Automated information fusion under uncertainty

Matt Duckham
John Stell
Mike Worboys
NCGIA, University of Maine
David’s example

Dictionary of terms

Structure of terms

Structure of terms

Structure of terms

marnda

bargu

plain

mountain
Land cover information fusion

- CORINE: Fine-grained thematic, coarse-grained spatial

- OS MasterMap: coarse-grained thematic, fine grained spatial
Information fusion: schema-level

- Process of integrating information from multiple sources, where the information derived is more valuable, more reliable or in some way more useful than the original information.
- Existing fusion techniques largely schema-driven.
- Semantics are difficult to precisely capture.
Rosetta Stone: C2\textsuperscript{nd} tablet inscribed with three copies of the same decree in Egyptian (text and hieroglyphs) and Greek (text)

C18\textsuperscript{th}/19\textsuperscript{th} Archaeologists used the stone to decipher Egyptian hieroglyphs

Successful because of structure on the (instance-level) inscriptions provided a basis for (schema-level) inferences
Land cover application

- CORINE: Fine-grained thematic, coarse-grained spatial
- OS MasterMap: coarse-grained thematic, fine grained spatial
An ST-layer is a distribution of elements of a thematic space over a spatial framework.

Formally:

\[ S \xrightarrow{f} T \]

To:

Assume that both layers are based on the same spatial framework \( S \). (If not, extend each layer to the union of frameworks).
ST-layer integration

- From:

\[ T_1 \xrightarrow{f_1} S \xrightarrow{f_2} T_2 \]
ST-layer integration

\[ S \] is connected to \( T_1 \) through \( f_1 \) and \( T_2 \) through \( f_2 \). The tensor product \( T_1 \otimes T_2 \) is connected to \( S \) through \( f_1 \otimes f_2 \). Connections to \( T_1 \otimes T_2 \) are labeled with \( p_1 \) and \( p_2 \).
Example 1: Unstructured thematic space

- Assume spatial framework:

- Assume ST-layers:
Integrated ST-layer

\[ T_1 = \{u,c\} \]
\[ T_2 = \{h,w\} \]
\[ T_1 \otimes T_2 = \{(u,h),(u,w),(c,h),(c,w)\} \]
\[ f_1 \otimes f_2 : s \mapsto (sf_1,sf_2) \]
Example 2: Partitioned thematic space

- Suppose we have a collection of landcover types:
  - c conifer woodland
  - b broad-leaved woodland
  - n natural grassland
  - m moorland
  - h heathland
  - o orchard
  - p pasture
  - a arable land

- Suppose two thematic spaces are given as:
  \[ T_1 = \{\{p,a,o\},\{c,b,n,m,h\}\} \]
  \[ T_2 = \{\{p,a,n\},\{m,h\},\{c,b,o\}\} \]

The integrated thematic space will be:

\[ T_1 \times T_2 = \{\{p,a\},\{o\},\{c,b\},\{n\},\{m,h\}\} \]
Spatio-thematic product

- **agricultural**

- **forest**

- **semi-natural**

- **herbaceous plant cover**

- **trees**
Spatio-thematic product

An incompatible region
Suppose that there is an underlying space $U$ of ‘atomic’ themes, and that $T_1$ and $T_2$ are partitions of $U$.

Suppose $f_1$ and $f_2$ are spatio-thematic layers:

\[ f_1 : S \rightarrow T_1 \]
\[ f_2 : S \rightarrow T_2 \]

Then

\[ T_1 \otimes T_2 = \{ t_1 \cap t_2 \mid t_1 \in T_1, t_2 \in T_2, t_1 \cap t_2 \neq \emptyset \} \]
\[ f_1 \otimes f_2 : S \rightarrow sf_1 \cap sf_2 \]

Notice that $f_1 \otimes f_2$ may not be defined, the layers may not be compatible.
... but we are interested in thematic hierarchies.
From this ...
... to this.
Layer or instantiation?

- An ST-layer is a distribution of elements of a thematic space over a spatial framework.

\[ S \xrightarrow{f} T \]

- A spatial instantiation of a classification is an association with each theme a part of the space.

\[ T \xrightarrow{g} R \]
Simple example: schema-level

[Diagram]

