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)
```

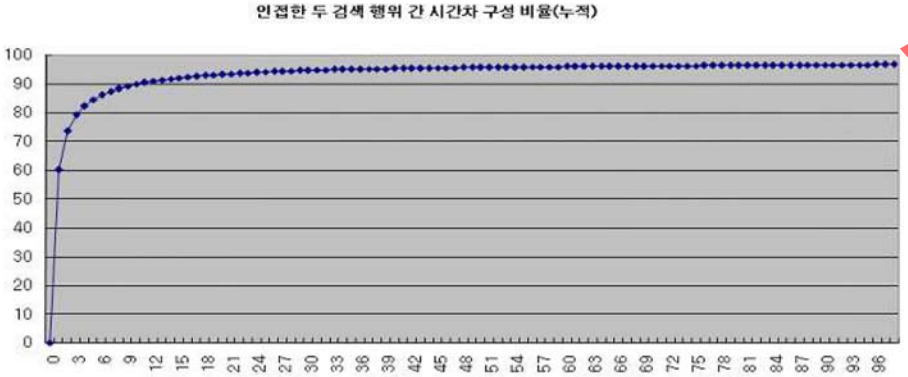
[그림 1.5] 웹로그 클라이언트 요청 URL의 패러미터 값을 통해 DB에 접근하는 사례

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