

비즈니스 인텔리전스를 위한
데이터마이닝

제**11**장 연관성규칙
(Association Rule)

INDEX

- 시장 바구니 분석이란?
- 연관성 규칙 적용분야
- 시장 바구니 분석 기본 개념
- 연관성 분석 측량화 방법
- Apriori 알고리즘
- 장단점
- 사용예제

시장 바구니 분석 (연관성 규칙 발견) 이란 ?

- 하나의 거래나 사건에 포함되어 있는 항목들의 경향을 파악해서 상호 연관성을 발견 하는 것

EX) Products in Shop Cart (One trip, Together)



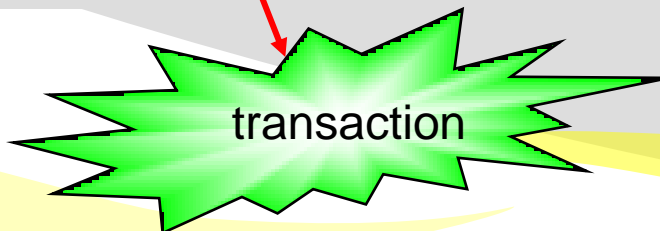
- 1) 구매자가 제품을 구매할 때 이웃의 영향이 있었는가?
- 2) 오렌지 주스와 청정재 구입시 윈도우 클리너를 같이 구입하는가?
- 3) 우유를 바나나 구입시 함께 구입하는가? 또한 구입 할 때 특정 브랜드를 구입 하는가?
- 4) 청정재를 어느 곳에 위치시켜야지만 판매고를 최대화하는가?

거래(transaction)와 항목(item)

- Market Basket Analysis는 하나 또는 여러 개의 product 나 service offering 의 거래와 이 거래에 대한 정보에서 시작.

Ex) 마크로의 Point-Of-Sale Transaction

customer	Set of products
1	오렌지 주스, 바나나
2	오렌지 주스, 우유
3	청정제, Window Cleaner



연관성 규칙 (Association Rule)

- 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

연관성 규칙 적용분야

- 교차 판매 (Cross Selling)
- 상품 진열 (Inventory Display)
- 부정탐지(fraud detection)
 - 상당히 높은 신뢰도를 갖는 규칙에 대해 특정 고객이 그 규칙이 적용이 안된다면 수상할 수 있음
- Catalog Design
 - 상품의 배치문제, 패키지 상품의 구성, 쿠폰 발행, 카탈로그의 구성, 신상품의 카테고리 선정

규칙의 활용 방법

- 전건의 내용을 중심으로 한 활용
 - ․ 전건에 포함된 특정 품목들만을 포함하는 규칙을 모아 제시함
 - ․ 못과 망치의 관계/슈퍼에서 담배를 파는 이유
- 후건의 내용을 중심으로 한 활용
 - ․ 후건의 내용에 관련이 있는 정보를 제공함
 - ․ 품목의 진열, **Cross Selling**
- 정확도(신뢰도)에 근거한 활용
 - ․ 단일 거래 규모가 큰 경우
 - ․ 금융 시장의 예
- 발생(적용)빈도에 근거한 활용
 - ․ 가장 대표적인, 쉽게 적용한 규칙 발견
- 가치에 근거한 활용
 - ․ 정확도가 높을수록, 적용빈도가 높을수록, 통상적인 상식의 틀을 벗어날수록 가치가 높음

연관성 규칙 결과 유형

- **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.

시장 바구니 분석의 기본 개념

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

Co-occurrence of Product(횟수)

	소주	맥주	콜라	주스	와인
소주	4	2	3	1	2
맥주	2	3	3	0	1
콜라	3	3	4	0	2
주스	1	0	0	2	0
와인	2	1	2	0	2

시장 바구니 분석의 기본 개념

1. 단순 패턴의 발견

- 소주와 콜라 , 맥주와 콜라가 다른 combination보다 많이 발생
- 주스는 맥주, 콜라, 와인 과는 결코 함께 구매되지 않는다.

연관성 규칙 발견

2. 연관성 규칙의 예

- 맥주를 구입한 사람들 모두는 콜라도 구매한다.

위에서 제시된 연관성 규칙은 얼마나 유용할까?

이 질문을 해결하기 위해 수치적으로 나타내는 것이 필요하고 ,
이 수치적인 계산에는 확률을 사용한다.

