Knowledge Discovery in Databases
SS 2016

Chapter 2: Data Representation

Lecture: Prof. Dr. Thomas Seidl

Tutorials: Julian Busch, Evgeniy Faerman, Florian Richter, Klaus Schmid
Overview

- Introduction
- Data presentation
- Data management, storage architecture
Objects carry information

- Objects and attributes
  - Entity-Relationship diagram (ER)
  - UML class diagram
  - Data tables (relational model)

```
<table>
<thead>
<tr>
<th>Name</th>
<th>sem.</th>
<th>major</th>
<th>skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ann</td>
<td>3</td>
<td>CS</td>
<td>Java, C, R</td>
</tr>
<tr>
<td>Bob</td>
<td>1</td>
<td>CS</td>
<td>Java, PHP</td>
</tr>
<tr>
<td>Charly</td>
<td>4</td>
<td>History</td>
<td>Piano, …</td>
</tr>
<tr>
<td>Debra</td>
<td>2</td>
<td>Arts</td>
<td>Painting, …</td>
</tr>
</tbody>
</table>
```
Overview of (attribute) data types

- Simple data types
  - Numerical, categorical, ordinal

- Composed data types
  - Sets, sequences, vectors

- Complex data types
  - Multimedia: images, videos, audio, text, documents, web pages, etc.
  - Spatial, geometric: shapes, molecules, geography, etc.
  - Structures: graphs, networks, trees, etc.

- Examples for complex objects
  - Molecules: shape + structure + physical-chemical properties + …
  - City maps: shapes + traffic networks + points of interests + …
  - Mechanical parts: shape + physical properties + production process descr.
Simple data types and comparisons

• Numeric data
  – Numbers: natural, integer, rational, real numbers
  – Examples: age, income, shoe size, height, weight
  – Comparison: difference
  – Example: 3 is more similar to 30 than to 3,000

• Generalization: metric data
  – Metric space \((O, d)\) consists of object set \(O\) and metric distance \(d\)
  – Comparison by (metric) distance \(d: O \times O \rightarrow \mathbb{R}_0^+\)
    • Symmetry: \(\forall p, q \in O: d(p, q) = d(q, p)\)
    • Identity of indiscernibles \(\forall p, q \in O: d(p, q) = 0 \iff p = q\)
    • Triangle inequality \(\forall p, q, o \in O: d(p, q) \leq d(p, o) + d(o, q)\)
  – Example: points in 2D space – Euclidean distance
Simple data types and comparisons

- Numeric data, metric data
- Categorical data
  - „Just identifiers“
    - Example occupation = \{ butcher, hairdresser, physicist, physician, … \}
    - Example subjects = \{ physics, biology, math, music, literature, history, EE, … \}

- Comparison: how to compare values ???
  - Trivial metric: \( d(p, q) = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{else} \end{cases} \)
    - Always works but is quite coarse
  - Generalization hierarchies can help
    - Path length seems appropriate
    - \( d(\text{music}, \text{literature}) = 2 \)
    - \( d(\text{music}, \text{biology}) = 4 \)
    - \( d(\text{music}, \text{music}) = 0 \)
Simple data types and comparisons

- **Numeric data, metric data**
- **Categorical data, hierarchical types**
- **Ordinal data**
  - Some data carry a (total) order \((O, \leq)\)
    - **Transitivity** \(\forall p, q, o \in O: p \leq q \land q \leq o \Rightarrow p \leq o\)
    - **Antisymmetry** \(\forall p, q \in O: p \leq q \land q \leq p \Rightarrow p = q\)
    - **Totality** \(\forall p, q \in O: p \leq q \lor q \leq p\)
  - **Examples**
    - **Numbers** \(3 \leq 30 \leq 3,000\)
    - **Words** \(\text{high} \leq \text{highschool} \leq \text{highscore}\) (i.e., lexicographical order)
    - **Frequencies** „How often did you sleep bad last year?“
      - never \(\leq\) seldom \(\leq\) rarely \(\leq\) occasionally \(\leq\) sometimes \(\leq\) often \(\leq\) frequently \(\leq\) regularly \(\leq\) usually \(\leq\) always
    - **(Vague) sizes** „How big was that problem?“
      - tiny \(\leq\) small \(\leq\) medium \(\leq\) big \(\leq\) huge
Composed data types

- Sets
  - Put individual values together
  - Example: $skills = \emptyset(\{\text{Java, C, Python, R, ...}\})$
  - Comparison
    • Symmetric set difference: $d(R, S) = (R - S) \cup (S - R) = (R \cup S) - (R \cap S)$
    • Jaccard distance:
      \[
      d(R, S) = \frac{(R \cup S) - (R \cap S)}{R \cup S}
      \]

- Bitvector representation of a set on a given, ordered base set
  - Sample base $B = \{\text{math, physics, chemistry, biology, music, arts, english}\}$
  - Example sets $S = \{\text{math, music, english}\} = \langle 1,0,0,0,1,0,1 \rangle$
    $R = \{\text{math, physics, arts, english}\} = \langle 1,1,0,0,0,1,1 \rangle$
  - Hamming distance = sum of different entries: $d(R, S) = 3$
    - Equals the symmetric set difference
Composed data types

- Sequences, vectors
  - Put $n$ values of a domain $D$ together
  - Order does matter: $I_n \rightarrow D$ for an index set $I_n = \{1, \ldots, n\}$

- Comparison of vectors: two steps
  - Determine individual differences or distances $d(o_i, q_i)$
  - Combine individual distances to overall distance $d(o, q)$

- Examples
  - (Simple) sum: $d_1(o, q) = \sum_{i=1}^{n} d(o_i, q_i)$ (Manhattan)
  - Root of sum of squares: $d_2(o, q) = \sqrt{\sum_{i=1}^{n} (o_i - q_i)^2}$ (Euclidean)
  - Maximum: $d_\infty(o, q) = \max_{i=1,\ldots,n} \{|o_i - q_i|\}$
  - General formula: $d_p(o, q) = \sqrt[p]{\sum_{i=1}^{n} |o_i - q_i|^p}$ (Minkowski)
  - Weighted Minkowski dist.: $d_{p,w}(o, q) = \sqrt[p]{\sum_{i=1}^{n} w_i \cdot |o_i - q_i|^p}$