

ARM (Association Rule Mining, 연관규칙탐사)

개념, 알고리즘 및 응용

연관규칙탐사(ARM)란 ?

- 연관규칙탐사(ARM: Association Rule Mining) : 하나의 거래나 사건에 포함되어 있는 항목들의 경향을 파악해서 상호 연관성을 발견 하는 것
EX) Products in Shopping Cart (One trip, Together)



- 1) 구매자가 제품을 구매할 때 이웃의 영향이 있었는가?
- 2) 오렌지 주스와 청정재 구입시 윈도우 클리너를 같이 구입하는가?
- 3) 우유를 바나나 구입시 함께 구입하는가? 또한 구입 할 때 특정 브랜드를 구입 하는가?
- 4) 청정재를 어느 곳에 위치시켜야지만 판매고를 최대화하는가?



연관규칙(Association Rule) (1/3)

- 어떤 Item 집합의 존재가 다른 Item 집합의 존재를 암시하는 것을 의미하며 다음과 같이 표시한다.

(Item set A) \Rightarrow (Item set B)

(if A then B : 만일 A 가 일어나면 B 가 일어난다.)

- 함께 구매하는 상품의 조합이나 서비스 패턴 발견하는데 이용
- 특정 제품 또는 사건들이 동시에 발생 하는 패턴을 파악하는데 이용
EX) 가정 용품 판매 기간 동안 같이 판매해야 하는 상품의 패턴 발견

연관규칙(Association Rule) (2/3)

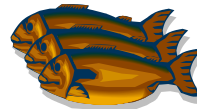
Buying Pattern



전항(Antecedent)



야채



생선

후항(Consequent)



포도주? 맥주?

전항(Antecedent)



핸드폰



잭 연결기

후항(Consequent)



이어폰? 메모리?

결론 \leq 전제(1) & 전제(2) & ... & 전제(m)
Consequent Antecedents

연관규칙(Association Rule) (3/3)



Pattern Miner System



연관규칙의 평가기준 (1/3)

■ 지지도 (Support)

- 전체 거래 중 항목 X와 항목 Y를 동시에 포함하는 거래가 어느 정도인가 ?

$$S = P(X \cap Y) = \frac{\text{품목X와 품목Y를 포함하는 거래 수}}{\text{전체 거래 수}(N)}$$

- 전체적 구매도에 대한 경향을 파악



연관규칙의 평가기준 (2/3)

■ 신뢰도 (Confidence)

- 항목 X를 포함하는 거래 중에서 항목 Y가 포함될 확률은 어느 정도인가 ?

$$C = P(Y | X) = \frac{P(X \cap Y)}{P(X)}$$

$$= \frac{\text{품목X와 품목Y를 포함하는 거래 수}}{\text{품목X를 포함한 거래 수}}$$

- 조건부확률
- 연관성의 정도
- not symmetric



연관규칙의 평가기준 (3/3)

■ 향상도 (Lift)

- 항목 X를 구매한 경우 그 거래가 항목 Y를 포함하는 경우와 항목 Y가 X와 무관하게 임의로 구매되는 경우의 비율

$$L = \frac{P(Y | X)}{P(Y)} = \frac{P(X \cap Y)}{P(X)P(Y)}$$

Lift	의 미	예
1	두 품목이 서로 독립적인 관계	과자와 후추
> 1	두 품목이 서로 양의 상관 관계	빵과 버터
< 1	두 품목이 서로 음의 상관 관계	지사제, 변비약

연관규칙탐사 예제

고객의 구매 상품 List

ID	판매 상품
1	소주 , 콜라 , 맥주
2	소주 , 콜라 , 와인
3	소주 , 주스
4	콜라 , 맥주
5	소주 , 콜라 , 맥주 , 와인
6	주스

지지도가 50% 이상인 연관성 규칙

지지도 50% 이상인 규칙	해당 Transaction	신뢰도
소주 => 콜라	1,2,5	75 %
콜라 => 맥주	1,4,5	75 %
맥주 => 콜라	1,4,5	100 %

- Lift = $P(\text{콜라}|\text{맥주}) / P(\text{콜라}) = 1 / (4/6) = 1.5$

*** 연관규칙 : 맥주를 구입한 사람들 모두는(100%) 콜라도 구매한다**

- 지지도: 그리고 이러한 경향을 가지는 사람들은 전체의 절반(50%) 정도이다.
- 리프트: 맥주 구매 시 콜라를 구입하게 될 가능성은 맥주 구매가 전제되지 않았을 경우보다 1.5배나 높아진다.



