ProbaMap: a scalable tool for discovering probabilistic mappings between taxonomies

Rémi Tournaire
Jean-Marc Petit - Marie-Christine Rousset - Alexandre Termier

AKBC, Grenoble

May 19, 2010
General problem addressed: ontology alignment

- Scalable automatic discovery of mappings between taxonomies → Major issue for the future Semantic Web
- Decentralized nature of the web → independent construction of personalized lightweight ontologies (often taxonomies) to annotate documents
- Mappings enable collaborative exchange of documents in the Web. In particular, *inclusion mappings* are the basis for query reformulation.

![Diagram of sub-taxonomies](image)

Recent Classical

![Diagram of sub-taxonomy relationships](image)
General problem addressed: ontology alignment

- Scalable automatic discovery of mappings between taxonomies
  → Major issue for the future Semantic Web
- Decentralized nature of the web
  → independent construction of **personalized** lightweight ontologies (often taxonomies) to annotate documents
- Mappings enable collaborative exchange of documents in the Web. In particular, *inclusion mappings* are the basis for query reformulation.

![Diagram showing mappings between taxonomies](image-url)
Main distinguishing points of our approach

Ontology alignment: a lot of existing methods

Surveys: [Schvaiko Euzenat 2005] [Rahm Bernstein 2001]

Yearly Ontology Alignment Evaluation Initiative (OAEI) contest.

Lot of remaining challenging issues.

**Focus of our work:**

1. Inclusion mappings discovery
2. Handling uncertainty with a clear semantics
3. Scalability
Modeling uncertainty with probabilities

Two probabilistic semantics for inclusion mappings

Considered probabilistic semantics extends the logical semantics.

1. \( P_c(A_i \sqsubseteq B_j) = P(B_j \mid A_i) = \frac{P(A_i \cap B_j)}{P(A_i)} \)

2. \( P_u(A_i \sqsubseteq B_j) = 1 - P(A_i \setminus B_j) = 1 - P(A_i) + P(A_i \cap B_j) \)

Bayesian estimators for \( P(A_i \cap B_j) \) and \( P(A_i) \) based on instances:

\[
P(A_i) = \frac{1 + |Ext(A_i, T_i)|}{2 + |Ext(T_i)|}
\]

\[
P(A_i \cap B_j) = \frac{1 + |Ext(A_i \cap B_j, T_i \cup T_j)|}{4 + |Ext(T_i \cup T_j)|}
\]

Usage of classifiers to compute \( Ext(A_i \cap B_j, T_i \cup T_j) \).
Relation between logical and probabilistic semantics

$P_u$ and $P_c$ are monotonous with respect to the logical implication

$T_i, T_j, m \models m' \Rightarrow P_u(m) \leq P_u(m')$

Weaker property for $P_c$

Example

\[ C_1 \sqsubseteq F_2, T_1, T_2 \models C_1 \sqsubseteq E_2 \]

\[ P_u(C_1 \sqsubseteq F_2) \leq P_u(C_1 \sqsubseteq E_2) \]

\[ P_c(C_1 \sqsubseteq F_2) \leq P_c(C_1 \sqsubseteq E_2) \]
The ProbaMap algorithm

A scalable algorithm to discover the most probable mappings between two taxonomies.

**Input**
- Two taxonomies $T_i$, $T_j$ and their instances
- Two thresholds $S_u$ and $S_c$

**Output**
All mappings $m$ between $T_i$ and $T_j$ for which $P_u \geq S_u$ and $P_c \geq S_c$.

**Principle overview**
- Generation of candidate mappings to be tested from the most general to the most specific ones, according to the logical implication.
- Pruning of the search space by exploiting the monotony properties.
Quantitative and qualitative experiments

*Scalability:* experiments on taxonomies larger than 3000 classes. ProbaMap handles a search space larger than 30 millions mappings.

Comparative experiment

Experiments conducted on Yahoo and Google subdirectories alignment. Comparison with SBI[Ichise et. al, IJCAI 2003] in term of accuracy of the alignment.

Results for the subdirectories

Google/Recreation/Autos vs. Yahoo/Recreation/Automotive
Conclusion & Perspectives

Conclusion

- 2 probabilistic semantics for inclusion mappings that extend and connect to the logical semantics
- Probabilities estimation with statistics on instances
- A scalable algorithm for mapping discovery: ProbaMap

Perspectives

Probabilistic query reformulation

- Based on discovered mappings
- In order to associate probabilities to answers
You are invited to come to see our poster.

Thanks for your attention!