Evaluating a confidence value for ontology alignment

Paulo Maio, Nuno Bettencourt, Nuno Silva, and João Rocha

GECAD - Knowledge Engineering and Decision Support Group
Institute of Engineering - Polytechnic of Porto
Porto, Portugal
{paom,nmgb,nps,jsr}@isep.ipp.pt

Abstract. Many methods for automatic and semi-automatic ontology alignment have been proposed, but they remain error prone and labor-intensive. This paper describes a novel generic process for evaluating the mappings' confidence value. This process uses rules extracted through inductive machine learning methods from the matching results proposed by others. Further, the precision and recall of the extracted rules are exploited in order to transform each rule into a mathematical formula that generates the mappings' confidence value. Mappings are then classified not as valid or invalid but through a quantitative confidence value that can be easily managed during the alignment process.

1 Introduction

Ontology alignment overcomes the information heterogeneity problem and provides mechanisms for each system to process data as if it was represented according to their internal model (ontology). The ontology alignment process aims to define an alignment between a source and target ontology [5].

The alignment specification is a very time consuming and knowledge demanding task, whose result is error prone even when domain experts are part of the process [2]. This problem is even bigger in scenarios where online alignment is required (*e.g.* e-business, e-commerce). Automatic mechanisms are necessary in order to supply the necessary consensus and speed up the interoperability process.

In last few years different methods for automatic ontology alignment have been proposed to overcome this gap, but there still remains the need to automatically combine multiple diverse and complementary alignment strategies of all indicators, *i.e.* extensional and intensional descriptions, in order to produce comprehensive, effective and efficient alignment methods [3]. Such methods need to be flexible to cope with different strategies for various application scenarios.

This paper presents a novel confidence value evaluation method based on machine learning techniques that can be easily integrated into general alignment methods like QOM [4] and PROMPT [7], or can be applied in relaxation processes required in distributed ontology alignment negotiation processes (*e.g.* [9]).

The document structure is as follows: the next section introduces our approach in order to evaluate mappings' confidence value. At the end, a brief discussion about the proposed approach is presented emphasizing the major contributions of the paper and suggesting further research and development directions.

2 Our approach

The adopted approach is based on inductive machine-learning methods. However, the extracted rules are not directly applied on classifying the mappings but serve as an input for the configuration of the system. We pursue a method that reflects the reliance of the rules upon the training dataset. In fact, the precision and recall of the extracted rules are often low, easily leading to many false positives and false negatives when applied to testing and running data sets.

The method comprises of three phases, described in the following sub-sections.

2.1 Extracting Rules

This section applies a set of machine-leaning methods (*e.g.* J48 and JRIP [1]) to the training dataset. This set (see Table 1) is comprised of several ontology mappings, in which are identified pairs of source and target ontologies' entities and the values generated by several matching algorithms [8]. The goal attribute of the learning process is the validity of the mapping.

Table 1. Partial training dataset example.

O	O'	Source Entity	Target Entity	Valid	$Matcher_1$	...	$Matcher_n$
O_1	O_2	Woman	Woman	Yes	1.00	...	0.50
O_1	O_2	HumanBeing	Hermaphrodite	No	0.13	...	0.00

The result is a set of extracted rules $SR = \{r_1, ..., r_n\}$ for each learner. A rule (r_i) can be of two types: (i) *Simple*, (Example 1) which exploits a single matching algorithm and (ii) *Complex* (Example 2), which exploits at least two different matching algorithms. Each complex rule can be split into sub-rules (sr_j). Each sub-rule establishes one and only one criteria through a unique matching algorithm.

Example 1. $valueof(StructurePlus) \geq 0.95$.

Example 2. $valueof(INRIA) \geq 0.84 \wedge valueof(Cano) \geq 0.31$.

2.2 Converting Rules into a Formula

The rules are therefore prepared to dichotomously map pairs of ontologies' entities (*i.e.* valid or invalid). This often leads to poor results evidenced by many false positives and false negatives.