연관성 규칙 측량화 방법

- 지지도 (Support), 발생빈도

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

$$S_{X \rightarrow Y}(X \rightarrow Y) = \frac{\text{품목X와 품목Y를 포함하는 거래 수}}{\text{전체 거래 수}(N)}$$

- 전체적 구매도에 대한 경향을 파악
- Reflexive(재귀 법칙) :

신뢰도(Confidence)

- 항목 X를 구입한 사람이 Y를 구입할 확률
- 항목 X를 포함하는 거래 중에서 항목 Y가 포함될 확률은 어느 정도인가 ?

$$C \approx P(Y | X) \approx \frac{P(X \wedge Y)}{P(X)}$$

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

- 조건부 확률
- 연관성의 정도
- not symmetric

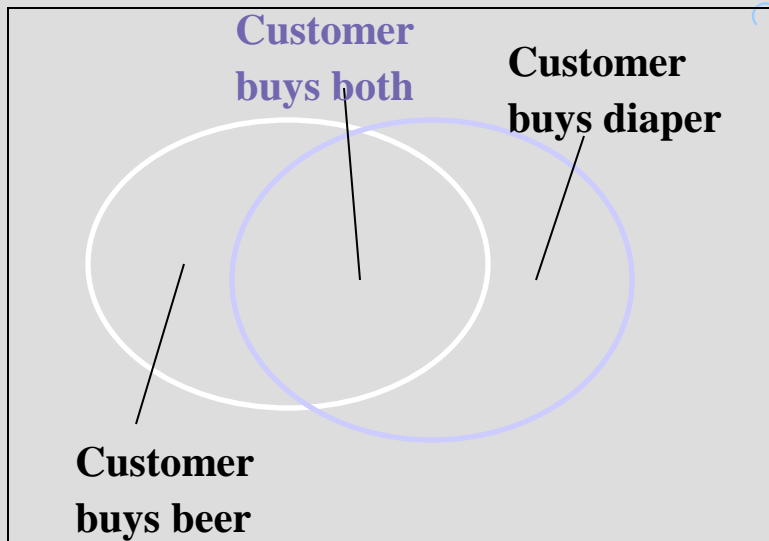
리프트 (Lift / improvement)

- 항목 X를 구매한 경우 그 거래가 항목 Y를 포함하는 경우와 항목 Y가 임의로 구매되는 경우의 비는 ?

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

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

Rule Measures: Support and Confidence



Find all the rules $X \& Y \Rightarrow Z$ with minimum confidence and support

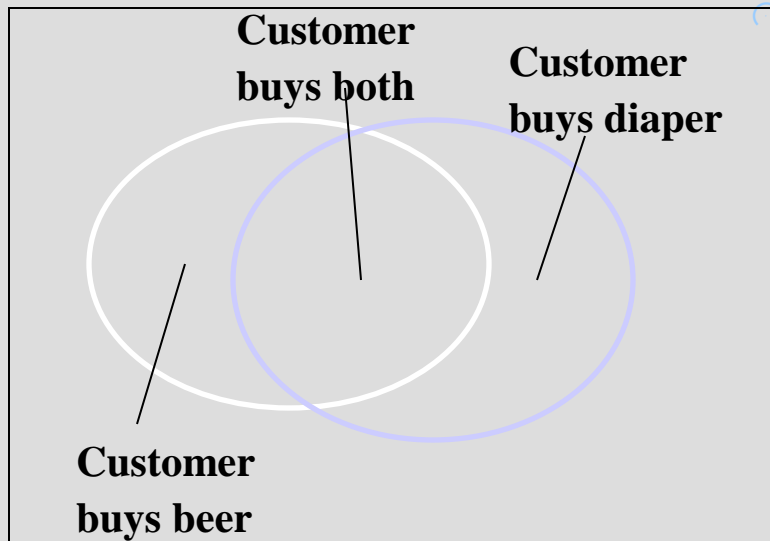
- | support, s , probability that a transaction contains $\{X \cap Y \cap Z\}$
- | confidence, c , conditional probability that a transaction having $\{X \cap Y\}$ also contains Z

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Let minimum support 50%, and minimum confidence 50%, we have

- $A \Rightarrow C$ (?, ?)
- $C \Rightarrow A$ (?, ?)

Rule Measures: Support and Confidence



Find all the rules $X \& Y \Rightarrow Z$ with minimum confidence and support

- | support, s , probability that a transaction contains $\{X \cap Y \cap Z\}$
- | confidence, c , conditional probability that a transaction having $\{X \cap Y\}$ also contains Z

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Let minimum support 50%, and minimum confidence 50%, we have

- $A \Rightarrow C$ (50%, 66.6%)
- $C \Rightarrow A$ (50%, 100%)

고려 사항

1. 신뢰도의 값이 크면 좋지만 신뢰도가 크다고 최선의 연관성 규칙이라고 볼 수는 없다.

