Data Mining

Frequent Pattern Mining

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

- Basic Concepts
- Frequent Itemset Mining Methods
- Which Patterns Are Interesting?—Pattern Evaluation Methods

Learning Outcomes

- Define the basic concepts in frequent pattern analysis
- Compute support and confidence for association rules
- Use Apriori & FPGrowth algorithms to generate frequent itemsets

What Is Frequent Pattern Analysis?

- Frequent pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
- Motivation: Finding inherent regularities in data
  - What products were often purchased together?— Juice and diapers?!
  - What are the subsequent purchases after buying a PC?
  - What kinds of DNAs are sensitive to this new drug?
  - Can we automatically classify web documents?
- Applications
  - Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

Basic Concepts: Frequent Patterns and Association Rules

- Let I $=\{I_1, I_2, \ldots, I_m\}$ *be a set of items* E.g., I={Coffee,Diaper,Eggs,Juice,Milk,Nuts}
- A transaction *T* is a set of items such that $T \subseteq I$ E.g., T={Nuts, Eggs, Milk}
- A transaction *T* is said to contain *A* if and only if $A \subseteq T$
- An association rule is an implication of the form $A \Rightarrow B$ where $A \subset I$, $B \subset I$ and $A \cap B = \phi$

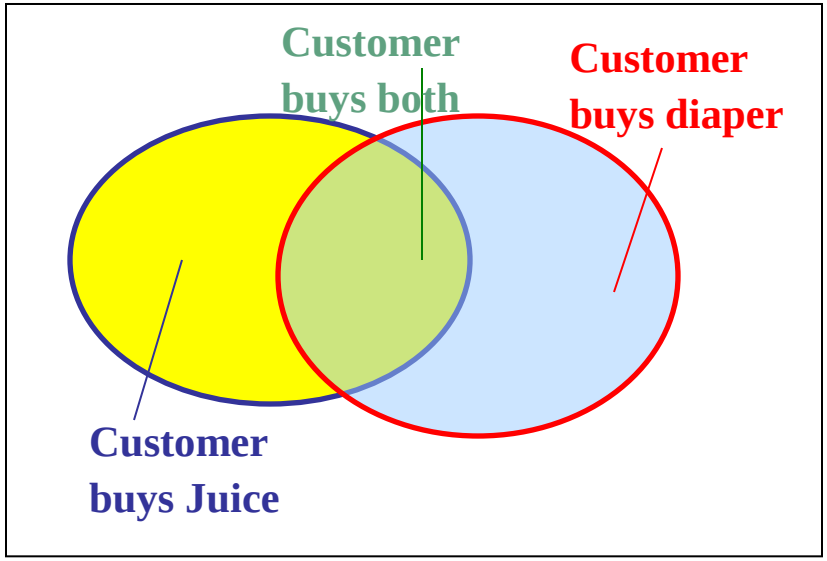
  E.g.,
  A={Nuts}
  B={Eggs, Milk}
  A=>B
  {Nuts} => {Eggs,Milk}
- Support is the percentage of transactions containing $A \cup B$
- Confidence is the percentage of transactions containing A that also contains B
- Itemset is a set of items
- An itemset containing k items is a k-itemset

  E.g.,
  A={Nuts}
  B={Eggs, Milk}
  A and B are itemsets
  A is 1-itemset
  B is 2-itemset

Basic Concepts: Frequent Patterns

E.g., I={Coffee,Diaper,Eggs,Juice,Milk,Nuts}

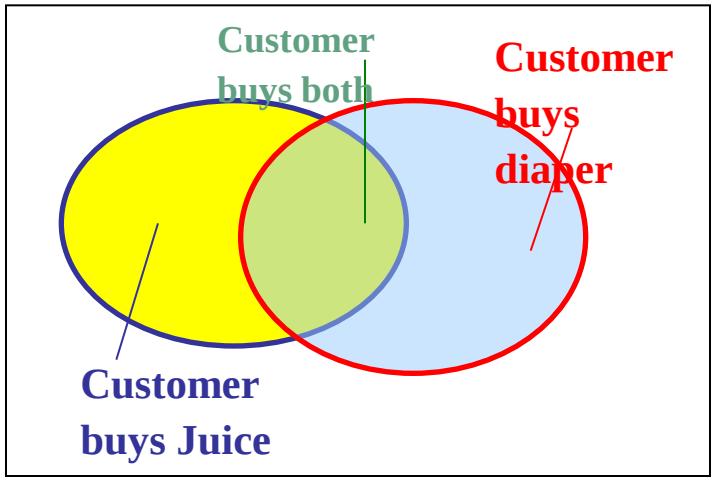
Tid	Items bought
10	Juice, Nuts, Diaper
20	Juice, Coffee, Diaper
30	Juice, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



