등을 바탕으로 사용자의 명성 및 영향력을 분석함
- 활용
  - ✓ 주로 마케팅을 위하여 소셜 네트워크 상에서 입소문의 중심이나 허브 역할을 하는 사용자를 찾음
  - ✓ 수학의 그래프 이론을 이용하여 소셜 네트워크의 연결 구조와 연결 강도 등을 바탕으로 사용자의 영향력을 측정함
- 텍스트 마이닝 기법에 의해 주로 이루어짐
- 확산된 내용과 함께 연결의 맥락을 파악하여 분석하는 기법

▣ 플랫폼을 활용한 빅데이터 분석방법론

1. 빅데이터 분석 플랫폼 활용

분석 용도에 따른 데이터 분석 기법

5 평판 분석(Sentiment Analysis)

- 오피니언 마이닝이라고도 불림
- 소셜미디어 등의 정형/비정형 텍스트의 긍정(Positive), 부정(Negative), 중립(Neutral) 의 선호도를 판별하는 기술
- 활용
  - ✓ 특정 서비스 및 상품에 대한 시장규모 예측
  - ✓ 소비자의 반응
  - ✓ 입소문 분석(Viral Analysis) 등
- 정확한 오피니언 마이닝을 위해서는 전문가에 의한 선호도를 나타내는 표현/단어 자원의 축적이 필요함

6 군집 분석(Cluster Analysis)

- 비슷한 특성을 가진 개체를 합쳐가면서 최종적으로 유사 특성의 군(Group)을 발굴하는데 사용함

  예
  - ✓ 트위터 상에서 주로 사진/카메라에 대해 이야기하는 사용자 군
  - ✓ 자동차에 대해 관심 있는 사용자 군

- 관심사나 취미에 따른 사용자 군을 군집 분석을 통해 분류 가능