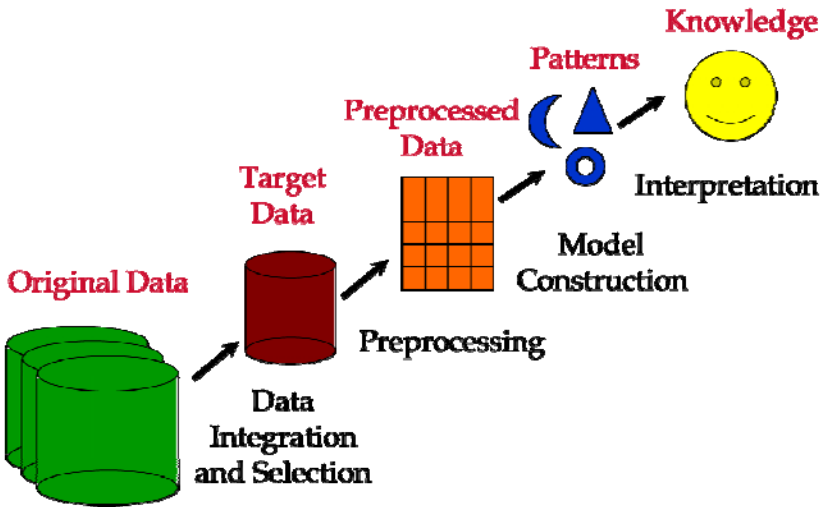
Ch.1 데이터전처리 개요

데이터전처리 개요

- 데이터 전처리 정의

  데이터 분석 작업을 하기 전에 데이터를 분석하기 좋은 형태로 만드는 과정을 총칭하는 개념



* 출처 : Ramakrishnan and Gehrke. Database Management Systems, 3rd Edition

[그림 1.1] 데이터마이닝 과정에서의 데이터 전처리

데이터전처리 개요

- 데이터 전처리가 필요한 이유
  - 실무 데이터는 분석 기법을 바로 적용하기 힘든 형태
  - 비어있음(missing value), 잡음(noise), 적합하지 않은 데이터구조
  - 낮은 품질의 데이터로는 좋은 분석결과를 얻기 힘듬
- 데이터 품질 저하의 원인
  - 불완전(incomplete)
    - 데이터가 비어 있어 있는 경우로 DB 테이블의 속성값이 NULL인 경우
  - 잡음(noisy)
    - 데이터에 오류(error)가 포함된 경우. 예) 나이 = -20
  - 모순된(inconsistent)
    - 데이터 간의 정합성(일관성)이 없는 경우. 예) 성별은 남자인데, 주민번호 뒷 7자리 중 첫 번째 자리가 2인 경우
- 고품질 데이터라 하더라도 전처리 필요
  - 실무에서 존재하는 데이터의 구조적 형태(format)가 분석목적이나 분석기법에 적합한 경우가 드물기 때문

데이터전처리 예 – 웹로그 분석

- 분석 환경 및 주제
  - 웹로그 : Apache 사의 access log ([그림 1.2])
  - 사이트 : 사진을 서비스하고 앨범을 만들어 주는 사이트로 가정
  - 분석주제 : 사진 조회 빈도 및 사진 간 연관조회 현황 분석

```
64.242.88.10 - - [07/Mar/2004:16:05:49 -0800] "GET /twiki/bin/edit/Main/Double_bounce_sender?topicparent=Main.ConfigurationVariables HTTP/1.1" 401 12846
64.242.88.10 - - [07/Mar/2004:16:06:51 -0800] "GET /twiki/bin/rdiff/TWiki/NewUserTemplate?rev1=1.3&rev2=1.2 HTTP/1.1" 200 4523
64.242.88.15 - - [07/Mar/2004:16:07:03 -0800] "GET /07T2KZone/PostPhotos/photo_read.asp?oid=3110&category=1&theme=t01&page_num=11&konum=3&kosm=m7_3 HTTP/1.1" 200 76137 Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.0.7) Gecko/20060909 Firefox/1.5.0.7
```

[그림 1.2] Apache access log의 예

데이터마이닝의 이론적 사례

다음과 같은 itemset을 갖는 트랜잭션 t_1, t_2, t_3가 있다.

t_1 = {a, b, c}, t_2 = {a, b}, t_3 = {c}

트랜잭션 t_1, t_2, t_3에 대해서, minimum support 0.6 이상인 itemset을 구하고, 이들을 대상으로 minimum confidence 0.8 이상인 연관규칙을 찾으시오.

[그림 1.3] 연관규칙에 대한 이론적 사례

- 실무에서는
  - t1, t2, t3와 같은 트랜잭션의 개념을 어떻게 잡아야 할까?
  - 사례처럼 간단명료한 항목집합(itemset)에 존재할까?

