Uncertainty in Semantic Schema Integration

Nikos Rizopoulos[1], Matteo Magnani[2], Peter McBrien[1], and Danilo Montesi[3]

[1] Department of Computing, Imperial College London
{nr600,pjm}@doc.ic.ac.uk
[2] Department of Computer Science, University of Bologna
matteo.magnani@cs.unibo.it
[3] Department of Mathematics and Informatics, University of Camerino
danilo.montesi@unicam.it

1 Introduction

In this paper we present a new method of semantic schema integration, based on uncertain semantic mappings. The purpose of semantic schema integration is to produce a unified representation of multiple data sources. First, *schema matching* [1] is performed to identify the *semantic mappings* between the schema objects. Then, an integrated schema is produced during the *schema merging* process [2] based on the identified mappings. If all semantic mappings are known, schema merging can be performed (semi-)automatically.

As an illustrative example, consider the schemas S_1 and S_2 in Figure 1. Schema S_1 models a data source of undergraduate students. Undergraduates are registered (reg) in courses that are taught (tch) by staff members. Schema S_2 models a data source of postgraduate students, which can also optionally register in fourth-year courses to refresh their knowledge or familiarize themselves with new subjects. Therefore, S_1.student and S_2.student are disjoint, while S_1.course subsumes S_2.course. Such semantic mappings drive the schema integration process. For example, the disjointness mapping between the student entities triggers schema transformations that rename the entities to make them distinct, *e.g.* into ug and pg, and add a union entity, *e.g.* student, that represents the union set of both undergraduate and postgraduate students.

In this example, we already know the semantics of the schema objects, thus we can specify their semantic mappings. However, this is not true in general. Manual schema matching is usually time consuming and automatic schema matching is uncertain because the semantics of schema objects cannot be directly compared. In order to take into account this uncertainty in the schema matching results, we extend the concept of semantic mapping.

We assume to have a finite amount of belief, that can be distributed to alternative semantic mappings. When we are certain about a mapping, we assign all our belief to it. This is implicitly done by most existing schema matching techniques [1]. A straightforward extension of this concept can be obtained by allowing several alternative mappings to be possible, and distributing our belief to them. For example, we may think that the two student entities in S_1 and S_2 are either disjoint, in the case they refer to undergraduates and postgraduates,

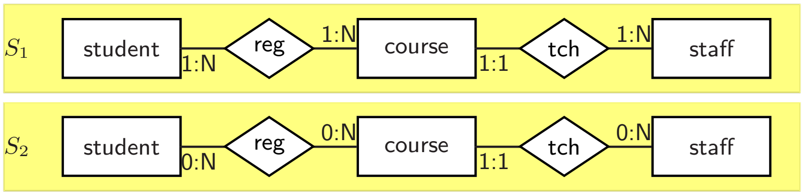

Fig. 1. Schema S_1 and S_2: undergraduate and postgraduate data sources

or equivalent. This legitimate uncertainty should not prevent the integration of the schemas. In fact, we can think of two possible integrations, the former corresponding to disjointness and the latter to equivalence. The uncertainty in the mapping between the two student entities propagates in the corresponding alternative integrated schemas. The final integrated schema is created by combining all the produced mappings.

2 A New Schema Integration Approach

2.1 Uncertain Semantic Relationships

As already discussed in the introduction, for every pair of schema objects we consider several alternative mappings, and our beliefs are distributed over them. The possible mappings are: equivalence ($\overset{S}{\equiv}$), subset-subsumption ($\overset{S}{\subset}$), superset-subsumption ($\overset{S}{\supset}$), intersection ($\overset{S}{\cap}$), disjointness ($\overset{S}{\not\cap}$), and incompatibility ($\overset{S}{\not\sim}$) [3]. We use Θ to refer to the set of all possible mappings. To represent beliefs, we have adopted Shafer's belief functions [4]. This choice is justified by the fact that Shafer's belief functions can represent the main kinds of uncertainty that are present in schema integration. An *uncertain semantic relationship* (USR) is defined by a special function m, that assigns a probability mass to sets of semantic mappings between schema objects. We omit the mathematical details, which can be found in [4].

