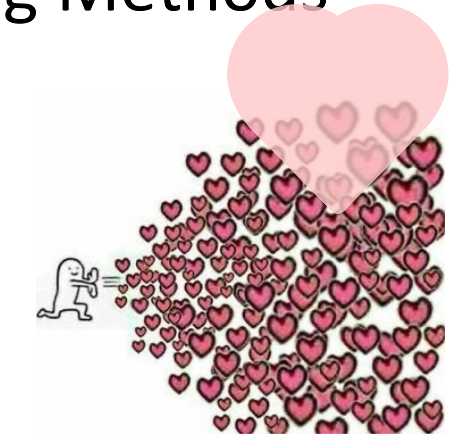
Chapter 6: Mining Frequent Patterns, Association and Correlations: Basic Concepts and Methods

- ❑ Basic Concepts
- ❑ Efficient Pattern Mining Methods
- ❑ Pattern Evaluation
- ❑ Summary



Mining frequent patterns

การหา patterns ที่ฝังซ่อนอยู่ใน Data

การค้นหารูปแบบ

What Is Pattern Discovery?

- ❑ What are patterns? ไอเทมหลายๆอย่างที่มักจะซื้อร่วมกัน (patterns การซื้อของคน), set of items มักจะเกิดขึ้นซ้ำๆ
  - ❑ Patterns: A set of items, subsequences, or substructures that occur frequently together (or strongly correlated) in a data set
  - ❑ Patterns represent intrinsic and important properties of datasets
- ❑ Pattern discovery: Uncovering patterns from massive data sets
- ❑ Motivation examples:
  - ❑ What products were often purchased together? สนค้าอะไรที่คนจะซื้อพร้อมกันเสมอ
  - ❑ What are the subsequent purchases after buying an iPad? หลังจากที่ซื้อสิ่งหนึ่งไปแล้ว ครั้งหน้าจะมาซื้ออะไร
  - ❑ What code segments likely contain copy-and-paste bugs? มี code ของคนที่ทำก่อนหน้ามา ใช้แก้ปัญหาได้ แต่จะมี bug นะ "ดูตัวแปรด้วย"
  - ❑ What word sequences likely form phrases in this corpus?

Basic Concepts: k-Itemsets and Their Supports

"ขั้นตอนการทำงาน"

เซ็ตของไอเท็มที่คนมัก จะซื้อร่วมกัน

- □ Itemset: A set of one or more items

ไอเทมเซ็ตที่มีสมาชิก k ตัว

- □ k-itemset: X = {x_1, …, x_k}
  - □ Ex. {Beer, Nuts, Diaper} is a 3-itemset

(แอบโซลูท) ซัพพอร์ต → จำนวนของ transaction ที่มา support เรา

ทำมือได้

- □ (*absolute*) *support* (*count*) of X, sup{X}: Frequency or the number of occurrences of an itemset X

  จำนวนทรานแซคชันว่ามีเบียร์กี่ทรานแซคชั่น

  - □ Ex. sup{Beer} = 3
  - □ Ex. sup{Diaper} = 4

  ทรานแซคชั่นแบบซื้อพร้อมกัน

  - □ Ex. sup{Beer, Diaper} = 3
  - □ Ex. sup{Beer, Eggs} = 1

Tid	Items bought
10	Beer, Nuts, Diaper ทรานแซคชั่น ID คือที่ขายต่อ 1 ใบเสร็จ
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

สัดส่วนของ transaction ที่ support ไอเท็มเซ็ตนั้นมีเท่าไร

- □ (*relative*) *support*, *s*{*X*}: The fraction of transactions that contains X (i.e., the probability that a transaction contains X)

  จากที่นับคือ 3 หาร 5 (จน.ทรานแซคชั่นทั้งหมด)

  - □ Ex. s{Beer} = $\underline{3/5}$ = 60%
  - □ Ex. s{Diaper} = 4/5 = 80%
  - □ Ex. s{Beer, Eggs} = 1/5 = 20%

Basic Concepts: Frequent Itemsets (Patterns)

- An itemset (or a pattern) X is *frequent* if the support of X is no less than a *minsup* threshold σ

