

Chapter 3

Association Mining

Associate Professor Yachai Limpiyakorn, Ph.D.

Basic Concepts

TID	Produce
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CEREAL
4	MILK, BREAD, SUGAR
5	MILK, CEREAL
6	BREAD, CEREAL
7	MILK, CEREAL
8	MILK, BREAD, CEREAL, EGGS
9	MILK, BREAD, CEREAL

- Given:
 - (1) database of transactions/ transactional database
 - (2) each transaction is a list of items purchased

- Find:

ความสัมพันธ์ที่น่าสนใจระหว่างไอเทมเซต (itemset) ในชุดข้อมูล ความสัมพันธ์ที่ได้เขียนอยู่ในรูปกฎความสัมพันธ์ (Association Rule) ของเซตของไอเทมที่เป็นเหตุ (Antecedent) ไปสู่เซตของไอเทมที่เป็นผล (Consequent)

{Cheese, Milk} → Bread [S=5%, C=80%]

80% of customers who buy cheese and milk also buy bread and 5% of customers buy all these products together




How can association rules be used?

Stories – Beer and Diapers

- ♦ Diapers and Beer. Most famous example of market basket analysis for the last few years. If you buy diapers, you tend to buy beer.
- T. Blischok headed Terradata's Industry Consulting group.
- K. Heath ran self joins in SQL (1990), trying to find two itemsets that have baby items, which are particularly profitable.
- Found this pattern in their data of 50 stores/90 day period.
- Unlikely to be significant, but it's a nice example that explains associations well.

Ronny Kohavi ICML 1998



Probably mom was calling dad at work to buy diapers on way home and he decided to buy a six-pack as well.

The retailer could move diapers and beers to separate places and position high-profit items of interest to young fathers along the path.

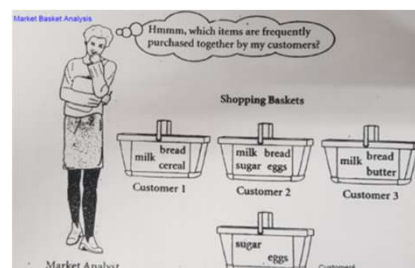
2110773-3 2/67

3

Application₁

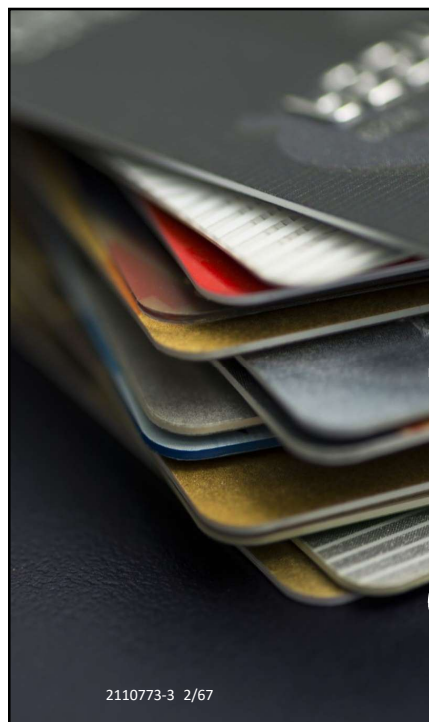
ส่วนใหญ่มักประยุกต์ใช้เทคนิคการทำเหมืองความสัมพันธ์กับการวิเคราะห์ทางการตลาด (Market Basket Analysis: MBA) ซึ่งเป็นรูปแบบการจัดกลุ่ม (Clustering) แบบหนึ่ง ที่ใช้เพื่อหากลุ่มสิ่งของที่น่าจะปรากฏร่วมกันในทรานแซกชันหนึ่งๆ มักเป็นทรานแซกชัน ณ จุดขาย (point-of-sale) ผลลัพธ์หรือแบบจำลองที่ได้สามารถแสดงได้ด้วยกฎซึ่งบอกความเป็นไปได้ของการซื้อผลิตภัณฑ์ต่างๆร่วมกัน การวิเคราะห์ทางการตลาดมีบทบาทสำคัญต่ออุตสาหกรรมการค้าปลีก (Retail industry) เพื่อให้ทราบถึงพฤติกรรมการซื้อสินค้าของลูกค้าซึ่งเป็นประโยชน์ในการ

