

8장. 데이터 마이닝 (Data Mining): 개념

2019년 6월 4일 (Tue)

송 병 호, 상명대학교

Prof. Byoungcho Song, Ph. D.

Sangmyung University



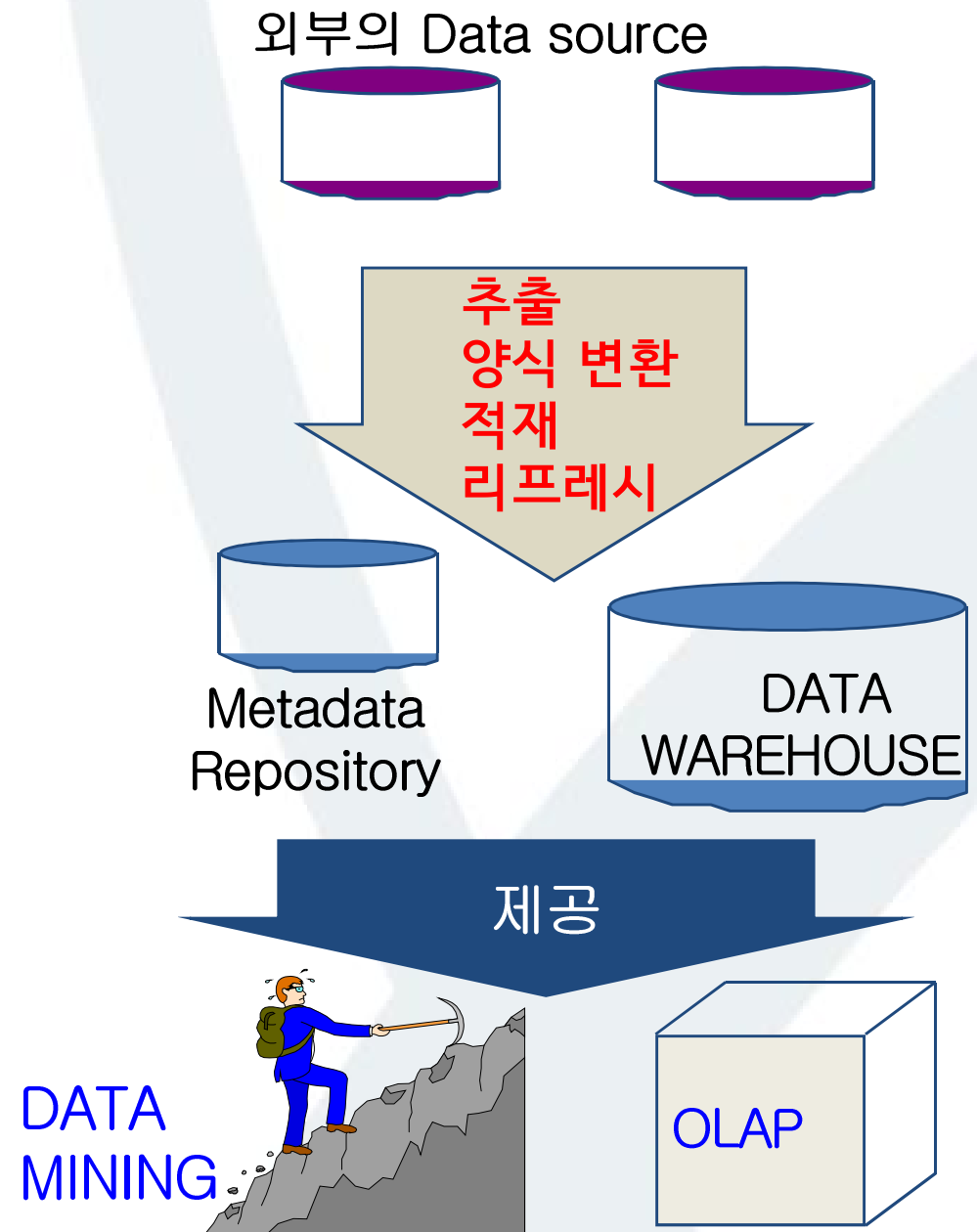
OLAP, DW, DM

OLAP

- OLTP
 - On-Line **Transaction** Processing
 - ⑮ 일상적인 업무 처리 측면
- OLAP
 - On-Line **Analytical** Processing
 - 그룹짓기 GROUPing, 집계 Aggregate Function
이 중심
 - 예: 실시간 매출 현황, 지점별 매출 비교 등
- 다차원 데이터로 진화
 - Multi-dimensional Data
 - Star Schema

Data Warehouse

- 장기간에 걸쳐 여러 Data source로부터 모은 Data와 그 요약 정보
- 보통은 단일 사이트에 구축
- Data Mart
 - 특정 분야에 초점



