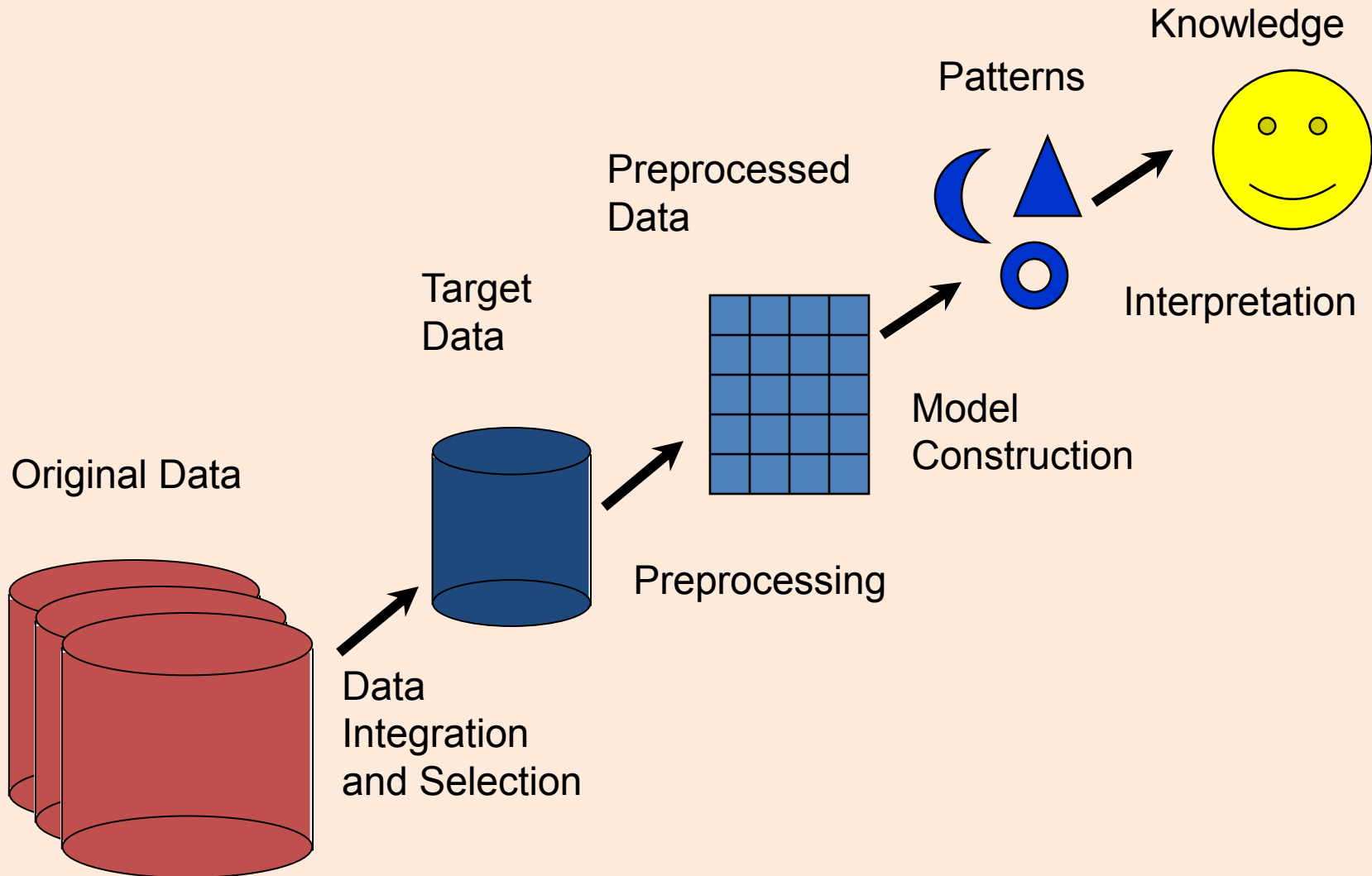


ICS 624 Spring 2013

Association Rule Mining

Asst. Prof. Lipyeow Lim
Information & Computer Science Department
University of Hawaii at Manoa

The Knowledge Discovery Process



Market Basket Analysis

- Consider shopping cart filled with several items
- Market basket analysis tries to answer the following questions:
 - Who makes purchases?
 - What do customers buy together?
 - In what order do customers purchase items?

