Combining Ontology Mapping Methods Using Bayesian Networks

Ondřej Šváb, Vojtěch Svátek

University of Economics, Prague, Dep. Information and Knowledge Engineering,
Winston Churchill Sq. 4, 130 67 Praha 3, Prague, Czech Republic
svabo@vse.cz, svatek@vse.cz

Abstract. Bayesian networks (BNs) can capture interdependencies among ontology mapping methods and thus possibly improve the way they are combined. Experiments on ontologies from the OAEI collection are shown, and the possibility of modelling explicit mapping patterns in combination with methods is discussed.

1 Introduction

Most existing systems for ontology mapping combine various methods for achieving higher performance in terms of recall and precision. Our approach relies on *Bayesian networks* (BNs) as well-known formal technique that can capture interdependencies among random variables. A Bayesian network (BN) [3] is a directed acyclic graph with attached local probability distributions. Nodes in the graph represent random variables with mutually exclusive and exhaustive sets of values (states). Edges in the graph represents direct interdependences between two random variables. We believe that this approach can bring additional benefits compared to ad hoc combination of methods, mainly resulting from better adaptability (training from data within a well-established formal framework).

Two approaches that use BNs for Ontology Mapping have recently been reported. The first is *OMEN* [4], which mainly serves for enhancing existing mappings. Its input are results of another mapping tool, while its output are more precise mappings as well as and new mappings. Nodes in the BN represent pairs of concepts that can potentially be mapped. Edges follow the taxonomy given in original ontologies. The network structure thus mimics that of ontologies themselves, though heuristics for graph pruning are employed in this transformation. For constructing conditional probability tables (CPTs) for each node meta-rules are used, such as : "if two nodes match and so do two arrows coming out of these nodes then the probability that nodes at the other end of the arrows match is increased". The second project, *BayesOWL* ([5]), is rather a framework for ontology mapping than a mapping method per se. The probabilistic ontological information is assumed to be learnt (in forms of probabilistic constraints) from web data using a text-classification-based learner; this information is translated to BNs. Mappings among concepts from two different ontologies then can be discovered using so-called evidential reasoning across two BNs.

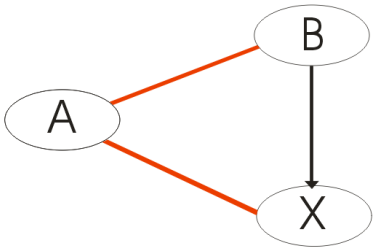

Fig. 1. Example of mapping pattern across two ontologies

2 Modelling Dependencies among Mapping Methods

Our approach differs from prior approaches in the sense that we don't apply BN modelling to ontologies or their mappings themselves but rather to different *mapping methods*. The BNs are assumed to contain nodes (or sub-networks) representing the results of individual methods plus one representing the final output. This will allow us not only to combine the methods (in the probabilistic framework) but also to talk about conditionally in/dependent methods, a minimal required subset of methods and the like. The mapping methods can have varying degree of granularity: we focus on low-level methods, understood as *mapping justifications*. Moreover, in the work-in-progress part of our research, we account for *mapping patterns* encompassing small structural fragments of ontologies. The patterns will capture, to some degree, similar information as OMEN meta-rules, we however prefer to model them directly within the BN formalism.

We distinguish among *families of methods* (string-based, linguistic-resource-based, graph-based, logic-based etc.) sharing some generic principle and input resources. Each family encompasses multiple low-level methods; for example, a string-based method can be built upon diverse string distance measures. We dedicate a separate node of the BN to each low-level method, viewed here as *mapping justification*. We believe that such methods are a meaningful target for BN modelling, as their statistical dependencies are likely to reflect plausible relationships even interpretable by a human.

The notion of *mapping pattern* is a natural counterpart to that of intra-ontology ('design') pattern [1]. Mapping patterns have been implicitly proposed by Ghidini & Serafini [2], who even consider mappings among different modelling constructs (such as concept-to-relation). A mapping pattern is, essentially, a structure containing some (at least one) constructs from each of the two (or more) ontologies plus some (candidate) mapping among them. The simplest mapping pattern only connects one concept from each of the two ontologies. An example of a bit more complex mapping pattern is in Figure 1. The left-hand side (class A) is from O_1 and the right-hand side (class B and its subclass X) is from O_2. We try to map class A simultaneously to class B and to class X.

