An Experience with Ontology Clustering for Information Integration

Pepijn R. S. Visser and **Valentina A. M. Tamma**
CORAL - Conceptualisation and Ontology Research at Liverpool
Department of Computer Science, University of Liverpool
PO Box 147, Liverpool, L69 7ZF
United Kingdom
(tel.) +44.151.794.3709
{pepijn,valli}@csc.liv.ac.uk

Abstract

This article presents a structure of multiple shared ontologies to integrate heterogeneous sources. This structure is intended to be easy to implement to maintain and to scale, and also close to the human model of conceptualisation. The structure has been investigate in a small scale experiment set in the domain of the international coffee preparation. The coffee-preparing domain is attractive as it serves to illustrate that different communities may share knowledge at different abstraction levels.

1 Introduction

In this article we discuss a small-scale experiment investigating an architectures for the integration of heterogeneous resources. In this architecture resources are clustered on the basis of the resemblance between their conceptualisations of their domains. One of the motivating ideas is that - as with inter-person interaction - resources with a similar conceptualisation can have more 'in-depth' conversations than those who share less of their conceptualisation. The architecture investigated is intended to be closer to human conceptual model, more convenient to implement and give better prospects for maintenance and scalability. This structure of ontologies builds on ideas illustrated in two previous papers [Sha97] [Vi98b] and has been investigated in a small scale experience set in the international coffee-preparing domain. In section 2 the background of the ontology-clustering idea is discussed while section 3 presents the motivating scenario. Section 4 then presents a so-called 'ontology cluster' architecture while section 5 illustrates how communication between resources is performed in this architecture. Finally, section 6 conclusion are drawn.

2 Multiple shared ontologies

The integration of heterogeneous knowledge sources has been addressed using different approaches, some of them integrate knowledge via *shared ontologies*. All the approaches, however, are based on the some functions performing the translations between the ontologies (shared or not). These functions are often called in the literature *mapping functions*: Concepts can be shared between different resources if an appropriate mapping function can be found that translates a concept understood by one resource into a concept that is understood by another resource. This is the minimal requirement for two resources to share knowledge.

The integration of heterogeneous sources can be accomplished without an intermediate ontology. This is the so-called 'one-to-one' approach, where for each ontology a set of translating functions is provided to allow the communication with the other ontologies. Such an approach would require in the worse case, that is if the mappings are not isomorphic, the definition of $(n^2 - n)$ mapping functions, if n ontologies are comprised in the structure. This is what happens in the system OBSERVER [Men96]. This approach only seems feasible only if there are a few ontologies (resources). It also would not be very scalable because if a new resource is added to the structure this approach requires the definition of n new mapping functions.

Many architectures to integrate resources comprise a single shared ontology, an example is given by InfoSleuth, [Bay97] and by the KRAFT project [Gra97] [Gra98]. Whether such approach is conceptually realistic is a matter of debate [Sha97]. The drawbacks of dealing with a single shared ontology are similar to those of any standard (see also: [Vi98b]). Often, standards are not very convenient to use since they have to be suitable for all potential uses. Also, the task of defining such standards is often lengthy and complicated. Moreover, committing to a standard restricts the degree of heterogeneity that may exist between those using the standards, and, last but not least, standards - by their nature - resist changes, partly due to the aforementioned reasons.

In contrast to an approach in which all resources share one body of knowledge here we propose to locate shared knowledge in multiple but smaller shared ontologies. This approach, which is thought to be more flexible and scalable, is referred to as ontology-based resource clustering, or shortly, ontology clustering [Sha97]. Resources no longer commit to one comprehensive ontology but they are clustered together on the basis of the similarities they show in the way they conceptualise the common domain. Ontology clusters are then organised in a hierarchical fashion. The structure of ontology cluster is described in section 4.

Concept	Description
Coffee ingredient	The substance derived from the coffee plant that is used to prepare coffee.
Kitchen appliance	A physical object that is used as kitchen tool.
Coffee drink	A drink produced by using some coffee ingredient, water and a kitchen appliance.
Coffee maker	A kitchen appliance that produces coffee drink and is composed by a filter component, a liquid container and a coffee holder that holds either some solid substance or some liquid substance.
Heating device	A kitchen appliance that is used to warm something.
Hot water	Is water with a specific temperature greater than 90 Celsius degrees.
Solid container	Some sort of substance container.
Liquid container	Some sort of substance container.

