Introducing myself

- Heterogeneous data management.
- Imperfect data management.
Dealing with Imperfection in Schema Integration

Some preliminary ideas

Matteo Magnani
Dept. of Computer Science
University of Bologna
Via Mura A.Zamboni, 7
40127 Bologna (Italy)

Nikos Rizopoulos
Dept. of Computing
180 Queen’s Gate
South Kensington Campus
Imperial College London
SW7 2AZ London (UK)
Outline

- Imperfect data: an overview.
Outline

- Imperfect data: an overview.
- Uncertainty in data integration based on semantic schema matching.
  - Semantic Schema Integration.
  - Adding uncertainty.
Imperfect Data: An Overview
A taxonomy of imperfection

To identify the main classes of imperfection, we use the following scenario:
(Complete) Absence
Absence

- All our belief is committed to the known set of alternatives.
- For example: “We do not know the age of John”.

1

Imprecision: Non-specificity

- Imprecision concerns the cardinality of our believes.
- When we believe in a crisp set, imprecision is called non-specificity.
- For example: “John is between 170 and 180 cm. tall.”
Imprecision: Vagueness

- If a set representing our belief is not crisp, imprecision is called vagueness.
- For example: “John is not very tall”
Uncertainty

- We have uncertainty when we do not commit all our belief.
- For example, “John should be 27 y.o.”.
### Examples

<table>
<thead>
<tr>
<th>Type</th>
<th>Abbr</th>
<th>Example (John’s tallness)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absence</td>
<td>ABS</td>
<td>Not known.</td>
</tr>
<tr>
<td>Non-Specificity</td>
<td>NS</td>
<td>Between 180 and 190 cm. 183 or 187 cm.</td>
</tr>
<tr>
<td>Vagueness</td>
<td>VAG</td>
<td>Not very tall.</td>
</tr>
<tr>
<td>Uncertainty</td>
<td>UN</td>
<td>Perhaps, 183 cm.</td>
</tr>
</tbody>
</table>
Uncertainty in Schema Integration based on Semantic Schema Matching

with Nikos Rizopoulos
Semantic Schema Integration

\[ S_1 \]
\begin{align*}
\text{student} & \xrightarrow{\text{res}} \text{course} \\
\end{align*}

\[ S_2 \]
\begin{align*}
\text{UG} & \xrightarrow{\text{res}} \text{course} \\
\end{align*}
Semantic Schema Integration

$S_1$

student \(\text{res}\) course

$S_2$

UG \(\text{res}\) course

= \cap \cup \subset \subseteq \cup \neq

SEMANTIC RELATIONSHIPS
Semantic Schema Integration

\[ S_1 \cap S_2 \neq \\emptyset \]

PARTIAL INTEGRATED SCHEMA

\[ \text{student} \]

\[ \text{UG} \]
Semantic Schema Integration

$S_1$

student \( \rightarrow \) res \( \rightarrow \) course

$S_2$

UG \( \rightarrow \) res \( \rightarrow \) course

\( \neq \)

PARTIAL INTEGRATED SCHEMA

\[ \subseteq \subseteq \subseteq \subseteq \subseteq \]

student \( \rightarrow \) course
Semantic Schema Integration

$S_1$

student \(\rightarrow\) res \(\rightarrow\) course

$S_2$

UG \(\rightarrow\) res \(\rightarrow\) course

\[\text{INTEGRATED SCHEMA}\]

student \(\rightarrow\) res \(\rightarrow\) course

\[\cup \subseteq \subseteq \subseteq \cup \neq\]
Dealing with uncertainty

\[ S_1 \]

\[ S_2 \]

\[ \text{student} \rightarrow \text{res} \rightarrow \text{course} \]

\[ \text{UG} \rightarrow \text{res} \rightarrow \text{course} \]

\[ \neq \]

?
Dealing with uncertainty

\[ S_1 \]

\[
\begin{array}{c}
\text{student} \\
\text{res} \\
\text{course}
\end{array}
\]

\[ S_2 \]

\[
\begin{array}{c}
\text{UG} \\
\text{res} \\
\text{course}
\end{array}
\]

\[ \neq \]

\[
\begin{array}{c}
\text{UG} \\
\text{student} \\
\text{UG}
\end{array}
\]

\[
\begin{array}{c}
\text{student} \\
\text{all-stu}
\end{array}
\]
Dealing with uncertainty

\[ S_1 \]

\[
\begin{array}{c}
\text{student} \\
\text{res} \\
\text{course}
\end{array}
\]

\[ S_2 \]

\[
\begin{array}{c}
\text{UG} \\
\text{res} \\
\text{course}
\end{array}
\]

?
Four main problems

- Identify uncertain relationships.
Four main problems

- Identify **uncertain relationships**.
- Produce **uncertain partial integrated schemas**.
Four main problems

- Identify **uncertain relationships**.
- Produce **uncertain partial integrated schemas**.
- Put together the uncertain partial integrated schemas, to obtain an **uncertain integrated schema**.
Four main problems

- Identify **uncertain relationships**.
- Produce **uncertain partial integrated schemas**.
- Put together the uncertain partial integrated schemas, to obtain an **uncertain integrated schema**.
- **Query** the database.
Uncertain relationships

- Dempster–Shafer’s theory to represent believes in relationships.

