

Association : $X \rightarrow Y$ \times ទិន្នន័យ ឱ្យ ការងារ \rightarrow រូប. សម្រាប់ការងារដែលអាចបង្កើតឡើង

គម្រោង = គម្រោងប៉ុណ្ណោះ

ចំណាំគម្រោង = គម្រោង

$$\frac{\text{ចំណាំគម្រោង}}{\text{total}} = \frac{\# \text{occurrence}}{\text{total}}$$

តាមតារាងនៃការងារ និងការងារដែលអាចបង្កើតឡើង
 $P \rightarrow Q$ តឹន្នន័យ
 $Q \rightarrow P$ តឹន្នន័យ
 $X \rightarrow Y$ ទិន្នន័យ

Chapter 3 Association Mining

Associate Professor Yachai Limpiyakorn, Ph.D.

ការរាយជាមួយគម្រោងសម្រាប់ការងារ

Association Mining \rightarrow Unsupervised L.

$P \rightarrow Q \neq Q \rightarrow P$
 Association Rules ក្នុងការងារ item set ឱ្យ \rightarrow item set ឱ្យ

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:

គម្រោងដែលនឹងបង្កើតឡើង ឬទេរងទៅគម្រោងដែលបានបង្កើតឡើង (Association Rule) ទៅការងារដែលបានបង្កើតឡើង (Antecedent) ឬទេរងទៅគម្រោងដែលបានបង្កើតឡើង (Consequent)

{Cheese, Milk} \rightarrow Bread [S=5%, C=80%] $5\% \text{ support}$ $80\% \text{ confidence}$

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.



ผู้ค้า
diaper → beer ✓
beer → diaper ✗ ไม่ต้องไปซื้อ
ไม่ใช่สิ่งเดียวกัน

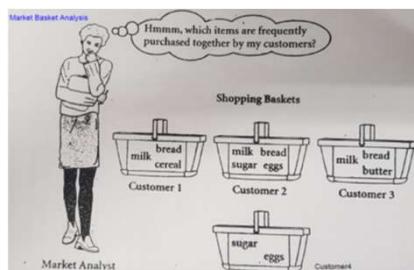
2110773-3 2/66

3

Application 1

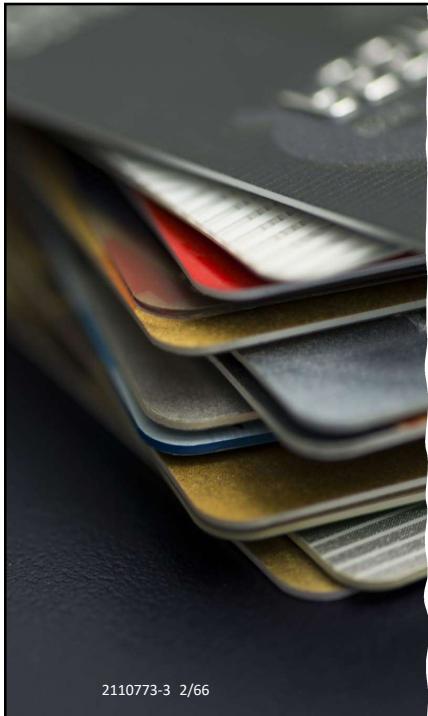
ส่วนใหญ่นักประยุกต์ใช้เทคนิคการทำเหมืองความสัมพันธ์กับการวิเคราะห์ทางการตลาด (Market Basket Analysis: MBA) ซึ่งเป็นรูปแบบการจัดกลุ่ม (Clustering) แบบหนึ่ง ที่ใช้เพื่อหากลุ่มสิ่งของที่น่าจะไปรากฐานกันในทรัพย์ เช่น ชุดของเสื้อผ้า หรือ ของใช้ในครัวเรือน ณ จุดขาย (point-of-sale) ผลลัพธ์หรือแบบจำลองที่ได้สามารถแสดงได้ด้วยภูมิปัญญาเชิงความเป็นไปได้ของการซื้อผลิตภัณฑ์ต่างๆร่วมกัน การวิเคราะห์ทางการตลาดมีบทบาทสำคัญต่ออุตสาหกรรมการค้าปลีก (Retail industry) เพื่อให้ทราบถึงพฤติกรรมการซื้อสินค้าของลูกค้าซึ่งเป็นประโยชน์ในการ