Exercises

- 아래의 트랜잭션에서 추출된 연관규칙 중 하나인 “B \rightarrow C”의 신뢰도(Confidence)는 얼마인가?
transaction #1 {A, B, C}
transaction #2 {A, B, D}
transaction #3 {A, E}
transaction #4 {B, C}
transaction #5 {A, B, C, D}
transaction #6 {E}
- 어느 할인매장의 생필품 판매내역으로부터 다음과 같은 결과를 얻었다. $\Pr(\text{세제})=0.4$, $\Pr(\text{식용유})=0.5$, $\Pr(\text{세제} \ \& \ \text{식용유})=0.3$. 이 값으로부터 연관성규칙 "세제 \rightarrow 식용유"의 향상도(Lift)를 구하면 얼마인가?



연관규칙탐사 프로세스

적절한 **Item Set** 결정 및 분석 수준 결정

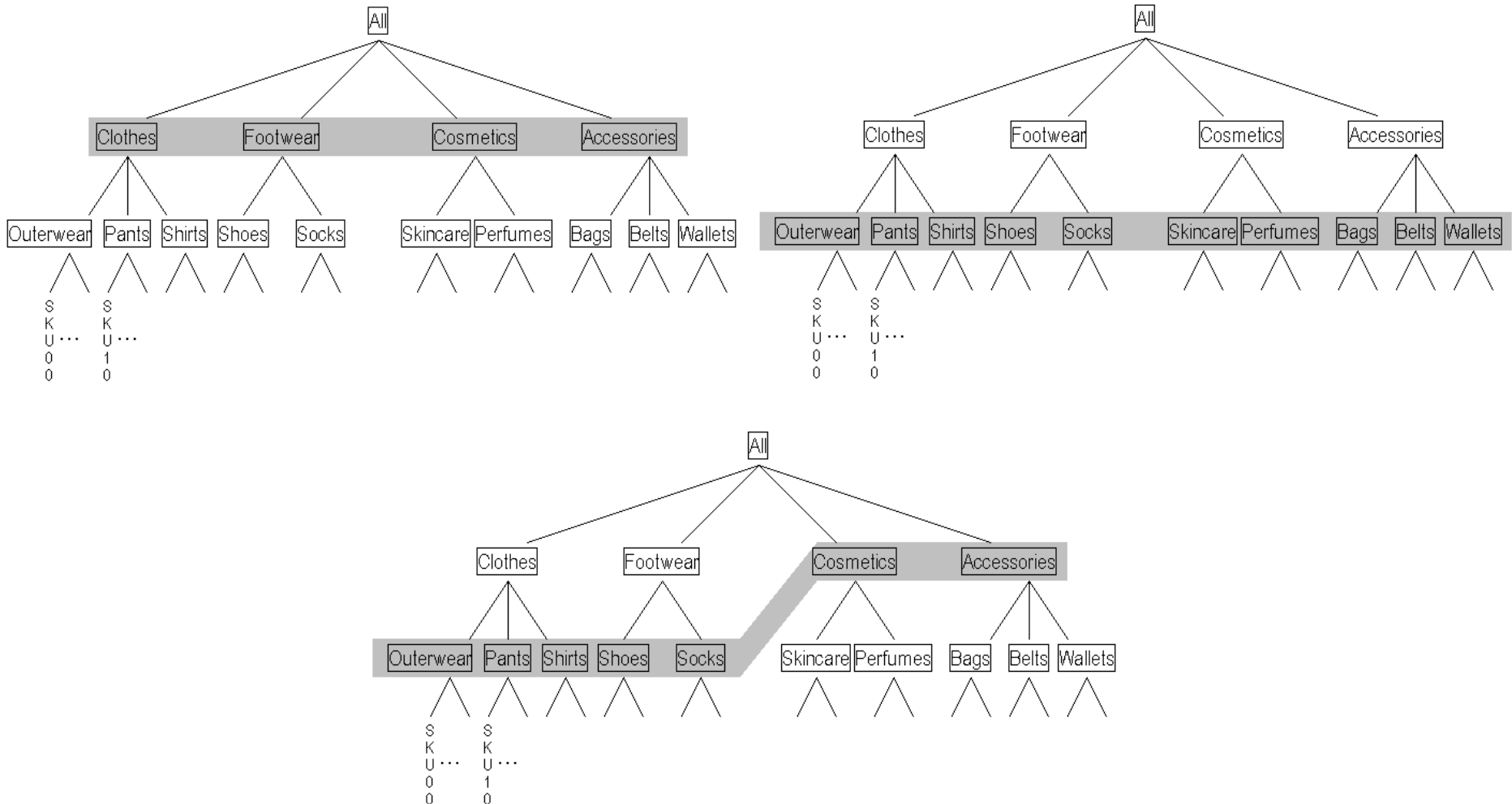
상품간 단순 패턴 발견

- 소주와 콜라, 맥주와 콜라는 타(他)상품의 경우보다 동시구매 횟수가 높다.
- 주스는 맥주, 콜라, 와인과 동시에 구매되지 않는다.