- $\Theta = \{=, \cap, \subseteq, \supseteq, \cup, \neq\}$
Uncertain relationships

- Dempster–Shafer’s theory to represent beliefs in relationships.

\[ \Theta = \{=, \cap, \subset, \supset, \cup, \neq \} \]

- \( m(\{=\}) = .4 \)
- \( m(\{\supset\}) = .3 \)
- \( m(\{\subset\}) = .3 \)
Uncertain relationships

- Dempster–Shafer’s theory to represent believes in relationships.

\[ \Theta = \{ =, \cap, \subseteq, \supseteq, \cup, \neq \} \]

- \[ m(\{ \cup, \neq \}) = 1 \]
Uncertain relationships

- Dempster–Shafer’s theory to represent believes in relationships.

\[ \Theta = \{=, \cap, \subset, \supset, \cup, \neq \} \]

- \( m(\{=\}) = 0.2 \)
- \( m(\Theta) = 0.8 \)
Uncertain relationships

- Dempster–Shafer’s theory to represent believes in relationships.

- \( \Theta = \{=, \cap, \subset, \supset, \cup, \neq\} \)

- \( m(\Theta) = 1 \)
Architecture

\[
\begin{align*}
exp_1 & \quad exp_2 & \quad \ldots & \quad exp_{n-1} & \quad exp_n \\
UR & \quad UR & \quad UR & \quad UR
\end{align*}
\]
Experts can be software agents or humans.

Dealing with Imperfection in Schema Integration – p. 17
Architecture

Scalable.
Architecture

- Scalable.
- Experts can be software agents or humans.
Example

$\text{Exp}_1$ (Cardinality):

$m(\{\cap, \subseteq, \cup, \neq\}) = 1$
Example

Exp$_1$ (Cardinality):
- $m(\{\cap, \cup, \#, \neq\}) = 1$

Exp$_2$ (Thesaurus):
- $m(\{\#\}) = .5$
- $m(\{=\}) = .2$
- $m(\Theta) = .3$
Example

Exp₁ (Cardinality):
- \( m(\{\cap, \supset, \cup, \neq\}) = 1 \)

Exp₂ (Thesaurus):
- \( m(\{\supset\}) = .5 \)
- \( m(\{=\}) = .2 \)
- \( m(\Theta) = .3 \)

Exp₃ (Human):
- \( m(\{=, \supset, \cup\}) = 1 \)
Example

\textbf{Exp}_1 \ (\text{Cardinality}):\n\begin{itemize}
  \item \( m(\{\cap, \cup, \forall, \neq\}) = 1 \)
\end{itemize}

\textbf{Exp}_2 \ (\text{Thesaurus}):\n\begin{itemize}
  \item \( m(\{\Downarrow\}) = .5 \)
  \item \( m(\{\Rightarrow\}) = .2 \)
  \item \( m(\Theta) = .3 \)
\end{itemize}

\textbf{Exp}_3 \ (\text{Human}):\n\begin{itemize}
  \item \( m(\{\Rightarrow, \Downarrow, \cup\}) = 1 \)
\end{itemize}

\textbf{\oplus} \ (\text{Combination}):\n\begin{itemize}
  \item \( m(\{\Downarrow\}) = \frac{5}{8} \)
  \item \( m(\{\Downarrow, \cup\}) = \frac{3}{8} \)
\end{itemize}
Unc. partial integrated schemas

Given two objects, some belief is committed to each possible relationship between them.

\[ \text{Bel: } = \quad \cap \quad \subset \quad \supset \quad \cup \quad \neq \]
Unc. partial integrated schemas

- Given two objects, some belief is committed to each possible relationship between them.
- From each possible relationship we can obtain a partial integrated schema.

\[
\text{Bel: } = \quad \cap \quad \subset \quad \supset \quad \cup \quad \neq
\]

\[
\downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow
\]

\[
\text{PI}(=) \quad \text{PI}(\cap) \quad \text{PI}(\subset) \quad \text{PI}(\supset) \quad \text{PI}(\cup) \quad \text{PI}(\neq)
\]
Given two objects, some belief is committed to each possible relationship between them.

From each possible relationship we can obtain a partial integrated schema.

The belief committed to each partial integrated schema is the same previously committed to the corresponding possible relationship.

\[
\begin{align*}
\text{Bel:} & \quad = \quad \cap \quad \subset \quad \supset \quad \cup \quad \neq \\
\downarrow & \quad \downarrow & \quad \downarrow & \quad \downarrow & \quad \downarrow & \quad \downarrow \\
\text{Bel:} & \quad \text{PI}(=) \quad \text{PI}(\cap) \quad \text{PI}(\subset) \quad \text{PI}(\supset) \quad \text{PI}(\cup) \quad \text{PI}(\neq)
\end{align*}
\]
Example

Uncertain relationship (Student-UG):
\[ m(\{\supset\}) = \frac{5}{8} \]
\[ m(\{\supset, \cup\}) = \frac{3}{8} \]

Uncertain partial integrated schema:
\[ m(\{\text{PI}(\supset)\}) = \frac{5}{8} \]
\[ m(\{\text{PI}(\supset), \text{PI}(\cup)\}) = \frac{3}{8} \]
Uncertain integrated schemas

Main idea: take all possible combinations of uncertain partial integrated schemas.

