Pattern mining

Basic concepts and methods

These slides are based on the slides created and provided by the authors of the course textbook Data Mining Concepts and Techniques

Pattern Mining: Basic Concepts and Methods

- Basic Concepts
- Frequent Itemset Mining Methods – Apriori Algorithm
- Which Patterns Are Interesting? - Pattern Evaluation Methods

Pattern Discovery: Basic Concepts

- **Basic Concepts**
  - What Is Pattern Discovery? Why Is It Important?
  - Bäsic Concepts: Frequent Patterns and Association Rules
  - Compressed Representation: Closed Patterns and Max-Patterns

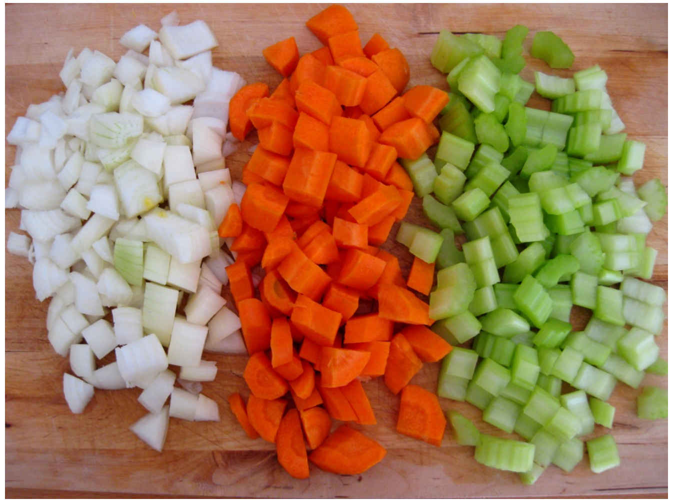
What Is Pattern Discovery?

Motivating examples:

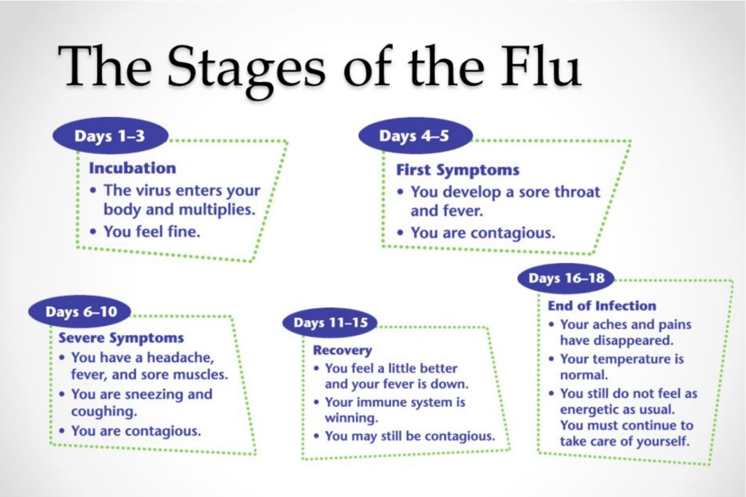
- What products were often purchased together?
- What are the subsequent purchases after buying an iPad?
- What code segments likely contain copy-and-paste bugs?
- What word sequences likely form phrases in this corpus?

What Are Patterns?

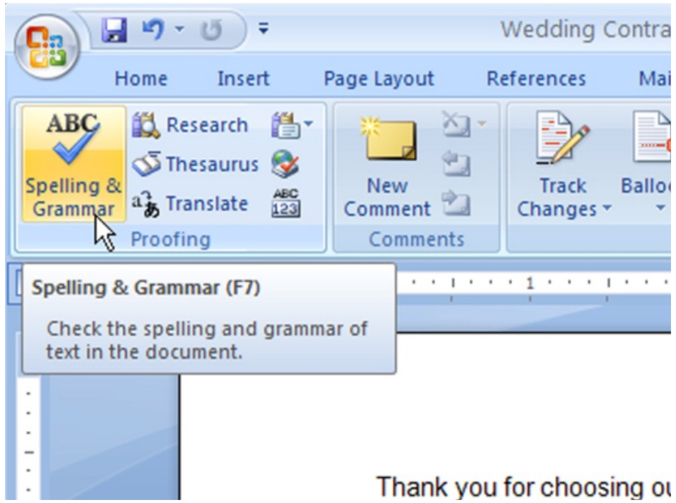
- What are patterns?
  - Patterns: A set of items, subsequences, or substructures that occur frequently together (or strongly correlated) in a data set
  - Patterns represent intrinsic and important properties of datasets