연관규칙 발견

- 지지도, 신뢰도, 리프트 값을 통한 연관규칙의 유용성 분석
- 유용한 연관규칙 결정

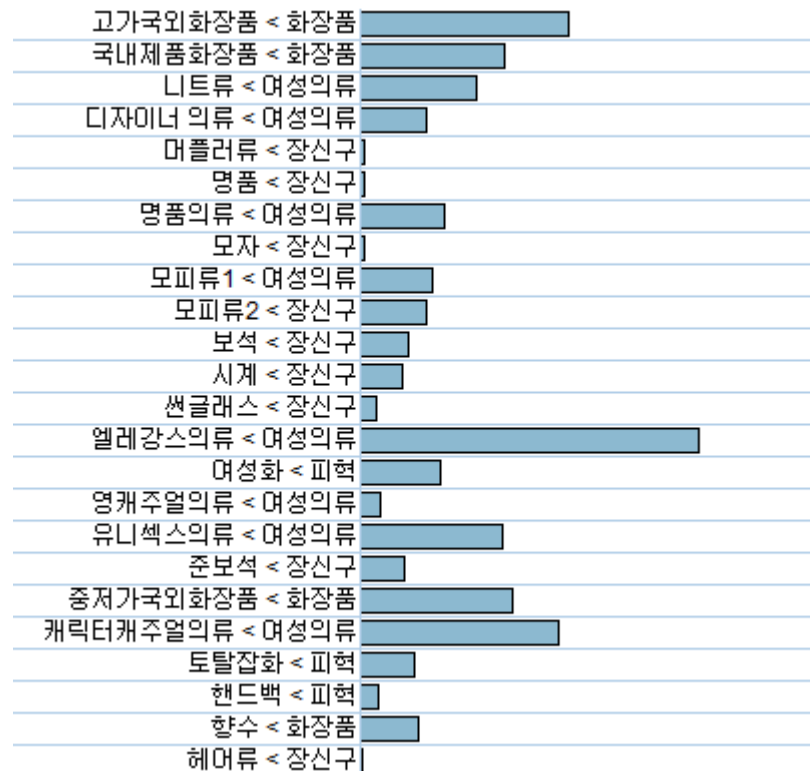
Item 분석수준(Grain) 결정의 예 (1/2)



Item 분석수준(Grain) 결정의 예 (2/2)

마리끌레르 < 명품의류 < 여성의류
마리나라날디 < 모피류1 < 여성의류
마리양주 < 캐릭터캐주얼의류 < 여성의류
마인 < 유니섹스의류 < 여성의류
막스마라 < 모피류1 < 여성의류
막스마라행사 < 모피류1 < 여성의류
막스앤스펜서 < 명품의류 < 여성의류
막스앤코 < 모피류1 < 여성의류
말로 < 모피류1 < 여성의류
머스트비 < 캐릭터캐주얼의류 < 여성의류
메세지 < 캐릭터캐주얼의류 < 여성의류
메세 < 여성화 < 피혁
메이컬포에버 < 중저가국외화장품 < 화장품
메키 < 니트류 < 여성의류
메트로씨티 < 여성화 < 피혁
모네 < 준보석 < 장신구
모라도 < 니트류 < 여성의류
모르간 < 캐릭터캐주얼의류 < 여성의류
모스키노 < 모피류1 < 여성의류
미끄마끄 < 명품의류 < 여성의류
미니멈 < 유니섹스의류 < 여성의류
미샤 < 유니섹스의류 < 여성의류
미세스정 < 영캐주얼의류 < 여성의류
미소니 < 모피류1 < 여성의류
미소페 < 여성화 < 피혁
미스로즈 < 니트류 < 여성의류
미스박 < 디자이너 의류 < 여성의류
미스식스티 < 엘레강스의류 < 여성의류
미스지 < 디자이너 의류 < 여성의류
미쏘니 < 모피류1 < 여성의류
미쓰제이 < 여성화 < 피혁
미오르제타 < 모피류2 < 장신구
미찌코런던 < 여성화 < 피혁
민수여성의류 < 유니섹스의류 < 여성의류
밀라노스토리무역 < 유니섹스의류 < 여성...
밀로스 < 니트류 < 여성의류
바닐라 < 캐릭터캐주얼의류 < 여성의류
바바라 < 니트류 < 여성의류
바비브라운 < 고가국외화장품 < 화장품
바이블랙 < 유니섹스의류 < 여성의류
박동준 < 디자이너 의류 < 여성의류
박순영니트 < 니트류 < 여성의류
박윤수 < 디자이너 의류 < 여성의류

H백화점 여성용품 Case





연관규칙탐사의 결과유형

■ Useful Result

- 마케팅 전략상 유용한 결과가 나온 경우
- EX) 주말을 위해, 목요일 소매점에 기저귀를 사러 온 아빠들은 맥주도 함께 사 간다. => 주말에 FOOTBALL을 보면서 마심

■ Trivial Result

- 기존의 마케팅 전략에 의해 연관성이 높게 나온 경우
- EX) 정비계약을 맺은 소비자들은 많은 설비를 구매 할 것 같다. => 정비계약은 대개의 경우 따로 맺어지는 것이 아니라, 많은 설비 구입시 함께 제시된다.