- ◆ จัดพื้นที่ร้านค้า (Store layout) ฟัน ร์ด ไบ แพปิ่ง Blackpink x AIS
- ◆ ทำตลาดเพื่อส่งเสริมการขายสินค้าหรือบริการซึ่งกันและกัน (Cross-marketing)
- ◆ ออกรูปแบบหนังสือแคตตาล็อกสินค้า (Catalog design)
- ◆ วางแผนการส่งเสริมการขายและการตั้งราคาผลิตภัณฑ์ (Product pricing and promotion) ว่องไวสูง แม่ค้า printer ใบราคากัน 50%



2110773-3 2/66

4



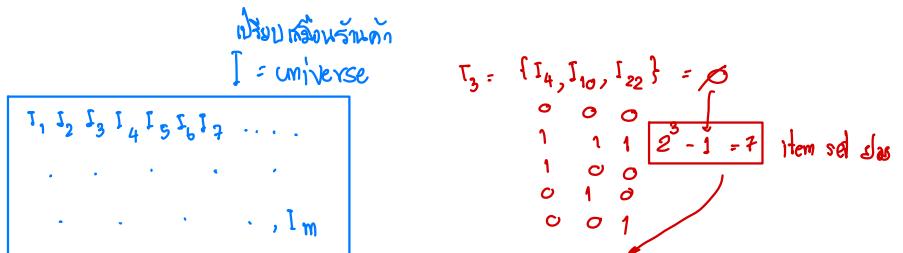
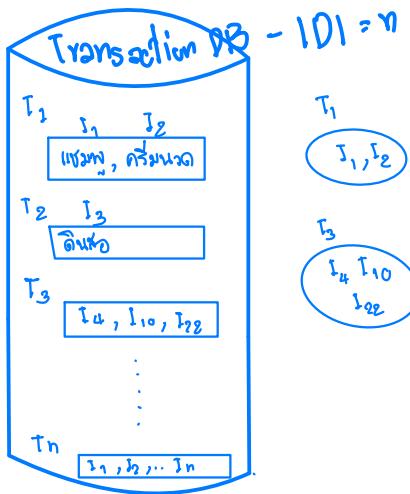
Application 2

นอกจากนี้ สามารถประยุกต์ใช้การวิเคราะห์ทางการตลาดกับกิจกรรมใกล้เคียงที่ลูกค้ามักกระทำด้วยกัน เพื่อก่อให้เกิดรายได้สูงสุดจากการจัดประเภทผลิตภัณฑ์หรือบริการเข้าด้วยกัน ได้แก่

- ◆ การใช้จ่ายผ่านบัตรเครดิตของลูกค้าในการเข้าพักริมแม่น้ำ เช่ารถ ทำให้สามารถทำนายค่าใช้จ่ายต่อไปของลูกค้า
- ◆ แพ็กเกจการให้บริการการสื่อสารโทรคมนาคม เพื่อก่อให้เกิดรายได้สูงสุด
- ◆ การให้บริการทางธนาคารที่ลูกค้ามักซื้อด้วยกัน เพื่อก่อให้เกิดประโยชน์สูงสุด เช่น **ประเภทบัญชีที่ลูกค้ามักเปิดด้วยกัน (account bundle)** การให้บริการการลงทุนครบวงจร และแพ็กเกจสินเชื่อการซื้อรถ เป็นต้น

นิยามพื้นฐาน การทำเหมือง ความสัมพันธ์

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



- ① $I_4 \rightarrow I_{10} I_{22}$
- ② $I_{10} \rightarrow I_4 I_{22}$
- ③ $I_{22} \rightarrow I_4 I_{10}$
- ④ $I_4 I_{10} \rightarrow I_{22}$
- ⑤ $I_4 I_{22} \rightarrow I_{10}$
- ⑥ $I_{10} I_{22} \rightarrow I_4$