- ♦ จัดพื้นที่ร้านค้า (Store layout)
- ♦ ทำตลาดเพื่อส่งเสริมการขายสินค้าหรือบริการซึ่งกันและกัน (Cross-marketing)
- ♦ ออกแบบหนังสือแคตตาล็อกสินค้า (Catalog design)
- ♦ วางแผนการส่งเสริมการขายและการตั้งราคาผลิตภัณฑ์ (Product pricing and promotion)



2110773-3 2/67

4



Application₂

นอกจากนี้ สามารถประยุกต์ใช้การวิเคราะห์ทางการตลาดกับกิจกรรมใกล้เคียงที่ลูกค้ามักกระทำด้วยกัน เพื่อก่อให้เกิดรายได้สูงสุดจากการจัดประเภทผลิตภัณฑ์หรือบริการเข้าด้วยกัน ได้แก่

- ◆ การใช้จ่ายผ่านบัตรเครดิตของลูกค้าในการเข้าพักโรงแรม เช่ารถ ทำให้สามารถทำนายค่าใช้จ่ายต่อไปของลูกค้า
- ◆ แพ็กเกจการให้บริการการสื่อสารโทรคมนาคม เพื่อก่อให้เกิดรายได้สูงสุด
- ◆ การให้บริการทางธนาคารที่ลูกค้ามักซื้อด้วยกัน เพื่อก่อให้เกิดประโยชน์สูงสุด เช่น ประเภอบัญชีที่ลูกค้ามักเปิดด้วยกัน (account bundle) การให้บริการการลงทุนครบวงจร และแพ็กเกจสินเชื่อการซื้อรถ เป็นต้น

2110773-3 2/67

5

นิยามพื้นฐาน การทำเหมือง ความสัมพันธ์

- ◆ ไอเทมเซต (itemset - I) คือเซตที่มีไอเทมทั้งหมดเป็นสมาชิก ซึ่งไอเทมในที่นี้อาจเป็นชื่อสินค้า หรือชื่อใดๆ ที่เป็นหน่วยพื้นฐานที่จะนำมาทำการเรียนรู้
- ◆ ทรานแซกชัน (transaction - T) เป็นเซตของไอเทม โดยที่ $T \subseteq I$
- ◆ เซตข้อมูล (data set - D) คือเซตที่มีทรานแซกชันทุกตัวเป็นสมาชิก
เรากล่าวว่าทรานแซกชัน T บรรจุไอเทมเซตย่อย X ก็ต่อเมื่อ $X \subseteq T$
เพราะฉะนั้นจึงนิยามกฎความสัมพันธ์ได้ว่า
- ◆ กฎความสัมพันธ์ (Association Rule) คือการอุปนัยในรูปแบบ $X \rightarrow Y$ เมื่อ $X \subset I, Y \subset I$ และ $X \cap Y = \emptyset$

2110773-3 2/67

6

Objective measures of rule interest

- Support
- Confidence or strength
- Lift or Interest or Correlation
- Conviction
- Leverage or Piatetsky-Shapiro
- Coverage

Association Rule

- Rule form

Antecedent \rightarrow Consequent [*support*, *confidence*]

Note: *support* and *confidence* are user defined measures of interestingness

- Examples

$\text{buys}(x, \text{"computer"}) \rightarrow \text{buys}(x, \text{"financial management software"})$ [0.5%, 60%]

