1.3 데이터 수집 및 저장 계획

개념 1

● 데이터 수집 방법

데이터 유형에 따른 데이터 수집 방법은 다음과 같다.

1. 정형 데이터

데이터 종류는 DBMS, 스프레드시트 이고 수집 방법은 ETL, FTP, Open API 이다.

2. 반정형 데이터

데이터 종류는 HTML, XML, JSON, 웹 문서, 웹 로그, 센서 데이터 이고 수집 방법은 웹크롤링, RSS, Open API, FTP 이다.

3. 비정형 데이터

소셜 데이터, 문서, 이미지, 오디오, 비디오, loT 이고 수집 방법은 웹 크롤링, RSS, Open API, Streaming, FTP 이다.

개념 2

● 정형 데이터와 비정형 데이터

비정형 데이터

정해진 구조가 없는 데이터를 말한다. 예를 들어, 동영상, 소셜 네트워크 댓글, 위치 데이터 등을 예시로 들 수 있다. 비정형 데이터는 크기고 크고 복잡하다.

정형 데이터

고정된 구조로 정해진 필드에 저장된 데이터를 의미한다. 엑셀, csv, RDBMS 형태가 대표적이고, 데이터로서 활용성이 가장 높다.

반정형 데이터

데이터와 메타데이터, 스키마 등을 포함하는 데이터를 의미한다. XML, HTML, JSOn등이 대표적이다.

개념 3

● 빅데이터 분석 절차

문제인식 ->연구조사 ->모형화 ->자료수집 ->분서결과 공유

개념 4

● 수집데이터

수집 데이터는 내부/외부 데이터로 나뉘며 이에 대한 정의는 다음과 같다.

1. 내부 데이터

내부 조직간 협의를 통한 데이터 수집, 수집이 용이한 정형 데이터

2. 외부 데이터

조직 외부에 데이터가 위치 , 수집이 어려운 비정형 데이터

개념 5

● 척도

척도란 수집된 데이터가 다른 데이터와 구분하기 위한 특성을 의미한다.

데이터는 측정 방법에 따라, 질적 자료와 양적 자료로 구분되고, 질적 자료는 명목척도와 순위척도 그리고 양적 자료는 구간 척도와 비율 척도로 구분된다.

개념 6

● 질적 자료에 대한 척도

명목 척도 : 순위가 없이 특정 범주에 존재하는 척도를 의미한다.

예시) 성별, 혈액형, 거주지역

순위 척도 : 순위가 있는 척도를 의미한다.

예시) 학년, 석차, 소득 수준

개념 7

● 양적 자료에 대한 척도

구간 척도 : 절대적인 원점이 존재하지 않는다. 즉, 0이라고 해서 값이 없다고 할 수 없는 값이다.

예시) 온도, 지수, 점수

비율 척도 : 절대적인 원점이 존재한다.

예시) 무게, 거리, 키, 나이

개념 8

● 데이터 변환(데이터 전처리 작업)

데이터 분석을 좀 더 효율적으로 처리하기 위해 데이터 변환을 한다.

개념 9

● 데이터 변환 방법

1. 평활화

이상치를 제거하는 방법을 의미한다.

2. 집계

그룹화 연산을 이용하여 데이터를 요약하는 방법이다. 예를 들어, 매일 발생하는 데이터를 월별 또는 연도별로 요약하는 방법이다.

3. 일반화

특정 구간에 속하는 값으로 스케일을 변화시키는 방법을 의미한다.

4. 정규화

데이터를 특정 구간 안에 들어가게 이상값을 변환하는 방법을 말한다. 예시로는 최소-최대 정규화, z-score 정규화, 소수 스케일링 정규화가 있다.

5. 범주화

데이터 통합을 위해 사위 레벨 개념의 속성이나 특성을 이용해 일반화하는 방법이다.

6. 데이터 축소, 차원 축소

데이터 축소 : 같은 정보량을 가지면서 데이터의 크기를 줄이는 방법

차원 축소 : 데이터 차원의 크기를 축소하는 방법

개념 10

● 데이터 비식별화

데이터에 개인을 식별할 수 있는 정보가 있는 경우 일부 또는 전체를 삭제하거나 일부를 대체 처리함으로써 특정 개인을 식별할 수 없게 하는 것을 말한다.

가명처리, 총계처리, 데이터 삭제, 범주화, 데이터 마스킹

개념 11

● 데이터 품질 검증 요소

1. 데이터 값 검증

업무 규칙 : 정형 데이터와 메타데이터를 대상으로 업무적으로 만족시킬 수 있는 운영, 정의, 제약사항 등의 기술 규칙이다. 검증 대상 데이터에 업무규칙을 적용해 준수 여부를 검증할 수 있다.

2. 데이터 프로파일링

통계 기법을 이용해서 패턴을 파악해서 데이터 품질검증을 하는 방법이다.

개념 12

● 데이터 구조 검증

데이터 모델링 관점으로 데이터 구조 검증이 이루어 진다. 데이터베이스의 구조 무결성, 데이터 구조 표준화 등등을 검증한다.

* 데이터 품질요소.

⇒ 정확성, 완전성, 적시성, 일관성

* 메타 데이터

⇒ 데이터에 관한 구조화된 데이터로서 다른 데이터를 설명해주는 데이터.

* 데이터 프로파일링

⇒ 데이터 현황 분석을 위한 자료수집을 통해 잠재적 오류 정후를 발견하는 방법.

절차 : 메타데이터 수집 및 분석

↓

대상 및 유형 설정

↓

프로 파일링 수행

↓

프로 파일링 결과 리뷰

↓

프로 파일링 결과 종합

* 품질 검증 기준

복잡성, 완전성, 유용성, 시간적 요소, 일관성, 타당성, 정확성.

↳ Data 수집과 전달 사이의 소요시간.

개념 13

● 빅데이터 품질 요소와 품질 전략

데이터 품질 요소 : 정확성, 완전성, 적시성, 일관성

개념 14

● 정형 데이터 품질 검증 기준

완전성 : 필수항목에 누락이 없어야 한다.

유일성 : 데이터는 중복되면 안된다.

유효성 : 데이터 유효범위 및 도메인을 충족해야 한다.

일관성 : 데이터의 형태가 일관되어야 한다.

정확성 : 실세계에 존재하는 객체의 표현 값이 정확하게 반영되어야 한다.