Objective measures of rule interest

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

7

2110773-3 2/66

Association Rule

- Rule form

Antecedent → Consequent [support, confidence]

Note: support and confidence are user defined measures of interestingness

- Examples

- * buys(x, "computer") → buys(x, "financial management software") [0.5%, 60%]
- age(x, "30..39") ∧ income(x, "42..48K") → buys(x, "car") [1%, 75%]

2110773-3 2/66

8

Rule basic Measures: Support and Confidence

$A \Rightarrow B [s, c]$
เพื่อสัมภารณ์ item set B

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

คำนวณหาค่าสนับสนุน ได้จากการจำนวนทรานแซกชันที่มีรายการ X และ Y เกิดร่วมกัน หารด้วยจำนวนทรานแซกชันทั้งหมด

$$\text{support} (X \rightarrow Y)$$

$$= P(X \cup Y)$$

$$= \text{tran_count} (X \cup Y) / \text{tran_count} (D)$$

• Confidence

คำนวณค่าความเชื่อมั่นได้จากการจำนวน ทรานแซกชันที่มีรายการ X และ Y เกิดร่วมกัน หารด้วยจำนวนทรานแซกชันที่มีรายการ X

$$\text{confidence} (X \rightarrow Y)$$

$$= P(Y | X)$$

$$= \text{tran_count} (X \cup Y) / \text{tran_count} (X)$$

$$S(A \rightarrow B) = P(A \cup B)$$

↑ เก็บว่า

$$A \cap B = \emptyset$$

item set A ไม่共存กับ item set B ใน transaction นั้นๆ
ดังความคืบหน้า ↓
un boundary

Support

$$S(A \rightarrow B) = P(A \cup B) = \frac{\# \text{ Transaction } (A \cup B)}{\text{total } \# \text{ Trans} = |D|}$$

Confidence

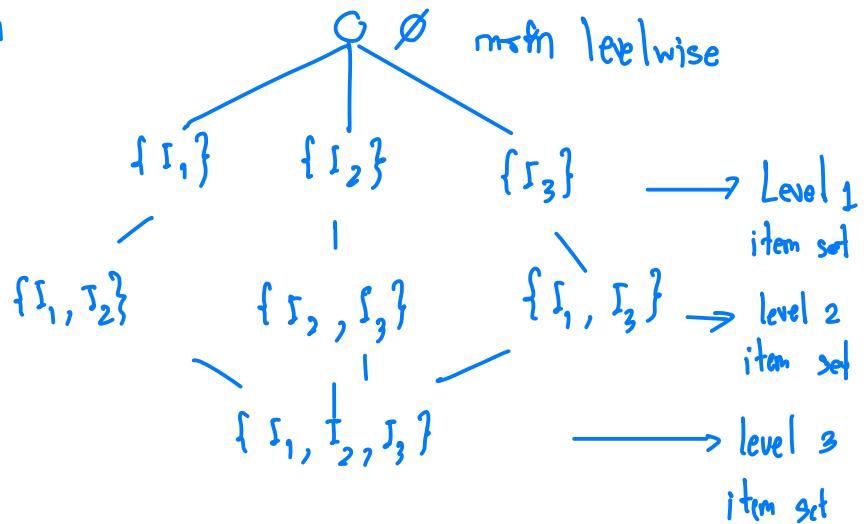
$$C(A \rightarrow B) = P(B | A) = \frac{P(B \cup A)}{P(A)}$$

$$= \frac{\# \text{ Trans } (B \cup A) / |D|}{\# \text{ Trans } (A) / |D|}$$

$$= \frac{\# \text{ Trans } (A \cup B)}{\# \text{ Trans } (A)}$$

① find frequent itemset $\geq \min \text{ support}$ $I = \{I_1, I_2, I_3\}$

ผลลัพธ์ที่เป็นไปได้ $\Rightarrow 2^{n-1}$



② ตีความรู้ไปตามสมมติฐาน

$$I_2 \rightarrow I_3$$

$$I_3 \rightarrow I_2$$

level k item set
= item set containing k items

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% อยากร้าบว่า

ความสัมพันธ์ $A \rightarrow C$ และ $C \rightarrow A$ ความสัมพันธ์ใดเป็นกฎความสัมพันธ์