➔ 실무에서 존재하는 [그림 1.2]와 같은 데이터를 분석이 용이하도록 [그림 1.3]과 같은 데이터 형태로 바꾸는 데이터전처리 필요

웹로그 분석을 위한 데이터 전처리 단계

1) 웹로그와 사진정보 DB 간의 매개 현황 파악
   - 웹로그의 클라이언트 요청 URL에 존재하는 모듈과 패러미터 파악
     - ✓ 사진조회 모듈 : photo_read.asp
     - ✓ 패러미터 : oid, category, theme

2) 사진조회와 직접적인 관련이 없는 로그 제거

64.242.88.15 - - [07/Mar/2004:16:07:03 -0800] "GET /07T2KZone/PostPhotos/photo_read.asp?oid=3110&category=1&theme=t01&page_num=11&konum=3&kosm=m7_3 HTTP/1.1" 200 76137 Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.0.7) Gecko/20060909 Firefox/1.5.0.7

[그림 1.4] 로그 필터링을 통해 추출한 분석대상 로그

웹로그 분석을 위한 데이터 전처리 단계

3) 로그 파싱(parsing)을 통한 패러미터 값 추출

- 사진정보 조회에 필요한 패러미터 값을 얻기 위해 클라이언트 요청 URL(CLIENT_FULL_RESQ 항목)을 다시 파싱

SEQ	1814718
CLIENT_IP	84.222.150.222
SERVER_IP	-
AUTH_NM	-
DATE_TIME	2008-04-01 17:06
CS_METHOD	GET
CLIENT_FULL_RESQ	http://chinese.tour2korea.com/07T2KZone/PostPhotos/photo_read.asp?oid=3110&category=1&theme=t01&page_num=11&konum=3&kosm=m7_3
FLAG	
URI_PROTOCOL	HTTP/1.1
SERVER_STAT	200
CONTENT_LENGTH	76137
REFERER	-
USER_AGENT	Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.0.7) Gecko/20060909 Firefox/1.5.0.7
COOKIE	-

[표 1.1] 웹로그 텍스트 파싱(parsing) 결과

웹로그 분석을 위한 데이터 전처리 단계

4) 추출한 패러미터 값을 조건으로 사진정보 DB를 조회하고 조회된 사진의 키 값으로 조회사진 항목집합(itemset)을 구성한다

```
-- t_picture: 사진정보 저장 테이블, t_pic_class: 사진 분류정보 저장 테이블
SELECT a.pic_no -- t_picture 테이블의 key 속성
FROM t_picture a, t_pic_class b
WHERE a.oid = :v_oid -- :v_oid = 웹로그 패러미터 oid 값 (3110)
AND a.pic_no = b.pic_no -- 조인조건 (b.pic_no는 a.pic_no를 참조하는 외래키
AND b.category = :v_category -- :v_category = 웹로그 패러미터 category 값 (1)
AND b.theme = :v_theme -- :v_theme = 웹로그 패러미터 theme 값 (t01)
```

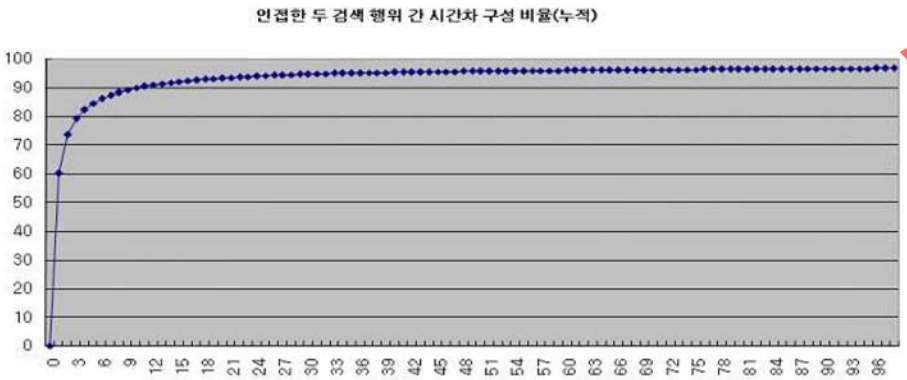
[그림 1.5] 웹로그 클라이언트 요청 URL의 패러미터 값을 통해 DB에 접근하는 사례