Table 1: Concepts shared by all agents

3 Motivating Scenario

The ontology clustering approach has been investigated in a small scale experience set in the domain of preparing coffee. Four agents from four different countries are hypothesised to tell each other what coffee means in their country. In the remainder the word agent refers to either a human or a software one. Software agents were not implemented in the experiment. The agents are *François* from France, *Nicola* from Italy, *Charles* from the United Kingdom and *Klaas* from the Netherlands. The agents share a basic understanding of the domain in that they know what the basic ingredients are and that the coffee powder (where powder refers either to the ground coffee or to the instant coffee grains) and hot water somehow need to be combined, but there are regional differences. Agent know how coffee is made in their nation, what the ingredients are and the tools necessary to prepare it, and what their name is. Stereotyping these nations a bit further we here assume that François only knows about cafetière coffee[1], Nicola only knows espresso coffee (prepared with an espresso coffee maker[2], Charles only knows about instant coffee (prepared with a kettle), and Klaas can only make coffee with an electric coffee maker. The shared concepts however, should guarantee that dialogues about the meaning of unfamiliar concepts are possible and it will be illustrated that agents who share more concepts can have more 'meaningful' conversations. Table 1 shows the most important shared concepts.

Besides the universally shared concepts, the agents also have a set of local concepts about coffee drinks, such as the concept of coffee maker in their countries or the type of coffee used in their nations to prepare a coffee drink. These concepts are related to more general knowledge such as that the coffee drink has water and coffee as ingredients, that a coffee maker is a kitchen tool etc. The local concepts for each agent are illustrated in Table 2.

Although all the agents share the basic concepts above those are not the only shared concepts. In fact some more knowledge is shared by a restricted number of agents. For example, from the description of the local concepts in table 1 it is possible to notice that the concept of Electric coffee maker, known by Klaas, is quite similar to the one of Espresso coffee maker, known by Nicola. Indeed both are coffee makers and both have a component where the water is put, a component to hold the coffee ingredient that also acts as filter during the coffee preparation and a component where the coffee drink is saved.

Communication between the agents centres on finding the similarities in the conceptualisations. That is, an agent tries to explain to another agent how coffee in his country is made, using his own concepts as starting point. He will try to understand what the other agent knows and explains unknown concepts in terms familiar to the other agent. To illustrate the process, we here give an example of the type of interactions between the agents in the experiment. This specific dialogue refers to a conversation between Nicola and Franois (disregarding their native languages to preserve clarity).

Nicola:	What do you use to make coffee?
François:	I use hot water, ground coffee powder, a kettle and a cafetière
Nicola:	How hot is the water?
François:	Hot water is a kind of water that has temperature higher than 90 degree Celsius
Nicola:	You use ground coffee powder, what is that?
François:	Ground coffee powder is the same as ground coffee
Nicola:	What is a kettle?
François:	A kettle is a heating device
Nicola:	What is a cafetière?
François:	A cafetière is a coffee maker that consists of a jug and a filter device
Nicola:	Does the cafetière have a water reservoir?
François:	A cafetière has a jug that is a substance container, where substance can be either solid or liquid. The jug can contain liquid that is hot water or the actual coffee drink
Nicola:	What is the ground coffee kept in?
François:	Ground coffee is kept in the jug, as it can also contain solids
Nicola:	Does the jug have a filter, then?
François:	No, the jug does not have a filter, but the cafetire has a filter device that is a filter component.

Nicola and François are struggling to understand each

[1] A cafetière is a jug in which ground coffee is placed. After pouring hot water on the coffee a filter is pressed through the jug.

[2] An espresso coffee maker has a water tank, a filter (that also holds the coffee), and a coffee reservoir. Coffee is made by forcing boiling water under pressure (in the water tank) through the ground coffee that is held in the filter.)

François	**Nicola**	**Charles**	**Klaas**
Ground coffee powder	Ground coffee ingredient	Instant coffee grains	Ground coffee
French coffee	Espresso	Instant coffee	Dutch coffee
Cafetière (composed by a jug, a filter device and producing French coffee)	Espresso coffee maker (composed by a water reservoir, a coffee reservoir, and a filter and coffee holder, and producing Espresso coffee)	Mug	Electric coffee maker (composed by a jug, a water reservoir and a filter and coffee holder constituent and producing Dutch coffee)
Kettle		Kettle	

Table 2: Concepts not shared by all agents

other because they share only very general concepts about coffee. Moreover, the dialogue does not completely explain the relationships between the meaning of terms. For example, Nicola will be able to understand that a jug can contain both liquids and solids, but he will not be able to fully infer that the jug in the cafetière corresponds to both the water reservoir and the coffee reservoir and the filter (in that it contains ground coffee) in the Espresso coffee maker. A conversation between Nicola and the Dutch agent Klaas will be less troublesome since these agents share more concepts.

