Association Rules Fundamentals

Elena Baralis, Silvia Chiusano, Tania Cerquitelli
Politecnico di Torino

Considering weight

- Items may be characterized by different importance within a transaction

  - Example: product quantity or price in transactions

- Transactions may be weighted

  - Example: discount on entire market basket

- Weighted dataset

  - Each item is assigned a weight measuring its relevance in the corresponding transaction

Weighted association rules

- Consider item/transaction weights during association rule extraction

- Extend rule quality measures
    - E.g., weighted support, weighted confidence

- Apply ad-hoc weight aggregation functions
    - E.g., min, max, avg

# Considering hierarchies

- **Generalization hierarchies**
  - Aggregation over attributes in a dataset
  - Typically user provided
- **Examples**
  - Time hierarchy
  - Product category
  - Location hierarchy
  - ...

Example of Hierarchical Category
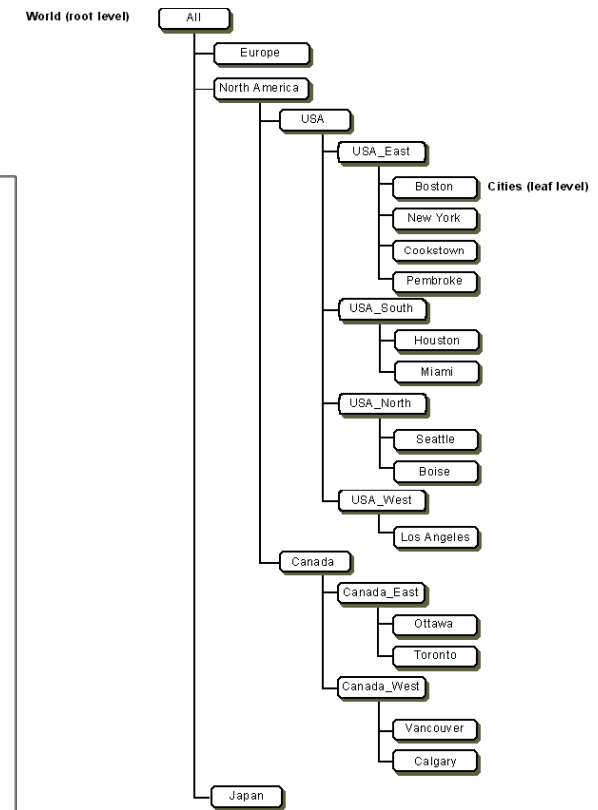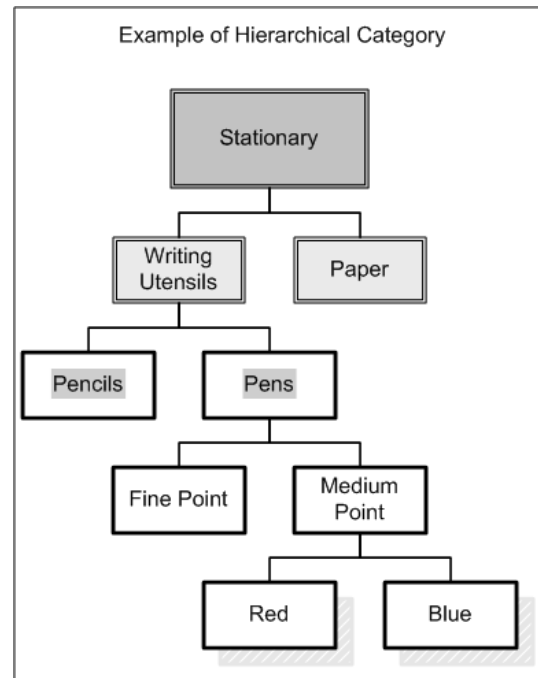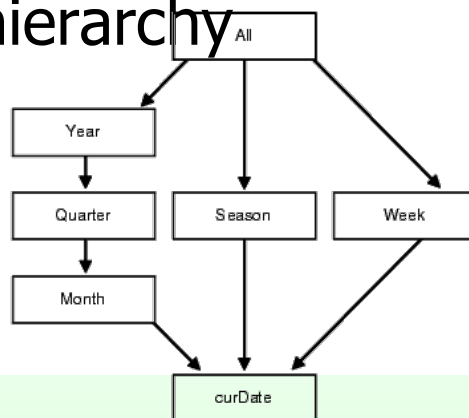
Taxonomy

- A taxonomy is a set of is-a hierarchies that aggregate data items into higher-level concepts

- Data item
    - Instance in the (transactional) dataset
    - Represents detailed concepts