$\text{age}(x, \text{"30..39"}) \wedge \text{income}(x, \text{"42..48K"}) \rightarrow \text{buys}(x, \text{"car"})$ [1%, 75%]

Rule basic Measures: Support and Confidence

$$A \Rightarrow B [s, c]$$

Support: denotes the frequency of the rule within transactions. A high value means that the rule involve a great part of database.

$$\text{support}(A \Rightarrow B) = p(A \cup B)$$

Confidence: denotes the percentage of transactions containing A which contain also B. It is an estimation of conditioned probability .

$$\text{confidence}(A \Rightarrow B [s, c]) = p(B|A) = \text{sup}(A,B)/\text{sup}(A)$$

Calculation of Support and Confidence

• Support

คำนวณค่าสนับสนุน ได้จากจำนวนทรานแซกชันที่มีรายการ X และ Y เกิดร่วมกันหารด้วยจำนวนทรานแซกชันทั้งหมด

$$\begin{aligned}\text{support}(X \rightarrow Y) &= P(X \cup Y) \\ &= \text{tran_count}(X \cup Y) / \text{tran_count}(D)\end{aligned}$$

• Confidence

คำนวณค่าความเชื่อมั่นได้จากจำนวน ทรานแซกชันที่มีรายการ X และ Y เกิดร่วมกันหารด้วยจำนวนทรานแซกชันที่มีรายการ X

$$\begin{aligned}\text{confidence}(X \rightarrow Y) &= P(Y|X) \\ &= \text{tran_count}(X \cup Y) / \text{tran_count}(X)\end{aligned}$$

Practice Calculating Support and Confidence

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

ก. ให้คำนวณหาค่า support และ confidence ของ
ความสัมพันธ์ $A \rightarrow C$ และ $C \rightarrow A$

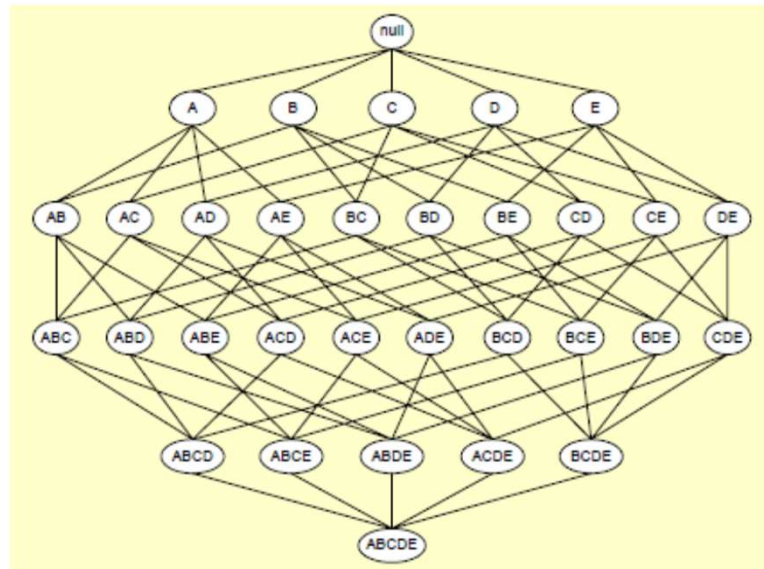
ข. กำหนดให้ minimum support = 50% และ
minimum confidence = 80% อยากทราบว่า
ความสัมพันธ์ $A \rightarrow C$ และ $C \rightarrow A$ ความสัมพันธ์ใด
เป็นกฎความสัมพันธ์