웹로그 분석을 위한 데이터 전처리 단계

5) 트랜잭션 설계

- 사진의 조회 연관성 분석을 위해서는 트랜잭션 설계가 필수적
  - ✓ 트랜잭션 내 조회사진 itemset에 대한 빈발항목을 구함
- 트랜잭션을 어떻게 설계할 것인가?
  - ✓ 트랜잭션 : 특정 사용자가 동시에 조회하는 사진 itemset
  - ✓ But, 웹로그는 트랜잭션의 개념이 없는 개별적인 웹서버 사용 기록
  - ➔ 로그 간 시간차 분석을 바탕으로 가상의 트랜잭션을 도출

예)

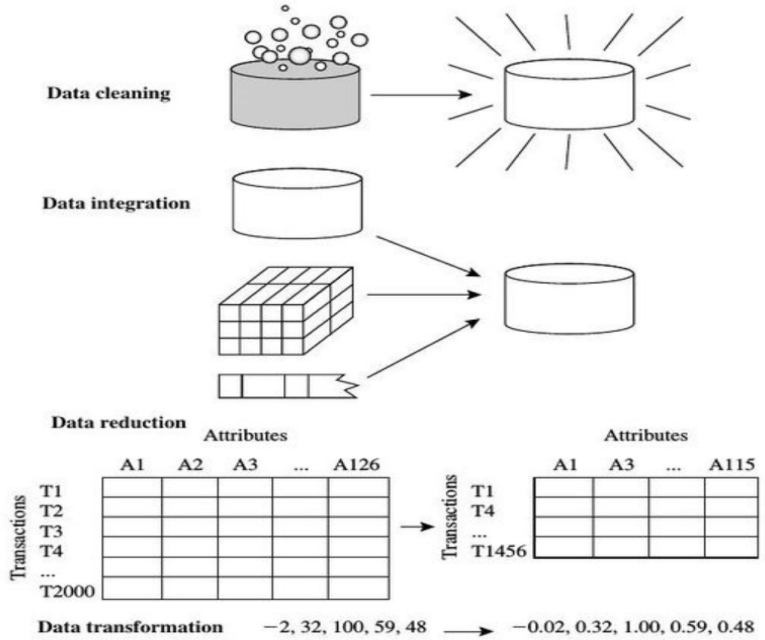


인접한 두 검색행위(로그) 시각 차가 OO분 이내이면 동일 트랜잭션으로 봄

데이터 전처리 주요 기법

- 데이터 정제(Data Cleansing)
  없는 데이터(missing values)는 채우고, 잡음(noisy data)는 제거하며, 모순된 데이터(inconsistent data)는 정합성이 맞는 데이터로 교정하는 작업
- 데이터 통합(Data Integration)
  여러 개의 데이터베이스(databases), 데이터큐브(data cubes), 또는 파일(files)을 통합하는 작업
- 데이터 축소(Data Reduction)
  샘플링(sampling) 등을 통해 데이터 볼륨(volume)을 줄이거나 분석대상 속성(차원)을 줄이는 작업
- 데이터 변환(Data Transformation)
  데이터 정규화(normalization) 또는 집단화(aggregation) 하는 작업
- 데이터 이산화(Data discretization)
  데이터 축소(data reduction)의 일종으로 연속적인 수치 데이터에 대한 구간화 작업 (예, 실제 나이를 10대, 20대, 30대 등으로 변환)

데이터 전처리 주요 기법



* 출처 : Jiawei Han and Micheline Kamber. Data Mining: Concepts & Techniques

[그림 1.6] 데이터 전처리 주요 기법의 개념 (concept)

학습의 방향

- 데이터 전처리 유형
  - 실무에 있는 데이터는 매우 다양한 형태로 존재하기 때문에 이를 특정 분석 목적에 맞게 가공하는 일은 사안마다 다름
  - 데이터 전처리를 개념적인 몇가지 분류 내로 국한시켜서는 분석 대상 데이터를 만들어내기 힘듬
  - 데이터 전처리는 개념적·이론적으로 제시되는 범위보다 훨씬 넓고 포괄적
- 학습의 방향
  - 모든 데이터 전처리 유형을 다루는 것은 애초부터 불가능 → 이론적으로도 제시되어 있으면서 실무에 자주 등장하는 전처리 유형을 중심으로 학습
  - 이론적 이해 + 실무 사례 적용을 통한 활용능력 배양
- 활용 도구(tools)
  - Python : 플랫폼 독립적이며 인터프리터식, 객체지향적, 동적 타이핑(dynamically typed) 대화형 언어로서 데이터프레임, 기계학습 등의 데이터분석을 위한 다양한 함수를 제공하여 데이터전처리에 적합한 도구
  - Oracle : Python의 SQL은 오라클의 SQL에서 제공하는 윈도우(window) 함수와 같은 데이터의 집합적 처리를 위한 강력한 기능은 제공되지 않은 부분이 있어서, 일부 사례에서는 오라클 SQL을 통해 해답을 제시