■ Inexplicable Result

- 의미를 발견하기 위해 많은 고민이 필요한 경우
- EX) 새로 철물점을 개업하면, 대개 화장실 문고리를 많이 사 간다.



의미 있는 연관규칙의 도출

- 지지도 값의 최소기준치를 미리 설정하여
- 최소기준치 이상의 지지도 값을 갖는 규칙을 생성한다.
- 생성된 규칙 중 높은 신뢰도를 갖는 규칙들을 의미 있는 연관규칙으로 선정한다.
- 자주 구매되는 상품에 대해서 지지도와 신뢰도가 우연히 높게 나올 수 있다.
 - 리프트 (>1)



연관규칙탐사의 장단점

■ 장점

- 전문지식이 필요치 않으며 결과에 대한 이해가 쉬움
- 도출된 규칙간의 상호비교, 평가가 쉬움
- Undirected Data 분석에 유용
- 다양한 크기의 데이터에 적합
- 신경망이나 유전자 알고리즘에 비해 단순

■ 단점

- 문제의 크기가 커질수록 지수적으로 증가
- 데이터 속성에 대한 제한적 지원
- 항목에 대한 올바른 수 결정의 어려움
- 희박한 항목에 대해서는 문제화
- 품목의 수에 비해 거래 수가 충분치 못하면 신뢰 확률이 낮은 연관규칙 발견 가능성



연관규칙탐사 활용 분야

교차판매 (Cross-Selling), 상승판매 (Up-Selling)

- ✓ 스펜서 존슨의 '누가 내 치즈를 옮겼을까?'라는 책을 구매한 고객에게 최인훈의 '상도' 연관 상품을 추천하는 데 활용

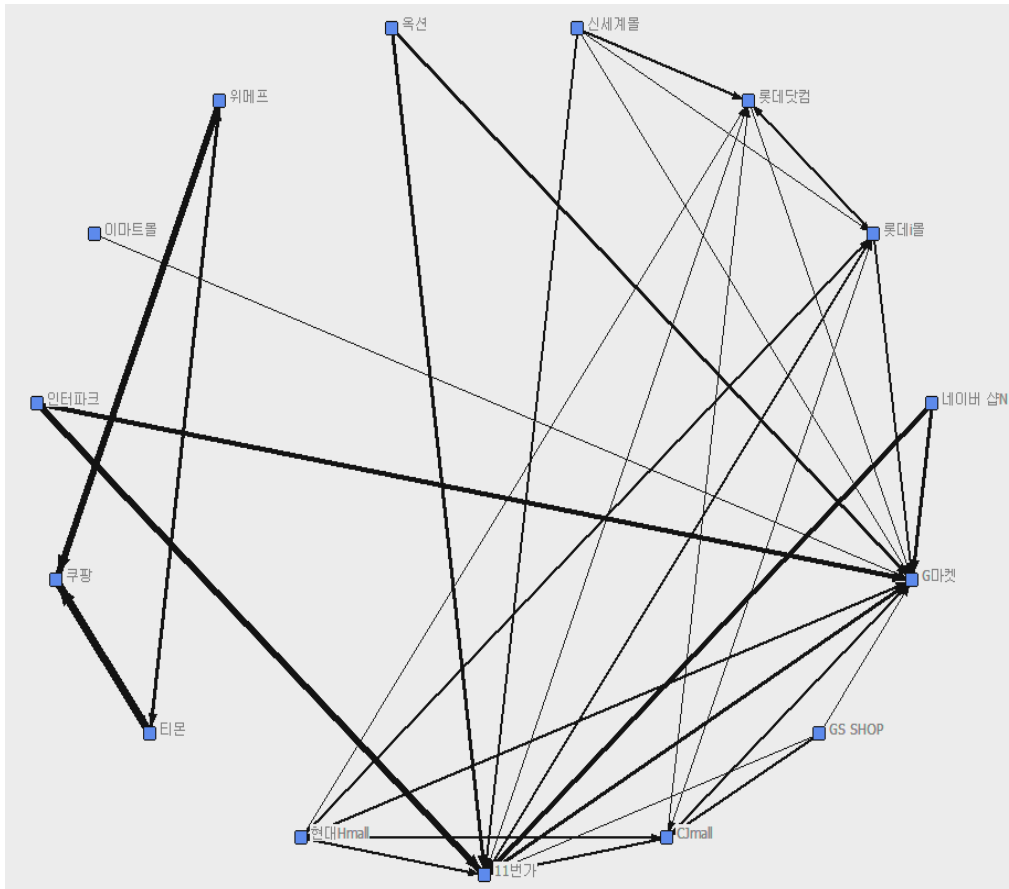
부정탐지 (Fraud Detection) : Negative Rule의 활용