- \((A, B)\), \(\neq (A, B)\)
- \((B, C)\), \(\neq (B, C)\)

<table>
<thead>
<tr>
<th></th>
<th>(A, B)</th>
<th>(B, C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(=)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(=)</td>
<td>(\neq)</td>
<td></td>
</tr>
<tr>
<td>(\neq)</td>
<td></td>
<td>(=)</td>
</tr>
<tr>
<td>(\neq)</td>
<td></td>
<td>(\neq)</td>
</tr>
</tbody>
</table>

Some issues:

- Compact representation.
- Dependencies.
Uncertain integrated schemas

Uncertain relationship (Student–UG):
- \( m(\{\subseteq\}) = \frac{5}{8}, m(\{\subseteq, \cup\}) = \frac{3}{8} \)

Uncertain relationship (res–res):
- \( m(\{=\}) = \frac{1}{3}, m(\{\neq\}) = \frac{2}{3} \)
Uncertain integrated schemas

- Uncertain partial integrated schema:
  \( m(\{\text{PI}(\sqsubset)\} \times \{\text{PI}(=)\}) = \frac{5}{24} \)
  \( m(\{\text{PI}(\sqsubset), \text{PI}(\sqcup)\} \times \{\text{PI}(=)\}) = \frac{3}{24} \)
  \( m(\{\text{PI}(\sqsubset)\} \times \{\text{PI}(\neq)\}) = \frac{5}{12} \)
  \( m(\{\text{PI}(\sqsubset), \text{PI}(\sqcup)\} \times \{\text{PI}(\neq)\}) = \frac{3}{12} \)
Uncertain integrated schemas

= (A, B), ≠ (A, B)

= (B, C), ≠ (B, C)

= (A, C), ≠ (A, C)
Uncertain integrated schemas

\[= (A, B), \not= (A, B)\]
\[= (B, C), \not= (B, C)\]
\[= (A, C), \not= (A, C)\]

<table>
<thead>
<tr>
<th></th>
<th>(A, B)</th>
<th>(B, C)</th>
<th>(A, C)</th>
<th>Allowed</th>
</tr>
</thead>
<tbody>
<tr>
<td>(=)</td>
<td>(=)</td>
<td>(=)</td>
<td>(=)</td>
<td>(=)</td>
</tr>
<tr>
<td>(=)</td>
<td>(=)</td>
<td>(=)</td>
<td>(\not=)</td>
<td>(\not=)</td>
</tr>
<tr>
<td>(=)</td>
<td>(\not=)</td>
<td>(=)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\not=)</td>
<td>(=)</td>
<td>(=)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\not=)</td>
<td>(=)</td>
<td>(\not=)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\not=)</td>
<td>(\not=)</td>
<td>(\not=)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\not=)</td>
<td>(\not=)</td>
<td>(=)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\not=)</td>
<td>(\not=)</td>
<td>(\not=)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dealing with Imperfection in Schema Integration – p. 23
Uncertain integrated schemas

\[ = (A, B), \neq (A, B) \]
\[ = (B, C), \neq (B, C) \]
\[ = (A, C), \neq (A, C) \]

<table>
<thead>
<tr>
<th></th>
<th>( A, B )</th>
<th>( B, C )</th>
<th>( A, C )</th>
<th>Allowed</th>
</tr>
</thead>
<tbody>
<tr>
<td>=</td>
<td>=</td>
<td>=</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>=</td>
<td>=</td>
<td>\neq</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>=</td>
<td>\neq</td>
<td>=</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>\neq</td>
<td>=</td>
<td>=</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>\neq</td>
<td>=</td>
<td>\neq</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>=</td>
<td>\neq</td>
<td>\neq</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>\neq</td>
<td>\neq</td>
<td>=</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>\neq</td>
<td>\neq</td>
<td>\neq</td>
<td>Y</td>
<td></td>
</tr>
</tbody>
</table>
Querying the database

- No idea . . .
- It should not be very difficult to define the semantics of a query.
- Efficiency problems.
  - Cardinality reduction.
  - Compact query plans.
Concluding remarks

Uncertainty is one of many possible types of imperfection/ignorance.

We start our investigation from a method of schema integration based on semantic schema matching.

In real cases of data integration, it can be difficult to identify certain semantic relationships.

We have presented some preliminary ideas on how to extend this method to deal with uncertainty.
Dealing with Imperfection in Schema Integration

Some preliminary ideas

Matteo Magnani
Dept. of Computer Science
University of Bologna
Via Mura A.Zamboni, 7
40127 Bologna (Italy)

Nikos Rizopoulos
Dept. of Computing
180 Queen’s Gate
South Kensington Campus
Imperial College London
SW7 2AZ London (UK)