Frequent item set



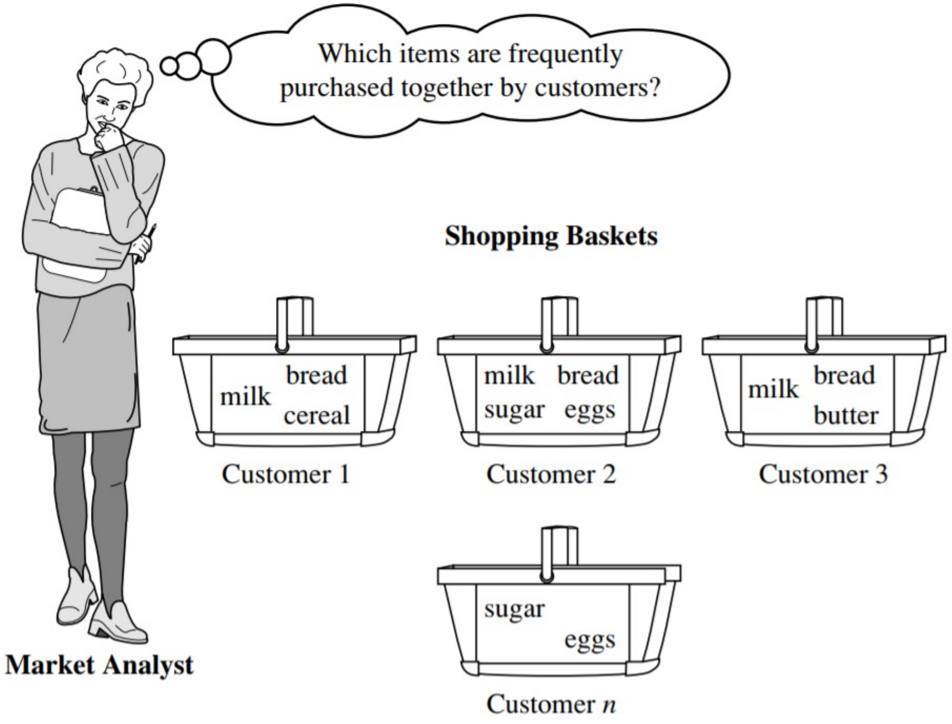
Frequent sequences



Frequent structures

What Is Pattern Discovery?

- Pattern discovery:
  Uncovering patterns from massive data sets
- It can answer questions such as:
  - What products were often purchased together?
  - What are the subsequent purchases after buying an iPad?



Pattern Discovery: Why Is It Important?

- 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
  - **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
  - Many types of data: spatiotemporal, multimedia, time-series, and stream data

Pattern Discovery: Basic Concepts

- **Basic Concepts**
    - What Is Pattern Discovery? Why Is It Important?
    - Bäsic Concepts: Frequent Patterns and Association Rules
    - Compressed Representation: Closed Patterns and Max-Patterns

Basic Concepts: Transactional Database

- Transactional Database (TDB)
  - Each transaction is associated with an identifier, called a TID.
  - May also have counts associated with each item sold

Tid	Items bought
1	Beer, Nuts, Diaper
2	Beer, Coffee, Diaper
3	Beer, Diaper, Eggs
4	Nuts, Eggs, Milk
5	Nuts, Coffee, Diaper, Eggs, Milk

Basic Concepts: k-Itemsets and Their Supports

- Itemset: A set of one or more items

$$I = \{ I_1, I_2, \cdots, I_m \}$$

- k-itemset: An itemset containing k items:

  $X = \{x_1, \ldots, x_k\}$

  - Ex. {Beer, Nuts, Diaper} is a 3-itemset

Tid	Items bought
1	Beer, Nuts, Diaper
2	Beer, Coffee, Diaper
3	Beer, Diaper, Eggs
4	Nuts, Eggs, Milk
5	Nuts, Coffee, Diaper, Eggs, Milk

- Absolute support (count)
  - sup{X} = occurrences of an itemset X
    - Ex. sup{Beer} = 3
    - Ex. sup{Diaper} = 4
    - Ex. sup{Beer, Diaper} = 3
    - Ex. sup{Beer, Eggs} = 1

- 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%

Basic Concepts: Frequent Itemsets (Patterns)