- ✓ 신용카드 회사와 같은 금융기관에서는 연관성 규칙을 이용하여 카드 도용과 같은 부정행위를 적발하는 데 활용
 - Negative Rule의 활용
 - Negative Rule은 조건과 결과에 'True' 뿐만 아니라 'False'를 포함한다. 예) $\sim A \Rightarrow B$, $A \Rightarrow \sim B$, $\sim A \Rightarrow \sim B$ 등

매장의 상품진열 (Shelf Planning)

- ✓ 「케이크 ► 와인」이라는 유용한 연관 규칙이 발견 되었다면, 케이크와 와인 상품을 나란히 진열하여 동시 구매를 유도하는 데 활용

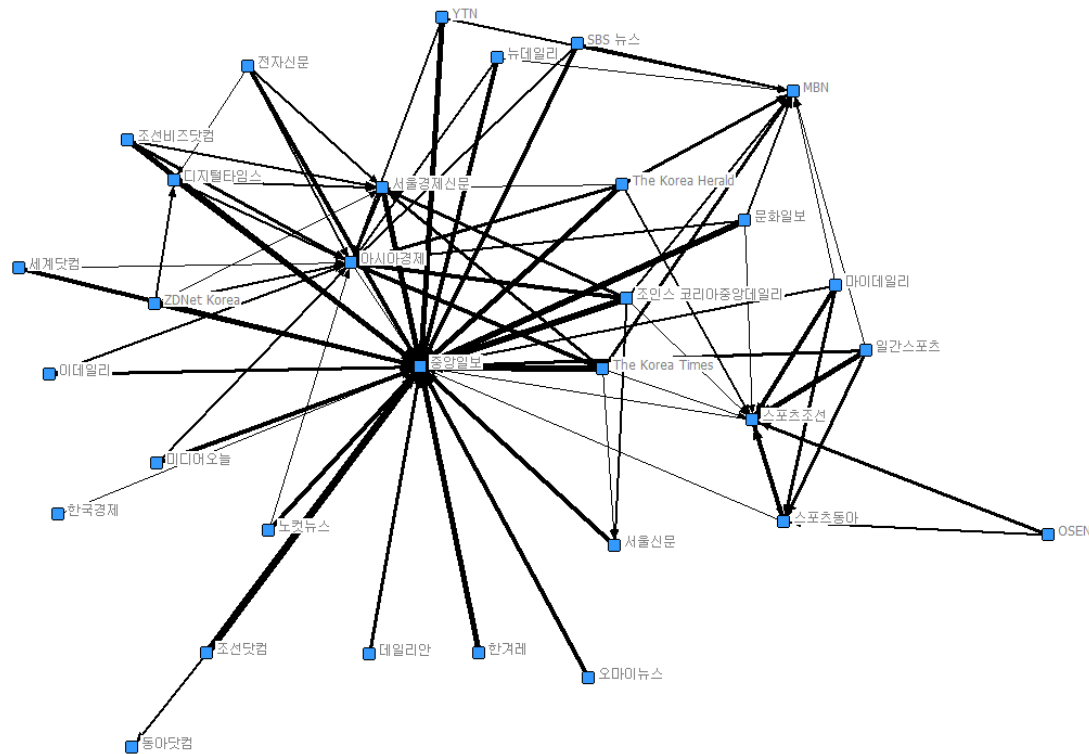
Case Study – 클릭스트림 분석 (1/2)



