CS 412 Intro. to Data Mining

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

Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017

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

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

What Is Pattern Discovery?

- What are patterns? การค้นหา patterns ที่ซ่อนอยู่ใน Data
  - 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

เอาไปใช้ประโยชน์อะไรได้บ้าง

- Pattern discovery: Uncovering patterns from massive data sets
- Motivation examples:
  - What products were often purchased together? สินค้าที่มักจะซื้อคู่กันเสมอ
  - What are the subsequent purchases after buying an iPad? เมื่อซื้อ IPad ไปแล้ว จะกลับมาซื้ออะไร
  - What code segments likely contain copy-and-paste bugs?
  - What word sequences likely form phrases in this corpus?

Pattern Discovery: Why Is It Important?

มันมีความสำคัญ เพราะเป็นปัญหาแรกเลยที่นำ Data mining มาแก้ไข

- Finding inherent regularities in a data set
- Foundation for many essential data mining tasks
  - Association, correlation, and causality analysis
  - Mining sequential, structural (e.g., sub-graph) patterns
  - Pattern analysis in spatiotemporal, multimedia, time-series, and stream data
  - Classification: Discriminative pattern-based analysis
  - Cluster analysis: Pattern-based subspace clustering
- Broad applications
  - Market basket analysis, cross-marketing, catalog design, sale campaign analysis, Web log analysis, biological sequence analysis

ศัพท์เฉพาะที่ต้องรู้

Basic Concepts: k-Itemsets and Their Supports

ชุดของ items ที่คนมักซื้อร่วมกัน

- Itemset: A set of one or more items

มี k ตัว

- k-itemset: $X = \{x_1, \ldots, x_k\}$ itemset ที่มีสมาชิกกี่ตัว
  - Ex. {Beer, Nuts, Diaper} is a 3-itemset

จำนวนของ transaction ที่มาสนับสนุนแนวคิด

- (*absolute*) *support* (*count*) of X, sup{X}: Frequency or the number of occurrences of an itemset X
  - Ex. sup{Beer} = 3
  - Ex. sup{Diaper} = 4
  - Ex. sup{Beer, Diaper} = 3
  - Ex. sup{Beer, Eggs} = 1

ถ้าเกิดซื้อ Beer จะมีกี่อันที่มา support ก็คือมี 3

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

- (*relative*) *support*, *s*{*X*}: The fraction of transactions that contains X (i.e., the probability that a transaction contains X)
  - Ex. s{Beer} = 3/5 = 60%
  - Ex. s{Diaper} = 4/5 = 80%
  - Ex. s{Beer, Eggs} = 1/5 = 20%

Beer 3 ทั้งหมด 5 = 60%

ดังนั้น support ของ Beer จะเป็น 60% ของ transaction ทั้งหมด

Basic Concepts: Frequent Itemsets (Patterns)

ดูความถี่

- An itemset (or a pattern) X is *frequent* if the support of X is no less than a *minsup* threshold σ
ค่าจัดแบ่ง ระหว่าง ใช่ หรือ ไม่ใช่
- Let σ = *50%* (σ: *minsup* threshold)

  For the given 5-transaction dataset
  - All the frequent 1-itemsets:
    - Beer: 3/5 (60%); Nuts: 3/5 (60%)
    - Diaper: 4/5 (80%); Eggs: 3/5 (60%)
  - All the frequent 2-itemsets:
    - {Beer, Diaper}: 3/5 (60%)
  - All the frequent 3-itemsets?
    - None

Coffee: $\frac{2}{5}$ (40%)

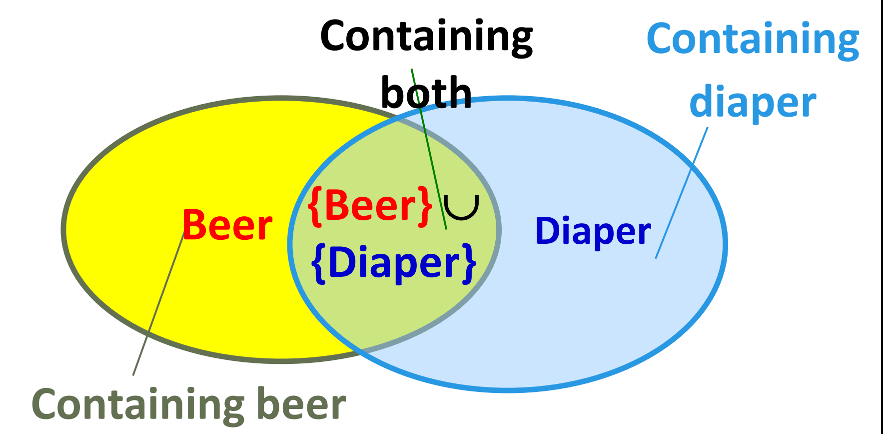
Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

- 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* ถ้าซื้อ Diaper มีแนวโน้มไปสู่การซื้อ Beer ด้วย
    - *Buying diapers may likely lead to buying beers*
- How strong is this rule? (support, confidence)
  - Measuring association rules: *X* → *Y* (s, c)
    - Both *X* and *Y* are itemsets
  - Support, *s*: The probability that a transaction contains X ∪ Y
    สัญลักษณ์ support ของ x และ Y
    - Ex. s{Diaper, Beer} = 3/5 = 0.6 (i.e., 60%)
  - Confidence, *c: The conditional probability* that a transaction containing X also contains *Y*
    - Calculation: *c* = sup(X ∪ Y) / sup(X)
    - $\frac{3/5}{4/5} = \frac{3}{4}$
    - Ex. *c* = sup{Diaper, Beer}/sup{Diaper} = ¾ = 0.75

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** ก่อนจะหา ต้องกำหนด
  - Given two thresholds: *minsup*, *minconf*
  - Find all of the rules, $X \rightarrow Y$ (s, c)
    - such that, s ≥ *minsup* and *c* ≥ *minconf*

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

- Let *minconf = 50%* สูตร $c = sup(X \cup Y) / sup(X)$
  - *Beer* → *Diaper* (60%, 100%)
  - *Diaper* → *Beer* (60%, 75%)

(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

ชิ่ง

- 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

หลักการสำคัญ ตัดแต่ง

- 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)

ขั้นตอนในการหา

Apriori: A Candidate Generation & Test Approach

- Outline of Apriori (level-wise, candidate generation and test)
  - Initially, scan DB once to get frequent 1-itemset
  - Repeat วนลูป
    - 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 (Pseudo-Code)

C_k: Candidate itemset of size k

F_k : Frequent itemset of size k

Pseudo-Code = เขียนโค้ดโปรแกรม เพื่อให้เรา สามารถนำ ไปแปลงเป็นภาษาที่เราใช้ได้

```
K := 1;
F_k := {frequent items};   // frequent 1-itemset
While (F_k != ∅) do {      // when F_k is non-empty
    C_k+1 := candidates generated from F_k;  // candidate generation
    Derive F_k+1 by counting candidates in C_k+1 with respect to TDB at minsup;
    k := k + 1
    }
return ∪_k F_k             // return F_k generated at each level
```

The Apriori Algorithm—An Example

① หา one item set ว่าสามารถเป็นได้ตัวอะไรบ้าง
หา support ของแต่ละตัว

minsup = 2

Database TDB

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

1st scan → C_1

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

ตัดเพราะไม่ถึง 2

→ F_1

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

เอามาจับคู่

C_2

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

2nd scan → C_2

ตัดอีก

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

→ F_2

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

ทำเป็น 3 item set

C_3

Itemset
{B, C, E}

3rd scan → F_3

Itemset	sup
{B, C, E}	2