Containment and Complementarity Relationships in Multidimensional Linked Open Data

Marios Meimaris and George Papastefanatos
Institute for the Management of Information Systems
Research Center “Athena”

{m.meimaris, gpapas}@imis.athena-innovation.gr
Multidimensional data

• Schema
  – Dimensions
  – Measures
  – Attributes
  – Code lists

• Data
  – Observations
Multidimensional Linked Data

- Origin of different source datasets
- LD recommendations and Best Practices provide **common grounds** across remote sources
- RDF Data cube\(^1\) provides a common meta-schema
- Re-use of:
  - Dimension properties
  - Measure properties
  - Code lists
  - **Hierarchies**
- In case of no re-use, mapping/alignment is needed

Problem tackled

• Relating points in multidimensional data spaces semantically
• Bulk detection and computation of containment and complementarity relationships between observations
  – in the same dataset or
  – in different datasets
• Observation relationships are useful for:
  – performing OLAP analytics over multidimensional, multi-dataset data spaces
  – computing similarities/distances between observations
  – Suggestion mechanisms for relevant statistics
  – Exploratory analysis and discovery

Meimaris@SEMSTATS2014
Observations are related

• We identify two (non-exhaustive) types of relationships:
  – Observation *containment*
  – Observation *complementarity*
Observation Complementarity

- Two observations complement each other when they provide different information for the same point in the data space.

\[ (P_a \subseteq P_b) \land (\forall p_i \in P_a \cap P_b : h^i_a = h^i_b) \land (\forall p_j \in P_b \setminus P_a : h^j_b = c_{\text{root}}) \]

- \( P_k \): the set of dimension properties for observation I
- \( p_i \): a single dimension property
- \( h^m_i \): the value of property m for observation I
- \( c_{\text{root}} \): the top (root) concept for all hierarchies
Observation Containment

• An observation contains another observation when it is a *partial* or *full* generalization of the latter w.r.t to their shared dimension values

• *Full* containment vs *Partial* containment
  – Full containment means that a contained/containing observation can be directly rolled-up/drilled-down to the containing/contained observation,
  – Partial containment means that both contained and containing observation must be *rolled-up on their disjoint dimensions* to complement each other

  full \( (\exists M_i \in M_a \cap M_b) \land (P_a \subseteq P_b) \land (\forall p_i \in P_a \cap P_b : h_a^i \succeq h_b^i) \)

  partial \( (\exists M_i \in M_a \cap M_b) \land (P_a \subseteq P_b) \land (\exists p_i \in P_a \cap P_b : h_a^i \succ h_b^i) \)
## Containment example

<table>
<thead>
<tr>
<th>location</th>
<th>time</th>
<th>sex</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italy</td>
<td>2012</td>
<td>Total</td>
<td>59,478,000</td>
</tr>
<tr>
<td>Riva del Garda</td>
<td>2012</td>
<td>Male</td>
<td>15,100</td>
</tr>
<tr>
<td>Trentino</td>
<td>2012</td>
<td>Female</td>
<td>248,400</td>
</tr>
</tbody>
</table>

Hierarchy is reflexive (i.e. a value is a parent of itself)
Computation

1. Build the feature space
2. Group by dimension / measure
3. Extract containment per dimension / measure
4. Compute overall containment scores and classify as full or partial
5. Compute complementarity scores
1. Build the feature space into an **occurrence matrix**
   - Each dimension value is a feature
   - Encoded is the hierarchy of features (1 for occurrence and all parents, 0 otherwise)

|     | refArea |           |           |           |           |           |           |           |           | refPeriod |           |           |           |           | M | F | T |
|-----|---------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|---|---|---|
| obs1 | WLD     | EUR       | AM        | GR        | IT        | Ath       | Rom       | US        | TX        | ALL       | 2001      | 2011      | Jan11     | Feb11     |   |   |   |
| obs2 | 1       | 1         | 0         | 1         | 0         | 1         | 0         | 0         | 0         | 1         | 1         | 1         | 1         |   |   |   |
| obs3 | 1       | 0         | 1         | 0         | 0         | 0         | 0         | 1         | 1         | 1         | 0         | 1         | 0         |   |   |   |
| obs4 | 1       | 0         | 1         | 0         | 0         | 0         | 0         | 0         | 0         | 1         | 0         | 1         | 0         |   |   |   |
| obs5 | 1       | 0         | 1         | 0         | 0         | 0         | 0         | 0         | 0         | 1         | 0         | 1         | 0         |   |   |   |
| obs6 | 1       | 0         | 1         | 0         | 0         | 0         | 0         | 0         | 0         | 1         | 0         | 1         | 0         |   |   |   |
| obs7 | 1       | 0         | 1         | 0         | 0         | 0         | 0         | 0         | 0         | 1         | 0         | 1         | 0         |   |   |   |
| obs8 | 1       | 0         | 1         | 0         | 0         | 0         | 0         | 0         | 0         | 1         | 0         | 1         | 0         |   |   |   |
| obs9 | 1       | 0         | 1         | 0         | 0         | 0         | 0         | 0         | 0         | 1         | 0         | 1         | 0         |   |   |   |
| obs10| 1       | 0         | 1         | 0         | 0         | 0         | 0         | 0         | 0         | 1         | 0         | 1         | 0         |   |   |   |

