Category Theory in Ontology Research: Concrete Gain from an Abstract Approach

Markus Krötzsch
Pascal Hitzler
Marc Ehrig
York Sure

Institute AIFB, University of Karlsruhe, Germany;
{mak,hitzler,ehrig,sure}@aifb.uni-karlsruhe.de

Abstract

The focus of research on representing and reasoning with knowledge traditionally has been on single specifications and appropriate inference paradigms to draw conclusions from such data. Accordingly, this is also an essential aspect of ontology research which has received much attention in recent years. But ontologies introduce another new challenge based on the distributed nature of most of their applications, which requires to relate heterogeneous ontological specifications and to integrate information from multiple sources. These problems have of course been recognized, but many current approaches still lack the deep formal backgrounds on which todays reasoning paradigms are already founded. Here we propose category theory as a well-explored and very extensive mathematical foundation for modelling distributed knowledge. A particular prospect is to derive conclusions from the structure of those distributed knowledge bases, as it is for example needed when merging ontologies.

1 The challenge of distributed knowledge bases

At first sight, the goals pursued in ontology research appear to be largely the same as in classical research on knowledge representation and reasoning. Indeed, the specification of knowledge and the possibility to automatically draw conclusions from such data is at the heart of many related investigations, and the well-known trade-off between semantic expressivity and computational feasibility becomes a central topic. Much progress was made in these areas, leading to numerous new logical languages which are specifically tailored to the needs of the new domain. Prominent examples include *description logics* and *frame logics*.

These issues alone, however, are not a fundamental novelty: previous decades saw extensive research on similar problems in related areas of artificial intelligence, most prominently in logic programming. Yet, ontology research introduces an elementary paradigm shift when compared to classical approaches. While it is still necessary for the single user to create logical specifications and to reason upon this data, additional emphasis is now put on the utilization of knowledge from other ontologies as well. Thus the user may gather specifications from various sources in order to draw conclusions from the combined knowledge.

Indeed, the utility value of typical applications like the semantic web lies in the fact that such distributed specifications are crafted in a rather independent fashion by many users and for different purposes. Now before one can even employ a reasoner on a single specification, one is faced with the challenge of integrating information from different distributed sources. In the first place, this requires to *compare* different ontologies, establishing how they relate to each other and to a user's local knowledge (that may typically also involve a representation of the question that the user wants to answer through the reasoning process). We will argue below that the operation of comparing two ontologies, sometimes referred to as *ontology mapping* [KS03], is indeed an atomic concept for many other tasks of distributed reasoning, in particular of the process of *ontology merging*.

2 Heterogeneity occurs on all levels

The integration of formally represented knowledge from different sources would be easy if all ontologies would be similar, using the same vocabulary for the same intended meaning, the same background assumptions, and the same logical formalism. This is not going to happen. The reason lies in the distributed way that ontologies are crafted. Different users have different preferences and assumptions that are adjusted to the requirements of the given domain. Although efforts have been made to establish standards for ontology languages and for single basic top-level ontologies, there is still a multitude of different approaches, each of which has its own merits over the others. Even in the unlikely situation that there would be global consensus in these matters, it would still be necessary to update the chosen techniques once in a while, which again would lead to non-trivial differences between various means of representing knowledge for a given purpose.

In consequence, one always has to face some amount of *heterogeneity* between available ontologies. Obviously, this heterogeneity may have different reasons and thus occurs on different levels. More than one classification for the different forms of heterogeneity has been proposed (see, e.g., [BEE$^+$04]). Possible heterogeneity emerges from different choices of underlying logic (e.g. description logic vs. frame logic), syntactic representation format (e.g. XML vs. RDF), terminological and conceptual assumptions (e.g. different languages, different top-level ontologies) and, finally, from different possible practical interpretations of the symbolic specifications.

Our current interest is mainly in situations where ontologies are based on the same or very similar logics, but where one finds substantial terminological and conceptual differences.

3 Bridging heterogeneity by relating ontologies

Heterogeneity of distributed ontologies raises the question how two given ontologies are related. This problem has many practical aspects, depending on the application domain and on the exact form of heterogeneity (see [KS03] for a survey). However, in any case one must decide on a suitable representation for a relationship between two ontologies, in order to set up a framework for specifying such relations. This task is very similar to the specification problems that are encountered when working with single ontologies. In fact, we argue that the definition of relations between ontologies can be viewed as another level of ontological engineering, which requires experts of the involved domains and specification languages to express interrelations.

Extending this view along the lines of knowledge representation and reasoning, one wonders whether there is also a possibility to *reason* with specifications of relationships. Depending on the exact definition of the considered relationships, this may already be possible for a single relation. For example, the user may specify some relationships between entities of two ontologies from which further connections can be derived. However, we want to go beyond this fundamental mechanism and draw inferences that involve numerous ontologies and relationships between them, like it is the case when collecting multiple specifications from the web.

Our approach requires one additional assumption: we consider relationships between ontologies to be *directed.* This seems to deviate from some existing proposals where one considers relationships such as binary relations, which do not have a preferred direction of interpretation. This is no real limitation, since our further investigations do not require that there is only one possible direction – we merely assume that one direction has been fixed for the given task.