▶ ARM과 SNA를 결합한 분석

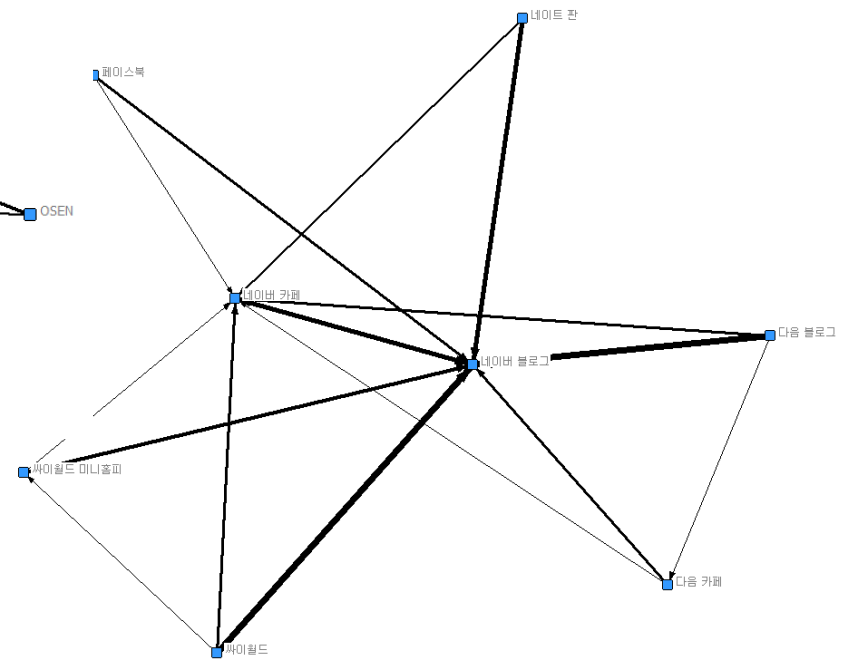
- 5000명의 1년간 쇼핑사이트 접속기록을 활용
- 한 사람이 하루 동안 동시에 접속하는 쇼핑사이트 리스트를 하나의 트랜잭션으로 간주
- 전향(Antecedents)의 항목 수를 1로 고정. 즉 $A \Rightarrow B$ 의 연관규칙만 생성되도록 함
- $A \Rightarrow B$ 라는 규칙이 발견되면 A에서 B로 링크를 연결하고 링크의 강도(굵기)는 신뢰도로 표현하는 Social Network를 구축
- 쇼핑사이트 간의 관계 및 각 쇼핑사이트의 브랜드 파워 차이를 시각적으로 파악할 수 있음

Case Study – 클릭스트림 분석 (2/2)



▶ 뉴스 사이트 관계분석

▶ 커뮤니티 사이트 관계분석



Case Study – POS 데이터 분석

■ 상승규칙 vs. 하향규칙

- 분기별 판매규칙을 기준으로 꾸준히 신뢰도가 상승 또는 하락하는 규칙

규칙	1분기	2분기	3분기	4분기	판정
치킨1⇒치킨2	25%	26%	28%	31%	상승 규칙
치킨1⇒짬닭	61%	52%	57%	60%	
짬닭⇒치킨1	76%	54%	43%	42%	하향 규칙
치킨2⇒소주	46%	36%	33%	45%	

■ 소멸규칙 vs. 새로운 규칙

- 기준 월(분기)과 비교대상 월(분기)을 정하여 비교했을 때 소멸 또는 새로 나타나는 규칙

규칙	1분기	2분기	3분기	4분기	판정
치킨1⇒치킨2	25%	26%	28%	0%	소멸 규칙
치킨1⇒짬닭	61%	52%	57%	60%	
짬닭⇒치킨1	34%	43%	43%	41%	
치킨1⇒소주	45%	46%	43%	45%	

규칙	1분기	2분기	3분기	4분기	판정
치킨1⇒치킨2	25%	26%	28%	32%	
치킨1⇒짬닭	61%	52%	57%	60%	
짬닭⇒치킨1	0%	0%	0%	23%	새로운 규칙
치킨1⇒소주	45%	46%	43%	45%	



Case Study – 주가 분석 (1/2)

■ Objectives

- investigate the forecasting movement of the KOSPI using the time series data of various interrelated world stock market indices.

■ Input variables

Index	Description
Kospi Up	Today's Kospi index is higher than that of the day before
Kospi Down	Today's Kospi index is lower than that of the day before
Dow Jones Up	Today's Dow Jones Industrial Average index is higher than that of the day before
Dow Jones Down	Today's Dow Jones Industrial Average index is lower than that of the day before
Nikkei Up	Today's Nikkei225 index is higher than that of the day before
Nikkei Down	Today's Nikkei225 index is lower than that of the day before
SSE Up	Today's SSE Composite Index is higher than that of the day before
SSE Down	Today's SSE Composite Index is lower than that of the day before
TSEC Up	Today's TSEC weighted index is higher than that of the day before
TSEC Down	Today's TSEC weighted index is lower than that of the day before
Hang Seng Up	Today's Hang Seng index is higher than that of the day before
Hang Seng Down	Today's Hang Seng index is lower than that of the day before
FTSE Up	Today's FTSE100 index is higher than that of the day before
FTSE Down	Today's FTSE100 index is lower than that of the day before
CAC Up	Today's CAC 40 index is higher than that of the day before
CAC Down	Today's CAC 40 index is lower than that of the day before
DAX Up	Today's DAX index is higher than that of the day before
DAX Down	Today's DAX index is lower than that of the day before

Case Study – 주가 분석 (2/2)

■ Transactions (2006.1 ~ 2008.12)

Day	Variable	Day	Variables	ID
...
06.06.09	KOSPI Up	06.06.08	Dow Jones Up, Nikkei225 Down, SSE Up, TSEC Down, Hang Seng Down, FTSE Up, CAC Up, DAX Up	101
06.06.10	KOSPI Up	06.06.09	Dow Jones Down, Nikkei225 Down, SSE Down, TSEC Down, Hang Seng Down, FTSE Up, CAC Down, DAX Up	102
06.06.11	KOSPI Down	06.06.10	Dow Jones Up, Nikkei225 Down, SSE Up, TSEC Down, Hang Seng Down, FTSE Up, CAC Up, DAX Down	103
...