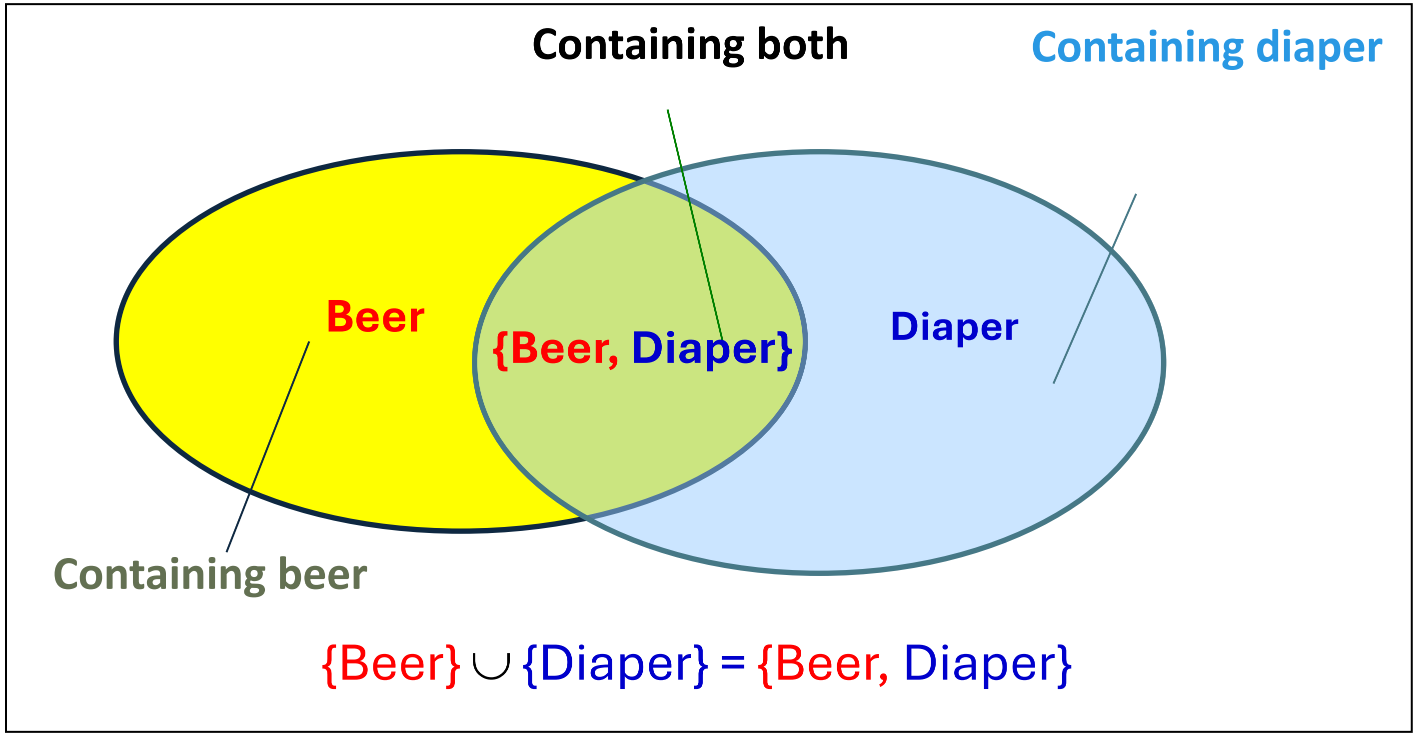
- An itemset (or a pattern) X is *frequent* if the support of X is no less than a *minsup* threshold σ
- Let σ = *50%* (σ: *minsup* threshold) for the given 5-transaction dataset
  - All the frequent 1-itemsets:
    - Beer: 3/5 (60%); Nuts: 3/5 (60%); Diaper: 4/5 (80%); Eggs: 3/5 (60%)
  - All the frequent 2-itemsets:
    - {Beer, Diaper}: 3/5 (60%)
  - All the frequent 3-itemsets?
    - None

Tid	Items bought
1	Beer, Nuts, Diaper
2	Beer, Coffee, Diaper
3	Beer, Diaper, Eggs
4	Nuts, Eggs, Milk
5	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

- Compared to itemsets, association rules can be more telling
  - Ex. *Diaper → Beer*
    - *Buying diapers may likely lead to buying beers*



Note: X ∪ Y: the union of two itemsets

■ The set contains both X and Y

From Frequent Itemsets to Association Rules

- How do we compute the strength of an association rule $X \rightarrow Y$ (Both X and Y are itemsets)?
- We first compute the following two metrics, s and c.
  - Support of $X \cup Y$
    - Ex. s{Diaper, Beer} = 3/5 = 0.6 (i.e., 60%)
  - Confidence of $X \rightarrow Y$
    - The *conditional probability* that a transaction containing X also contains *Y:*

      $c = \sup(X, Y) / \sup(X)$
    - Ex. *c* = sup{Diaper, Beer}/sup{Diaper} = ¾ = 0.75

Tid	Items bought
1	Beer, Nuts, Diaper
2	Beer, Coffee, Diaper
3	Beer, Diaper, Eggs
4	Nuts, Eggs, Milk
5	Nuts, Coffee, Diaper, Eggs, Milk

- In pattern analysis, we are often interested in those rules that dominate the database, and these two metrics ensure the popularity and correlation of 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%*
  - Freq. 1-itemsets: Beer: 3, Nuts: 3, Diaper: 4, Eggs: 3
  - Freq. 2-itemsets: {Beer, Diaper}: 3

- Let *minconf = 50%*
  - *Beer* → *Diaper* (60%, 100%)
  - *Diaper* → *Beer* (60%, 75%)

(Q: Are these all the rules satisfying the two conditions?)

Tid	Items bought
1	Beer, Nuts, Diaper
2	Beer, Coffee, Diaper
3	Beer, Diaper, Eggs
4	Nuts, Eggs, Milk
5	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

Association Rule Mining: two-step process

1. **Find all frequent itemsets**:
   - By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, min_sup.
   - This step is computationally expensive
2. **Generate strong association rules from the frequent itemsets:**
   - By definition, these rules must satisfy minimum support and minimum confidence.
   - This step is computationally **in**expensive

Because of this, the overall performance is determined by step 1