$$S(A \rightarrow C) = P(A \cup C) = \frac{\# Trans(A \cup C)}{Total} = \frac{2}{4}$$

$$S(C \rightarrow A) = S(A \rightarrow C) = \frac{1}{4}$$

$$C(A \rightarrow C) = \frac{\# Trans(A \cup C)}{\# Trans(A)} = \frac{2}{3}^{11} = \frac{2}{3}$$

$$C(C \rightarrow A) = \frac{\# Trans(C \cup A)}{\# Trans(C)} = \frac{2}{2} = 100\%$$

2110773-3 2/66

Association Mining

เป็นปัญหาการค้นหากฎความสัมพันธ์ นิยามได้ดังนี้

- การค้นหากฎความสัมพันธ์ คือ การหากฎความสัมพันธ์ทั้งหมดในทรานแซคชันทุกตัวของเซตข้อมูลที่กำหนดให้ โดยกฎความสัมพันธ์ที่หาได้ทั้งหมดจะต้องมีค่าสนับสนุน (support) ไม่ต่ำกว่าค่าสนับสนุน้อยสุด (minimum support) ที่ผู้ใช้กำหนดไว้ และมีค่าความเชื่อมั่น (confidence) ไม่ต่ำกว่าค่าความเชื่อมั่น้อยสุด (minimum confidence) ที่ผู้ใช้กำหนดไว้ ~ **กำหนดก่อนสร้างกฎ**
- การค้นหากฎความสัมพันธ์สามารถแบ่งย่อยได้เป็นสองขั้นตอน คือ
 - ค้นหาเซตของไอเทมประภัย (frequent itemset) หรือไอเทมเซตที่มีค่าสนับสนุนไม่ต่ำกว่าค่าสนับสนุน้อยสุดที่กำหนดให้ ผ่านเกณฑ์ลังท์ $\geq \text{minimum support}$
 - นำไอเทมเซตประภัยเหล่านั้นมาสร้างเป็นกฎความสัมพันธ์ต่อไป

2110773-3 2/66

12

I [A B C D E]