■ Results of ARM

Rule	Condition
Rule 1	If Nikkei225 index is Down and Dow Jones and DAX indices are up, then KOSPI index is Up
Rule 2	If Hang Seng index is Down and Dow Jones and DAX indices are up, then KOSPI index is Up
Rule 3	If Hang Seng index is Down and Dow Jones and FTSE indices are up, then KOSPI index is Up
Rule 4	If Hang Seng index is Down and Dow Jones and CAC indices are up, then KOSPI index is Up

연관규칙탐사 알고리즘

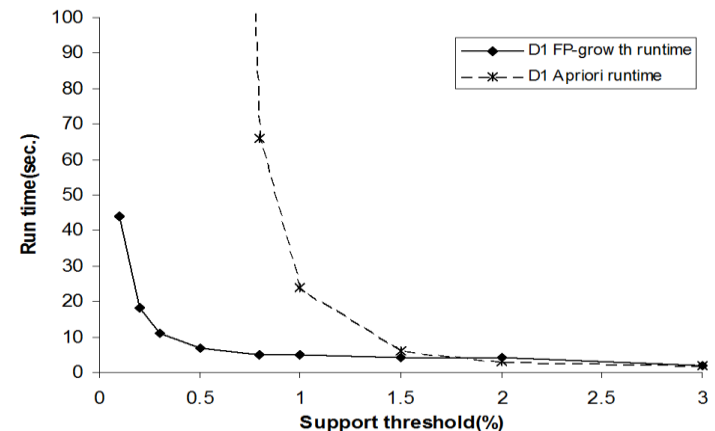
Apriori

- ✓ 최소 규칙 지지도(Support), 최대 규칙 신뢰도(Confidence), 최대 전항값 수(Antecedent)로 규칙 생성
- ✓ 품목필드가 이분형(flag) 또는 범주형(set)인 경우에 적용 가능
- ✓ 결합(join)과 가지치기(prune)의 두 과정으로 구성
- ✓ 아이템의 수에 따라 런타임이 기하급수적으로 증가

FP-Growth

- ✓ Apriori의 단점인 DB 스캔 횟수를 2회로 제한
- ✓ Candidate itemset을 만들지 않고 FP-tree라는 자료구조를 사용
- ✓ 트랜잭션과 아이템의 수에 따라 런타임이 선형적으로 증가

File	Apriori	FP-Growth
Simple Market Basket test file	3.66 s	3.03 s
"Real" test file (1 Mb)	8.87 s	3.25 s
"Real" test file (20 Mb)	34 m	5.07 s
Whole "real" test file (86 Mb)	4+ hours (Never finished, crashed)	8.82 s





Apriori 알고리즘 (1/3)

■ 빈발항목집합(frequent itemset)

- 최소지지도 이상을 갖는 항목집합
- 트랜잭션에 나타나는 모든 항목들의 집합을 $I = \{i_1, i_2, \dots, i_m\}$ 라 할 때, 모든 가능한 부분집합의 수는 $2^m - 1$ (공집합 제외)
- k개의 항목으로 이루어진 집합을 k-항목집합이라 함
- 원시적으로 연관규칙을 찾기 위해서는 모든 가능한 부분집합에 대해 전체 트랜잭션에 대한 지지도를 계산해야 함