As in [3], the comparison of schema objects is performed by a pool of experts, each one specialized on some features. However, to support the inherent uncertainty of schema matching, experts produce USRs. The mapping between any two schema objects is computed by aggregating the results of all the available experts. The aggregation of USRs is easily achieved by using Dempster's combination rule [4].

As an example, consider the comparison of the two student entities in schemas S_1 and S_2, by means of three experts. The USR produced by one expert might be: $m_1(\{\overset{S}{\cap}, \overset{S}{\supset}, \overset{S}{\not\cap}, \overset{S}{\not\sim}\}) = 1$. Another expert might produce the following USR: $m_2(\{\overset{S}{\equiv}\}) = .7, m_2(\{\overset{S}{\subset}, \overset{S}{\supset}, \overset{S}{\not\cap}, \overset{S}{\cap}\}) = .2, m_2(\{\Theta\}) = .1$. Finally, the third expert's output could be: $m_3(\{\overset{S}{\cap}, \overset{S}{\not\cap}\}) = .8, m_3(\{\Theta\}) = .2$. Due to space limitations, we cannot provide additional details about the experts. The combination of m_1, m_2, and m_3 is obtained by applying Dempster's rule, and produces the following USR: $m(\{\overset{s}{\cap},\overset{s}{\not\cap}\}) = 4/5,\ m(\{\overset{s}{\cap},\overset{s}{\supset},\overset{s}{\not\cap}\}) = 2/15,\ m(\{\overset{s}{\cap},\overset{s}{\supset},\overset{s}{\not\cap},\overset{s}{\nsim}\}) = 1/15$.

2.2 Schema Merging

In the previous example we compared student entities, obtaining a set of possible semantic relationships between them, with a corresponding representation of our belief distribution. Similarly, we can compare the reg ER relationships, and all the other objects.

#	S_1.stud.,S_2.stud.	S_1.reg,S_2.reg	S_1.course,S_2.course	S_1.staff,S_2.staff	S_1.tch,S_2.tch
(a)	$\overset{s}{\not\cap}$	$\overset{s}{\not\cap}$	$\overset{s}{\supset}$	$\overset{s}{\equiv}$	$\overset{s}{\supset}$
(b)	$\overset{s}{\not\cap}$	$\overset{s}{\nsim}$	$\overset{s}{\supset}$	$\overset{s}{\equiv}$	$\overset{s}{\supset}$
(c)	$\overset{s}{\cap}$	$\overset{s}{\not\cap}$	$\overset{s}{\supset}$	$\overset{s}{\equiv}$	$\overset{s}{\supset}$
(d)	$\overset{s}{\cap}$	$\overset{s}{\cap}$	$\overset{s}{\supset}$	$\overset{s}{\equiv}$	$\overset{s}{\supset}$
(e)	$\overset{s}{\cap}$	$\overset{s}{\supset}$	$\overset{s}{\supset}$	$\overset{s}{\equiv}$	$\overset{s}{\supset}$
(f)	$\overset{s}{\cap}$	$\overset{s}{\nsim}$	$\overset{s}{\supset}$	$\overset{s}{\equiv}$	$\overset{s}{\supset}$
(g)	$\overset{s}{\supset}$	$\overset{s}{\not\cap}$	$\overset{s}{\supset}$	$\overset{s}{\equiv}$	$\overset{s}{\supset}$
(h)	$\overset{s}{\supset}$	$\overset{s}{\cap}$	$\overset{s}{\supset}$	$\overset{s}{\equiv}$	$\overset{s}{\supset}$
(i)	$\overset{s}{\supset}$	$\overset{s}{\supset}$	$\overset{s}{\supset}$	$\overset{s}{\equiv}$	$\overset{s}{\supset}$
(j)	$\overset{s}{\supset}$	$\overset{s}{\nsim}$	$\overset{s}{\supset}$	$\overset{s}{\equiv}$	$\overset{s}{\supset}$
(k)	$\overset{s}{\nsim}$	$\overset{s}{\nsim}$	$\overset{s}{\supset}$	$\overset{s}{\equiv}$	$\overset{s}{\supset}$

Table 1. Possible combinations of semantic relationships in the integrated schema

Now assume that the comparison of S_1.reg and S_2.reg provides the same USR as in the comparison of the student entities. Also, assume that S_1.course $\overset{s}{\supset}$ S_2.course, S_1.tch $\overset{s}{\supset}$ S_2.tch, and S_1.staff $\overset{s}{\equiv}$ S_2.staff are certain. We can build a table (Table 1), representing all possible combinations of semantic relationships between all pairs of schema objects.