The input to the process of BN *training* for ontology mapping are positive and negative examples with results of individual methods ('mapping justifications'), and possibly also the network structure, unless we want to learn it as well. The positive examples correspond to pairs for which mapping has previously been established, while the negative ones are (all or a subset of) pairs that have been identified as non-matching. Then CPTs and possibly the structure are learnt. In the phase of using the trained BN, the mapping justifications for unseen cases (pairs of concepts) are counted and inserted into the BN as evidence. The result of alignment is calculated via propagation of this evidence.

3 Experiments

For experiments we choose ontologies from the *OntoFarm* collection (`http://nb.vse.cz/~svabo/oaei2006/`), which is currently part of the OAEI 2006 setting. It models the domain of conference organisation; individual ontologies were designed independently by different people and based on different resources: personal experience with conference organisation, conference web pages or conference organisation support tools.

We restricted the first experiments to ten string distance measures implemented in the *SecondString* library (`http://secondstring.sourceforge.net/`: Levenshtein, Jaro, Jaccard, Char-Jaccard, Smith-Waterman, Monge-Elkan, SLIM, TokenFelligiSunter, UnSmoothedJS and TFIDF. Because of the local nature of distance string measures, capturing context by means of mapping patterns does not seem to bring great benefits; we thus only focused on the combination of low-level methods. We extracted classes from two ontologies (`ekaw.owl` and `ConfOf.owl`). Our training data consist of 798 pairs, of which 149 were manually labelled as positives and 649 as negatives. They were 'semi-randomly' picked from different parts of the ontology; the overall number of possible pairs would be about 2500 (the product of concept counts in both ontologies). The results were transformed from the $[0, 1]$ scale to two categories: 'true' if the value is over 0.5 and 'false' if the value is lower or equal to 0.5.

To learn the BN we use the *Hugin* tool (`http://www.hugin.com/`): the structure was trained using the NPC method and CPTs were trained using the EM algorithm. We learnt two Bayesian networks in this way. The first one has been enforced the *naive Bayesian structure*, which assumes independence of methods; only the CPTs were learned from data. For the second network, we also *learnt the structure*; in this way we could also explore interdependencies among low-level methods. The learnt structure is in Figure 2. From the structure and the defintion of so-called *Markov blanket* [3] we can conclude that if we know the mapping justifications of TFIDF, Smith-Waterman, Jaccard, Jaro, and SLIM, other methods do not matter. Methods unrelated to some other method (TokenFelligiSunter and UnSmoothedJS) are not in the BN at all.

To evaluate the performance of each proposed Bayesian classifier we used the *one-leave-out method*. For the naive Bayesian classifier, we got the best result with probability threshold 80%: 73% precision, 60% recall (F-measure was then

