Learning mappings and queries

Marie Jacob
University Of Pennsylvania

DEIS 2010
Schema mappings

- Denote relationships between schemas
- Relates source schema $S$ and target schema $T$
- Defined in a query language like Datalog or first-order logic.

**Data Integration**

*Med schema*

$S_1$ $S_2$ $S_3$

**Data Warehousing**

ETL

**Data Sharing**

Participant $p_2$

Participant $p_3$

Warehouse
Informal problem

• Two variants:
  – Given schemas S and T, and instances I and J, find a set of s-t mappings that “naturally” translate S to T.
  – Given a set of schemas, $S_1, S_2, \ldots, S_n$, find an integrated schema that best reflects combination of all source schemas, and their corresponding s-t mappings.
Mapping Tasks

- iMap [Dhamankar et al, 2004]

Integrated schema

Mapping creation

Schema matching

[Chiticariu et al, 2008]
[Das Sarma et al, 2008]
[Clio [Miller et al, 2000]
Mapping Tasks

S-Match [Giunchigla et al, 2007]
iMap [Dhamankar et al, 2004]

Schema matching

Integrated schema

Mapping creation

[Chiticariu et al, 2008]
[Das Sarma et al, 2008]

Clio [Miller et al, 2000]
Schema matching

• Determine if two attributes relate to each other.
  – Is Employee(id) the same as Emp(eid)?

• Challenges:
  – Heterogeneity.
  – Types of relationships.
  – Complex matches.
S-Match

[Giunchiglia et al, 2007]

- Matches elements in source and target tree-structured models (e.g. XML)
- Abstracts labels into high-level concepts, encoded in description logic.
- Label A has concept $C_A$
- Classifies pairwise concepts, $C_A$, $C_B$:
  - $C_A = C_B$ (equivalent)
  - $C_A \sqsubseteq C_B$ (less general)
  - $C_A \sqsupseteq C_B$ (more general)
  - $C_A \perp C_B$ (disjoint)
S-Match
[Giunchiglia et al, 2007]

1. Compute concept of labels

\[(C_{\text{history}} \cup C_{\text{philosophy}}) \cap C_{\text{science}}\]
2. Compute concepts at nodes:

\[ C_{\text{classes}} \cap (C_{\text{arts}} \cup C_{\text{science}}) \cap C_{\text{college}} \cap (C_{\text{history}} \cup C_{\text{philosophy}}) \cap C_{\text{science}} \]
3. Compute relations between atomic concepts

- $C_{\text{classes}} = C_{\text{courses}}$
- $C_{\text{economics}} \subseteq C_{\text{microeconomics}}$
- $C_{\text{Mathematics}} = C_{\text{Math}}$
4. Compute relationships between nodes
   - Is $C_{\text{classes}} \cap C_{\text{math}}$ the same as $C_{\text{courses}} \cap C_{\text{mathematics}}$?
   - Construct logical implication formula axioms $\rightarrow \text{rel}(C_A, C_B)$
   - If negation is unsatisfiable, $\text{rel}(C_A, C_B)$ holds.

<table>
<thead>
<tr>
<th>$\text{rel}(C_A, C_B)$</th>
<th>Translation to prop. logic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_A = C_B$</td>
<td>$C_A \leftrightarrow C_B$</td>
</tr>
<tr>
<td>$C_A \subseteq C_B$</td>
<td>$C_A \rightarrow C_B$</td>
</tr>
<tr>
<td>$C_A \supseteq C_B$</td>
<td>$C_B \rightarrow C_A$</td>
</tr>
<tr>
<td>$C_A \perp C_B$</td>
<td>$\neg (C_A \land C_B)$</td>
</tr>
</tbody>
</table>
S-Match
[Giunchiglia et al, 2007]

4. Compute relationships between nodes

Is $C_{classes} \cap C_{math} = C_{courses} \cap C_{mathematics}$?

