〔과제1. 데이터 분석과 빅데이터 분석의 차이점 4가지 〕

▷ 데이터 분석과 빅데이터 분석의 차이점 4가지

1인 PC세대에는 데이터 분석으로 처리할 수 있는 일들이 모바일장치의 확산, 소셜미디어의 일상화, 스마트폰 시대가 도래하면서 빅데이터 분석이 필요하게 되었다.

1. 첫번째 차이점 : 사전처리(Pre-Processing)와 사후처리(Post-Processing)

정보관리 시스템이 충분히 데이터를 처리할 수 있도록 필요한 정보를 추려내고 필요 없는 정보를 제거하는 사전처리 보다는 데이터에서 의미를 찾아내기 위해 데이터 유형에 따라 적합한 저장관리, 품질관리, 보안관리를 수행하는 사후처리 역할이 더욱 중요해졌다.

2. 두번째 차이점 : 표본조사와 전수조사

유의미한 시간에 대규모 전수 데이터를 정제 및 분석함으로써 패턴과 같은 정보 제공이 필요해 졌으며, 이로 인해 조사의 대상이 되는 자료 일부만을 택해 조사하여 전체를 추측하는 표본조사보다는 조사의 대상이 되는 자료 전체를 빠짐없이 조사하는 전수조사를 하여 표본조사가 주지 못하는 패턴과 같은 정보를 제공한다는 점이다.

3. 세번째 차이점 : 질과 양

분석할 수 있는 데이터가 많을수록 결과의 정확성은 증가하게 되며, 데이터의 양이 증가함에 따라 사소한 몇몇 오류는 분석 결과에 큰영향을 미치지 않게 됨으로써, 데이터의 질보다 양이 더 강조되고 있다.

4. 네번째 차이점 : 인과관계와 상관관계

비즈니스의 상황에서 중요한 것은 인과관계보다도 상관관계 파악이다. 빠르고, 저렴한 비용으로 상관관계의 파악이 가능해짐에 따라, 인과관계를 증명한 후 행동하기에는 비용대비 효과가 나오지 않을 때가 훨씬 더 많기 때문에, 빅데이터 분석을 통한 다양한 상관관계를 도출할 수 있게 되었다.

〔과제2. 데이터 전처리 기술 3가지가 무엇인지 명칭을 작성하고, 그 개념에 대해 설명 〕

▷ 데이터 전처리 기술 3가지 명칭

데이터 필터링, 데이터 변환, 데이터 정제

▷ 데이터 전처리 기술 3가지 개념 설명

데이터 전처리를 정의하자면 수집 데이터를 저장소에 적재하기 위해 처리하는 작업이며, 데이터 유형과 분석목적을 고려하여 적절한 데이터 처리기법을 선정하는 것을 데이터 전처리라 하며, 데이터 전처리를 하는 이유로는 일반적으로 일관성있는 데이터 형태로 통합하여 분석에 소요되는 시간과 노력을 절약하고, 의미파악이 어려운 비정형 데이터는 분석이 가능한 형태로 변환하기 위해서이다. 데이터 전처리 기술 3가지에 대해서 설명하겠다.

1. 첫번째 : 데이터 필터링

데이터 필터링이란 오류를 발견하고 보정 및 삭제하며 중복성 검사 등을 수행하는 것을 의미한다. 필터링 처리시 기술고려사항으로는 생성된 파일의 중복성을 확인할 수 있도록 파일명, 확장자등 필터링 기능을 제공해야 하며, 데이터 필터링시 사전 정의된 기준에 의거하여야 하며, 오류에 대한 이력을 저장해야 한다. 실제 사전 테스트에서 필터링 과정을 수행하게 되고, 필터링 기준을 최적화하여 활용한다. 데이터 종류가 비정형 데이터인 경우에는 데이터 마이닝을 사용.

2. 두번째 : 데이터 변환

다양한 형식으로 수집된 데이터를 분석에 용이하도록 일관성 있는 형식으로 변환하는 것을 데이터 변환이라고 한다. 데이터 변환 기술 5가지로는 평활화(Smoothing), 집계(Aggregation), 정규화(Normalization), 일반화(Generalization), 속성생성(Attribute/Feature Construction)이 있다. 첫번째, 평활화란 데이터로부터 잡음제거를 위해 데이터 추세에 벗어나는 값들을 변환하는 기법이며 데이터 집합에 존재하는 잡음으로 인해 거칠게 분포된 데이터를 매끄럽게 만들기 위해 구간화, 군집화 등의 기법을 사용한다. 두번째, 집계란 복수의 속성을 하나로 줄이거나 유사한 데이터 객체를 줄이고 스케일을 변경하여 데이터를 요약하는 기법이다. 세번째, 정규화란 데이터를 정해진 구간 내에 들도록하는 기법이며 데이터에 대한 통계적 기법으로는 최소-최대 정규화, Z-스코어 정규화, 소수 스케일링이 있다. 네번째, 일반화란 특정 구간에 분포하는 값으로 스케일

을 변화시키며, 범용적인 데이터에 적합한 모델을 만드는 기법이며, 일반화가 잘 되어 있다면, 이상값이나 노이즈가 들어오더라도 크게 흔들리지 않는다. 다섯번째, 속성 생성이란 데이터 통합을 위해 새로운 속성이나 특징을 만드는 기법이며, 주어진 여러 데이터 분포를 대표할 수 있는 새로운 속성이나 특징으로 대체하여 데이터를 변경 처리한다. 데이터 변환시 기술 고려 사항으로는 알고리즘 함수 또는 변환구조를 정의하는 기능이 제공되어야 하며, 데이터 변환시 사용자가 지정한 변환 형식에 준하여 변환이 이루어졌는지 확인할 수 있는 기능이 제공되어야 하고, 데이터 변환 실패 부분에 대해 재시도할 수 있는 기능을 제공하거나 신규 변환 데이터가 생성을 취소할 수 있는 기능을 제공해야 된다.

3. 세번째 : 데이터 정제

수집된 데이터의 불일치성을 교정하기 위한 것이 데이터 정제이다. 데이터 정제 기술 2가지로는 결측치(Missing Value) 처리방법, 잡음(Noise) 처리방법이 있다. 첫번째, 결측치 처리방법은 해당 레코드 무시하기, 자동으로 채우기, 담당자 수작업으로 입력하기가 있다. 두번째, 잡음 처리방법은 구간화, 회기값 적용, 군집화가 있으며, 잡음이란 랜덤 에러나 측정된 변수의 변형된 값을 의미하며, 잡음 발생원인으로는 센서의 작동 실패, 데이터 인트리 문제, 데이터 전송문제, 기술적인 한계, 데이터 속성 값의 부정확성등이 있다.