Schema-level information

T

Forest  Built-up area

T

Woodland  Urban
Simple example: schema-level

Schema-level information

Human expertise

Forest → Urban

Woodland ← Sub-urban

Built-up area ← Urban

T

T
Simple example: schema-level

Schema-level information

Instance-level information

Forest

Built-up area

T

Forest

Built-up area

Woodland

Suburban

Built-up area

Woodland

Urban

Urban

T

Forest

Built-up area

Woodland

Urban

T

Woodland

Urban

Forest

Built-up area

Woodland

Urban

Woodland

Woodland

Urban
Simple example: schema-level

Schema-level information

Forest

Built-up area

T

Forest

Urban

Woodland

Sub-urban

Built-up area

Woodland

Urban

T

Instance-level information

Forest

Built-up area

Woodland

Sub-Urban

Built-up area

Woodland

Urban

Woodland

Woodland
Schema-level information fusion

- Databases
  - schema integration (eg Kim & Sea, 1992)
  - mediators (eg Wiederhold, 1992)
- Interoperability (eg Sheth, 1999)
- Semantic web (eg Berners-Lee et al., 2001)
- Data warehousing and IR (eg Widom, 1995)
- Ontology-based IS (Fonseca et al, 2002)
- Knowledge representation (Calvanese et al., 1998)
Conventional schema-driven information fusion

Schema A

Shared ontological structure

Human Domain Expertise

Schema B

- Concentrating solely on the schema ignores
  - a potentially valuable source of examples
  - the mismatch between how information is defined and structured and how it is actually used

- High levels of human domain expertise are normally needed to fuse information
Instance-based information fusion

Conventional schema-driven information fusion

Schema-level information

Schema A

Shared schematic structure

Human Domain Expertise

Schema B

Shared geometrical, topological, hierarchical, or metric structure

Instance-level fusion

Information source A

Automation

Information source B
Instance-level information fusion

- Probabilistic approaches in semantic web (Doan et al, 2002)
- Neural nets in databases (Li and Clifton, 2000)
- Machine learning in databases (Berlin and Motro, 2001)
Assume we have “good” classifications that induce a partition on observable phenomena.

- Devise rule to relating instance-level structure to schema-level structure, e.g.
  - If the extension of a category $x$ is a proper part of the extension of another category $y$ then category $x$ is subsumed by category $y$. 
Simple example: instance-level

Schema-level information

Instance-level information

T
Forest Built-up area

T
Woodland Urban

Forest
Built-up area

Woodland
Urban
Woodland
Simple example: instance-level

Schema-level information

- T
  - Forest
  - Built-up area

Instance-level information

- Forest
- Built-up area

Mereological structure:
- PP(Woodland, Forest)
- PP(Built-up area, Urban)
- DR(Woodland, Built-up area)

- Woodland
- Urban
Simple example: instance-level

Schema-level structure:
- Woodland \leq Forest
- Built-up area \leq Urban
- Woodland \land Built-up area = \emptyset

Mereological structure:
- PP(Woodland, Forest)
- PP(Built-up area, Urban)
- DR(Woodland, Built-up area)
Simple example: instance-level

Schema-level information

- Forest
- Built-up area

Instance-level information

- Forest
- Built-up area

T

Forest

Urban

Woodland

Forest & Urban

Built-up area

T

Woodland

Urban
Simple example: instance-level
Assume $T_1$ and $T_2$ are meet semilattices.

Given two functions $g_1: T_1 \rightarrow \mathbb{R}$, $g_2: T_2 \rightarrow \mathbb{R}$, the first step is to build the fused hierarchy.

Take $T_1 \times T_2$ as the set of ordered pairs, where:

- $(s_1, s_2) \land (t_1, t_2) = (s_1 \land s_2, t_1 \land t_2)$
- (we also have to worry a little about $\bot$)

How does this instantiate?

- $g(t_1, t_2) = g_{t_1} \land g_{t_2}$ (in the lattice of partitions)

Factor $T_1 \times T_2$ by a set of constraints

- e.g. $t =< t'$ iff $g_t =< g_{t'}$
Real data sets rarely match up so nicely, as a result of uncertainty

- Vagueness
- Imprecision
- Inaccuracy
# Logics of fusion under uncertainty

<table>
<thead>
<tr>
<th>fuse</th>
<th>1</th>
<th>?</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>0</td>
<td>?</td>
<td>?</td>
<td>0</td>
</tr>
</tbody>
</table>

Vague region
no errors
conservative

<table>
<thead>
<tr>
<th>fuse</th>
<th>1</th>
<th>?</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>X</td>
</tr>
<tr>
<td>?</td>
<td>1</td>
<td>?</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>X</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Crisp region viewed with imprecision
no errors
weighted towards definite
Inaccurate schema

- We want to incorporate schema-level knowledge in fusion process.
- This knowledge may conflict with instance-level information.
Inaccurate schema