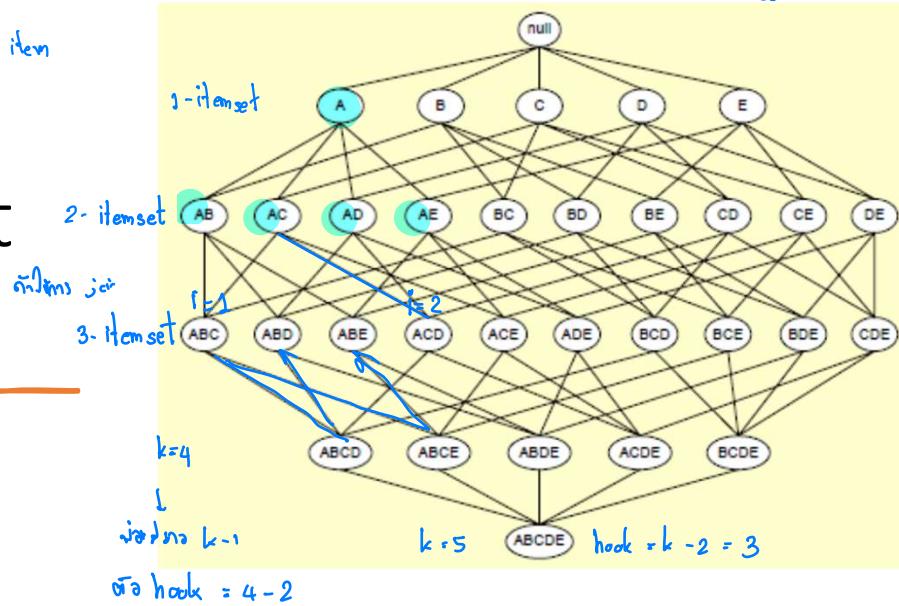
1. itemset តើខ្លះដែលមាន 1 item

Itemset Lattice



2110773-3 2/66

ឡាយ → Possible set = $2^n - 1$ ^(C) = $2^5 - 1 = 31$



13

Apriori Principle

Anti-monotone property ឬ $P(X) < \text{min_support}$
Superset $P(X \cup S) < \text{min_support}$

Any subset of a frequent itemset must also be frequent

No superset of any infrequent itemset should be generated or tested superset តើអ្វីដែរបាន និតាដោយ generate

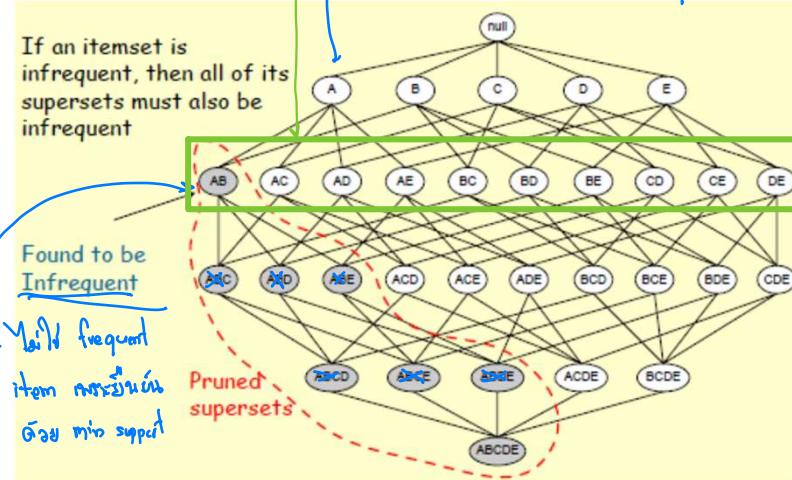
- Many item combinations can be pruned

2110773-3 2/66

14

$$\text{Cartesian Product } k=2, \# \text{ 2 item-sets} = \binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{5 \times 4 \times 3 \times 2}{2 \cdot 1 \cdot 3 \cdot 2} = 10$$

Apriori Principle for Pruning Candidates



2110773-3 2/66

15

Association Mining: 2 key steps

- Find all Frequent Itemsets: the sets of items that pass minimum support
 - ❖ Apriori Algorithm
 - มีการจัดเรียงลำดับของไอเทมในแต่ละท่านแยกขั้นก่อนประมวลผล
 - การสร้างไอเทมเซ็ตจะสร้างตามระดับขั้น จากขั้นที่ k, k+1, k+2, ...
 - ใช้ความรู้ก่อนหน้าคือคุณสมบัติของไอเทมเซ็ตเกิดบ่อยในการตัดเลือก
- 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 $\geq \text{min_confidence}$

Prequisite
ส่วน itemset ที่ต้อง^{มี} minimum support

2110773-3 2/66

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. *minimum min support stellieren*

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)