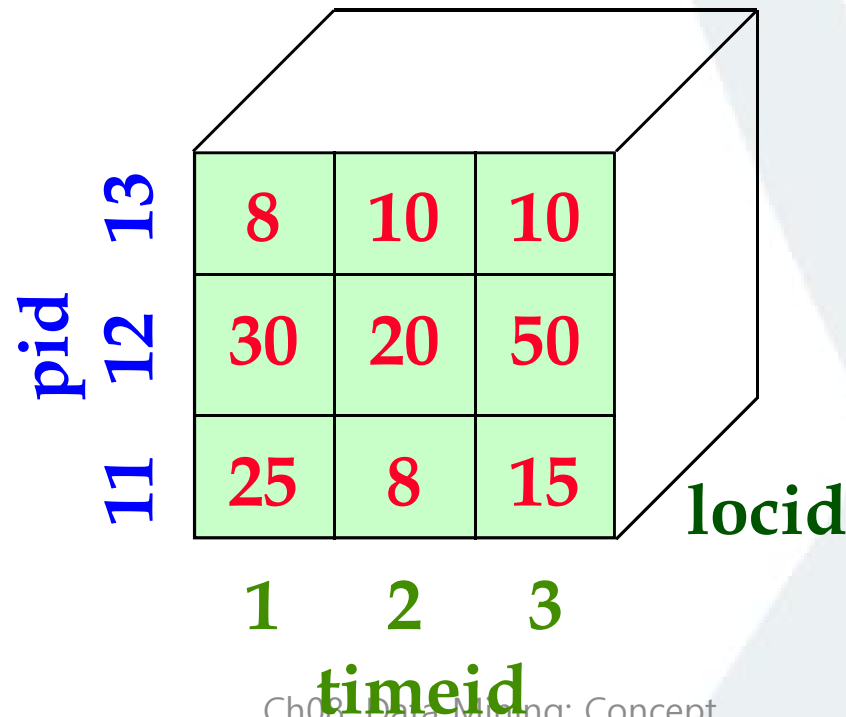
Warehousing Issues

- **원천들이 이질적임** Heterogeneous Sources:
다양한 포맷과 등록소의 데이터를 받아야.
- **의미측면의 통합** Semantic Integration:
화폐단위, 스키마 등 불일치를 해소.
- **데이터 적재** Load, **주기적 리프레시** Refresh, **오래된 데이터의 제거** Purge
- **메타데이터 관리** Metadata Management **가 필요함**:
모든 데이터 각각에 대하여 원천이 어딘지, 적재
시점 등의 정보 관리

Multidimensional Data Model

- 여러 차원 dimension으로 표현된 공간 속 각 셀_{cell}에 측정값 measure 기록.
 - 예) 측정값 **Sales**, 차원(제품, 지역, 시간)
Product (key: pid), **Location** (locid),
 and **Time** (timeid).

슬라이스
locid=1
의 예:



pid	timeid	locid	sales
11	1	1	25
11	2	1	8
11	3	1	15
12	1	1	30
12	2	1	20
12	3	1	50
13	1	1	8
13	2	1	10
13	3	1	10
11	1	2	35

MOLAP vs ROLAP

- 다차원 데이터 저장방법:
 - 물리적으로도 다차원으로 저장: **MOLAP** systems.
 - ⑮관계 DB로 저장: **ROLAP** systems.
- 중심이 되는 릴레이션: 팩트 테이블 (**fact table**).
- 각 차원축에 대해 자세한 설명: 차원 테이블 (**dimension table**).
 - E.g., **Products(pid, pname, category, price)**
 - 예: 제품(제품번호, 제품명, 카테고리, 가격)
 - 팩트 테이블은 차원 테이블보다 매우 매우 크다.

차원 계층구조 Dimension Hierarchies

- ϵ 차원축은 다시 다단계로 구성 가능:

PRODUCT

카테고리
|
제품명

TIME

연도
|
분기/계절
/
주차 월차
\
날짜

LOCATION

국가
|
지방
|
도시

OLAP Queries

- SQL과 spreadsheet를 흉내.
- 주로 사용하는 연산은 다양한 차원축 조합의 집계 aggregate .
 - 전체 매출.
 - 도시별 (지역별) 매출.
 - 매출 상위 5개 top 5 제품.
- 롤업 Roll-up: 차원축 상위 레벨로 올라가는 집계.
 - 도시별 매출을 롤업 하면, 지방별, 국가별 매출이 된다 (더 큰 단위로 집계)

OLAP Queries

- 드릴다운 Drill-down: 롤업의 반대.
 - 지방별 매출합계를 드릴다운 하면, 도시별 매출합계가 된다.
 - 지역별 매출에 대해, 다른 차원축으로 드릴다운해서 제품별로 세분할 수도 있다.

- 피벗 Pivoting: 차원축을 하나 이상 정하고 그에 따른 집계.

- 지역축 Location 과 시간축 Time으로 피벗해서 만든 교차집계표

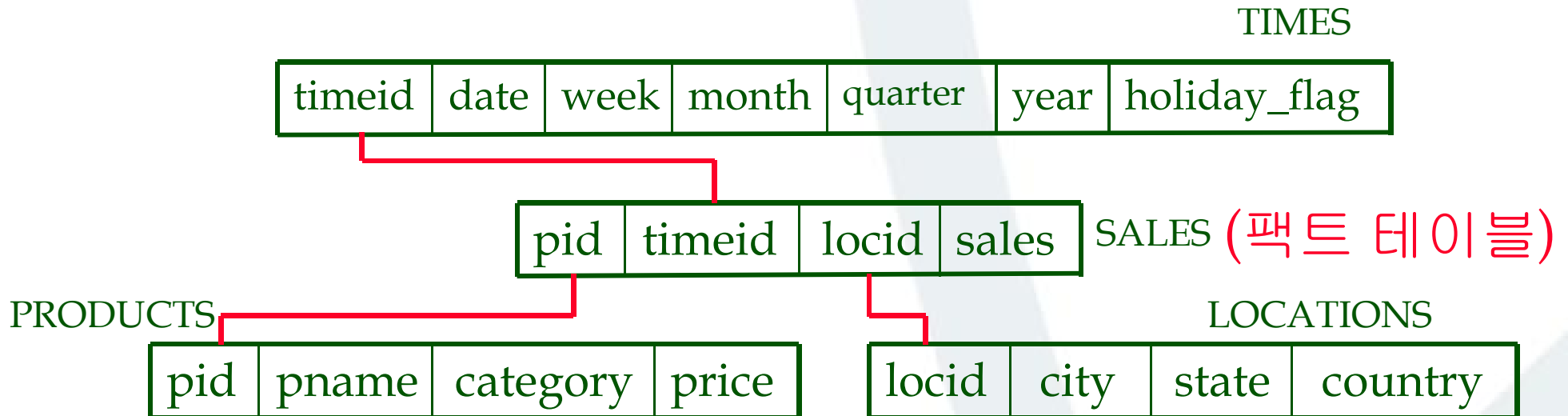
cross-tabulation:

v 슬라이싱 Slicing: 축 값에서
자른 단면

v 다이싱 Dicing: 축 범위로
떼어낸 부분

	WI	CA	Total
1995	63	81	144
1996	38	107	145
1997	75	35	110
Total	176	223	339

DB 스키마 모델링



- A ⌘ 테이블은 BCNF; 차원 테이블들은 정규화 안된 모습.
 - 차원 테이블은 작고, 갱신은 별로 발생하지 않는다; 따라서 갱신 이상의 단점보다는 검색질의 속도가 더 중요함.
- 이런 형태는OLAP에서 매우 흔히 나타나는데, 보통 스타 스키마 **star schema** 라고 한다; 이 테이블들을 모두 조인하는 연산을 스타 조인 **star join** 이라고 한다.

유의사항: 다차원 팩트 테이블

- N-진 relationship을 테이블로 표현한 것과 같다!

슬라이스
locid=1
의 예:

pid	timeid	locid	sales
11	1	1	25
11	2	1	8
11	3	1	15
12	1	1	30
12	2	1	20
12	3	1	50
13	1	1	8
13	2	1	10
13	3	1	10
11	1	2	35

pid	timeid	locid	sales
11	1	1	25
11	2	1	8
11	3	1	15
12	1	1	30
12	2	1	20
12	3	1	50
13	1	1	8
13	2	1	10
13	3	1	10
11	1	2	35

Data Mining

- 데이터로부터 지식 발견 knowledge discovery from data
- 방대한 데이터 모음에서 흥미로운 (non-trivial, implicit, previously unknown and potentially useful) 패턴이나 지식을 추출
- 다른 이름도 있다
 - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Data Mining



정의

대량의 데이터를 탐사하고 분석하여 유효하고 valid, 새로우며 novel, 쓸모가 있을것 같고 potentially useful, 궁극적으로는 이해가능한 ultimately understandable 패턴을 찾아내기.

패턴의 예 (미국 인구통계국 자료):

If (관계 = husband), then (성별 = male). 99.6%

Definition (Cont.)

대량의 데이터를 탐사하고 분석하여 유효하고 valid, 새로우며 novel, 쓸모가 있을것 같고 potentially useful, 궁극적으로는 이해가능한 ultimately understandable 패턴을 찾아내기.

Valid: 일반성.

Novel: 이전에는 알려지지 않았던.

Useful: 액션 고안이 가능한.

Understandable: 해석 가능하고 이해 가능한.

데이터 마이닝을 현재 사용하는 이유?

사람의 분석으로는 감당이 안됨:

- 데이터의 양과 다차원성
- 데이터 증가 곡선이 가파르다

이런 것들이 가용 available하게 됨:

- Data
- Storage
- Computational power
- Off-the-shelf software
- 전문적 지식

풍부한 데이터

- 마켓 결제 POS data
- 사람들의 신용카드 선호 패턴
- 신용카드 거래 자료
- 이메일 수발신
- 콜센터 누적자료
- 현금인출기 ATM machines
- 인구통계 데이터 Demographic data
- 센서 네트워크 (IoT)
- 카메라
- 웹서버 로그
- 고객 웹 사이트 추적값 Customer web site trails

데이터베이스 기술의 발전

- 1960년대: 계층모델, 네트워크 모델
- 1970년대: 관계 데이터 모델, 최초의 관계 DBMS 구현제품
- 1980년대: RDBMS의 속성, 특정 응용을 위한 DBMS, (공간 spacial data, 과학 scientific data, 이미지 image data, etc.), 객체지향 object-oriented DBMS
- 1990년대: 고성능 RDBMS 기술, 병렬 parallel DBMS, 테라바이트급 대용량 data warehouses, 객체관계 object-relational DBMS, 미들웨어 등 웹 기술 middleware and web technology
- 2000년대: 고가용성 High availability, 제로관리 zero-administration, 비즈니스 프로세스와 자연스런 통합 seamless integration into business processes
- 2010: 센서 DB 시스템 Sensor database systems, 내장형 DB databases on embedded systems, P2P DB 시스템, 대용량 Publish-subscribes, ???

데이터 마이닝을 현재 사용하는 이유?

경쟁의 압박!

성공의 비결이라면, 무언가 남들은 모르는 것을 아는 것.

선박왕 오나시스

- 서비스 경쟁은 가격 외의 요인들이 있다 (은행, 통신회사, 호텔 체인, 렌터카)
- 고객 관리 Personalization, CRM
- 실시간 회사 The real-time enterprise
- “전방위 청취 Systemic listening”
- 보안, 국가안보

사례연구: Bank

- **비즈니스 목표:** 주택담보대출 home equity loans 실적
- **현행 모델:**
 - 대학연령의 자녀가 있으면 등록금 때문에 주택을 담보로 대출을 받는다
 - 소득이 안정적이지 않으면 소득 충당을 위해 주택담보대출을 이용한다
- **데이터:**
 - 대규모 데이터 웨어하우스
 - 가동중인 42개 데이터 소스로부터 데이터를 받아서 통합

Case Study: Bank (Contd.)

1. 고객정보 중에서 주택담보대출 권유를 받아본 고객정보만 뽑아보기

- 대출 거절한 고객 Customers who declined
- 대출에 응답한 고객 Customers who signed up

1년수입	자녀수	수표발행 상한액	...	응답
\$40,000	2	\$1500		Yes
\$75,000	0	\$5000		No
\$50,000	1	\$3000		No
...

Case Study: Bank (Contd.)