- 두 항목의 기본적인 구매율이 어느 정도 수준이 되어야만 의미가 있다. 즉, 지지도가 일정 수준에 도달 해야만 한다.

2. 신뢰도와 지지도는 자주 구매되는 항목에 대해서는 연관성 때문이 아니라 우연하게 높게 나올 수도 있다

- Lift를 본다.

3. 신뢰도가 높을 경우에는 $X \rightarrow Y$ 에서 항목 Y의 확률이 커야지 이 연관성 규칙에 의미가 있다.

- Lift 값이 1보다 커야 유용한 정보가 된다.

The Basic Steps in Market Basket Analysis

choosing the right set of item and right level
- taxonomy(관리도)를 이용



co-occurrence matrix 작성과
확률 (지지도, 신뢰도, Lift) 계산



확률 분석과 유용한 연관성 규칙 결정

장 단점

● 장점

1. 결과가 명확하고 이해하기 쉽다.
2. 자료구조와 계산과정이 간단하다.

● 단점

1. 항목의 수를 결정하기가 어렵다.
2. 드물게 발생하는 항목에 대해서 처리가 어렵다.
3. DBMS등과 같은 전산화 작업이 없을 시는 동일한 거래를 추적하기가 힘들다.
4. 항목의 수가 증가하면 계산시간이 급격히 증가한다.

시장 바구니 분석 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 %

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

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

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

Sequences Association Rules Discovery

$A \longrightarrow B$: A라는 사건이 발생한 후 B가 발생

예제 : 새 컴퓨터를 구입한 사람 중 25%는 그 다음날에 레이저 프린터를 구입할 것이다.

Dissociation Rules Discovery

If $\sim A$ and $\sim B$ then $\sim C$

If $\sim A$ and $\sim B$ then C

If $\sim A$ and B then $\sim C$

If $\sim A$ and B then C

If A and $\sim B$ then $\sim C$

If A and $\sim B$ then C

Apriori 알고리즘

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Min. support 50%
Min. confidence 50%

Frequent Itemset	Support
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

For rule $A \Rightarrow C$:

$$\text{support} = \text{support}(\{A \cap C\}) = 50\%$$

$$\text{confidence} = \text{support}(\{A \cap C\}) / \text{support}(\{A\}) = 66.6\%$$

The Apriori principle:

Any subset of a frequent itemset must be frequent

Apriori 알고리즘

- Find the *frequent itemsets*: the sets of items that have minimum support
 - A subset of a frequent itemset must also be a frequent itemset
 - i.e., if $\{AB\}$ is a frequent itemset, both $\{A\}$ and $\{B\}$ should be a frequent itemset
 - Iteratively find frequent itemsets with cardinality from 1 to k (k -itemset)
- Use the frequent itemsets to generate association rules.

Apriori 알고리즘

- Join Step: C_k is generated by joining L_{k-1} with itself
- Prune Step: Any $(k-1)$ -itemset that is not frequent cannot be a subset of a frequent k -itemset
- Pseudo-code:

C_k : Candidate itemset of size k

L_k : frequent itemset of size k

$L_1 = \{\text{frequent items}\};$

for ($k = 1; L_k \neq \square; k++$) **do begin**

C_{k+1} = candidates generated from L_k ;

for each transaction t in database **do**

increment the count of all candidates in C_{k+1}

that are contained in t

L_{k+1} = candidates in C_{k+1} with min_support

end

return $\bigcup_k L_k$;