- itemset: A set of one or more items
- k-itemset $X = \{x_1, \ldots, x_k\}$
- *(absolute) support*, or, *support count* of X: Frequency or occurrence of an itemset X
- *(relative) support*, *s*, is the fraction of transactions that contains X (i.e., the probability that a transaction contains X)
- An itemset X is *frequent* if X's support is no less than a *minsup* threshold

Basic Concepts: Association Rules

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



- Find all the rules $X \rightarrow Y$ with minimum support and confidence
  - support, *s*, probability that a transaction contains X U Y
  - confidence, *c*, conditional probability that a transaction having X also contains *Y*

Let minsup = 50%, minconf = 50%

Freq. Pat.: Juice:3, Nuts:3, Diaper:4, Eggs:3, {Juice, Diaper}:3

- Association rules: (many more!)
  - *Juice* → *Diaper* (3/5=60%, 3/3=100%)
  - *Diaper* → *Juice* (3/5=60%, ¾=75%)

Exercise 1

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

- List any three association rules
  - 1.
  - 2.
  - 3.
- Find the support of these rules
  - 1.
  - 2.
  - 3.
- Find the confidence of these rules
  - 1.
  - 2.
  - 3.
- If the minsup is 40%, which association rules are frequent?

Exercise 1 – Sol

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

- List any three association rules
  - 1. {Milk, Eggs} => {Juice}
  - 2. {Coffee} => {Diaper}
  - 3. {Juice, Nuts, Milk} => {Coffee, Eggs}
- Find the support of these rules
  - 1. {Milk, Eggs} => {Juice} : 0/5 = 0%
  - 2. {Coffee} => {Diaper} :2/5 = 40%
  - 3. {Juice, Nuts, Milk} => {Coffee, Eggs} : 0
- Find the confidence of these rules
  - 1. {Milk, Eggs} => {Juice} : 0
  - 2. {Coffee} => {Diaper} : 2/2 = 100%
    - {Diaper} => {Coffee} : 2/4 = 50%
  - 3. {Juice, Nuts, Milk} => {Coffee, Eggs} : 0
- If the minsup is 40%, which association rules are frequent?

Association Rule Mining

- Association rule mining can be viewed as a two-step process
  - Find all frequent itemsets: items satisfying minimum support
  - Generate strong association rules from the frequent itemsets: these rules must satisfy minimum support and minimum confidence
- Second step is much less costly than the first
- Overall performance of mining association rules is determined by the first step

Computational Complexity of Frequent Itemset Mining

- How many itemsets are potentially to be generated in the worst case?
  - The number of frequent itemsets to be generated is senstive to the minsup threshold
  - When minsup is low, there exist potentially an exponential number of frequent itemsets
  - The worst case: 2^N (number of subsets) where N: # distinct items
- The worst case complexty vs. the expected probability
  - Ex. Suppose Walmart has 10^4 kinds of products
  - Total number of possible itemsets = 2^{10^4}

Chapter 5: Mining Frequent Patterns, Association and Correlations: Basic Concepts and Methods

- Basic Concepts
- Frequent Itemset Mining Methods
- Which Patterns Are Interesting?—Pattern Evaluation Methods

Scalable Frequent Itemset Mining Methods

- Apriori: A Candidate Generation-and-Test Approach
- Improving the Efficiency of Apriori
- FPGrowth: A Frequent Pattern-Growth Approach

The Downward Closure Property and Scalable Mining Methods

- The downward closure property of frequent patterns
  - Any subset of a frequent itemset must be frequent
  - If **{juice, diaper, nuts}** is frequent, so is **{juice, diaper}**
  - i.e., every transaction having {juice, diaper, nuts} also contains {juice, diaper}
- Scalable mining methods: Three major approaches
  - Apriori (Agrawal & Srikant@VLDB'94)
  - Freq. pattern growth (FPgrowth—Han, Pei & Yin @SIGMOD'00)
  - Vertical data format approach (Charm—Zaki & Hsiao @SDM'02)

Apriori: A Candidate Generation & Test Approach

- Apriori pruning principle: If there is any itemset which is infrequent, its superset should not be generated/tested!
  Method:
  - Initially, scan DB once to get frequent 1-itemset
  - Generate length (k+1) candidate itemsets from length k frequent itemsets
  - Test the candidates against DB
  - Terminate when no frequent or candidate set can be generated