Association Mining

เป็นปัญหาการค้นหากฎความสัมพันธ์ นิยามได้ดังนี้

- **การค้นหากฎความสัมพันธ์** คือ การหาความสัมพันธ์ทั้งหมดในทรานแซกชันทุกตัวของเซตข้อมูลที่กำหนดให้ โดยกฎความสัมพันธ์ที่หาได้ทั้งหมดจะต้องมีค่าสนับสนุน (support) ไม่ต่ำกว่าค่าสนับสนุนน้อยสุด (minimum support) ที่ผู้ใช้กำหนดไว้ และมีค่าความเชื่อมั่น (confidence) ไม่ต่ำกว่าค่าความเชื่อมั่นน้อยสุด (minimum confidence) ที่ผู้ใช้ได้กำหนดไว้
- การค้นหากฎความสัมพันธ์สามารถแบ่งย่อยได้เป็นสองขั้นตอน คือ
 1. ค้นหาเซตของไอเทมปรากฏบ่อย (frequent itemset) หรือไอเทมเซตที่มีค่าสนับสนุนไม่ต่ำกว่าค่าสนับสนุนน้อยสุดที่กำหนดให้
 2. นำไอเทมเซตปรากฏบ่อยเหล่านั้นมาสร้างเป็นกฎความสัมพันธ์ต่อไป

Itemset Lattice



2110773-3 2/67

13

Apriori Principle

Any subset of a frequent itemset must also be frequent

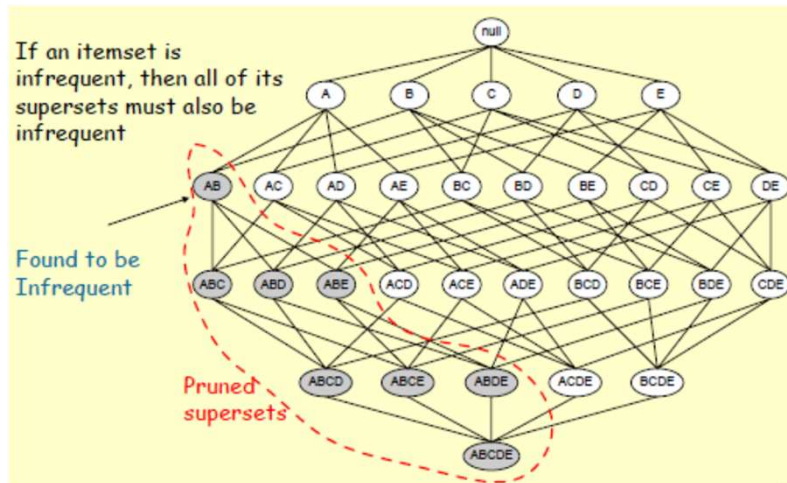
No superset of any infrequent itemset should be generated or tested

- Many item combinations can be pruned

2110773-3 2/67

14

Apriori Principle for Pruning Candidates



2110773-3 2/67

15

Association Mining: 2 key steps

1. Find all Frequent Itemsets: the sets of items that pass minimum support
 - ❖ Apriori Algorithm
 - มีการจัดเรียงลำดับของไอเทมในแต่ละทรานแซกชันก่อนประมวลผล
 - การสร้างไอเทมเซตจะสร้างตามระดับชั้น จากชั้นที่ k , $k+1$, $k+2$
 - ใช้ความรู้ก่อนหน้าคือคุณสมบัติของไอเทมเซตเกิดบ่อยในการตัดเล็ม
2. For every frequent itemset X , generate all non-empty subset S of X

$$S \rightarrow (X-S)$$

Output the rule $S \rightarrow (X-S)$
If confidence $\geq \text{min_confidence}$

2110773-3 2/67

16

Apriori Algorithm

Algorithm: Apriori. Find frequent itemsets using an iterative level-wise approach based on candidate generation.

Input:

- D , a database of transactions;
- min_sup , the minimum support count threshold.

Output: L , frequent itemsets in D .