The Apriori Algorithm — Example

Database D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Scan D

C_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

L_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3

C_2

itemset	sup
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

Scan D

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

L_2

itemset	sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_3

itemset
{2 3 5}

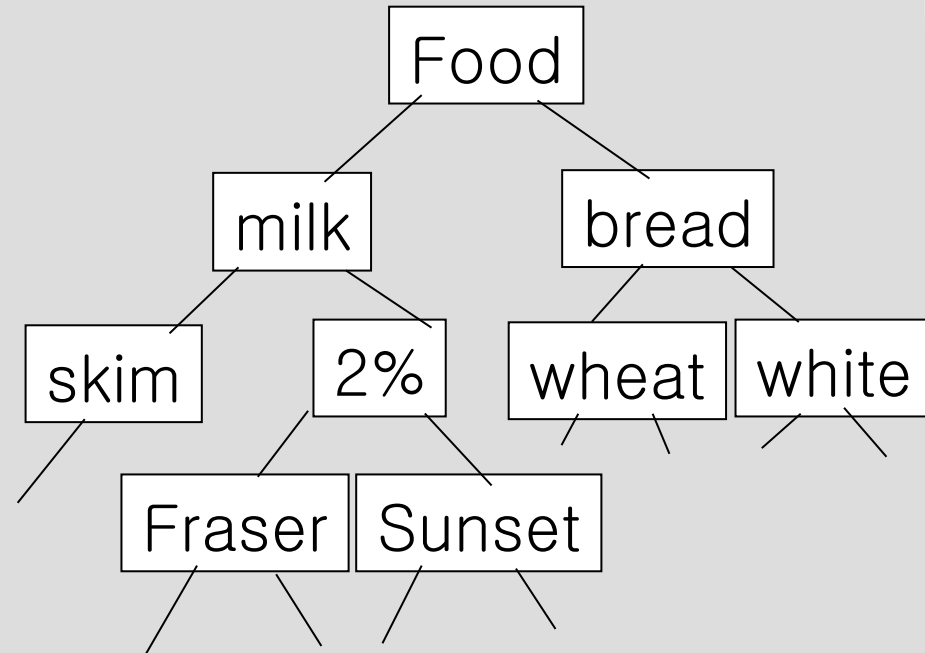
Scan D

L_3

itemset	sup
{2 3 5}	2





Multiple-Level Association Rules

- Items often form hierarchy.
- Items at the lower level are expected to have lower support.
- Rules regarding itemsets at appropriate levels could be quite useful.
- Transaction database can be encoded based on dimensions and levels
- We can explore shared multi-level mining



TID	Items
T1	{111, 121, 211, 221}
T2	{111, 211, 222, 323}
T3	{112, 122, 221, 411}
T4	{111, 121}
T5	{111, 122, 211, 221, 413}

Mining Multi-Level Associations

- A top_down, progressive deepening approach:
 - | First find high-level strong rules:
milk  bread [20%, 60%].
 - | Then find their lower-level “weaker” rules:
2% milk  wheat bread [6%, 50%].
- Variations at mining multiple-level association rules.
 - | Level-crossed association rules:
2% milk  *Wonder* wheat bread
 - | Association rules with multiple, alternative hierarchies:
2% milk  *Wonder* bread

Multi-level Association: Uniform Support vs. Reduced Support

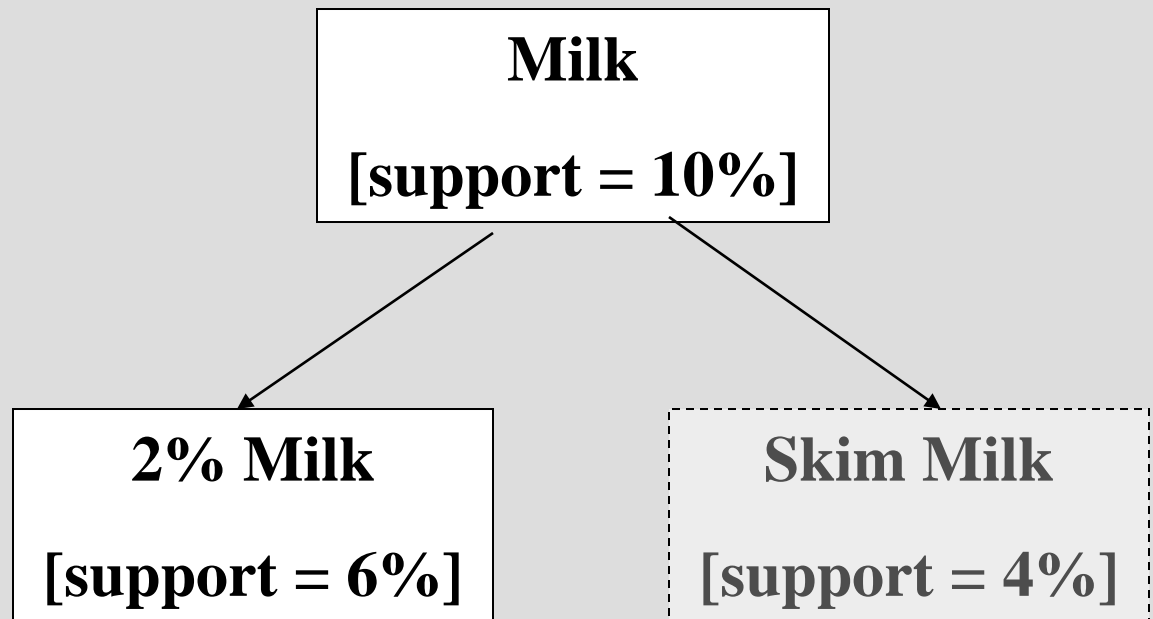
- Uniform Support: the same minimum support for all levels
 - + One minimum support threshold. No need to examine itemsets containing any item whose ancestors do not have minimum support.
 - Lower level items do not occur as frequently. If support threshold
 - too high \square miss low level associations
 - too low \square generate too many high level associations
- Reduced Support: reduced minimum support at lower levels
 - There are 4 search strategies:
 - Level-by-level independent
 - Level-cross filtering by k-itemset
 - Level-cross filtering by single item
 - Controlled level-cross filtering by single item