4 Ontology Clusters

Ontology clustering is based on the similarities between the concepts known to the different agents. Since in this application all agents are assumed to be familiar with concepts such as coffee beans, water, and kitchen appliances, we group these concepts in a so-called application (specific) ontology, rooted at the top of the hierarchy of ontologies. The ontology on top of the hierarchy describes the specific domain and so it is not reusable. For this reason, and following Van Heijst approach [Van97] it was named application ontology.
The concept definitions in this application ontology are derived from an existing top-level ontology, which is here chosen to be WordNet [Mil90].
The application ontology contains a relevant subset of WordNet concepts. For each concept a sense is selected, depending on the domain, from those provided by WordNet. If some agents share concepts that are not shared by other agents then there is a reason to create a new ontology cluster. A new ontology cluster here is a child ontology that defines certain new concepts using the concepts already contained in its parent ontology. The Italian and Dutch agent, for instance, share the concept of a "coffee-maker device" that has a water container a filter that also holds coffee and a coffee container. This concept is unknown to the French and English agents. Ultimately, the agents are likely to have concepts that are not shared with any other agent. In our ontology structure, we then create a separate, agent-specific ontology as sub ontology of the cluster in which the agent resides. We refer to these ontologies as mirror ontologies since they mirror the local agent ontologies. The mirror ontologies are the leaf nodes of our ontology hierarchy. Since the local ontologies are expressed in the agent's mother tongue the language heterogeneity (due to the use of different languages and different vocabulary) occurs between the local and the mirror ontologies. To overcome this kind of heterogeneity the local ontologies are translated

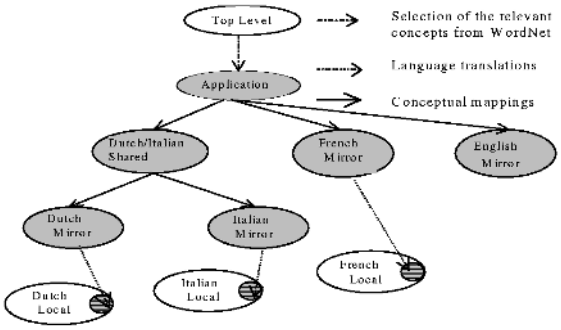

Figure 1: The hierarchy of multiple shared ontologies

in one common language, here English. In figure 1 the ontology hierarchy together with mappings between local and mirror ontologies are presented.

In each of the ontologies in the structure concepts are described in terms of attributes and inheritance relations holding in the ontology's structure. Concepts are hierarchically organised and the inheritance allows the passing down of information through the hierarchy.
Each sibling cluster specialises the concepts that are in its parent cluster, therefore the lower level clusters have more precise concept definitions than the higher levels, making the latter more abstract. Since different siblings can extend their parent cluster concepts in different ways the cluster hierarchy permit the co-existence of heterogeneous (sibling) ontologies.

Concepts are expressed in terms of *inherited* and *distinguishing* attributes. Inherited attributes are those expressing the similarities between a parent concept and its siblings (the parent concept can be defined in the ontology itself or in a parent ontology). They describe the main characteristics of a concept that are also present in its sub-concepts. A concept that specialises a more general one inherits all the attributes from its parent concept.
To the set of inherited attributes other attributes are added to distinguish the specific concept from the more general one. These attributes describe the characteristic differences between a concept and its siblings. The distinguishing attributes are used to map concepts from a source ontology into a destination ontology preserving the meaning of the concept.

5 Communication between resources

In the ontology structure communication between resources is performed via mapping functions (section 2). In this experiment mappings can be either partial or total and are not necessarily isomorphic; that is if a mapping function exists from a resource A to a resource B this does not imply that the opposite mapping from the resource B to the resource A exists.
The remainder of this section outlines how we envisage that communication between the resources in the ontology structure is performed. Two kinds of translations between ontologies are distinguished:

1. Translations of the first type are those mapping concepts from the agent's local ontology onto a corresponding concept within the 'mirror' ontology. This is a language translation and will largely imply a direct word-by-word mapping although common language-translation problems occur here (e.g. [Mah95]). This first step resolves the heterogeneity due to differences in the language and terminology used to represent the conceptualisation.
2. Translations of the second type are those that will be encoded in functions mapping concepts between the ontologies composing the structure, thus translating concepts from one ontology, possibly repeatedly, into its child or into its parent ontology. The aim of this step is to resolve ontology heterogeneity, that is ontological differences. Concepts belonging to one of the 'mirror' ontologies are mapped into concepts of another 'mirror' ontology via one or more shared ontologies. The remainder of this section will focus on this type of translations.