- Generalized item
    - Aggregation in higher-level concepts
    - Represents abstractions on instances
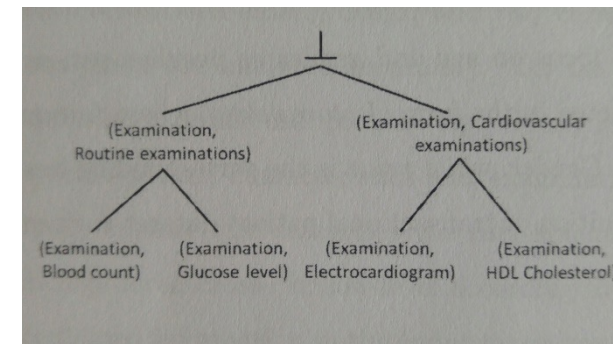
Generalized itemsets

- Sets of items at different generalization levels
    - May contain data items together with generalized items defined in the taxonomy
    - Summarize knowledge represented by a set of lower-level descendants
        - Both frequent and infrequent
- A generalized itemset covers a transaction when all
    - its generalized items are ancestors of items included in the transaction
    - its data items are included in the transaction
- Generalized itemset support
    - ratio between number of covered transactions and dataset cardinality

Context-aware data analysis

- Context data provided by different, possibly heterogeneous, sources
  - Mobile devices provide information on
    - the user context (e.g., GPS coordinates)
    - the supplied services
      - temporal information
      - service description
      - duration
  - Additional information available
    - demographics of the user requesting the service

- A very low support

  user: John, **time:** 6.05 p.m., **service:** Weather
  (s = 0.005%)

  - The itemset may be discarded

  - By generalizing
    - the time attribute on a time period
    - the user on a user category

  user: employee, **time:** 6 p.m. to 7 p.m., **service:** Weather
  (s = 0.2%)

  - May discover interesting properties generalizing *infrequent* items

Generalized association rules

- Extension of "classical" association rules

$$X \rightarrow Y$$

- X and Y are either generalized or not generalized itemsets
  - Extract associations among data items at multiple abstraction levels
  - Support, confidence and lift are defined accordingly

Patient data analysis

- **Analysis of multiple level correlations on patient treatment historical data**
  - Dataset collected by an Italian Local Health Center
    - Diabetes complications at various severity levels
    - 95K records, 3.5K patients
  - Features
    - Prescribed examinations (26 examinations, 7 categories)
    - Prescribed drugs (200 drugs, 14 categories)
    - Census patient data (gender, age discretized in age groups)

- **Sparse dataset**
  - Difficult setting of support threshold
    - Low: generates too many rules
    - High: interesting information at lower levels of abstraction may remain hidden
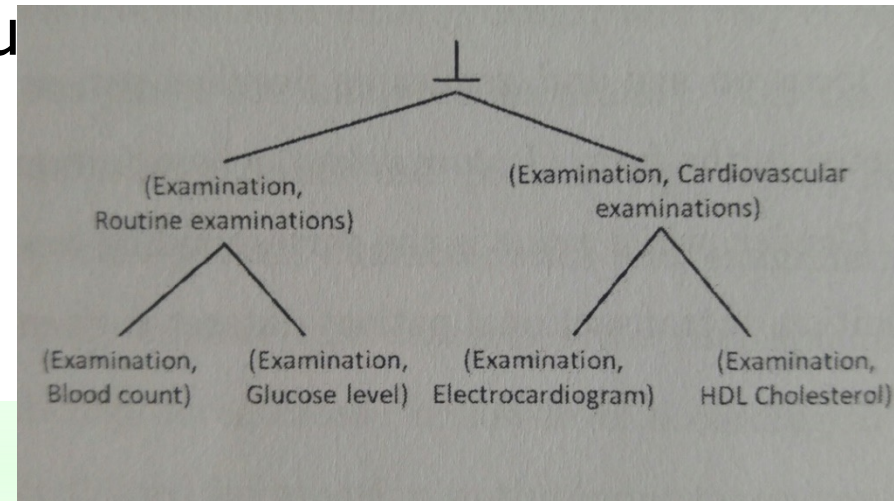
High level rules

- **Only generalized itemsets**
  - Represent general knowledge
    - May be too high level to perform targeted analyses

- **Example**

  (Examination, Liver) -> (Examination, Kidney)

  - Frequently prescribed together
  - May be used for examination schedu

Cross level rules

- **Different abstraction levels (generalized items and data items)**
    - Combine detailed and general information
- **Example**

    (Examination, Liver) -> (Examination, Uric acid)

    - Insight into specific kidney examinations correlated with liver examinations
        - Confidence: 74.8%

Low-lewel rules

- Only not generalized itemsets (only data items)
  - Very detailed knowledge
    - Covered by high and cross-level rules
  - Large rule set
    - Challenging exploration task
  - Drill down exploration based on formerly extracted high and cross-level rules