For this, the proposed process disregards the rule itself, but instead evaluates its precision *i.e.* $prec(r_i)$ and recall *i.e.* $reca(r_i)$ when applied to the training set. Similar values are evaluated for each sub-rule, *i.e.* $prec(sr_j)$ and $reca(sr_j)$.

The process proceeds by combining precision and recall into a reliance value for each rule/sub-rule. For this purpose, different functions can be used, *e.g.* harmonic average $fmeasure$ (see Equation 1) and the weighted average (see Equation 2), where α allows us to trade-off between precision and recall.

$$f_\alpha(sr_j) = \frac{(1+\alpha).prec(sr_j).reca(sr_j)}{\alpha.prec(sr_j) + reca(sr_j)} : \alpha \geq 0; \tag{1}$$

$$w_\alpha(sr_j) = \alpha.prec(sr_j) + (1-\alpha).reca(sr_j) : 0 \leq \alpha \leq 1 \tag{2}$$

Furthermore, in order to (i) Abstract from the combination function and (ii) Normalize the reliance value of the sub-rule in respect to the overall rule, the $p_\alpha(sr_j)$ is defined:

$$p_\alpha(sr_j) = \begin{cases} f_\alpha(sr_j) / \sum\limits_{k=1}^{n} f_\alpha(sr_k) \\ w_\alpha(sr_j) / \sum\limits_{k=1}^{n} w_\alpha(sr_k) \\ \ldots \end{cases} \tag{3}$$

An equivalent formula is defined for evaluating the reliance of each rule (r) in respect to the set of rules (SR).

Therefore, applying rule r and p_α it is possible to evaluate the confidence value of a mapping m_i through the following function:

$$u^r_{p_\alpha}(m_i) = \sum_{j=1}^{n} valueof_{m_i}(matcherof(sr_j)).p_\alpha(sr_j) \tag{4}$$

where, $matcherof(sr_j)$ returns the name of the matching algorithm used as criteria at sub-rule j, $valueof_{m_i}(MatcherName)$ represents the matching algorithm's similarity value for m_i and n is equal to the number of sub-rules of the rule r.

2.3 Aggregating and Applying Formulas

Because each learner extracts several rules (see 2.1), several valid $u^r(m_i)$ might exist, *i.e.* there is one different $u^r(m_i)$ for each rule in the extracted set of rules. In that sense, mappings have one distinct confidence value for each rule, given by $u^r(m_i)$.

Consequently, it is necessary to choose or evaluate an unique confidence value, *i.e.* $u(m_i)$ based on all available $u^r(m_i)$. With that purpose, an aggregation function (agg) is used. A preference list over the existing rules based on the learners' additional information (*e.g.* percentage of error) and the maximum, minimum, linear average or weighted average (*e.g.* using $p_\alpha(r_i)$) are some possible agg functions. In that sense, the confidence value of a mapping, *i.e.* $u(m_i)$ is evaluate by the function presented in Equation 5:

$$u(m_i) = agg[u^{r_1}_{p_\alpha}(m_i), ..., u^{r_n}_{p_\alpha}(m_i)] \tag{5}$$

Thus, despite the mappings classification (valid or invalid), formulas deliver a quantitative confidence value ($[0 - 1]$).

This allows constraint and relaxation of the alignment requirements by changing the acceptance/rejection threshold (t_r). Therefore, given two ontologies, the suggested mappings will be those where $u(m_i) \geq t_r$, where t_r is the acceptance/rejection threshold (Example 3).

Example 3. Having $t_r = 0.8$, $u^{r_1}_\alpha(m_i) = 0.9$ and $u^{r_2}_\alpha(m_i) = 0.7$ which means that $u(m_i) = agg[0.9, 0.7]$. Thus, using maximum function as agg, m_i is considered accepted ($0.9 \geq 0.8$). On the other hand, using minimum function as agg, m_i is considered rejected ($0.7 < 0.8$).