2. 주택담보대출 권유에 고객이 긍정 응답할 것을 예측하는 규칙성을 찾아보기

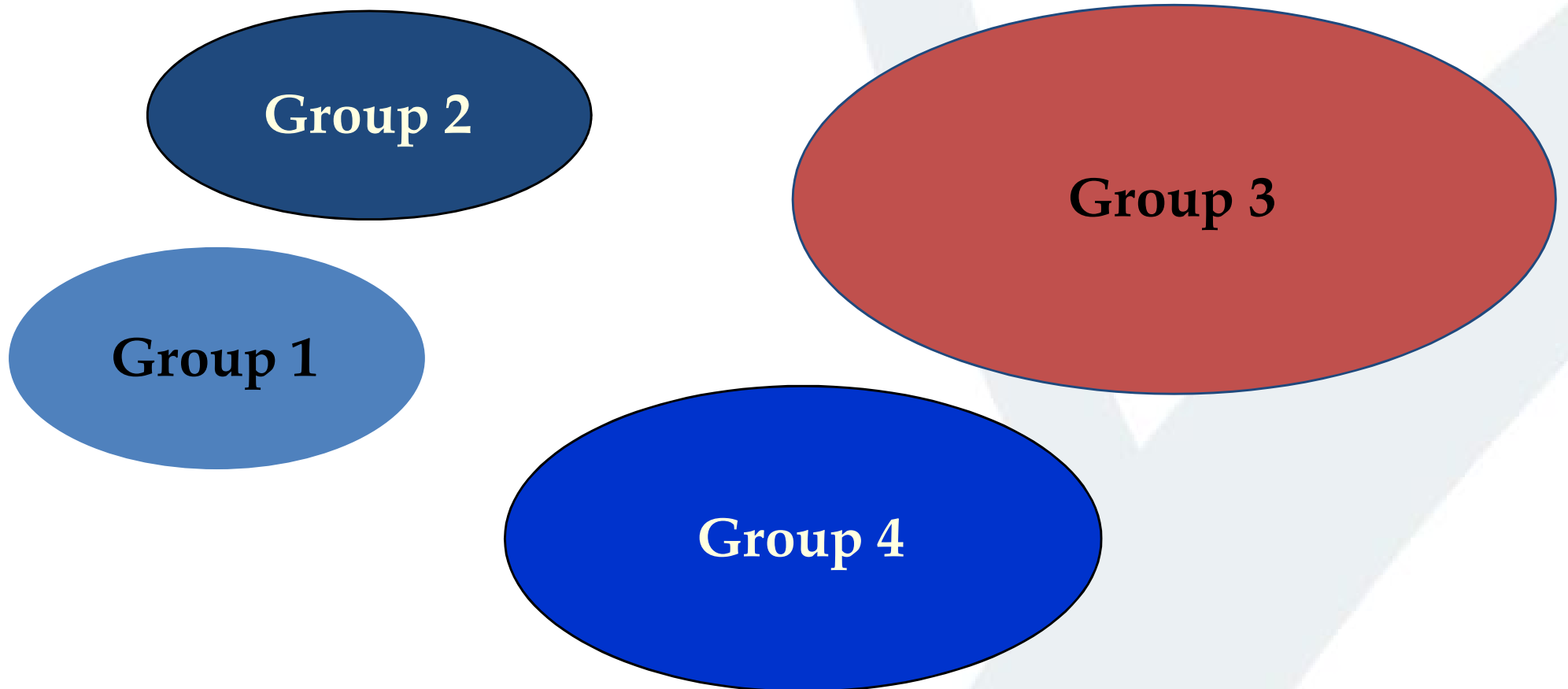
IF (수입 < 40k) and
(자녀수 > 0) and
(첫째자녀나이 > 18 and 첫째자녀나이 < 22)

THEN YES

...

Case Study: Bank (Contd.)

3. 고객을 묶는 클러스터(고객군)를 이리 저리 만들어 보고 하나하나 검토하기



Case Study: Bank (Contd.)



4. 검토 결과:

- 대부분의 클러스터는 “흥미롭지” 않았다.
- 클러스터 하나가 흥미로웠다! 업무보는 계좌와 개인 용 계좌를 따로 가지고 있는 고객군; 긍정 응답율이 눈에 띄게 높았음

Example: Bank (Contd.)

액션:

- 대출 마케팅을 바꾼다

결과:

- 주택담보대출 응답률이 2배로 향상

사례 연구: 사기 탐지

- **관련 산업:** 의료, 판매, 신용카드, 통신, 기업간 거래
- **Approach:**
 - 이력 historical 데이터를 이용해서 사기 행위 모델을 구축
 - 이 모델을 도입 deploy 해서 사기성 행위를 식별

사기 탐지 (계속)

- 예:

- 자동차 보험: 보험금을 노린 사고를 일으키는 사람들이 어떤 그룹인지 탐지
- 의료 보험: 부정 보험청구
- 돈 세탁: 의심스러운 돈 트랜잭션을 탐지 (금융감독원 네트워크)
- 통신 산업: 정상 범위를 이탈하는 호출패턴을 탐지 (호출의 송신자와 수신자, 통화시간, 하루중 몇 시, 한주의 어느 요일)

사례 연구: 스포츠

IBM Advanced Scout 는 미국 농구 NBA의 경기 통계를 분석하였다

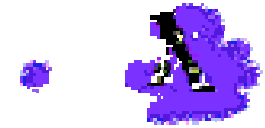
- 슛 블록
- 어시스트
- 파울

- Google: “IBM Advanced Scout”



Advanced Scout

- 패턴의 예: 뉴욕 킥스와 샬럿데 호넷의 어떤 경기를 분석해보니, "*Glenn Rice가 슈팅 가드로 출전했을때 점프슛 5/6 (83%)을 기록했다.*"
- 패턴이 흥미로운 이유:
이 게임에서 샬럿데 호넷 팀 전체의 슈팅율은 54%였다.