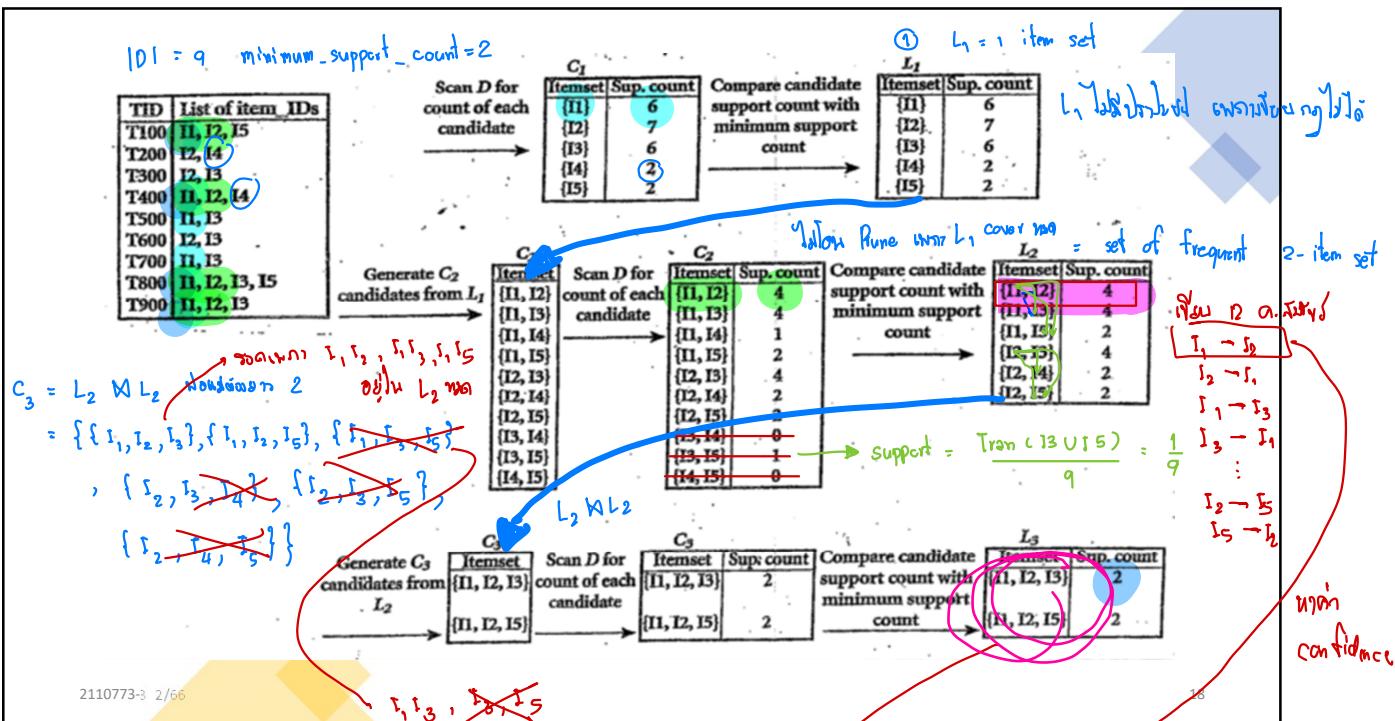
```

(1) for each itemset  $I_1 \in L_{k-1}$  inner loop
(2)   for each itemset  $I_2 \in L_{k-1}$ 
(3)     if  $(I_1[1] = I_2[1]) \wedge (I_1[2] = I_2[2]) \wedge \dots \wedge (I_1[k-2] = I_2[k-2]) \wedge (I_1[k-1] < I_2[k-1])$  then {
(4)        $I_c = I_1 \bowtie I_2$ ; // join step: generate candidates
(5)       outer loop if has_infrequent_subset( $c, L_{k-1}$ ) then zu b. Wörter
(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;
```



Apriori Pseudocode

C_k = Set of candidate k -itemsets

$$= \{ \{c_{11}, c_{12}, c_{13}, \dots, c_{1k}\}, \{c_{21}, c_{22}, \dots, c_{2k}\}, \dots \}$$
$$\quad \quad \quad C_1 \quad \quad \quad C_2$$

L_k = Set of frequent k -itemsets $\Rightarrow C$ has min. support

$$= \{ \{l_{11}, l_{12}, \dots, l_{1k}\}, \{l_{21}, l_{22}, \dots, l_{2k}\}, \dots \}$$