$(C_{classes} \Leftrightarrow C_{courses}) \land$
$(C_{Mathematics} \Leftrightarrow C_{Math})$
$\Rightarrow$
$(C_{Classes} \land C_{Mathematics}) \Leftrightarrow$
$(C_{Courses} \land C_{Math})$?
S-match

- Linguistic techniques a useful approach as attribute names/labels are described using natural language.
- Takes into account source structure.
- Would miss application-specific attribute namings (e.g. eid)
- Does not use type information
iMap [Dhanmankar, 04]

• System for determining complex matches between schemas.
  – Eg. concat(S.fname, S.lname) → T.name

• Searches a space of possible matches:
  – Employing learning techniques
  – Employing domain knowledge

• Designed to be flexible, “plug-in” type architecture
STEP 1: Find match candidates

STEP 2: Generate similarity scores

STEP 3: Select best candidates

S, T, I, J

Custom searchers

Combining scores

Using domain constraints
• Finding candidate matches:

name = concat(fname, lname)

Birth-date = b-day/b-month/b-year

list-price= price * (1+taxrate)
• Searchers:
  – Search strategy: Keep only k-highest scoring candidates for each combination size n.

- \( \text{concat}(\text{S.fname}, \text{S.lname}) = \text{T.name} \)  
  Score = 0.9

- \( \text{S.fname} = \text{T.name} \)  
  Score = 0.3

- \( \text{S.lname} = \text{T.name} \)  
  Score = 0.3
Exploiting Domain Knowledge

- Domain constraints (e.g. name & email-address unrelated)
- Overlap data: Test matches on overlapping data
- External resources: thesaurus

Generally higher accuracy when given domain knowledge.
Schema matching

• Discovering relationships between source and target attributes.

• Variety of work
  – Using instance-based approaches.
  – Using linguistic techniques.
  – Using structural constraints of schemas

• Survey on schema matching
  [Rahm, Bernstein, 2001]
Mapping Tasks

iMap [Dhamankar et al, 2004]

Schema matching

Integrated schema

[Chiticariu et al, 2008]  [Das Sarma et al, 2008]

Mapping creation

Clio [Miller et al, 2000]
Integrated Schemas

• Given:
  – A set of source schemas $S_1, S_2, \ldots, S_n$
  – A set of pairs of source and target attributes (weighted correspondences)

• Find:
  – A unified target schema $T$ best representing source schemas.
Integrated schemas
[Chitacariu et al, 08]

• Interactive Generation of Integrated schemas
1: Construct concept graph

- each relation is a node
- each edge denotes parent-child or key-foreign key relationship
Matching graph

- Step 2: Form matching edges \( X_i \)
- Step 3: Find assignments of boolean variables \( X_i \)
- Step 4: For every edge \( x_i \) set to true, merge concepts
Integrated schema

Assignment: \( x_1 = x_2 = x_3 = x_5 = 0, \)
\( x_0 = x_4 = x_6 = x_7 = 1 \)
Integrated schemas
[Chitacariu et al, 08]

• Different assignments can lead to the same schema
  – Add constraints to boolean variables
  – Find satisfying assignments for a set of Horn clauses

• Source-to-target mapping generation:
  – Use a variant of the chase in order to preserve source foreign key constraints in the target.
Integrated schemas
[Das Sarma et al, 08]

• Key idea: build $probabilistic$ schemas
  – Models uncertainty behind merging concepts
• Considers single relation source, target schemas
• Each attribute is a concept
• Attribute correspondences have weights
Integrated schemas [Das Sarma et al, 08]

• Algorithm
  1. Construct weighted graph using correspondences
  2. Remove edges with weight below $T$
  3. Each connected component forms cluster
Integrated schemas [Das Sarma et al, 08]

• Algorithm
  1. Construct weighted graph using correspondences
  2. Remove edges with weight below $T$
  3. Each connected component forms cluster

![Diagram of integrated schemas with nodes and weights]