사례 연구: 천문 탐사 Sky Survey

- 입력 데이터: 6년 이상 모은 3 TB 의 이미지 데이터. 20억 개 이상의 천문 물체.
- 목표: 물체들을 모두 타입으로 분류해서 카탈로그로 만들기.
- 방법: 결정트리를 이용해서 데이터 마이닝
- 결과:
 - 천문 물체 예측 클래스의 정확도 94%.
 - 가짜 물체를 분류해내는 수가 300% 증가.
 - 천문학자 지원팀이 16개의 새로운 고준위 적색편이 퀘이사를 발견했는데, 관측시간을 10배로 줄이고도 가능했다.

마이닝(지식발견) 프로세스

순서:

- | 비즈니스 문제 식별
- | 데이터 마이닝
- | 액션 (실행)
- | 평가 및 측정
- | 도입(배치) 및 비즈니스 과정 속에 융합

데이터 마이닝 순서 상세

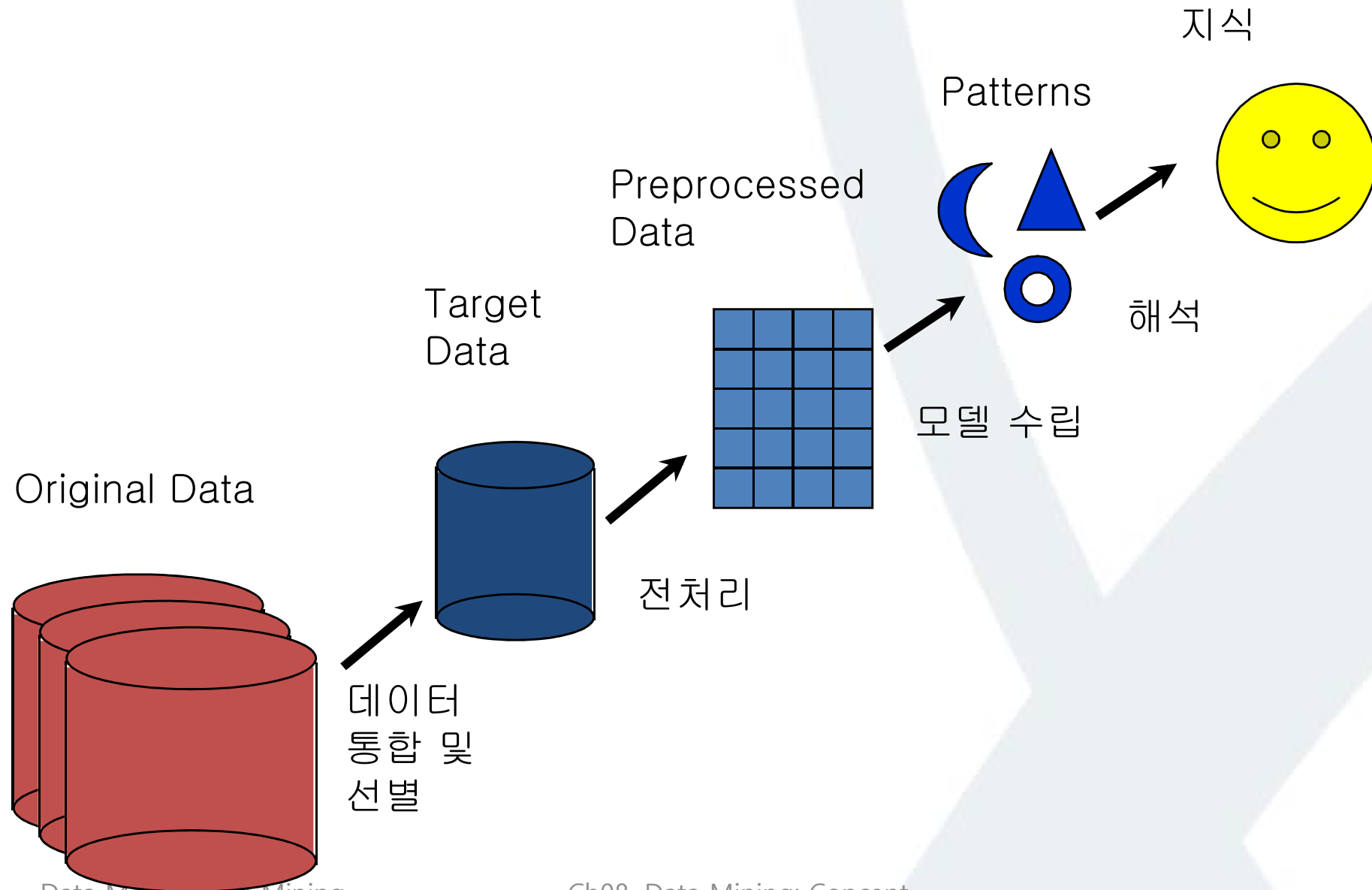
2.1 전처리 Data preprocessing

- 선별 Data selection: 목표 데이터군 target datasets과 관련 필드들을 파악한다
- 정리 Data cleaning
 - 'A' 값 제거 Remove noise and outliers
 - 형식 변환 Data transformation
 - 단위 통일 Create common units
 - 필요하다면 애트리뷰트 추가 Generate new fields

2.2 모델 수립 Data mining model construction

2.3 모델 평가 Model evaluation

Preprocessing and Mining



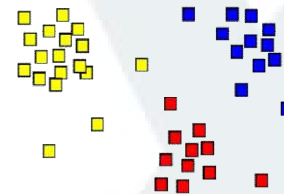
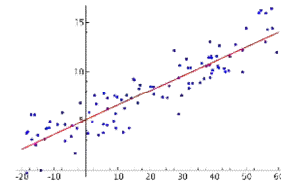
데이터 마이닝 모델이란?