① Generate $C_k = L_{k-1} \bowtie L_{k-1}$

② Prune by Apriori property

③ $L_k = \{c \in C_k \mid c \geq \text{min_support}\}$

การสร้างกฎความสัมพันธ์จากเซตของไอтемปราภกฎบอย

เมื่อได้ไอ템เซตปราภกฎบอยแล้ว จำเป็นต้องหากฎความสัมพันธ์จากไอ้มง เซตปราภกฎบอยนั้น โดยกฎความสัมพันธ์ที่ได้จะขึ้นว่าค่าความเชื่อมั่นไม่ต่ำ ก่าค่าความเชื่อมั่นน้อยสุดที่กำหนดให้ ถือกฎความสัมพันธ์ังกล่าวว่า

Strong Association Rules

ภายหลังจากที่ได้ไอ้มงเซตปราภกฎบอยทั้งหมดแล้ว จะสร้างเขตกฎ ความสัมพันธ์จากแพลตฟอร์มของไอ้มงเซตปราภกฎบอย | โดยสร้างทุกเขตโดยที่ไม่ ว่างของ | กด้าวีอุ ทุกเขตโดยที่ไม่ว่าง S ของ | แสดงกฎความสัมพันธ์ r → (I-S) ถ้าดูว่าส่วนระหว่างจำนวนทรานแซกชันของ | ต่อจำนวน ทรานแซกชันของ S | ไม่น้อยกว่าค่าความเชื่อมั่นที่กำหนด

- ผลลัพธ์ของไอ้มงเซตปราภกฎบอยที่ได้จากการจัดเรียงตามความเชื่อมั่น เป็นความสัมพันธ์ X → Y พร้อมคำนวนหาค่าความเชื่อมั่น (ในที่นี้ แสดงเพียงสมาชิก $\{I_1, I_2, I_5\} \in L_3$)

$$\begin{aligned} I_1, I_2 &\rightarrow I_5 \quad [\text{confidence} = 2/4 = 50\%] \\ I_1, I_5 &\rightarrow I_2 \quad [\text{confidence} = 2/2 = 100\%] \\ I_2, I_5 &\rightarrow I_1 \quad [\text{confidence} = 2/2 = 100\%] \\ I_1 &\rightarrow I_2, I_5 \quad [\text{confidence} = 2/6 = 33\%] \\ I_2 &\rightarrow I_1, I_5 \quad [\text{confidence} = 2/7 = 29\%] \\ I_5 &\rightarrow I_1, I_2 \quad [\text{confidence} = 2/2 = 100\%] \end{aligned}$$

- กำหนดค่าความเชื่อมั่นต่ำสุดเท่ากับ 70% จะได้ว่า ความสัมพันธ์ที่สร้างจาก L_3 ที่เป็น Strong Association Rules ประกอบด้วย

$$I_1, I_5 \rightarrow I_2; I_2, I_5 \rightarrow I_1; I_5 \rightarrow I_1, I_2$$

↑
support, confidence

Improving Apriori

Challenge

- every pass goes over whole data
- multiple scans of transaction database
- huge number of candidates
- one transaction may contain many candidates
หนักหนะ, chores
- 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

ข้อพิจารณาเกี่ยวกับค่าสนับสนุนและค่าความเชื่อมั่น

- ◆ การค้นหากฎความสัมพันธ์อาจล้มเหลว ถ้ากำหนดค่าสนับสนุนและค่าความเชื่อมั่นสูงเกินไป
- ◆ ถ้ากำหนดค่าสนับสนุนและค่าความเชื่อมั่นต่ำเกินไป อาจได้ความสัมพันธ์ระหว่างผลิตภัณฑ์หลากหลายเกินไปที่เราไม่ต้องการ
- ◆ กฎความสัมพันธ์ที่มีค่าสนับสนุนและค่าความเชื่อมั่นสูง แสดงระดับความเกี่ยวข้อง (degree of relevance) มากกว่ากฎความสัมพันธ์ที่มีค่าสนับสนุนและค่าความเชื่อมั่นต่ำ
- ◆ ค่าสนับสนุนแสดงความถี่ของจำนวนทรานแซกชันของการเกิดร่วมกันของผลิตภัณฑ์ โดยที่นำไปจะให้น้ำหนักความสำคัญแก่ทรานแซกชันที่เกิดบ่อย แต่บางครั้งทรานแซกชันที่มีค่าสนับสนุนต่ำอาจเป็นประโยชน์ต่อการค้นหากฎความสัมพันธ์บางอย่าง
- ◆ ค่าความเชื่อมั่นเพียงอย่างเดียวไม่อาจบอกได้ว่าการเกิดร่วมกันของผลิตภัณฑ์ A และ B เป็นไปโดยบังเอิญหรือไม่ ซึ่งเราสามารถใช้กฎความสัมพันธ์ระหว่างผลิตภัณฑ์ที่ไม่ได้เกิดขึ้นโดยความบังเอิญมากกว่า