- Schema-level information:
  - Forest
  - Built-up area

- Instance-level information:
  - Forest
  - Built-up area

- T
  - Woodland
  - Urban
Inaccurate schema

Schema-level structure:
Woodland ≤ Forest
Built-up area ≤ Urban
Woodland ∧ Built-up area = ∅
+ Forest ∧ Urban = ∅

Mereological structure:
PP(Woodland, Forest)
PP(Built-up area, Urban)
DR(Woodland, Built-up area)
Inaccurate schema

Schema-level information

Forest → Built-up area

Urban → Forest

Woodland → Built-up area

Built-up area → Woodland

Inconsistent

Instance-level information

Forest

Built-up area

Woodland

Built-up area

Woodland

Urban
Inaccurate instances

- What if one of the three texts had a typo?
- With example data sets, we don’t normally rely solely on a single instance to draw conclusions
- So if *most* instances agree, perhaps we could ignore errant instances like a typo
Inaccurate instances

Schema-level information

T

<table>
<thead>
<tr>
<th>Forest</th>
<th>Built-up area</th>
</tr>
</thead>
</table>

Instance-level information

T

<table>
<thead>
<tr>
<th>Woodland</th>
<th>Urban</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Built-up area</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Woodland</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban</td>
</tr>
<tr>
<td>Woodland</td>
</tr>
</tbody>
</table>
Inaccurate instances

Schema-level information

T

Forest Built-up area

Instance-level information

Forest

Built-up area

Mereological structure:
PO(Woodland, Forest)
PO(Built-up area, Urban)
PO(Woodland, Built-up area)
PO(Urban, Forest)

T

Woodland Urban

Woodland

Urban

Woodland
Inaccurate instances

- **Schema-level information**
  - Forest
  - Built-up area

- **Instance-level information**
  - Forest
    - Built-up area
  - Woodland
  - Urban

- **Set intersection symbols**:
  - $T$
  - $W \cap F$
  - $W \cap B$
  - $F \cap U$
  - $U \cap B$
  - $W \cap B$
Inaccurate instances

- Assume that extensions that overlap to a relatively small extent are due to inaccuracies.
- In our example, include information about geometric structure.

Eg, if $\frac{\text{Area}(A \text{ overlaps } B)}{\text{Area}(A)} < \delta$ and $\frac{\text{Area}(A \text{ overlaps } B)}{\text{Area}(B)} < \delta$ ignore overlap during inference process.
Inaccurate instances

Schema-level information

- Forest
- Built-up area

Instance-level information

- Forest
- Built-up area

Mereological & geometric structure:
- PP(Woodland, Forest)
- PP(Built-up area, Urban)
- DR(Woodland, Built-up area)

T

Woodland
Urban

T
Inaccurate instances

Schema-level information:
- Forest
- Built-up area

Instance-level information:
- Forest
  - Built-up area
- Woodland
  - Forest & Urban
  - Built-up area
- Woodland
  - Urban
What if one of the three texts had more detail than the others?

If we have qualitative knowledge about the relative granularities of text, perhaps we can allow for more detail, and not try to use everything in the more detailed text for inferring schema structure.
Imprecise instances

- Assume that extensions that overlap to a widely different extent are due to different levels of precision in information sources.

  Eg, if $\text{Area}(A \text{ overlaps } B)/\text{Area}(A) < \delta$ 
  and $\text{Area}(A \text{ overlaps } B)/\text{Area}(B) > 1-\delta$, ignore overlap.

  Eg, B="Hamlet" in fine grained data set, A="Woodland" in coarse grained data set.
Prototype implementation

Land Cover Data (eg DNF&CORINE)
Prototype implementation

Land Cover Data (eg DNF&CORINE)
Prototype implementation

Land Cover Data (eg DNF&CORINE)

Fusion constructor
Prototype implementation

Land Cover Data (e.g., DNF & CORINE)

Fusion constructor
Prototype implementation

Land Cover Data (e.g., DNF & CORINE)

Fusion constructor
Other structures: instance-level

- Other instance-level structures: hierarchical and similarity
- E.g. “Internet as Thesaurus”
- Use hierarchical structure of web pages
- Use similarity structure of text
- Use similarity structure of images

Diagram:
- Database
- User query
  - Heterogeneous schema
  - Question mark

User query
Heterogeneous schema
Other structures: instance-level

- Other instance-level structures: hierarchical and similarity
- E.g. “Internet as Thesaurus”
- Use hierarchical structure of web pages
- Use similarity structure of text
- Use similarity structure of images
Conclusions

- Trying to take advantage of the way that categories are used to inform the arrangement of schemas.
- Granularity needs special attention.
- Problems with conflation of mereological and subsumption structures?
- Preference relations are important, as this is really a problem of belief revision.