Notice that usually t_r is exclusively defined by the user, but the training stage might provide some good hints to be used when defining the threshold.

3 Discussion

This paper presents a novel process for calculating the mappings' confidence value for ontology alignments, using the rules extracted through machine learning methods. The basic idea is to convert the extracted rules into formulas that reflect the reliance of each rule. Rules are further combined in order to generate a single value on the mapping.

Matching algorithms and their results therefore play a relevant role. Thus, a careful and strict matching algorithms selection phase is required in order to include diversification of abilities (*e.g.* hierarchy, semantics, data types and instances analysis) according to the training dataset's characteristics. Their virtues and limitations have positive and negative influence in the results obtained. Notice that due to matching algorithms' internal configuration, the similarity value between the same pair of source and target entities might be different.

Machine-learning methods play another important role in the system. In fact, they are responsible for efficiently combining matching algorithms and finding out the relevant threshold for that combination. From all initially available matching algorithms, only a few of them are combined into rules. That selection is automatically performed based on matching algorithms acting capabilities which implies no user information is needed. Extracted rules can also be updated automatically when new matching algorithms or mapped ontologies are added.

The use of more than one learner is recommended as tests show that different learners can extract different rules with similar results. Therefore, this fact can be seen as an advantage when used to avoid errors or disambiguate results.

Also, the training data set's ontologies have a special relevance. Because each ontology is related to one main knowledge domain, if training ontologies are only concerned with one matching knowledge domain (*e.g.* health care) then the learned rules and information should only be used in a similar domain. However, preliminary tests showed that learning rules and information are independent of ontologies characteristics (*e.g.* flattened hierarchy vs expanded hierarchy).

It is our conviction that the proposed approach can be easily integrated with existing automatic and semi-automatic ontology alignment tools. Also, correctness of generated alignments can be improved when combining this approach with other existing techniques as, for instance, debugging alignments with logical reasoning [6]. While evaluation results are not conclusive, they are encouraging. Ongoing research is focused in (i) the systematization of the application of the *agg* functions, α parameter depending on the requirements of the mapping scenario, and (ii) on the generalization of the proposed approach to ontology attributes and relations.

4 Acknowledgments

This work is partially supported by the Portuguese MCT-FCT project EDGAR (POSI/EIA/61307/2004).

References

[1] Auknomi. Categorical machine learning algorithms. `http://www.auknomi.com/categorical_learners.html`, 2006.

[2] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Ontology matching: A machine learning approach. Handbook on Ontologies, 2004.

[3] M. Ehrig, S. Staab, and Y. Sure. Bootstrapping ontology alignment methods with APFEL. In *Proceedings of ISWC*, pages 186–200, Galway, Ireland, November 2005.

[4] Marc Ehrig, Steffen Staab, and York Sure. Qom - quick ontology mapping. volume 3298 of *Lecture Notes in Computer Science*, pages 683–697, Hiroshima, Japan, Nov 2005. Springer.

[5] Jérôme Euzenat. An API for ontology alignment. In *Proceedings of ISWC*, pages 698–712, 2004.

[6] C. Meilicke, H. Stuckenschmidt, and A. Tamilin. Improving automatically created mappings using logical reasoning. In *Ontology Mapping Workshop at ISWC*, Athens, GA, USA, 2006.

[7] Natalya F. Noy and Mark A. Musen. The prompt suite: Interactive tools for ontology merging and mapping. International Journal of Human-Computer Studies,59(6):983–1024, 2003.

[8] Pavel Shvaiko and Jérôme Euzenat. A survey of schema-based matching approaches. *Journal on data semantics*, 4:146–171, 2005.

[9] Nuno Silva, Paulo Maio, and João Rocha. An approach to ontology mapping negotiation. Banff (Alberta), Canada, Oct 2005.