The Apriori Algorithm—An Example

$Sup_{min} = 2$

Database TDB

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

$\xrightarrow{1^{st}\text{ scan}}$

C_1

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

$\longrightarrow$

L_1

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

C_2

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

$\xleftarrow{2^{nd}\text{ scan}}$

C_2

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

$\longleftarrow$

L_2

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

C_3

Itemset
{B, C, E}

$\xrightarrow{3^{rd}\text{ scan}}$

L_3

Itemset	sup
{B, C, E}	2

Exercise 2

- Find all the frequent itemsets with minsup = 3
- Create 3 association rules from the frequent itemsets
- Find support and confidence of the association rules

Tid	Items
1	A, B, D
2	B, C, D, E
3	A, B, D, F
4	B, C, D, F
5	A, B, C, D
6	B, D, E

Exercise 2

Tid	Items
1	A, B, D
2	B, C, D, E
3	A, B, D, F
4	B, C, D, F
5	A, B, C, D
6	B, D, E

- Find all the frequent itemsets with minsup = 3

A	3
B	6
C	3
D	6
E	2
F	2

AB	3
AC	1
AD	3
BC	3
BD	6
CD	3

ABD	3
BCD	3

  - A,B,C,D,AB,AD,BC,BD,CD,ABD,BCD
- Create 3 association rules from the frequent itemsets
  - A=>B, B=>A, A=>BD, …
- Find support and confidence of the association rules

A=>B (3/6=50%, 3/3=100%)
B=>A (3/6=50%, 3/6=50%)
A=>D (…)
…

A=>BD (3/6=50%, 3/3=100%)
B=>AD
D=>AB
AB=>D

Assignment

TID	Items
10	A,B,C,D,E
20	B,D,E
30	C,D
40	B,C,D,E
50	D,E
60	C,D,E

1. Use Apriori algorithm to find all the frequent itemsets with minimum support count of 2.
2. Find the support and confidence of the following rules

   A=>D,E

   D=>A,E

   A=>D

   C=>D

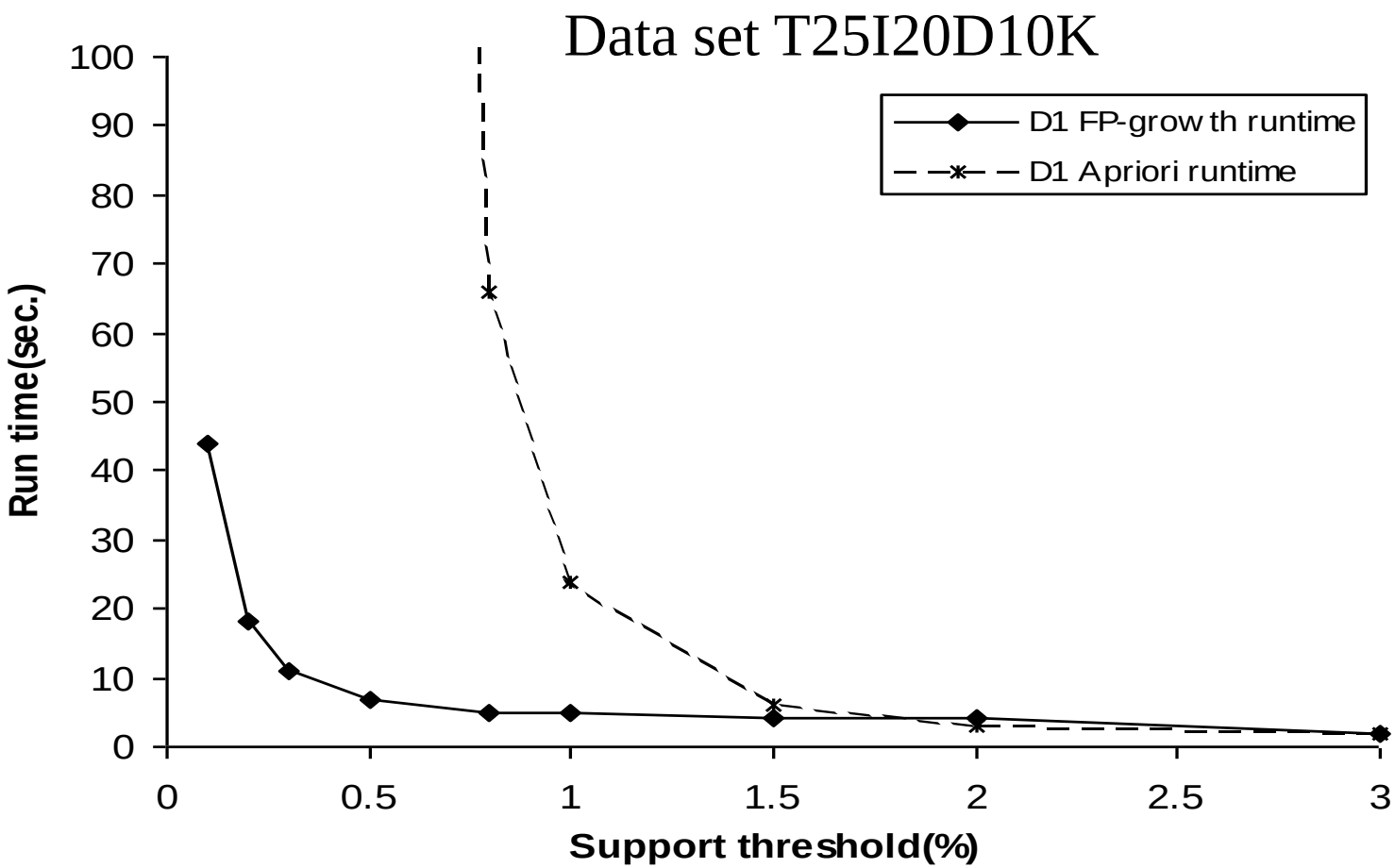
Scalable Frequent Itemset Mining Methods