Market Basket Analysis: Data

Given:

- A database of customer transactions
- Each transaction is a set of items
- Example:
Transaction with TID 111 contains items {Pen, Ink, Milk, Juice}

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	201	5/10/99	Pen	1
113	201	5/10/99	Milk	1
114	201	6/1/99	Pen	2
114	201	6/1/99	Ink	2
114	201	6/1/99	Juice	4
114	201	6/1/99	Water	1

Market Basket Analysis: “Queries”

- Co-occurrences

- 80% of all customers purchase items X, Y and Z together.



“Itemset”

- Association rules

- 60% of all customers who purchase X and Y also buy Z.

- Sequential patterns

- 60% of customers who first buy X also purchase Y within three weeks.

Frequent Itemsets

- An **itemset** (aka co-occurrence) is a set of items
- The **support** of an itemset $\{A,B,\dots\}$ is the fraction of transactions that contain $\{A,B,\dots\}$
 - $\{X,Y\}$ has support **s** if $P(XY) = s$
- **Frequent itemsets** are itemsets whose support is higher than a user specified minimum support *minsup*.
- The ***a priori* property**: Every subset of a frequent itemset is also a frequent itemset.

Frequent Itemset Examples

- {Pen, Ink, Milk}
 - Support: 50%
- {Pen, Ink}
 - Support: 75%
- {Ink, Milk}
 - Support: 50%
- {Pen, Milk}
 - Support: 75%
- {Milk, Juice}
 - support: ?

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	201	5/10/99	Pen	1
113	201	5/10/99	Milk	1
114	201	6/1/99	Pen	2
114	201	6/1/99	Ink	2
114	201	6/1/99	Juice	4
114	201	6/1/99	Water	1

Finding Frequent Itemsets

- Find all itemsets with support $> 75\%$

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	201	5/10/99	Pen	1
113	201	5/10/99	Milk	1
114	201	6/1/99	Pen	2
114	201	6/1/99	Ink	2
114	201	6/1/99	Juice	4
114	201	6/1/99	Water	1

A Priori Algorithm

- Foreach item
 - Check if it is a frequent itemset
- $k = 1$
- Repeat
 - Foreach new frequent itemset I_k with k items
 - Generate all itemsets I_{k+1} with $k+1$ items, $I_k \subset I_{k+1}$
 - Scan all transactions once and check if the generated $(k+1)$ -itemsets are frequent
 - $k = k+1$
- Until no new frequent itemsets are identified

Association Rules

- Rules of the form: LHS \Rightarrow RHS
- Example: {Pen} \Rightarrow {Ink}
 - “if pen is purchased in a transaction, it is likely that ink is also purchased in the same transaction”
- **Confidence** of a rule:
 - $X \rightarrow Y$ has confidence **c** if $P(Y|X) = c$
- **Support** of a rule:
 - $X \rightarrow Y$ has support **s** if $P(XY) = s$

Example

- $\{\text{Pen}\} \Rightarrow \{\text{Milk}\}$
 - Support: 75%
 - Confidence: 75%
- $\{\text{Ink}\} \Rightarrow \{\text{Pen}\}$
 - Support: 75%
 - Confidence: 100%
- $\{\text{Milk}\} \Rightarrow \{\text{Juice}\}$
 - support: ?
 - Confidence: ?

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	201	5/10/99	Pen	1
113	201	5/10/99	Milk	1
114	201	6/1/99	Pen	2
114	201	6/1/99	Ink	2
114	201	6/1/99	Juice	4
114	201	6/1/99	Water	1

Finding Association Rules

- Can you find all association rules with support $\geq 50\%$?

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	201	5/10/99	Pen	1
113	201	5/10/99	Milk	1
114	201	6/1/99	Pen	2
114	201	6/1/99	Ink	2
114	201	6/1/99	Juice	4
114	201	6/1/99	Water	1

Association Rule Algorithm

Goal: find association rule with given support *minsup* and given confidence *minconf*

- Step 1: Find frequent itemsets with support *minsup*
- Step 2: Foreach frequent itemset,
 - Foreach possible split into LHS=>RHS
 - Compute the confidence as $\text{support(LHS,RHS)} / \text{support(LHS)}$ and compare with *minconf*

Variations

- Association rules with isa hierarchies
 - Items in transactions can be grouped into subsumption hierarchies (like dimension hierarchies)
 - Items in itemsets can be any node in the hierarchy
 - Example:
 - $\text{Support}(\{\text{Ink}, \text{Juice}\}) = 50\%$
 - $\text{Support}(\{\text{Ink}, \text{Beverage}\}) = 75\%$
- Association rules on time slices
 - Eg. Find association rules on transactions occurring on the first of the month
 - Confidence and support within these “slices” will be different than over the entire data set.