A Tool For Mapping Concepts Between Two Ontologies

Sushama Prasad, Yun Peng, and Tim Finin
University of Maryland Baltimore County
{sprasa2, ypeng, finin}@csee.umbc.edu

We describe ongoing work which combines the recently emerging semantic markup language DAML+OIL (for ontology specification), the text-based classification technology (for similarity information collection), and Bayesian reasoning (for similarity synthesis and final mapping selection), to provide ontology mapping between two classification hierarchies. This work supports an interactive system used to semi-automatically build a mapping from one topic hierarchy into another. This system will be used as part of the ITTALKS [1] system to allow multiple topic ontologies to be used to describe the subjects of talks and the interests of users. The ontology maps developed by our tools can then be used to recognized that a talk described by terms in one ontology might be of interest to a users who has described her interests using terms drawn from another. A more complete description of this work is available at [2].

Ontologies. The two hierarchies we used as examples are ACM topic ontology and a small ITTALKS topic ontology which organizes classes of IT related talks in a way different from ACM classification. Both ontologies, as well as the output mappings, are marked up in DAML+OIL. Each concept/class in an ontology is associated with a set of *exemplar*s, which are URLs to the locations of text documents thought to belong to that class.

Text-based classification. The Rainbow text classifier [3] is used to generate similarity scores between concepts in the two ontologies based on their associated exemplar documents. First, a *model* is built for each ontology, which primarily contains statistical information about the exemplars associated with each concept in that ontology. Then, the similarity score $S_a(A_i, B_j)$ from concept B_j in ontology B to concept A_i in ontology A can be obtained by comparing the exemplars of B_j against the model of ontology A. In essence, $S_a(A_i, B_j)$ measures similarity between exemplars associated with B_j and those with A_i.

Bayesian subsumption. B_j may (partially) match more than one concept in A, each with a different similarity score. Also since a non-leaf node is a superclass of its children, its exemplars should include both those associated with it and those with all of its descendants in the hierarchy. Therefore, non-leaf nodes need to synthesize scores from their descendants before the final mapping can be selected. This is accomplished by a Bayesian extension of the subsumption operation of description logics. In this approach, we assume that all leaves in a hierarchy form a mutually exclusive and exhaustive set, and take the score $S_a(A_i, B_j)$ as $P(A_i|B_j)$ if B_j is a leaf.[1]

Then, the concept A^* is said to be the best mapping of a concept B_j if (1) $P(A * |B_j) > 0.5$, and (2) none of A^*'s children A_k has $P(A_k|B_j) > 0.5$. These two conditions together give A^* the flavor of the most specific subsumption in description logics.

The algorithm of finding A^* is a simple procedure that takes two passes over the hierarchy of A:
Bottom-up: Synthesize the probability for each non-leaf node A_i as $P(A_i \mid B_j) = \sum_{A_k \in child(A_i)} P(A_k \mid B_j)$.
Top-down: Recursively search for A^*, starting from A_{root}. If it does not satisfy the two conditions, then move down to its most probable child.

Experiment results. Preliminary experiments have been conducted over the two topic ontologies for a set of selected concepts. The resulting mappings were ranked by their respective probabilities, and were given to five people knowledgeable about computer science for evaluation. For the top 5%,10%, 15%, and 20% ranked mappings, acceptable rates[2]XS were 0.8, 0.7, 0.68, 0.65, respectively. Encouraged by these results, we plan to continue this work along several directions, including relaxing the mutual exclusive assumption, utilizing properties associated with individual classes, and conducting additional experiments of larger scales.

Acknowledgment: This research was supported in part by DARPA contract F30602-97-1-0215.

References

[1] R. S. Cost et. al., “ITTALKS: A Case Study in DAML and the Semantic Web”, IEEE Intelligent Systems, 17:1, 2002

[2] S. Prasad, Y. Peng, and T. Finin, ”A Tool For Mapping Concepts Between Two Ontologies”, Technical Report, CSEE, UMBC, 2002. http://daml.umbc.edu/papers/mapper/

[3] http://www-2.cs.cmu.edu/ mccallum/bow/rainbow/

[1] If B_j is a non-leaf, $P(A_i|B_j) = \sum_{B_k \in B_j} P(A_i|B_k) \cdot P(B_k)/P(B_j)$ can be computed from the scores $S_a(A_k, B_j)$ if $P(B_k)/P(B_j)$ are available.

[2] A mapping is acceptable if 4 of the 5 evaluators agreed