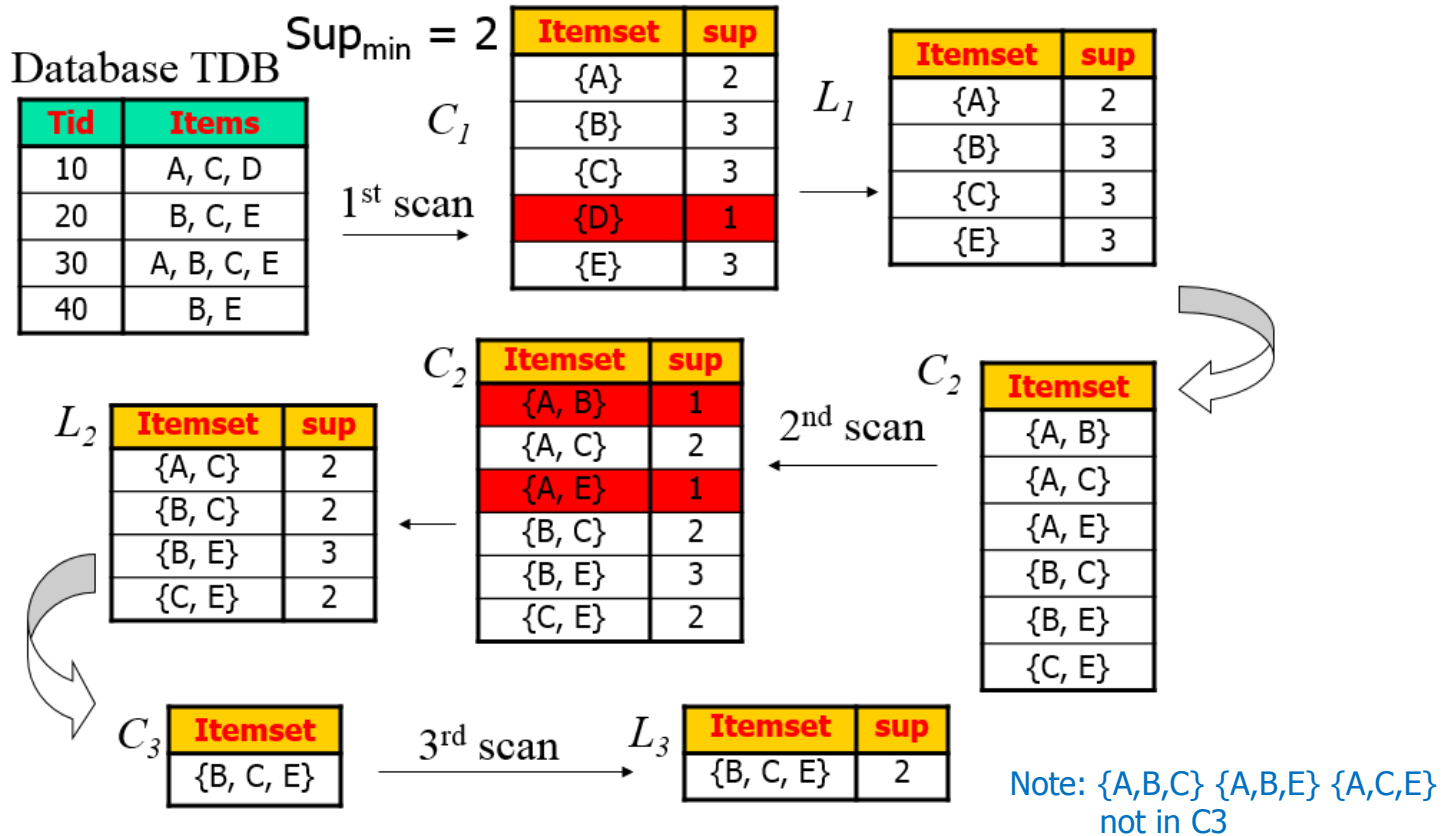
■ 선형적 규칙(Apriori Principle)

- 모든 항목집합에 대한 지지도를 계산하지 않고 원하는 빈발항목집합을 찾아내는데 이용되는 선형적 규칙:
 - 한 항목집합이 빈발하다면, 이 항목집합의 모든 부분집합 역시 빈발항목집합
 - 한 항목집합이 非빈발하다면, 이 항목집합의 모든 부분집합 역시 非빈발항목집합
- 이 사실을 이용하면 최소 지지도 기준을 넘지 못하는 항목집합들을 쉽게 가지치기 할 수 있는데, 이를 선형적 규칙을 이용한 빈발항목집합 추출 알고리즘(Apriori algorithm)이라 함

Apriori 알고리즘 (2/3)

■ 빈발항목집합 탐사과정

- 결합(join)과 가지치기(prune)의 두 과정으로 구성



Apriori 알고리즘 (3/3)

■ 연관규칙 추출과정

- 모든 빈발항목집합 L 에 대하여 L 의 모든 공집합이 아닌 부분집합들을 탐색
- 각각의 부분집합 A 에 대하여,
만약 $\text{Support}(A)$ 에 대한 $\text{Support}(L)$ 의 비율이 적어도 최소 신뢰도 이상이면 $A \Rightarrow (L-A)$ 의 형태의 규칙을 출력

$n(L)$	L	Rules ($\text{Conf}_{\min} = 0.7$)
2	$\{A, C\}$	$A \Rightarrow C (2/2)$; $C \Rightarrow A (2/3)$
	$\{B, C\}$	$B \Rightarrow C (2/3)$; $C \Rightarrow B (2/3)$
	$\{B, E\}$	$B \Rightarrow E (3/3)$; $E \Rightarrow B (3/3)$
	$\{C, E\}$	$C \Rightarrow E (2/3)$; $E \Rightarrow C (2/3)$
3	$\{B, C, E\}$	$B \Rightarrow C, E (2/3)$; $C \Rightarrow B, E (2/3)$; $E \Rightarrow B, C (2/3)$, $B, C \Rightarrow E (2/2)$; $B, E \Rightarrow C (2/3)$; $C, E \Rightarrow B (2/2)$