Method:

```

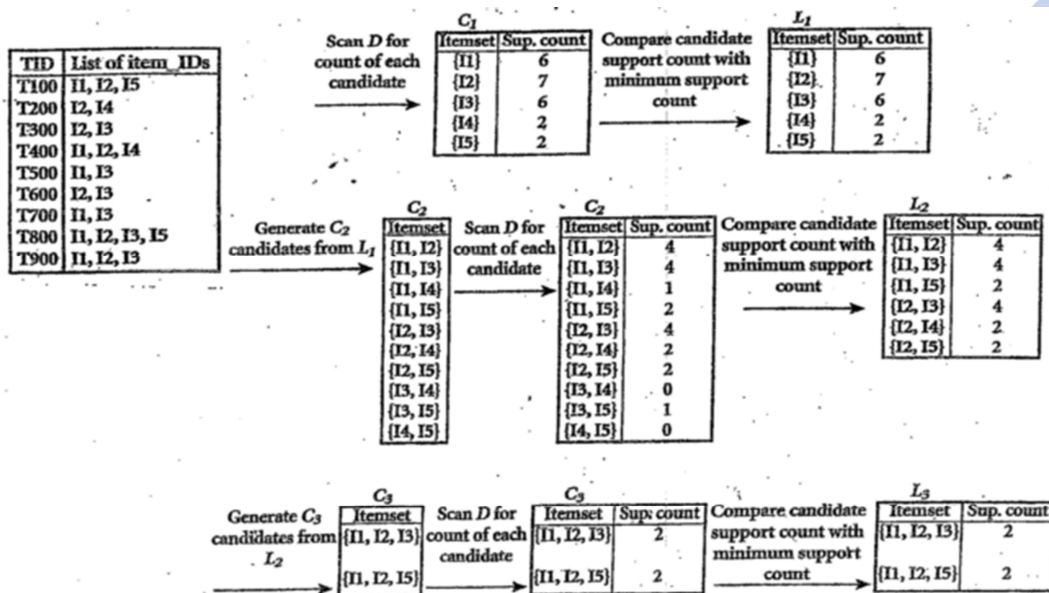
(1)  $L_1 = \text{find\_frequent\_1-itemsets}(D)$ ;
(2) for  $(k = 2; L_{k-1} \neq \emptyset; k++)$  {
(3)    $C_k = \text{apriori\_gen}(L_{k-1})$ ;
(4)   for each transaction  $t \in D$  { // scan  $D$  for counts
(5)      $C_t = \text{subset}(C_k, t)$ ; // get the subsets of  $t$  that are candidates
(6)     for each candidate  $c \in C_t$ 
(7)        $c.\text{count}++$ ;
(8)   }
(9)    $L_k = \{c \in C_k | c.\text{count} \geq min\_sup\}$ 
(10) }
(11) return  $L = \cup_k L_k$ ;

procedure apriori_gen( $L_{k-1}$ : frequent  $(k-1)$ -itemsets)
(1) for each itemset  $l_1 \in L_{k-1}$ 
(2)   for each itemset  $l_2 \in L_{k-1}$ 
(3)     if  $(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge \dots \wedge (l_1[k-2] = l_2[k-2]) \wedge (l_1[k-1] < l_2[k-1])$  then {
(4)        $c = l_1 \bowtie l_2$ ; // join step: generate candidates
(5)       if  $\text{has\_infrequent\_subset}(c, L_{k-1})$  then
(6)         delete  $c$ ; // prune step: remove unfruitful candidate
(7)       else add  $c$  to  $C_k$ ;
(8)     }
(9) return  $C_k$ ;

procedure has_infrequent_subset( $c$ : candidate  $k$ -itemset;
                                $L_{k-1}$ : frequent  $(k-1)$ -itemsets; // use prior knowledge
(1) for each  $(k-1)$ -subset  $s$  of  $c$ 
(2)   if  $s \notin L_{k-1}$  then
(3)     return TRUE;
(4) return FALSE;
  
