Background knowledge in ontology matching

Fausto Giunchiglia   Pavel Shvaiko   Mikalai Yatskevich

Department of Information and Communication Technology
University of Trento
38050, Povo, Trento, Italy
{fausto|pavel|yatskevi}@dit.unitn.it

1 Introduction

Ontology matching is a critical operation in many well-known metadata intensive applications, such as data integration and peer-to-peer information sharing. Typically, heterogeneity in these applications is reduced in two steps: (*i*) matching ontologies to determine correspondences and (*ii*) executing correspondences according to an application needs (e.g., data translation). In this position statement paper we focus only on the first, i.e., matching step.

In particular, we think of matching as an operation that takes two graph-like structures, such as lightweight ontologies [7], and produces a set of correspondences between the nodes of the graphs that correspond semantically to each other [8, 11].

Many diverse solutions of matching have been proposed so far, see [17, 19] for surveys. Also, recently this topic has been given a book account in [4]. It is worth noting that, on the one side, schema matching is usually performed with the help of techniques trying to guess the meaning encoded in the schemas. On the other side, ontology matching systems primarily try to exploit knowledge explicitly encoded in the ontologies. In real-world applications, various data and conceptual models usually have both well defined and obscure terms, and contexts in which they occur, therefore, solutions from both problems would be mutually beneficial [19]. Similar ideas of cross-fertilization among databases and artificial intelligence in the field of matching were also put forward in [16, 18].

Let us discuss one of the challenges in the matching area, which is the *lack of background knowledge* in matching tasks [9]. We believe that this challenge can be best tackled from the multi-disciplinary viewpoint, by building on top of the experiences in various communities, including databases, artificial intelligence and semantic web.

2 Lack of background knowledge

Recent evaluations of matching systems, such as those conducted by the Ontology Alignment Evaluation Initiative - OAEI[1] [3] as well as individual evaluations in [1, 9] show that lack of background knowledge, most often domain specific knowledge, is one of the key problems of matching systems. In fact, for example, should PO match Purchase Order or Project Officer?

Let us consider classifications, such as Google, Looksmart, Yahoo! and an evaluation dataset, called *TaxME* [1], which has been built out of them, see Table 1 for some indicators of its complexity.

Table 1: Some indicators of the *TaxME* dataset complexity.

	#nodes	**max depth**	**#labels per tree**
Google vs. Looksmart	706/1081	11/16	1048/1715
Google vs. Yahoo!	561/665	11/11	722/945
Yahoo! vs. Looksmart	74/140	8/10	101/222

As match quality measures we concentrate here on *recall*, which is a completeness measure. It varies in the [0 1] range, the higher the value, the smaller the set of correct correspondences which have not been found. For example, in OAEI-2005, on the *TaxME* dataset the best recall results were around 30%. In turn, in OAEI-2006 there has been shown some progress by the matching systems and the best recall results were around 45%. Similar results were also obtained in [1, 9], see Figure 1 and Figure 2 for the evaluation summary on the *TaxME* dataset in 2005 [1, 5] and 2006 [3, 9], respectively.

Notice that *TaxME* involves large matching tasks and, as from Figures 1 and 2, all the considered systems showed low recall values, while most of these systems for the cases of small examples, usually reported the recall around 90%. Also, contributing to this problem, the work in [15] shows that complex matching solutions, requiring months of algorithms design and development, on big tasks may perform as badly as a baseline matcher requiring one hour burden.

There are multiple strategies to attack the problem

[1] OAEI is a coordinated international initiative that organizes the evaluation of the increasing number of ontology matching systems. The main goal of OAEI is to be able to compare systems and algorithms on the same basis and to allow anyone for drawing conclusions about the best matching strategies. See `http://oaei.ontologymatching.org/` for details.

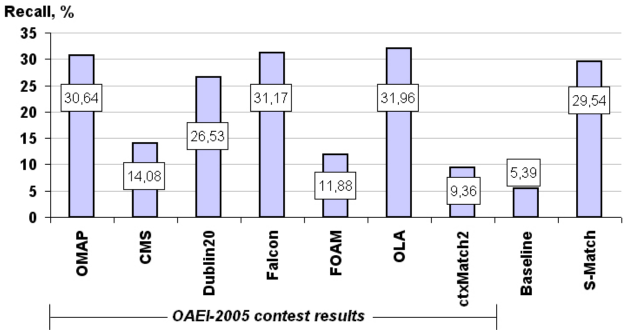

Figure 1: Recall: analytical comparative evaluation of 2005.

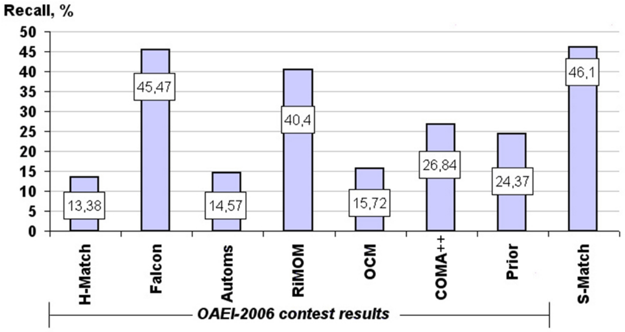

Figure 2: Recall: analytical comparative evaluation of 2006.

of the lack of background knowledge. Some of the plausible strategies include:

- Declaring the missing axioms manually as a pre-match effort, see, e.g., [2, 14].
- Reusing previous match results, see, e.g., [2].
- Querying the web, see, e.g., [12].
- Using domain specific corpus, see, e.g., [13].
- Using domain specific ontologies, see, e.g., [20].