Notice that not all combinations are possible: the intersection relationship between the two reg ER relationships specifies that there is at least one common instance between S_1.reg and S_2.reg, *i.e.* there is a common instance of S_1.student and S_2.student that is associated with a common instance of S_1.course and S_2.course. Therefore, whenever student entities are disjoint (do not have any instances in common) reg ER relationships cannot be intersecting. Thus only eleven of the sixteen possible integrated schemas are considered in Table 1.

Each row of this table corresponds to a possible integrated schema, where each semantic relationship defines a partial integrated schema. For example, in the possible integrated schema (a) of Table 1 student entities are disjoint, while in the possible integrated schema (c) S_1.student intersects S_2.student. The schema corresponding to row (a) of Table 1 is illustrated in Figure 2.

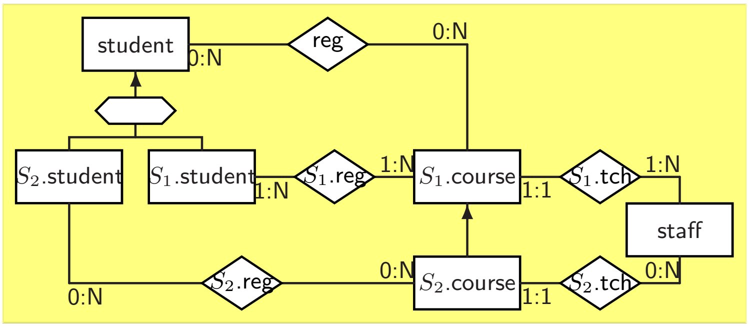

Fig. 2. One of the final alternative integrated schemas generated by our approach, corresponding to row (a) of Table 1

The belief distribution obtained as a combination of the aforementioned USRs is defined by[4]: $m\{(a), (c), (d)\} = \frac{16}{25}$, $m\{(a), (c)–(e)\} = \frac{8}{75}$, $m\{(a)–(f)\} = \frac{4}{75}$, $m\{(a), (c), (d), (g), (h)\} = \frac{16}{75}$, $m\{(a), (c)–(e), (g)–(i)\} = \frac{4}{225}$, $m\{(a)–(j)\} = \frac{2}{225}$, $m\{(a), (c), (d), (g), (h), (k)\} = \frac{4}{75}$, $m\{(a), (c)–(e), (g)–(i), (k)\} = \frac{2}{225}$, $m\{(a)–(k)\} = \frac{1}{225}$. The set $\{(a)–(k)\}$, together with the belief distribution m, is called an *uncertain integrated schema*, and is the final product of our schema integration approach on our working example. To reduce the cardinality of the possible integrated schemas, we can decide to keep only the rows to whom an amount of probability mass over a given threshold is assigned. For example, we could dispose of all rows of Table 1 but (a), (c), and (d).

3 Conclusion

In this paper we have presented a new method of semantic schema integration. Our approach differs from existing methods in that it handles the inherent uncertainty in (semi-)automatic schema matching, and supports six kinds of semantic relationships between schema objects. These features are essential to cope with real schema integration tasks, where many semantic relationships are possible, and it is very unlikely to know all of them with certainty.

References

1. Rahm, E., Bernstein, P.: A survey of approaches to automatic schema matching. VLDB Journal **10** (2001) 334–350
2. Bernstein, P.: Applying model management to classical meta data problems. In: Proc. CIDR, 2003. (2003) 209–220
3. Rizopoulos, N.: Automatic discovery of semantic relationships between schema elements. In: ICEIS (1). (2004) 3–8
4. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)

[4] With (b)–(d) we indicate all the rows between (b) and (d), *i.e.* (b), (c), (d).