데이터군에 대한 특정 측면의 기술(묘사)

a description of a specific aspect of a dataset. 입력값들이 주어지면, 출력값들을 만들어 낸다.

모델의 예:

- 선형 회귀 모델 Linear regression model
- \hat{y} x 모델 Classification model
- 클러스터링 (군집화) Clustering



데이터 마이닝 모델이란? (계속)

데이터 마이닝 모델 하나는 2가지 레벨별로 설명할 수 있다:

- 기능 Functional 레벨:
 - 용도에 따른 설명.
예: 분류 모델, 클러스터링 모델
- 표현 Representational 레벨:
 - 표현 방식에 따른 설명.
예: 로그 선형 Log-linear 모델, 분류 트리 classification tree, 최근접점 nearest neighbor 기법.
- 블랙박스 Black-box 모델 versus 투명 모델 transparent model

Data Mining: Types of Data

- Relational data and transactional data
 - Spatial and temporal data, spatio-temporal observations
 - Time-series data
 - Text
 - Images, video
 - Mixtures of data
 - Sequence data
-
- Features from processing other data sources

변수(애트리뷰트)의 타입

- 수치형 *Numerical*. Domain is ordered and can be represented on the real line (e.g., age, income)
- 카테고리형 *Nominal* or *categorical*. Domain is a finite set without any natural ordering (e.g., occupation, marital status, race)
- 순서형 *Ordinal*. Domain is ordered, but absolute differences between values is unknown (e.g., preference scale, severity of an injury)

Data Mining 기법 종류

- 지도학습 Supervised learning
 - 분류, 회귀
- 비지도학습 Unsupervised learning
 - 클러스터링
- 연관관계 모델링 Dependency modeling
 - 연관 Associations, 요약 summarization, 인과관계 causality
- 이상치, 편차 탐지 Outlier detection
- 변화 탐지 Trend analysis and change detection

분류의 예

- 학습 데이터베이스 training database의 예
 - 예측 속성 predictor attribute 2개 : Age 와 Car-type (Sport, Minivan and Truck)
 - Age 는 순서형 ordered, Car-type 은 카테고리형 categorical
 - Class 라벨: 사람이 제품을 샀는지 아닌지를 표시
 - 종속 속성 dependent attribute 은 카테고리형

Age	Car	Class
20	M	Yes
30	M	Yes
25	T	No
30	S	Yes
40	S	Yes
20	T	No
30	M	Yes
25	M	Yes
40	M	Yes
20	S	No

회귀의 예

- 학습 데이터베이스 training database의 예
 - 예측 속성 predictor attribute 2개 : Age 와 Car-type (Sport, Minivan and Truck)
 - Spent : 사람이 최근의 웹사이트 방문에서 얼마를 썼는지를 표시
 - 종속 속성 Dependent attribute 은 수치/형

Age	Car	Spent
20	M	\$200
30	M	\$150
25	T	\$300
30	S	\$220
40	S	\$400
20	T	\$80
30	M	\$100
25	M	\$125
40	M	\$500
20	S	\$420

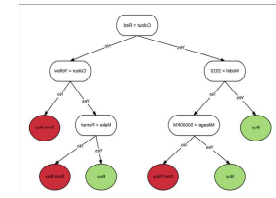
분류

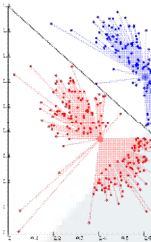
응용의 예: 텔레마케팅 telemarketing

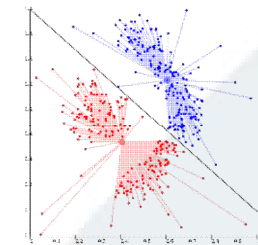


Copyright © 1997 United Feature Syndicate, Inc.
Redistribution in whole or in part prohibited

분류 (계속)



- 결정 트리 Decision tree를 접근방식으로 사용
 - 다른 접근방식 approach 들도 있다
 - 선형 판별 분석 Linear Discriminant Analysis
 - *K-최근접점* *k*-nearest neighbor methods
 - 로지스틱 회귀 Logistic regression
 - 신경망 Neural network
 - 서포트 벡터 머신 Support Vector Machine
- 