```

2110773-3 2/67

17



2110773-3 2/67

18

การสร้างกฎความสัมพันธ์จากเซตของไอเทมปรากฏบ่อย

เมื่อได้ไอเทมเซตปรากฏบ่อยแล้ว จำเป็นต้องหาความสัมพันธ์จากไอเทมเซตปรากฏบ่อยนั้น โดยกฎความสัมพันธ์ที่ได้จะต้องมีความเชื่อมั่นไม่ต่ำกว่าค่าความเชื่อมั่นน้อยสุดที่กำหนดให้ เรียกกฎความสัมพันธ์ดังกล่าวว่า **Strong Association Rules**

ภายหลังจากที่ได้ไอเทมเซตปรากฏบ่อยทั้งหมดแล้ว จะสร้างเซตกฎความสัมพันธ์จากแต่ละไอเทมเซตปรากฏบ่อย I โดยสร้างทุกเซตย่อยที่ไม่ว่างของ I กล่าวคือ ทุกเซตย่อยที่ไม่ว่าง S ของ I แสดงกฎความสัมพันธ์ $S \rightarrow (I-S)$ ถ้าอัตราส่วนระหว่างจำนวนทรานแซกชันของ I ต่อจำนวนทรานแซกชันของ S ไม่น้อยกว่าค่าความเชื่อมั่นที่กำหนด

- ผลลัพธ์ไอเทมเซตปรากฏบ่อยที่ได้จากโจทย์ตัวอย่าง ประกอบด้วยสมาชิกในเซต $L2$ และ $L3$ ซึ่งจะถูกนำมาเขียนเป็นความสัมพันธ์ $X \rightarrow Y$ พร้อมคำนวณหาค่าความเชื่อมั่น (ในที่นี้ แสดงเพียงสมาชิก $\{I1, I2, I5\} \in L3$)

$I1, I2 \rightarrow I5$	[confidence = $2/4 = 50\%$]
$I1, I5 \rightarrow I2$	[confidence = $2/2 = 100\%$]
$I2, I5 \rightarrow I1$	[confidence = $2/2 = 100\%$]
$I1 \rightarrow I2, I5$	[confidence = $2/6 = 33\%$]
$I2 \rightarrow I1, I5$	[confidence = $2/7 = 29\%$]
$I5 \rightarrow I1, I2$	[confidence = $2/2 = 100\%$]

- กำหนดค่าความเชื่อมั่นต่ำสุดเท่ากับ 70% จะได้ว่าความสัมพันธ์ที่สร้างจาก $L3$ ที่เป็น Strong Association Rules ประกอบด้วย

$I1, I5 \rightarrow I2$; $I2, I5 \rightarrow I1$; $I5 \rightarrow I1, I2$

2110773-3 2/67

19

Improving Apriori

Challenge

- every pass goes over whole data
- multiple scans of transaction database
- huge number of candidates
- one transaction may contain many candidates
- tedious workload of support counting for candidates

General ideas for improvement

- shrink number of candidates
- facilitate support counting of candidates, e.g. hash tree
- Transaction reduction: A transaction that does not contain any frequent k -itemset is useless in subsequent scans