(เกณฑ์) ค่าจัดแบ่งว่าเอาไม่เอา

ดูความบ่อย ไม่เกิน 50%

- Let σ = *50%* (σ: *minsup* threshold)

For the given 5-transaction dataset

- All the frequent 1-itemsets:

  Beer มี 3 จาก 5 transaction คือ 60%

  - Beer: 3/5 (60%); Nuts: 3/5 (60%)
  - Diaper: 4/5 (80%); Eggs: 3/5 (60%)
- All the frequent 2-itemsets:

  มี Beer กับ Diaper คู่กันอยู่ 3 ใน 5 transaction = 60%

  - {Beer, Diaper}: 3/5 (60%)
- All the frequent 3-itemsets?
  - None

□ coffee : 2/5 (40%) ⇒ ไม่ผ่านเกณฑ์

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

- Why do these itemsets (shown on the left) form the complete set of frequent *k*-itemsets (patterns) for any *k*?
- **Observation**: We may need an efficient method to mine a complete set of frequent patterns

From Frequent Itemsets to Association Rules

- ❑ Comparing with itemsets, rules can be more telling
  - ❑ Ex. *Diaper* → *Beer* คนซื้อ Diaper นำไปสู่การซื้อ Beer
    - ❑ *Buying diapers may likely lead to buying beers*

ยิ่งสูงกฎยิ่งน้อยแต่สูงมากจะไม่มีอะไรให้ดูเลยก็ได้

- ❑ How strong is this rule? (support, confidence)
  - ❑ Measuring association rules: *X* → *Y* (s, c)
    - ❑ Both *X* and *Y* are itemsets

เริ่มจากหา Support ของ X และ Y

  - ❑ Support, *s*: The probability that a transaction contains X ∪ Y
    - ❑ Ex. s{Diaper, Beer} = 3/5 = 0.6 (i.e., 60%) หา sup ได้ 0.6
  - ❑ Confidence, *c: The conditional probability* that a transaction containing X also contains *Y*

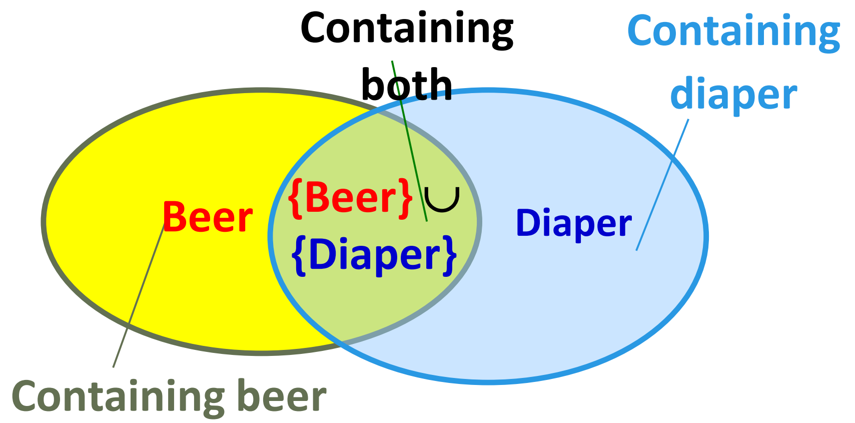
หา sup ของ 2อันหารด้วย sup ตัวหน้า

    - ❑ Calculation: *c* = sup(X ∪ Y) / sup(X)
    - ❑ Ex. *c* = sup{Diaper, Beer}/sup{Diaper} = $\frac{3}{4}$ = 0.75

Diaper / Beer มี

มี Diaper ทั้งหมด

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



{Beer} ∪ {Diaper} = {Beer, Diaper}

Note: X ∪ Y: the union of two itemsets

- ■ The set contains both X and Y

Mining Frequent Itemsets and Association Rules

- ❑ **Association rule mining**
  - ❑ Given two thresholds: *minsup, minconf*
  - ❑ Find all of the rules, *X* → *Y* (s, c)
    - ❑ such that, s ≥ *minsup* and *c* ≥ *minconf*