분류의 예

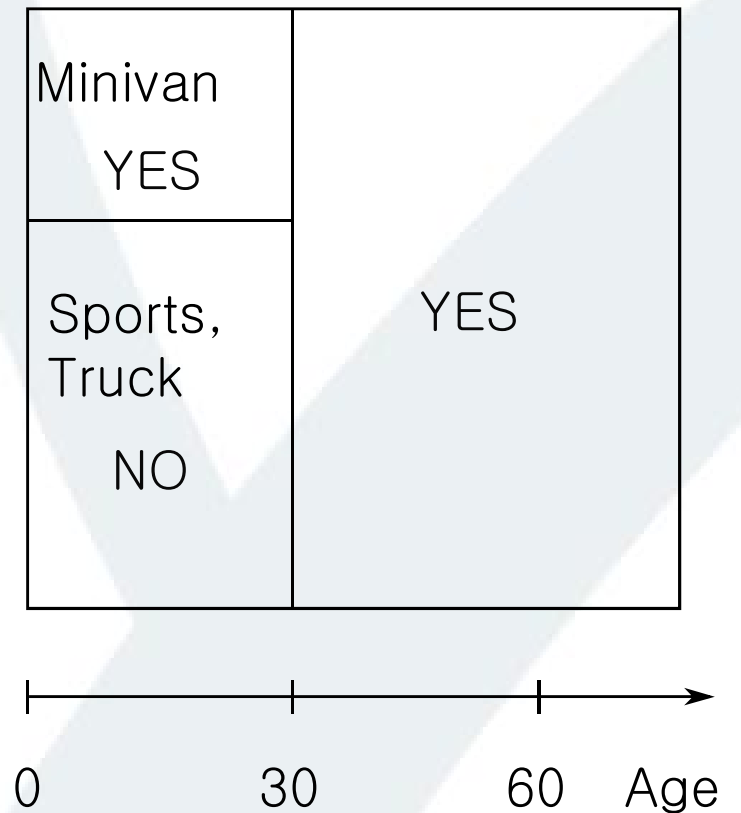
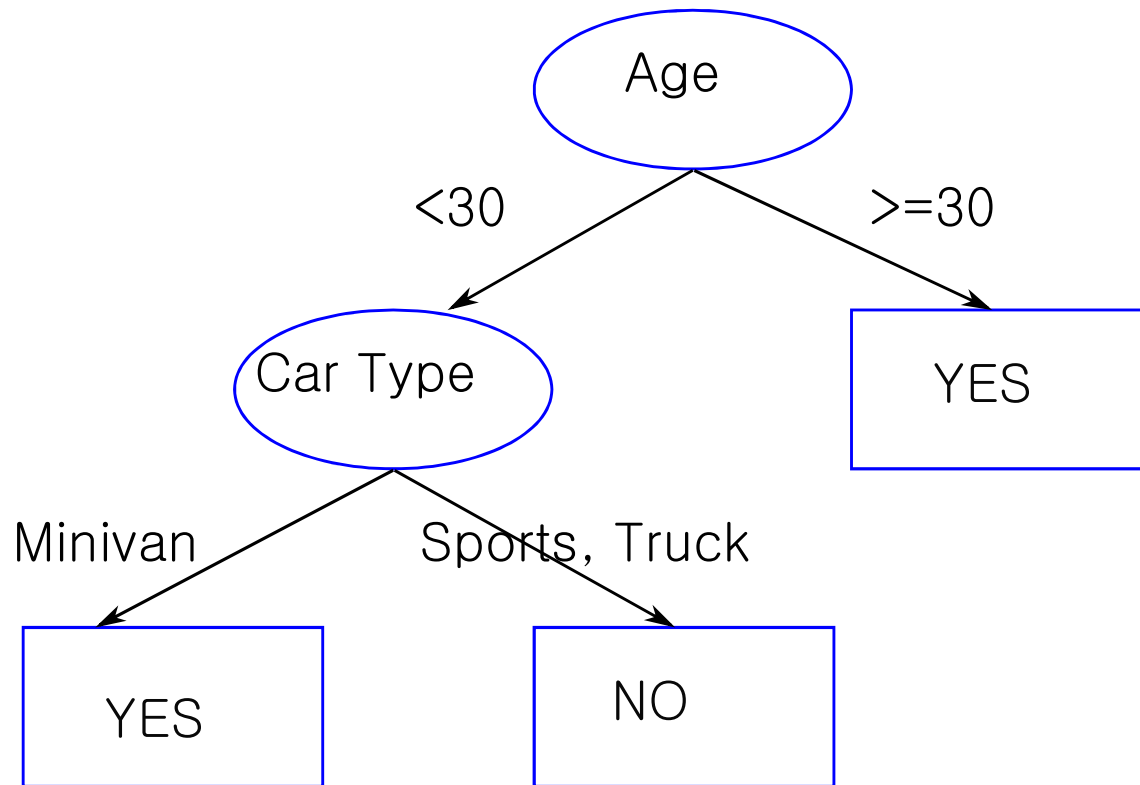
- 학습 데이터베이스 training database의 예
 - 예측 속성 predictor attribute 2개 : Age 와 Car-type (Sport, Minivan and Truck)
 - Age 는 순서형 ordered, Car-type 은 카테고리형 categorical
 - Class 라벨: 사람이 제품을 샀는지 아닌지를 표시
 - 종속 속성 dependent attribute 은 카테고리형

Age	Car	Class
20	M	Yes
30	M	Yes
25	T	No
30	S	Yes
40	S	Yes
20	T	No
30	M	Yes
25	M	Yes
40	M	Yes
20	S	No

목표와 요건

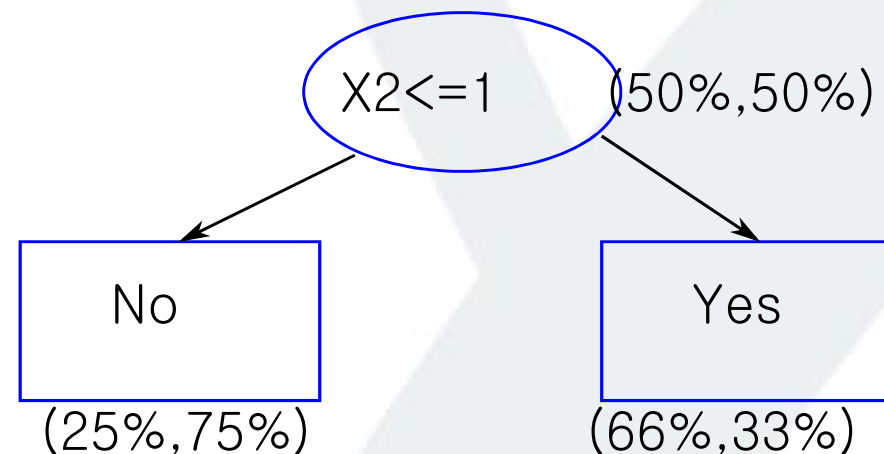
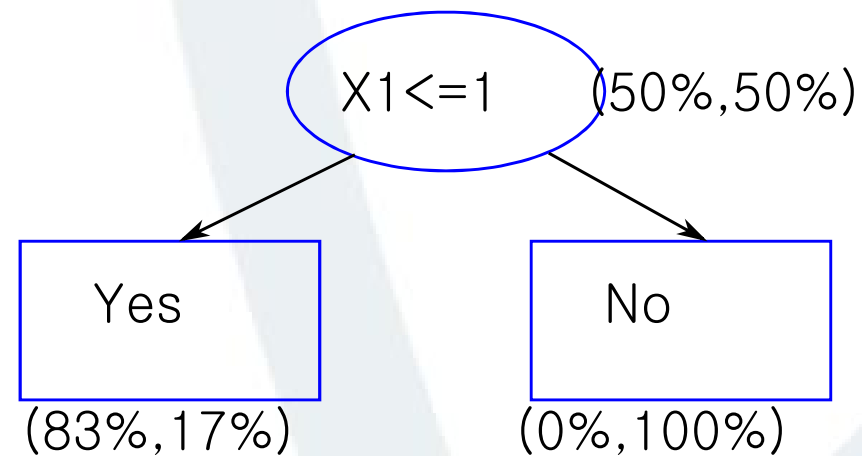
- 목표:
 - 정확한 분류기 classifier/회귀함수 regression function를 생성
 - ⑮주어진 문제의 구조를 이해
- 모델의 요건 Requirements:
 - 높은 정확성
 - 사람이 이해할 수 있고 해석할 수 있도록 interpretable
 - 대규모의 학습 데이터베이스를 빠르게 구축할 수 있어야