2110773-3 2/67

20

ข้อพิจารณาเกี่ยวกับค่าสนับสนุนและค่าความเชื่อมั่น

- ◆ การค้นหาความสัมพันธ์อาจล้มเหลว ถ้ากำหนดค่าสนับสนุนและค่าความเชื่อมั่นสูงเกินไป
- ◆ ถ้ากำหนดค่าสนับสนุนและค่าความเชื่อมั่นต่ำเกินไป อาจได้ความสัมพันธ์ระหว่างผลิตภัณฑ์หลากหลายเกินไปที่เราไม่ต้องการ
- ◆ กฎความสัมพันธ์ที่มีค่าสนับสนุนและค่าความเชื่อมั่นสูง แสดงระดับความเกี่ยวข้อง (degree of relevance) มากกว่ากฎความสัมพันธ์ที่มีค่าสนับสนุนและค่าความเชื่อมั่นต่ำ
- ◆ ค่าสนับสนุนแสดงถึงความถี่ของจำนวนทรานแซกชันของการเกิดร่วมกันของผลิตภัณฑ์ โดยทั่วไปมักจะให้น้ำหนักความสำคัญแก่ทรานแซกชันที่เกิดบ่อย แต่บางครั้งทรานแซกชันที่มีค่าสนับสนุนต่ำอาจเป็นประโยชน์ต่อการค้นหาความสัมพันธ์บางอย่าง
- ◆ ค่าความเชื่อมั่นเพียงอย่างเดียวไม่อาจบอกได้ว่าการเกิดร่วมกันของผลิตภัณฑ์ A และ B เป็นไปโดยบังเอิญหรือไม่ ซึ่งเราน่าจะสนใจความสัมพันธ์ระหว่างผลิตภัณฑ์ที่ไม่ได้เกิดขึ้นโดยความบังเอิญมากกว่า

2110773-3 2/67

21

Dependent Framework

- ◆ การทำเหมืองกฎความสัมพันธ์ โดยใช้ค่าสนับสนุน-ความเชื่อมั่นเป็นที่แพร่หลายในหลายแอปพลิเคชัน แต่ในบางครั้ง กรอบค่าสนับสนุน-ความเชื่อมั่นอาจทำให้ผู้ใช้เข้าใจผิดเกี่ยวกับความน่าสนใจของกฎที่ค้นพบ $A \rightarrow B$ เนื่องจากความจริงแล้วการเกิดเหตุการณ์ A ไม่ได้สื่อนัย (imply) การเกิดของเหตุการณ์ B ก่อให้เกิดคำถามว่า **Strong Association Rules** ที่ค้นพบน่าสนใจหรือไม่
- ◆ กรอบความขึ้นต่อกัน (Dependent Framework) สามารถใช้วัดความน่าสนใจของกฎที่ค้นพบในแง่ของค่าสหสัมพันธ์ (correlation) ของเหตุการณ์

2110773-3 2/67

22

Correlation/ Lift/ Interest

- Correlation/ Lift/ Interest

$$\begin{aligned}\text{Lift}(A \rightarrow B) &= P(B/A)/P(B) \\ &= \frac{P(A \cup B)}{P(A)P(B)}\end{aligned}$$

- $P(A \cup B) = P(B)*P(A)$, ถ้า A และ B เป็นเหตุการณ์อิสระต่อกัน
- ถ้าค่าสหสัมพันธ์มีค่าน้อยกว่า 1 แล้ว A และ B มีความสัมพันธ์เชิงลบ (negatively correlated) หรือในทิศทางตรงกันข้าม มิเช่นนั้น A และ B มีความสัมพันธ์เชิงบวก (positively correlated) หรือการเกิดขึ้นของเหตุการณ์หนึ่งมีผลต่อการเกิดของอีกเหตุการณ์ ตัวอย่างเช่น ค่าสหสัมพันธ์ของกฎ $A \Rightarrow B$ เท่ากับ 1.3 หมายความว่า การเกิดขึ้นของเหตุการณ์ A สามารถทำนายโอกาสที่ B จะปรากฏในทรานแซกชันเดียวกันได้แม่นยำกว่าเป็น 1.3 เท่าของความน่าจะเป็นที่ B จะเกิดขึ้นแบบสุ่ม

2110773-3 2/67

23

Example Calculation of Lift

X	1	1	1	1	0	0	0	0
Y	1	1	0	0	0	0	0	0
Z	0	1	1	1	1	1	1	1

Rule	Support	Confidence	Lift
$X \rightarrow Y$			
$X \rightarrow Z$			
$Y \rightarrow Z$			

2110773-3 2/67

24

Criticism to Support and Confidence

■ Example 1: (Aggarwal & Yu, PODS98)