- ❑ Let *minsup = 50%* → เป็นไอเทมเซ็ตที่มีอยู่เป็นส่วนใหญ่ของทรานแซ็กชั่น
  - ❑ Freq. 1-itemsets: Beer: 3, Nuts: 3, Diaper: 4, Eggs: 3
  - ❑ Freq. 2-itemsets: {Beer, Diaper}: 3

เกิดขึ้นซ้ำๆกันมากแค่ไหน

- ❑ Let *minconf* = 50% Sup(Beer/Diaper)/Sup(Beer)

  Sup conf
  - ❑ *Beer* → *Diaper* (60%, 100%) → เอา Diaper ไปวางชั้นขาย Beer
  - ❑ *Diaper* → *Beer* (60%, 75%)

กฎ 2 ตัว

(Q: Are these all rules?)

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

- ❑ **Observations:**
  - ❑ Mining association rules and mining frequent patterns are very close problems
  - ❑ Scalable methods are needed for mining large datasets

Efficient Pattern Mining Methods

- ❑ The Downward Closure Property of Frequent Patterns
- ❑ The Apriori Algorithm
- ❑ Extensions or Improvements of Apriori
- ❑ Mining Frequent Patterns by Exploring Vertical Data Format
- ❑ FPGrowth: A Frequent Pattern-Growth Approach
- ❑ Mining Closed Patterns

Apriori Pruning and Scalable Mining Methods

ตัดแต่ง

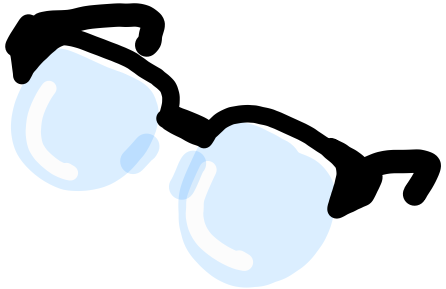
: ถ้าเกิดว่าไอเทมเซ็ตที่ต่ำกว่า ไม่ผ่าน minsup ไอเทมเซ็ตที่สูงกว่าก็ไม่มีทางผ่าน Minsup (ให้ดูคู่กัน)

- ❑ Apriori pruning principle: If there is any itemset which is infrequent, its superset should not even be generated! (Agrawal & Srikant @VLDB'94, Mannila, et al. @ KDD' 94)
- ❑ Scalable mining Methods: Three major approaches
  - ❑ Level-wise, join-based approach: Apriori (Agrawal & Srikant@VLDB'94)
  - ❑ Vertical data format approach: Eclat (Zaki, Parthasarathy, Ogihara, Li @KDD'97)
  - ❑ Frequent pattern projection and growth: FPgrowth (Han, Pei, Yin @SIGMOD'00)

Apriori: A Candidate Generation & Test Approach

การเขียนเป็นขั้นตอน

- ❑ Outline of Apriori (level-wise, candidate generation and test)
  - ❑ Initially, scan DB once to get frequent 1-itemset สแกนเริ่มจากดู 1 ไอเทมเซทก่อน
  - ❑ Repeat
    - ❑ Generate length-(k+1) candidate itemsets from length-k frequent itemsets
    - ❑ Test the candidates against DB to find frequent (k+1)-itemsets
    - ❑ Set k := k +1
  - ❑ Until no frequent or candidate set can be generated
  - ❑ Return all the frequent itemsets derived



The Apriori Algorithm—An Example

กำหนด minsup = 2

minsup = 2

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

มี 4 ทราน

1^{st} scan

สแกน

หา 1 ไอเทมเซ็ต

C_1

ตาราง oneItemset

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

F_1

ได้ตารางใหม่

คำตอบที่ 1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

จับคู่ Itemset

ได้

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2^{nd} scan

C_2

แสกนดูว่าตรงกับ minsup หรือไม่

Itemset	sup
~~{A, B}~~	~~1~~
{A, C}	2
~~{A, E}~~	~~1~~
{B, C}	2
{B, E}	3
{C, E}	2

จะได้

F_2

ตารางใหม่

คำตอบที่ 2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

สร้าง tree Itemset

Itemset
{B, C, E}

จริงๆได้ 8 ตัว แต่เขานับมาแล้ว

3^{rd} scan

F_3

คำตอบสุดท้าย

Itemset	sup
{B, C, E}	2