Uniform Support

Multi-level mining with uniform support

Level 1
min_sup = 5%

Level 2
min_sup = 5%

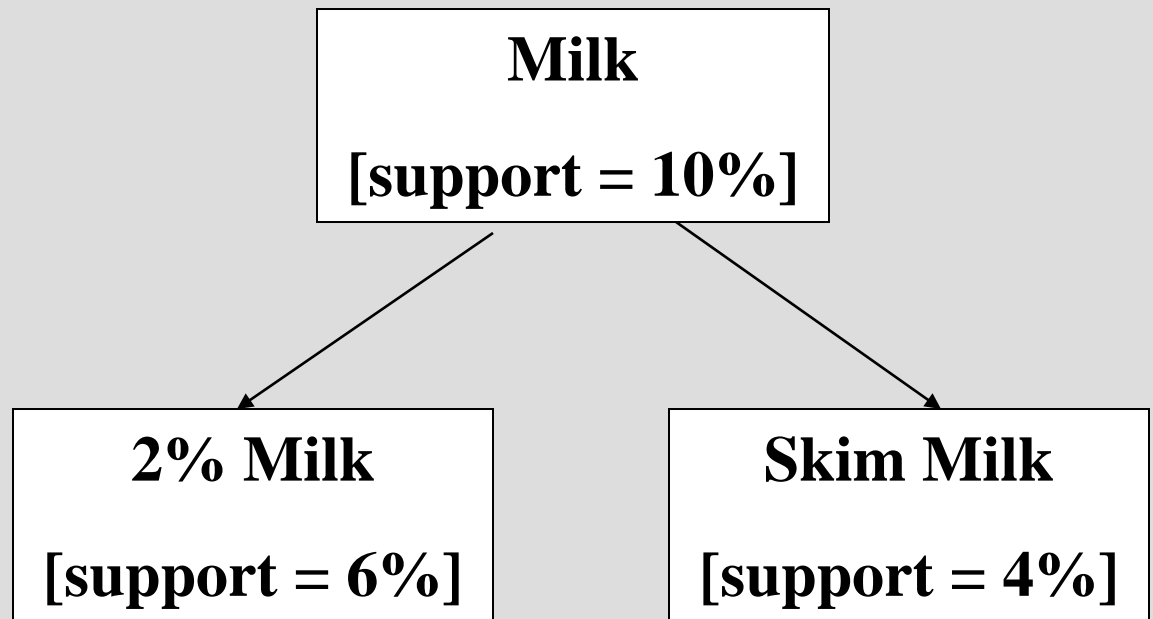


Reduced Support

Multi-level mining with reduced support

Level 1
min_sup = 5%

Level 2
min_sup = 3%



Multi-level Association: Redundancy Filtering

- Some rules may be redundant due to “ancestor” relationships between items.
- Example
 - | milk \square wheat bread [support = 8%, confidence = 70%]
 - | 2% milk \square wheat bread [support = 2%, confidence = 72%]
- We say the first rule is an ancestor of the second rule.
- A rule is redundant if its support is close to the “expected” value, based on the rule’s ancestor.

Multi-Dimensional Association: Concepts

- Single-dimensional rules:
 $\text{buys}(X, \text{"milk"}) \sqcap \text{buys}(X, \text{"bread"})$
- Multi-dimensional rules: 2 dimensions or predicates
 - Inter-dimension association rules (*no repeated predicates*)
 $\text{age}(X, \text{"19-25"}) \sqcap \text{occupation}(X, \text{"student"}) \sqcap \text{buys}(X, \text{"coke"})$
 - hybrid-dimension association rules (*repeated predicates*)
 $\text{age}(X, \text{"19-25"}) \sqcap \text{buys}(X, \text{"popcorn"}) \sqcap \text{buys}(X, \text{"coke"})$
- Categorical Attributes
 - finite number of possible values, no ordering among values
- Quantitative Attributes
 - numeric, implicit ordering among values