$T = 0.2$
Integrated schemas [Das Sarma et al, 08]

- Algorithm
  1. Construct weighted graph using correspondences
  2. Remove edges with weight below $T$
  3. Each connected component forms cluster
Integrated schemas
[Das Sarma et al, 08]

- Partition edges into **certain** and **uncertain** edges
- Each uncertain edge with weight between $T+\varepsilon$ and $T-\varepsilon$
- Create new schema by including/excluding uncertain edges.
Integrated schemas
[Das Sarma et al, 08]

• Partition edges into **certain** and **uncertain** edges
• Each uncertain edge with weight between T+\(\epsilon\) and T-\(\epsilon\)
• Create new schema by including/excluding uncertain edges.
Partition edges into **certain** and **uncertain** edges.

Each uncertain edge with weight between $T + \varepsilon$ and $T - \varepsilon$.

Create new schema by including/excluding uncertain edges.
Integrated schemas
[Das Sarma et al, 08]

• Probability of schema $M_i$:

$$\Pr(M_i) = \frac{c_i}{\sum_{j=1}^{l} c_j}$$

where $c_i = \text{number of sources consistent with } M_i$

• Source is consistent if no two distinct attributes are grouped together
  – Models uncertainty in grouping real-world concepts

• Also consider p-mappings: probabilistic mappings.
Mapping Tasks

- iMap [Dhamankar et al, 2004]

- Schema matching
- Integrated schema
- Mapping creation

- [Chiticariu et al, 2008]
- [Das Sarma et al, 2008]
- Clio [Miller et al, 2000]
Mapping generation

[Miller et al, 2000]

• Goal:
  – to discover mappings between independently created source and target schemas

• Given:
  – Source schema S, single-table target schema T, set of value correspondences.
  – Value correspondence \((f_i, p_i)\) where \(f_i\) is a function:
    \[ f_i: \text{dom}(A_1) \times \cdots \times \text{dom}(A_q) \to \text{dom}(B) \]
  and \(p_i\) is a predicate over the source attributes:
    \[ p_i: \text{dom}(A_1) \times \cdots \times \text{dom}(A_q) \to \text{boolean} \]
Value Correspondences

- Example:
  - $f_1: \text{PayRate}(\text{HrRate}) \times \text{WorksOn}(\text{Hrs}) \rightarrow \text{Personnel}(\text{Sal})$
Mapping generation

Sample mapping queries:

\( \{ (i, n, s, a) \mid Professor(i, n, s) \land Address(i, a) \} \)

\( \{ (i, n, s, a) \mid \exists r, h, y, x \ Student(n, g, y) \land PayRate(y, h) \land WorksOn(n, p, x, r) \land i = null \land a = null \land s = h \times x \} \)
Algorithm

1. Input Value Correspondences

2. Group Correspondences into candidate sets:
   - At most one correspondence per target attribute for each candidate set

Candidate sets $\rightarrow$ $\{\{f_1, f_2\}, \{f_2, f_3\}, \{f_1\}, \{f_2\}, \{f_3\}\}$
Algorithm

3. Prune candidate sets if they do not map to good queries
   - For set \( \{f_1:S1.A \rightarrow T.C, f_2:S2.A \rightarrow T.D\} \) prune if no way to join S1 and S2

4. Select covers
   - Cover: Subset of candidate sets with each correspondence in at least one set

5. Rank covers
   - According to number of candidate sets
Conclusions

• Deriving mappings consists of several tasks:
  – Schema matching
  – Generation of Integrated schemas
  – Generation of mappings

• In general, lots of uncertainty
  – No way to exactly know semantic relationships
  – Tackle through probabilistic models
  – Learn from user feedback
Bibliography


• L. Chiticariu, P.G. Kolatis, and L. Popa: *Interactive generation of integrated schemas*. SIGMOD'08


• R.J. Miller, L.M. Haas, and M.A. Hernandez. *Schema Mapping as Query Discovery*. VLDB'00