결정 트리란?



직관: Impurity Function

X1	X2	Class
1	1	Yes
1	2	Yes
1	2	Yes
1	2	Yes
1	2	Yes
1	1	No
2	1	No
2	1	No
2	2	No
2	2	No



클러스터링

클러스터링: 비지도학습

- □ □ , 참고:
 - 데이터군 Data Set D (학습 셋)
 - 유사도 Similarity/거리값 distance metric/정보 information
- 찾는다:
 - 데이터 사이의 분할 Partitioning of data
 - 유사한/가까운 항목들을 그룹짓기 Groups of similar/close items

유사도란 Similarity?

- 유사한 고객 그룹
 - 인구통계특성의 유사성
 - 〃 〃 행위의 유사성
 - ⑮건강의 유사성
- 유사한 상품 그룹
 - 가격의 유사성
 - ⑮기능의 유사성
 - 매장의 유사성
 - ...
- 유사도는 도메인/문제가 무엇이냐에 따라 달라지는게 일반적이다.

Association Analysis

(연관 분석)

장바구니 분석 Market Basket Analysis

- 장바구니에 상품 item 들이 들어있다고 하자.
- 장바구니 분석은 다음 질문에 답하려는 노력이다.
 - 어떤 사람이 구매하는가?
 - ⑮고객이 어떤 것들을 같이 사는가?
 - 고객이 상품(아이템)을 어떤 순서로 사는가?

장바구니 분석

주어진 것:

- 고객 거래 customer transaction 데이터베이스
- ϵ 트랜잭션은 아이템의 집합이다.
- 1 :
TID 111 트랜잭션에는
아이템 {Pen, Ink,
Milk, Juice} 가 있다.

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

유의사항: 장바구니 분석

- 고객 거래 customer transaction 데이터베이스
- 원본 데이터 (조인) 테이블 형태가 된다!

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

장바구니 분석 (계속)

- 동시발생
 - 고객의 80%가 X, Y, Z를 같이 구매한다.
- $\downarrow \ni \square \dot{P}$ Association rules
 - X와 Y를 구매한 고객의 60%가 Z도 구매한다.
- $\downarrow ,$ 패턴 Sequential patterns
 - X를 구매한 고객의 60%가 그 후에 Y를 구매한다.

입증도, 지지도

다음 2가지 기준을 통과한 연관 규칙들만 남긴다.

- 입증도 Confidence :
 - $X \rightarrow Y$ 의 입증도 c if $P(Y|X) = c$
 - X 가 나온 트랜잭션 중에, Y 도 있는 트랜잭션의 비율
- 지지도 Support :
 - $X \rightarrow Y$ 의 지지도 s if $P(XY) = s$
 - ⑮ 전체 트랜잭션 중에, X 와 Y 가 있는 트랜잭션의 비율

We can also define

- 아이템셋(동시출현) itemset XY 의 지지도 Support :
 - s if $P(XY) = s$

예

예:

- {Pen} => {Milk}
Support: 75%
Confidence: 75%
- {Ink} => {Pen}
Support: 75%
Confidence: 100%

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

예

- support $\geq 75\%$ 인
아이템집합을 모두 찾는다면?



























TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

예

- support $\geq 50\%$ 인
연관 규칙들을 모두
찾는다면?

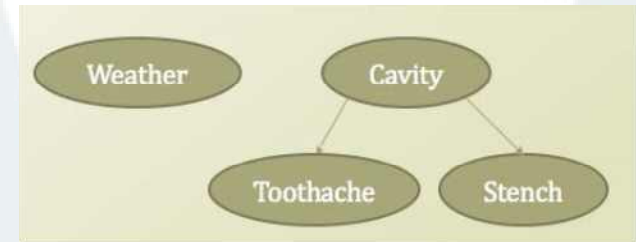
TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

장바구니 분석: 응용

-                          

빈발 아이템 집합 Frequent Itemset의 응용

- 장바구니 분석
- 연관 규칙
- 분류 (특히: 텍스트, 희귀 클래스 rare classes)
- 베이지안 네트워크 Bayesian Network 수립의 기초
- 로그 Web log 분석
- 협업 필터링 Collaborative filtering



A		✓	✗	✓	✓
B			✓	✗	✗
C		✓	✓	✗	
D		✗		✓	
E		✓	✓	?	✗



The End..

Thanks to
Jiawei Han, and
Raghu Ramakrishnan