In the reminder we will use the term source ontology to denote the ontology containing the concept that is to be translated, whereas we use the term destination ontology to denote the ontology the concept has to be translated to.
The ontologies in the structure are hierarchically organised, and for this reason translating from the source ontology into the destination ontology may generally consist of two types of translation steps. The first type of is generalisation (from the concept to its hypernym in the same or in a parent shared ontology). The second type is specialisation (from the concept in the parent shared ontology to its hyponym in the same or in another ontology). However, the mere translation of a concept through a generalisation and a subsequent specialisation is not enough; indeed such a translation is guaranteed to preserve the meaning only if the concept to translate has a synonym in the local destination ontology. If this is not the case the concept will be mapped into a more general one, and thus it will be an approximation. This is what happens in the SIMS project [Are96] where a query is reformulated as the union of its more general concepts using the relationship holding between a class of concepts and its super-class. To preserve the meaning, however, some constraints can be added.
The translation between local ontologies can be summarised by the following steps:

a) The concept needing to be translated is identified.
b) Once identified, the concept is translated into the terms of the shared ontology immediately above the source ontology. If a direct translation does not exist the first hypernym of the concept is found such that a translation exists between the hypernym and a concept in the shared ontology immediately above. The same translation process is applied to all the concepts in the destination ontology.
c) The hypernym of the concept is then located in the shared ontology.
d) The attributes of the concept in the source ontology are compared with the attributes of the hypernym just found to select the distinguishing features;
e) Then the concept expressed in terms of the shared ontology, (that is the relationships holding between concepts in the structure are identified) together with its distinguishing attributes is passed to the parent shared ontology;
f) If in the destination local ontology there is a concept that is a specialisation of the one passed to the shared ontology, then for this local concept a mapping can be defined between the original local concept and the one just selected. If not the procedure is recursively applied, climbing up a level to the more general shared ontology.

This kind of translation obtained by subsequent generalisation and specialisation steps is effective only if the source and the destination concepts have a common ancestor that is not too high in the hierarchy, otherwise the information loss due to the generalisation is too high, and the translation obtained might be a trivial one.
To avoid the loss of information that is intrinsic of a generalisation, attributes and relations linking concepts play a crucial role. In fact they not only allow the identification of the hypernym of a concept (either in the same or in a shared ontology) but they also allow to "attach" some characterising information to each concept thus giving a distinction between the concept itself and its parent.
An example showing how translation is performed in this structure can be found in [Vis99].

6 Conclusion

In this article we reported on a small experiment in the integration of heterogeneous information sources. The aim of our experiment is to investigate the feasibility of using a set of related ontologies rather than one over-arching ontology or several independent ontologies. We discussed a proposal for an agent architecture with a hierarchical ordering of ontologies. Ontologies lower in the structure contain more refined concepts than the ontologies higher in the structure and since different branches of ontologies may extend on their concepts in different ways, the structure allows heterogeneous ontologies. The coffee-preparing domain is attractive as it serves to illustrate that different communities may share knowledge at different abstraction levels. Since all communities share the 'coffee basics', there will always be a way to explain unknown concepts in known terms, albeit that this may cause loss some of information.

Although the idea of using abstract and more refined ontologies is not a novelty, the idea to use a structured set of heterogeneous ontologies simultaneously in a distributed architecture has not received much attention. In such architectures we hope to combine the advantages of having abstract ontologies (general applicability) and refined ontologies (more meaningful communication). Unfortunately, we also inherit some disadvantages. One important disadvantage of ontology structures such as the one proposed is that translations are required between the ontologies in the structure. In the article we have shown the role of inherited and distinguishing attributes in such translations. We think the disadvantage can be outweighed by the benefits of having a more flexible and maintainable way of dealing with communication standards. Ongoing experiments will focus on the evaluation of the translations obtained with such approach, and on extending the approach in the case of real life applications with several definitions. These experiments aim at giving us more insights regarding the circumstances under which advantages and disadvantages take manifest.
This research is partly conducted as a PhD project of the second author who will continue to explore the possibilities of ontology structures and their implications on agent communication.