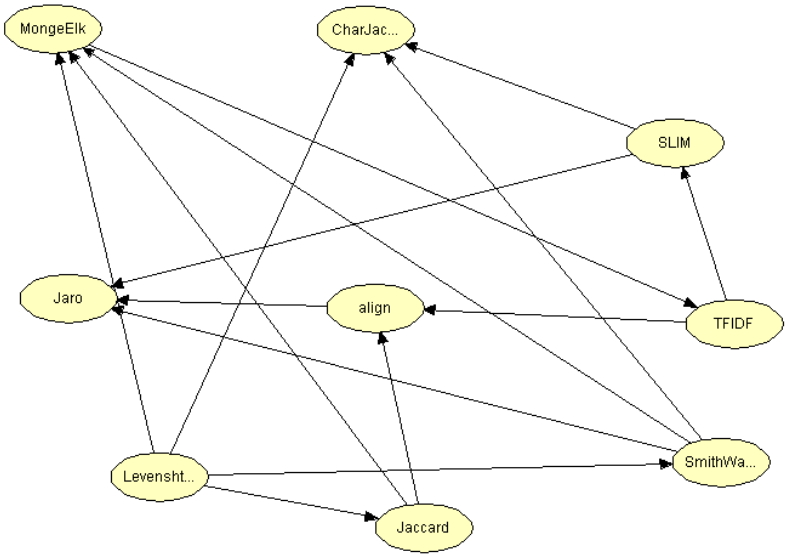

Fig. 2. BN - automatically learnt structure

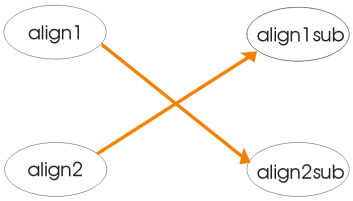

Fig. 3. Fragment of BN reflecting the mapping pattern from Fig. 1

0.66) and 88% accuracy. For the Bayesian classifier with learnt structure we got 84% precision, 53% recall (F-measure was 0.65) and 89% accuracy as best result, for whatever threshold between 40% and 70%. Both our classifiers outperform trivial classifiers that always predict true or false, respectively. Overall, the Bayesian classifier with learnt structure outperformed the naive Bayesian classifier. On the other hand, the best individual method (Jaccard) performed the same as the Bayesian classifier with learnt structure (84% precision, 53% of recall and 89% accuracy) with threshold around 50%. By this result, we can say that the combination (using BN) of string distance measures does not bring a direct benefit. However, the (second) Bayesian classifier is less sensitive to the change of threshold, while Jaccard moves towards 100% precision but rather low recall of 23% as soon as the threshold increases to 60%.

4 Conclusions and Future Work

We suggested to use low-level methods as 'mapping justifications' in order to train a Bayesian network on a sample of mappings to produce new mappings. Results of preliminary experiments with string distance measures as low-level methods are not entirely convincing in terms of performance, which can be explained by strong correlation among these methods; this correlation was actually discovered when learning the BN structure. The main role of this initial phase of research was to gain deeper insight into the problems addressed. The possibility to model explicit mapping patterns in combination with methods was also studied but not yet reflected in experiments.

In the future, we plan to employ, in the role of *mapping justifications*, not only string-based (low-level) techniques, but also e.g. graph-based or thesauri-based techniques. A more challenging task is however to design BNs reflecting the structure of *patterns*. Each method (and the final result) will be represented with a *set of nodes* corresponding to the given pattern. For example, a fragment of BN reflecting the mapping pattern from Fig. 1 is depicted in Fig. 3. It considers not only the equivalence relation but also the (proper) subsumption relation, and has four nodes that represent the alignment of each pair and each relation (equivalence of A and B, equivalence of A and X, subsumption of A and B and subsumption of A and X). `align1` represents the equivalence mapping between A and B. `align1sub` represents the subsumption mapping between A a B ($B \supset A$). `align2` represents the equivalence mapping between A a X. Finally, `align2sub` represents the subsumption mapping between A a X ($A \supset X$). Edges then should automatically be learnt for the pairs of nodes `align1` and `align2sub`, and `align2` and `align1sub`, respectively, due to strict dependencies.

We thank Jiří Vomlel for his assistance with Bayesian Networks. The research was partially supported by the IGA VSE grants no.26/05 "Methods and tools for ontological engineering", no.12/06 "Integration of approaches to ontological engineering: design patterns, mapping and mining", and by the Knowledge Web Network of Excellence (IST FP6-507482).

References

1. W3C Semantic Web Best Practices and Deployment Working Group. Ontology Engineering and Patterns Task Force (OEP). Online at `http://www.w3.org/2001/sw/BestPractices/OEP/`
2. Ghidini C., Serafini L.: Reconciling concepts and relations in heterogeneous ontologies. In: Proc. ESWC 2006, Budva, Montenegro, 2006.
3. Jensen F. V.: Bayesian Networks and Decision Graphs. Springer, 2001.
4. Mitra P., Noy N. F., Jaiswal A. R.: OMEN: A Probabilistic Ontology Mapping Tool. In: Workshop on Meaning coordination and negotiation at the Third International Semantic Web Conference (ISWC-2004), Hiroshima, Japan, 2004.
5. Pan R., Ding Z., Yu Y., Peng Y.: A Bayesian Network Approach to Ontology Mapping. In: Proceedings ISWC 2005.