In [9] we have proposed an automatic approach to deal with the lack of background knowledge in matching tasks by using *semantic matching iteratively.* In particular, in semantic matching as implemented in the S-Match system the key idea is that the relations (e.g., $=$, $\sqsubseteq$, $\sqsupseteq$) between nodes are determined by (*i*) expressing the entities of the ontologies as logical formulas and (*ii*) reducing the matching problem to a logical validity problem. Specifically, the entities are translated into logical formulas which explicitly express the concept descriptions as encoded in the ontology structure and in external resources, such as WordNet [6]. This allows for a translation of the matching problem into a logical validity problem, which can then be efficiently resolved using sound and complete state of the art satisfiability (SAT) solvers [8, 11]. In iterative semantic matching, in turn, the key idea is to identify critical points (hard matching tasks) in the matching process and attack them by exploiting additional sophisticated matchers which use, for example, WordNet glosses. Then, taking into account the newly discovered knowledge as additional axioms, we re-run SAT solver on a critical task.

3 Future directions

Future work proceeds at least in the following directions: (*i*) design and development of the new matchers, which for example, ask agents available on the web for the missing knowledge, (*ii*) involving user within the matching process, where his/her input is maximally useful and (*iii*) conducting further large and extensive real-world evaluations, see, e.g., [10].

References

[1] P. Avesani, F. Giunchiglia, and M. Yatskevich. A large scale taxonomy mapping evaluation. In *Proceedings of ISWC*, pages 67–81, 2005.

[2] H.-H. Do and E. Rahm. COMA – a system for flexible combination of schema matching approaches. In *Proceedings of VLDB*, pages 610–621, 2002.

[3] J. Euzenat, M. Mochol, P. Shvaiko, H. Stuckenschmidt, O. Šváb, V. Svátek, W. van Hage, and M. Yatskevich. Results of the ontology alignment evaluation initiative 2006. In *Proceedings of the ISWC workshop on Ontology Matching*, pages 73–95, 2006.

[4] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer, Heidelberg (DE), 2007.

[5] J. Euzenat, H. Stuckenschmidt, and M. Yatskevich. Introduction to the ontology alignment evaluation 2005. In *Procedings of the K-CAP workshop on Integrating Ontologies*, pages 61–71, 2005.

[6] C. Fellbaum. *WordNet: an electronic lexical database.* The MIT Press, Cambridge (MA US), 1998.

[7] F. Giunchiglia, M. Marchese, and I. Zaihrayeu. Encoding classifications into lightweight ontologies. *Journal on Data Semantics*, VIII:57–81, 2007.

[8] F. Giunchiglia and P. Shvaiko. Semantic matching. *The Knowledge Engineering Review*, 18(3):265–280, 2003.

[9] F. Giunchiglia, P. Shvaiko, and M. Yatskevich. Discovering missing background knowledge in ontology matching. In *Proceedings of ECAI*, pages 382–386, 2006.

[10] F. Giunchiglia, M. Yatskevich, and P. Avesani. A large scale dataset for the evaluation of matching systems. In *Posters of ESWC*, 2007.

[11] F. Giunchiglia, M. Yatskevich, and P. Shvaiko. Semantic matching: algorithms and implementation. *Journal on Data Semantics*, IX:1–38, 2007.

[12] R. Gligorov, Z. Aleksovski, W. ten Kate, and F. van Harmelen. Using google distance to weight approximate ontology matches. In *Proceedings of WWW*, 2007.

[13] J. Madhavan, P. Bernstein, A. Doan, and A. Halevy. Corpus-based schema matching. In *Proceedings of ICDE*, pages 57–68, 2005.

[14] J. Madhavan, P. Bernstein, and E. Rahm. Generic schema matching with Cupid. In *Proceedings of VLDB*, pages 48–58, 2001.

[15] B. Magnini, M. Speranza, and C. Girardi. A semantic-based approach to interoperability of classification hierarchies: Evaluation of linguistic techniques. In *Proceedings of CoLing*, pages 1133–1139, 2004.

[16] N. Noy, A. Doan, and A. Halevy. Semantic integration. *AI Magazine*, 26(1):7–10, 2005.

[17] E. Rahm and P. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350, 2001.

[18] A. Rosenthal and L. Seligman. Pragmatics and open problems for inter-schema constraint theory. In *Proceedings of ICDEW*, page 1, 2006.

[19] P. Shvaiko and J. Euzenat. A survey of schema-based matching approaches. *Journal on Data Semantics*, IV:146–171, 2005.

[20] S. Zhang and O. Bodenreider. Experience in aligning anatomical ontologies. *International Journal on Semantic Web and Information Systems*, 3(2):1–26, 2007.