Acknowledgements

This research is in part funded by BT Research Laboratories in the United Kingdom. The authors wish to thank Floriana Grasso for her invaluable contribution and would like to express their gratitude to Mike Shave, Trevor Bench-Capon, Ian Finch.

References

[Are96] Y. Arens, C. Hsu, and C. A. Knoblock Query processing in the SIMS Information Mediator. *Advanced Planning Technology*, Austin Tate (Ed.), AAAI Press, Menlo Park, CA, 1996.

[Bay97] R. J. Bayardo, W. Bohrer, R. Brice, A. Cichocki, G. Fowler, A. Helal, V. Kashyap, T. Ksiezyk, G. Martin, M. Nodine, M. Rashid , M. Rusinkiewicz, R. Shea, C. Unnikrishnan, A. Unruh, and D. Woelk InfoSleuth: Agent-Based Semantic Integration of Information in Open and Dynamic Environments. *ACM SIGMOD Record Vol. 26, No. 2 (June 1997), SIGMOD '97. Proceedings ACM SIGMOD International Conference on Management of Data*, Tucson, Arizona, USA, pp. 195-206, 1997.

[Gra97] P. D. M. Gray, A. Preece, N. J. Fiddian, W. A. Gray, T. J. M. Bench-Capon, M. J. R. Shave, N. Azarmi, M. Wiegand, M. Ashwell, M. Beer, Z. Cui, B. Diaz, S. M. Embury, K. Hui, A. C. Jones, D. M. Jones, G. J. L. Kemp, E. W. Lawson, K. Lunn, P. Marti, J. Shao, and P. R. S. Visser KRAFT: Knowledge Fusion from Distributed Databases and Knowledge Bases. *Proceedings of Database and Expert System Applications Conference (DEXA' 97)*, Toulouse, France, 1997.

[Gra98] P. D. M. Gray, Z. Cui, S. M. Embury, W. A. Gray, K. Hui, A. Preece An Agent-Based System for Handling Distributed Design Constraints. *Workshop on Agent-Based Manufacturing at Agents'98 International Conference*, Minneapolis, USA, 1998.

[Gru92] T. R. Gruber A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, Vol. 2 Nr. 5, pp. 199-220, 1992.

[Kim91] W. Kim, and J. Seo, J. Classifying Schematic and Data Heterogeneity in Multidatabase Systems . *IEEE Computer*, Vol. 24, pp. 12-18, December 1991.

[Mah95] K. Mahesh, and S. Nirenburg Semantic Classification for Practical Natural Language Processing. *Proceedings of the Sixth ASIS SIG/CR Classification Research Workshop: An Interdisciplinary Meeting*, Chicago IL, USA, 1995.

[Mil90] G. A. Miller Nouns in WordNet: a lexical inheritance system. *International journal of Lexicography*, Vol. 3, Nr. 4, pp. 245-264, 1990.

[Men96] E. Mena, V. Kashyap, A. Shet, A., and A. Illarramendi OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies. *Proceedings First IFCIS International Conference on Cooperative Information Systems (CoopIS '96)*, Brussels, Belgium, IEEE Computer Society Press, pp. 14-25, 1996.

[Sha97] M. J. R. Shave Ontological Structures for Knowledge Sharing. *The New Review of Information Networking*, Vol 3, pp. 125-134, 1997.

[Van97] G. Van Heijst, T. Schreiber, and B. Wielinga Using Explicit Ontologies in KBS. *International Journal of Human-Computer Studies*, Vol. 46, Nr. 2/3, pp. 183-292, 1997.

[Vi98a] P. R. S. Visser, D. M. Jones, T. J. M. Bench-Capon, and M. J. R. Shave. Assessing Heterogeneity by Classifying Ontology Mismatches. *Proceedings Conference on Formal Ontology (FOIS'98)*, N. Guarino , Ed., Trento, Italy, pp. 148-162. An earlier version of this article appeared at the *AAAI 1997 Spring Symposium on Ontological Engineering*, Stanford, CA, USA, 1996.

[Vi98b] P. R. S. Visser, and Z. Cui On Accepting Heterogeneous Ontologies in Distributed Architectures. *Proceedings of the ECAI'98 workshop on Applications of Ontologies and Problem-solving methods*, Brighton, UK, 1998.

[Vis99] P. R. S. Visser and V. A. M. Tamma An experience with Ontology-based Agent Clustering. *Proceedings of the IJCAI'99 workshop on Ontologies and Problem Solving Methods: Lessons Learned and Future Trends*, Stockholm, Sweden, 1999.