Many practical kinds of relationships between ontologies immediately allow for such a directed interpretation. For instance, in order to overcome terminological heterogeneity, it might be sufficient to *translate* the concepts from one ontology into the terminology of another ontology. The relationship then is a translation function, which is clearly directed. More complex forms of conceptual heterogeneity may not allow for such a functional translation. However, one might still be able to *deduce* some knowledge in the given ontology from the knowledge that is expressed in the conceptual framework of the other. In general, relationships are often directed (even if the representation of a relationship is not) since the user is interested in expressing distributed knowledge in terms of her local specification framework, which in itself is an ontology. This relates to the intuition that the directions of the relationships between ontologies indicate a *flow of information*, a theme that has been further elaborated in the theory of Information Flow [BS97].

These ideas lead to another simple form of inference on relationships. Given a relationship from ontology A to ontology B and another relationship from ontology B to ontology C, one would like to construct a relationship from A to C. This composition operation should be well behaved in the sense that, when composing a longer chain of relationships via composing pairs of adjacent relations, the order of the compositions does not affect the result: composition should be *associative.* Thus one obtains a huge directed graph of ontologies and composable relationships between them. With very few additional assumptions[1] such a graph is exactly what is called a *category* [LR03] in mathematics. In this context, the connections that we sloppily called “relationships” are known as “morphisms.”

4 E pluribus unum: merging ontologies

Recognizing that categories arise naturally in ontology research, one can draw from additional results of elementary category theory. Especially, category theory suggests further inferences based on the structure of the morphisms between ontologies.

A typical example is *ontology merging*, a process which tries to integrate the information from many ontologies into one single comprehensive specification. In order to do so, one must be given some ontologies together with specifications of their interrelations. Depending on the expressivity of the chosen type of morphisms (i.e. primitive relationships between ontologies), it may or may not be possible to express all interrelations directly via mor-

[1] Namely that there are identity relationships that express the obvious relationship of an ontology to itself.

phisms between the given ontologies. In some cases it may be necessary to introduce an auxiliary ontology which relates to the ontologies which ought to be merged. An example of this is the case where ontologies derive from a given top-level ontology. In any case, one obtains a small graph, consisting of the ontologies to be merged, possibly some auxiliary (top-level) ontologies, and the morphisms that specify the mutual relationships.

Given such a sub-graph of a category, one can employ categorical methods to obtain a new ontology that integrates all knowledge from the given categories without introducing any additional information. Omitting the technical details, this so-called *(co-)product* of the given graph is best circumscribed as the compilation of all sound and complete inferences that can be drawn from the given ontologies and their known relationships. Finally, co-products can be constructed stepwise using so-called *pushout* constructions that merge two ontologies at a time.

However, the mentioned constructions are not possible in all categories, basically due to the fact that the ontology language might not be able to represent the merging. Very roughly speaking, the more expressivity one allows for the description of relationships between ontologies (morphisms), the more expressiveness is required in the underlying ontology language. Currently, the only ontologically motivated investigations of these issues – the *information flow framework* [Ken00] and *institution theory* [GB92] – are based on the same notion of *infomorphisms* which supports many desired categorical properties, although intitutions provide a much more general framework [Gog04]. Yet one might as well consider other more expressive types of mophisms; [KHZ05] studies examples from *formal concept analysis*.

5 Future prospects of the categorical approach

Category theory is often compared to set theory, since it provides a rather general framework for many if not all mathematical disciplines [LR03]. In contrast, ontology research is a very concrete discipline which requires only a small amount of elementary category theory. The challenge for coming investigations is to bring both approaches together, leading the abstract – but quite accurate – ideas of category theory to practical implementations. The prospect of this approach is to obtain algorithms and representations which, in addition to yielding practically relevant results, are founded on a sound theoretical basis, that guides the choice of specification languages and that helps to verify the correctness of concrete implementations.

Finally, the tools of category theory might even provide techniques for overcoming the gaps between the different logical formalisms that are used in practice. Institution theory [GB92] appears to be the first attempt to create such a unified framework for the categorical representation and comparison of a broad range of logics.

References

[BEE$^+$04] P. Bouquet, M. Ehrig, J. Euzenat, E. Franconi, P. Hitzler, M. Krötzsch, L. Serafini, G. Stamou, Y. Sure, and S. Tessaris. Specification of a common framework for characterizing alignment. KnowledgeWeb deliverable D2.2.1v2, 2004.

[BS97] J. Barwise and J. Seligman. *Information flow: the logic of distributed systems*, volume 44 of *Cambridge tracts in theoretical computer science*. Cambridge University Press, 1997.

[GB92] J. Goguen and R. Burstall. Institutions: abstract model theory for specification and programming. *Journal of the ACM*, 39, 1992.

[Gog04] J. Goguen. Information integration in institutions. Draft of paper for Jon Barwise memorial volume edited by Larry Moss, available online at `www.cs.ucsd.edu/users/goguen/projs/inst.html`, 2004.

[Ken00] R. Kent. The information flow foundation for conceptual knowledge organization. In *Proceedings of the Sixth International Conference of the International Society for Knowledge Organization*, 2000.

[KHZ05] M. Krötzsch, P. Hitzler, and G.-Q. Zhang. Morphisms in context. Technical report, AIFB, Universität Karlsruhe, 2005. Available from `www.aifb.uni-karlsruhe.de/WBS/phi/pub/KHZ05tr.ps`.

[KS03] Y. Kalfoglou and M. Schorlemmer. Ontology mapping: the state of the art. *The Knowledge Engineering Review*, 18:1–31, 2003.

[LR03] F. W. Lawvere and R. Rosebrugh. *Sets for mathematics.* Cambridge University Press, 2003.