- Apriori: A Candidate Generation-and-Test Approach
- Improving the Efficiency of Apriori
- FPGrowth: A Frequent Pattern-Growth Approach

Further Improvement of the Apriori Method

- Major computational challenges
  - Multiple scans of transaction database
  - Huge number of candidates
  - Tedious workload of support counting for candidates
- Improving Apriori: general ideas
  - Reduce passes of transaction database scans
  - Shrink number of candidates
  - Facilitate support counting of candidates

FP-Growth vs. Apriori: Scalability With the Support Threshold



Chapter 5: Mining Frequent Patterns, Association and Correlations: Basic Concepts and Methods

- Basic Concepts
- Frequent Itemset Mining Methods
- Which Patterns Are Interesting?—Pattern Evaluation Methods

Interestingness Measure: Correlations (Lift)

- Measure of dependent/correlated events: lift

A ⇒ B [support, confidence, lift].
Lift<1 => negatively correlated, means occurrence of one likely leads to absence of the other
=1 independent,
Lift>1 => positively correlated

$$lift = \frac{P(A \cup B)}{P(A)P(B)}$$

$$lift(i, W) = \frac{2000/5000}{3750/5000 * 3000/5000} = 0.88$$

$$lift(i, \neg W) = \frac{1750/5000}{3750/5000 * 2000/5000} = 1.16$$

	WhatsApp	No WhatsApp	Sum (row)
iPhone	2000	1750	3750
No iPhone	1000	250	1250
Sum(col.)	3000	2000	5000

Exercise 3

- Find the lift of the rule "noIphone => noWhatsApp" based on the previous slide

$$lift(\neg i, \neg W) = \frac{\frac{250}{5000}}{\left(\frac{1250}{5000}\right) * \left(\frac{2000}{5000}\right)}$$

- Find the lift of the 3 association rules found in Exercise 2

A=>B, B=>A, A=>BD

Tid	Items
1	A, B, D
2	B, C, D, E
3	A, B, D, F
4	B, C, D, F
5	A, B, C, D
6	B, D, E

Exercise 3

- Find the lift of the 3 association rules found in Exercise 2

A=>B

$$lift(A,B) = \frac{\frac{3}{6}}{\left(\frac{3}{6}\right) * \left(\frac{6}{6}\right)}$$

B=>A

$$lift(B,A) = \frac{\frac{3}{6}}{\left(\frac{6}{6}\right) * \left(\frac{3}{6}\right)}$$

A=>BD

$$lift(A,BD) = \frac{\frac{3}{6}}{\left(\frac{3}{6}\right) * \left(\frac{6}{6}\right)}$$

$$lift = \frac{P(A \cup B)}{P(A)P(B)}$$

Tid	Items
1	A, B, D
2	B, C, D, E
3	A, B, D, F
4	B, C, D, F
5	A, B, C, D
6	B, D, E