5
Conclusions and directions for future work

In this thesis, we proposed hybrid matching techniques based on instance values and on schema information, such as datatypes, cardinality and relationships. The techniques uniformly apply similarity functions to generate matchings and are grounded on the interpretation, traditionally accepted, that “terms have the same extension when true of the same things” (Quine 1968). In our context, two concepts match if they denote similar sets of objects. The techniques essentially differ on the nature of the sets to be compared and on the similarity functions adopted. For example and in a very intuitive way, two classes match if their sets of observed instances are similar, two terms from different thesauri match if the sets of instances they classify are similar, properties match if their sets of observed values are similar.

The assumptions that the database schemas to be matched are described in OWL and that the data obtained from the databases is available as sets of RDF triples facilitated the construction of matching techniques. However, the techniques introduced in the thesis can be directly applied to conceptual schemas described in other database models, such as the relational model. In conjunction, these assumptions permitted us to concentrate on a strategy to unveil the semantics of the database schemas to be matched, without being distracted by syntactical peculiarities. In fact, we see as a good practice to provide OWL descriptions of the export schemas of data sources providers. In conjunction with WSDL descriptions of the Web Services encapsulating the backend databases, this measure facilitates the interoperability of databases.

In chapter 3 we addressed the catalogue schema matching problem. We introduced a matching approach, which applies to pairs of thesauri and to pairs of lists of properties. We then described matchings based on co-occurrence of data and introduced variations that explore certain heuristics. We also called attention to the fact that properties may have alternative representations, which impact the computation of the EMI matrix. We noted that the co-occurrence matrix (EMI) approach was not the similarity model with the best overall performance (*f*). The Contrast Model (CM) proved to be more efficient in detecting matching elements (Leme et al. 2008b), but it requires a training process. The co-occurrence matrix, on the other hand, can be used without this expensive process and performed well in the experiments.

In chapter 4 we focused on the more complex problem of matching two schemas that belong to an expressive OWL dialect. We decomposed the problem of OWL schema matching into the problem of vocabulary matching and the problem of concept mapping. We also introduced sufficient conditions guaranteeing that a vocabulary matching induces a correct concept mapping. We adopted the contrast model (Tversky and Gati 1978) as similarity function, which proved to efficiently capture the notion of similarity in this context, and described heuristics that led to practical OWL matchings.

Contrasting with (Doan et al. 2001, Madhavan et al. 2005), we did not use machine learning techniques to acquire knowledge about matchings. Instead, we captured semantic similarity by adopting similarity functions and heuristics that depend on the schema concepts. We consider this strategy to be more general because it can identify matching candidates which were not in the training corpus.

Unlike any of the instance-based techniques previously defined (see Section 1.2), the OWL schema matching process we described uses similarity functions to induce vocabulary matchings in a non-trivial way. The results showed good performances with a precision of 80% and recall of 80%.

Contrasting with (Brauner et al. 2007b, Wang et al. 2004), which measure the similarity between concepts only by the commonalities between sets of values, we used similarity functions which take into account not only the commonalities, but also the differences between concepts

Contrasting with (Bilke and Naumann 2005), we overcame the limitation of representing an instance by a string constructed out of all its property values, by representing an instance by a string constructed out of the values only of those properties that match, in a first approximation.

The work reported in this thesis appears in several publications by the author. An agent architecture for database mediation appears in (Leme et al. 2009a), the catalogue matching technique was published in (Leme et al. 2008a), and the complex schema matching technique will appear in (Leme et al. 2009b). A detailed evaluation of similarity functions and heuristics appears in a technical report (Leme et al. 2008b).

As future work, we suggest three broad areas. First, further work is required on techniques to gradually construct the matchings as new data becomes available, which is typical of a query mediation environment. We refer the reader to (Brauner et al. 2006, Brauner et al. 2008) for discussions about this issue. Second, belief revision techniques should be investigated to adjust the mediated schemas as new data sources join the mediated environment. Third, implementation issues are pending, although (Gazola 2008, Gazola et al. 2007) is a step in this direction.

In summary, unlike previous approaches, we proposed hybrid matching techniques that are uniformly grounded on similarity functions to generate matchings between simple catalogue schemas and between more complex OWL schemas. We introduced the idea of decomposing the problem of schema matching into the problems of vocabulary matching and concept mapping, which are often confused in the literature. We also showed when a vocabulary matching induces a concept mapping which is correct with respect to the integrity constraints of the schema, which is also frequently overlooked in the literature.