Structure-based Filtering for Ontology Alignment

Bi Chen, He Tan, Patrick Lambrix

Department of Computer and Information Science
Linköpings universitet

Outline

- Introduction
  - Ontology alignment
  - Ontology alignment systems
  - Structure-based alignment algorithm
  - Filtering algorithm
- Structure-based filtering
  - Basic ideas
  - Algorithms
  - Implementation
- Evaluation
- Conclusion

Ontology Alignment

- Many ontologies have been developed
- Many of them have overlapping information
- Use of multiple ontologies
  - e.g. custom-specific ontology + standard ontology
- Bottom-up creation of ontologies
  - experts can focus on their domain of expertise
- Important to know the inter-ontology relationships

Ontology Alignment Systems

- Many systems have been developed for ontology alignment

SIGNAL-ONTOLOGY (SigO)

- Immune Response
  - Allergic Response
  - Antigen Processing and Presentation
  - B Cell Activation
  - B Cell Development
  - Complement Signaling
  - Cytokine Response
  - Immune Suppression
  - Inflammation
  - Intestinal Immunity
  - Leukotriene Response
  - Leukotriene Metabolism
  - Natural Killer Cell Response
  - T Cell Activation
  - T Cell Development
  - T Cell Selection in Thymus

GENE ONTOLOGY (GO)

- Immune response
  - Acute-phase response
  - Anaphylaxis
  - Antigen presentation
  - Antigen processing
  - Cellular defense response
  - Cytokine metabolism
  - Cytokine biosynthesis
  - Regulation of cytokine biosynthesis
  - B-cell activation
  - B-cell differentiation
  - B-cell proliferation
  - Cell cycle regulation
  - T-cell activation
  - Activation of natural killer cell activity

- To define the relations between the terms in different ontologies
Structure-based Alignment Algorithm

- The intuition is that the similarity of concepts is influenced by their structural environment.
  - often requires previous similarity results
  - intuitively appealing, but poor results (in our experiments)

Filtering Algorithm

- Threshold algorithm
  e.g. retains pairs with the similarity above or equal to a threshold

\[ O_1: \{1, 2, 3, 4, 5\} \]
\[ O_2: \{A, B, C, D, E, F\} \]
\[ (2, B) \]
\[ (3, F) \]
\[ (6, D) \]
\[ (4, C) \]
\[ (5, C) \]
\[ (5, E) \]
\[ \ldots \]

Filtering Algorithm

- When the threshold is high, the precision is high, but the recall can be low.
- When the threshold decreases, the recall increases, but the precision can decrease.

\[ O_1: \{1, 2, 3, 4, 5\} \]
\[ O_2: \{A, B, C, D, E, F\} \]
\[ (2, B) \]
\[ (3, F) \]
\[ (6, D) \]
\[ (4, C) \]
\[ (5, C) \]
\[ (5, E) \]
\[ \ldots \]

Outline

- Introduction
- Structure-based filtering
  - Basic ideas
  - Algorithms
  - Implementation
- Evaluation
- Conclusion

Basic ideas

- Retain the pairs: \( \text{sim} \geq \text{upper-th} \)
- Filter out the pairs: \( \text{upper-th} > \text{sim} \geq \text{lower-th} \)
- Build a filter using the structural information

\[ O_1: \{1, 2, 3, 4, 5\} \]
\[ O_2: \{A, B, C, D, E, F\} \]
\[ \text{upper-th} \]
\[ \text{lower-th} \]
Algorithms

- Find a consistent suggestion group from the pairs with similarity values higher than or equal to the upper threshold
- Use the consistent suggestion group to partition the original ontologies
- Filter the pairs with similarity values between the lower and upper thresholds using the partition

Consistent suggestion group

- each concept occurs at most once in the group
- for \((A, A)\) and \((B, B)\): \(A \subseteq B \iff A \subseteq B\)

Partition ontologies

CSG: \{ (2, B), (3, F), (6, D) \}

Partition: ( {5}, {E} ), ( {1, 4, 7, 8}, {A, C} )

Filtering suggestions

\[
\begin{align*}
(2, B) & \quad \text{upper-th} \quad (5, C) \\
(3, F) & \quad (6, D) \\
\end{align*}
\]

ontology partition: ( {5}, {E} ), ( {1, 4, 7, 8}, {A, C} )

\( (5, E) \) is retained as a suggestion, but \((5, C)\) is discarded.

Implementation

- Find a (maximal) consistent suggestion group
  - A genetic algorithm is used
- Partition original ontologies
  - Based on a binary tree built from ontologies

Outline

- Introduction
- Structure-based filtering
- Evaluation
- Conclusion
**Evaluation**

- **GO vs. SigO**
  - GO: 70 terms
  - SigO: 15 terms
  - "immune defense"

- **MA vs. MeSH**
  - MA: 112 terms
  - MeSH: 45 terms
  - "eye"
  - MA: 15 terms
  - MeSH: 18 terms
  - "nose"
  - MA: 77 terms
  - MeSH: 39 terms
  - "ear"

**Evaluation**

- The precision is increased after filtering.
  - a linguistic alignment algorithm using WordNet
  - the upper threshold is 0.8

- The recall is constant in most cases after filtering.
  - a linguistic alignment algorithm using WordNet
  - the upper threshold is 0.8

**Evaluation**

- Time for filtering is only a fraction of the time for the similarity computation

<table>
<thead>
<tr>
<th>Case</th>
<th>B</th>
<th>ID</th>
<th>log2</th>
<th>corr</th>
<th>cyc</th>
</tr>
</thead>
<tbody>
<tr>
<td>TermWN</td>
<td>10.2</td>
<td>35.8</td>
<td>7.7</td>
<td>35.5</td>
<td>60.7</td>
</tr>
<tr>
<td>Filter</td>
<td>1.4</td>
<td>1.8</td>
<td>1.0</td>
<td>4.9</td>
<td>5.0</td>
</tr>
</tbody>
</table>

**Outline**

- Introduction
- Structure-based filtering
- Evaluation
- Conclusion

**Conclusion**

- Proposed a novel structure-based method for filtering alignment suggestions
- The method gives good results in the experiment.

- In the future, we will investigate the use of the partition to increase the recall.