- Among 5000 students
 - 3000 play basketball
 - 3750 eat cereal
 - 2000 both play basketball and eat cereal

	basketball	not basketball	sum(row)	
cereal	2000	1750	3750	75%
not cereal	1000	250	1250	25%
sum(col.)	3000	2000	5000	
	60%	40%		

play basketball \Rightarrow *eat cereal* [40%, 66.7%]

misleading because the overall percentage of students eating cereal is 75% which is higher than 66.7%.

play basketball \Rightarrow *not eat cereal* [20%, 33.3%]

is more accurate, although with lower support and confidence

2110773-3 2/67

25

Lift of a Rule

■ Example 1 (cont)

- *play basketball* \Rightarrow *eat cereal* [40%, 66.7%]

$$\text{LIFT} = \frac{\frac{2000}{5000}}{\frac{3000}{5000} \times \frac{3750}{5000}} = 0.89$$

- *play basketball* \Rightarrow *not eat cereal* [20%, 33.3%]

$$\text{LIFT} = \frac{\frac{1000}{5000}}{\frac{3000}{5000} \times \frac{1250}{5000}} = 1.33$$

	basketball	not basketball	sum(row)
cereal	2000	1750	3750
not cereal	1000	250	1250
sum(col.)	3000	2000	5000

2110773-3 2/67

26

Interestingness Measurements

- Are all of the strong association rules discovered interesting enough to present to the user?
- How can we **measure the interestingness** of a rule?
- Subjective measures
 - A rule (pattern) is interesting if
 - it is **unexpected** (surprising to the user); and/or
 - **actionable** (the user can do something with it)
 - (only the user can judge the interestingness of a rule)

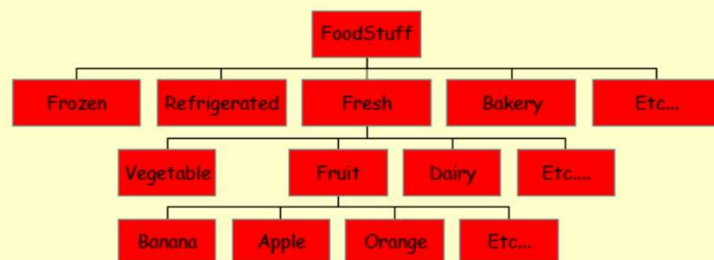
Objective measures of rule interest

- Support
- Confidence or strength
- Lift or Interest or Correlation
- Conviction
- Leverage or Piatetsky-Shapiro
- Coverage

2110773-3 2/67

27

Multiple-Level Association Rules



- **Fresh \Rightarrow Bakery [20%, 60%]**
- **Dairy \Rightarrow Bread [6%, 50%]**

Items often form hierarchy.
Flexible support settings: Items at the lower level are expected to have lower support.
Transaction database can be encoded based on dimensions and levels explore shared multi-level mining

2110773-3 2/67

28

Application Difficulties

- Wal-Mart knows that customers who buy Barbie dolls (it sells one every 20 seconds) have a 60% likelihood of buying one of three types of candy bars. What does Wal-Mart do with information like that?
- 'I don't have a clue,' says Wal-Mart's chief of merchandising, Lee Scott.

Some Suggestions

- By increasing the price of Barbie doll and giving the type of candy bar free, wal-mart can reinforce the buying habits of that particular types of buyer
 - Highest margin candy to be placed near dolls.
 - Take a poorly selling product X and incorporate an offer on this which is based on buying Barbie and Candy. If the customer is likely to buy these two products anyway then why not try to increase sales on X?
 - Probably they can not only bundle candy of type A with Barbie dolls, but can also introduce new candy of Type N in this bundle while offering discount on whole bundle. As bundle is going to sell because of Barbie dolls & candy of type A, candy of type N can get free ride to customers houses. And with the fact that you like something, if you see it often, Candy of type N can become popular.
- Suggest candies should be manufactured in the shape of Barbie dolls
 - Packaging Barbie, candy and perhaps other products together