Dependent Framework

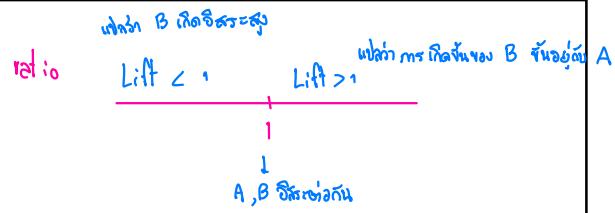
นั่งๆ กดๆ อยู่ตรงนี้ก็ได้เงินบังเขียวหรือไม่!

- ◆ การทำเหมือนกฎความสัมพันธ์ โดยใช้ค่าสนับสนุน-ความเชื่อมั่นเป็นที่แพร่หลายในหลายแอปพลิเคชัน แต่ในบางครั้ง ครอบคลุมค่าสนับสนุน-ความเชื่อมั่นอาจทำให้ผู้ใช้เข้าใจผิดเกี่ยวกับความน่าสนใจของกฎที่ค้นพบ $A \rightarrow B$ เนื่องจากความจริงแล้วการเกิดเหตุการณ์ A ไม่ได้ส่อนาย (imply) การเกิดของเหตุการณ์ B ก่อให้เกิดคำถามว่า **Strong Association Rules** ที่ค้นพบน่าสนใจหรือไม่
- ◆ **ครอบความเชื่อมั่นต่อกัน** (Dependent Framework) สามารถใช้วัดความน่าสนใจของกฎที่ค้นพบในแต่ละค่าสหสัมพันธ์ (correlation) ของเหตุการณ์

Correlation/ Lift/ Interest

- Correlation/ Lift/ Interest

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



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

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
$X \rightarrow Y$	$\frac{\#(X \cup Y)}{10} = \frac{2}{8}$	$\frac{\#(X \cap Y)}{\#X} = \frac{2}{4}$	$\frac{\text{confidence}}{P(Y)} = \frac{2/4}{2/8} = 2$
$X \rightarrow Z$	$\frac{\#(X \cup Z)}{8} = \frac{3}{8}$	$\frac{\#(X \cap Z)}{\#X} = \frac{3}{4}$	$\frac{\text{confidence}}{P(Z)} = \frac{3/4}{7/8} = \frac{3}{7}$
$Y \rightarrow Z$	$\frac{\#(Y \cup Z)}{8} = \frac{1}{8}$	$\frac{\#(Y \cap Z)}{\#Y} = \frac{1}{2}$	$\frac{\text{confidence}}{P(Z)} = \frac{1/2}{7/8} = \frac{4}{7}$

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%		

ผลลัพธ์ที่ 40%
 $\text{play basketball} \Rightarrow \text{eat cereal}$ [40%, 66.7%] $\frac{2000}{3000}$

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

$\text{play basketball} \Rightarrow \text{not eat cereal}$ [20%, 33.3%]

is more accurate, although with lower support and confidence

Lift of a Rule

- Example 1 (cont)

▪ $\text{play basketball} \Rightarrow \text{eat cereal}$ [40%, 66.7%]

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

▪ $\text{play basketball} \Rightarrow \text{not eat cereal}$ [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

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

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

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. ~~no candy, simuwq māq barbies~~
- 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
- Offering 'affinity program' that give Barbie accessories in exchange for proofs of purchase
- Packaging Barbie, candy and perhaps other products together