Meimaris@SEMSTATS2014
2. For N observations, compute one NxN containment matrix $C_{M_p^m}$ for each dimension $p_m$ in the set of all datasets. Then cell $[i,j]$ becomes:
- 1 if values of dimension are parent-child for observations i and j, or
- 0 otherwise

Function $sf$ to determine this for observations $o_a$ and $o_b$ and dimension $p_m$:

$$sf(o\downarrow a, o\downarrow b)\downarrow p\downarrow m = \begin{cases} 1, & (a \text{ AND } b) = b \text{ otherwise} \\ 0, & \text{otherwise} \end{cases}$$

where a and b are the bit vectors of observations
Containment relationships

3. Adding all containment matrices $CM_{pm}$ yields full and partial containment relationships in an overall containment matrix $OCM$:

$$OCM = \Sigma_{i=1}^{k} u_{i} CM_{i} / \Sigma_{i=1}^{k} u_{i}$$

For observations $o_a$ and $o_b$:

- $o_a$ cont$\_full$ $o_b$ iff $OCM[o_a, o_b]=1$
- $o_a$ cont$\_part$ $o_b$ iff $0 < OCM[o_a, o_b] < 1$
Complementarity relationships

4. Complementarity is computed as follows:

\[ cf(o \downarrow a, o \downarrow b) = \begin{cases} 1, & (sf(o \downarrow a, o \downarrow b) | \downarrow P = 1) \text{ AND } (a = b) \text{ otherwise} \\ 0, & \text{otherwise} \end{cases} \]

where \( P \) the occurrences of dimension properties and \( a, b \) the bit vectors of \( o_a \) and \( o_b \) in the occurrence matrix.

For observations \( o_a \) and \( o_b \):

- \( o_a \) compl\_full \( o_b \) iff \( \text{OCM}[o_a, o_b] > 0 \)

Containment is transitive, complementarity is symmetric
Data Cube Extension
Experimental Evaluation

• Datasets:
  • Population (Eurostat, Worldbank)
  • Internet households (Eurostat)
  • Poverty (Eurostat, Worldbank)
• 6 dimension properties
• 3 measure properties

<table>
<thead>
<tr>
<th>#of obs.</th>
<th>refArea</th>
<th>refPeriod</th>
<th>sex</th>
<th>unit</th>
<th>age</th>
<th>poverty</th>
<th>internet</th>
<th>population</th>
</tr>
</thead>
<tbody>
<tr>
<td>D₁ (539)</td>
<td>85 regions, 20 countries</td>
<td>2004-2011</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>D₂ (1693)</td>
<td>293 regions, 33 countries</td>
<td>2003-2010</td>
<td>N/A</td>
<td>Yes</td>
<td>N/A</td>
<td>Yes</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>D₃ (629)</td>
<td>42 regions, 3 countries</td>
<td>2009-2013</td>
<td>M, F, Total</td>
<td>Yes</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>Yes</td>
</tr>
<tr>
<td>D₄ (316)</td>
<td>65 regions, 7 countries</td>
<td>2009-2013</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>Yes</td>
<td>N/A</td>
</tr>
</tbody>
</table>
## Results - Discussion

- Most new relationships are partial containments (~27% of possible relationships)
- Complementarity is the strictest relationship (0.03% of the total possible observation pairs)
- Relatedness of complementarity to partial/full containment
- ~1.3 million new links between observations

<table>
<thead>
<tr>
<th></th>
<th>D₁</th>
<th>D₂</th>
<th>D₃</th>
<th>D₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>D₁</td>
<td>647 (0.31%) full 34.3k (16.32%) partial N/A compl</td>
<td>N/A full N/A partial N/A compl</td>
<td>N/A full N/A partial N/A compl</td>
<td>N/A full N/A partial N/A compl</td>
</tr>
<tr>
<td>D₂</td>
<td>605 (0.02%) full 605k (14.83%) partial 1238 (0.04%) compl</td>
<td>3370 (0.14%) full 378k (14.83%) partial N/A (complement)</td>
<td>N/A full N/A partial 204 (0.004%) compl</td>
<td>N/A full N/A partial N/A compl</td>
</tr>
<tr>
<td>D₃</td>
<td>N/A full N/A partial N/A compl</td>
<td>N/A full N/A partial N/A compl</td>
<td>1k (0.26%) full 261k (65.9%) partial N/A compl</td>
<td>N/A full N/A partial N/A compl</td>
</tr>
<tr>
<td>D₄</td>
<td>N/A full N/A partial 328 (0.05%) compl</td>
<td>N/A full N/A partial 218 (0.005%) compl</td>
<td>N/A full N/A partial 592 (0.07%) compl</td>
<td>437 (0.17%) full 22.2k (22.3%) partial N/A compl</td>
</tr>
</tbody>
</table>
Future Work

• Suggestion mechanisms based on computed relationships, conduct user studies to evaluate

• Faster and more efficient computations (now $O(N^2)$)
  – Better feature extraction
  – Dimensionality reduction

• Extracting *latent datasets* based on containment and complementarity relationships
Support

• **DIACHRON**
  Managing the Evolution and Preservation of the Data Web

• **KRIPIIS: SODAMAP Project**

• **linked-statistics.gr**