Ontology-Based Information Systems

Ian Horrocks

An ontology is model of (some aspect of) the world – it introduces vocabulary relevant to the domain and specifies the meaning (or semantics) of this vocabulary. For example, an anatomy ontology might introduce terms for kinds of object, such as Heart, Vein, Artery, and so on, as well as terms for relationships between objects, such as isPartOf, isConnectedTo, isCausedBy, and so on. The meaning of each term is specified by combining other terms using some formal language, typically a logic. For example the meaning of "Heart" might be given as "muscular organ that is part of the circulatory system" and formalised in First Order Logic as:

$$\mathsf{Heart} \sqsubseteq \mathsf{MuscularOrgan} \sqcap \exists \mathsf{isPartOf}.\mathsf{CirculatorySystem}$$

The most widely used ontology language is the Web Ontology Language OWL. OWL was developed by the World Wide Web Consortium (W3C), and was motivated by their vision of an enhanced Web in which content would be annotated so as to indicate its meaning, so making it more accessible to automated process. The terms for such annotations are provided by ontologies and so come with a well defined meaning. OWL was based on earlier languages, such as RDF, OIL and DAML+OIL, and became a W3C recommendation on 10th February 2004.

OWL is based on a Description Logic (DL) called $\mathcal{SHOIN}$ (an acronym derived from the various constructors available in the language). DLs are simply fragments of First Order Logic with desirable computational properties. In particular, they are decidable. It is also desirable that reasoning problems (such as query answering) have low complexity, as this guarantees that efficient implementation is possible. Achieving this is, however, very difficult, and inevitably means making compromises with respect to the expressive power of the language.

Compared to First Order Logic, DLs have a more succinct syntax that does not require the explicit use of variables. A DL Knowledge Base consists of two parts: the ontology (or TBox as it is known in DLs) defines the terminology, and is closely related to the conceptual schema in a database setting; the instance data (or ABox as it is known in DLs) consists of a set of ground facts describing particular individuals, and corresponds closely to the data in a database setting.

We might ask ourselves why the formal semantics of a DL are important. In the first place, ontologies are used as a computer model of situations in the real world, and we need to establish a precise relationship between the model and the reality that is being modelled. In the second place, we want to build software systems that can answer queries with respect to ontologies and instance data. In order to implement and test such a system we need a precise specification of the intended behaviour, which is itself dependent on the meaning that we assign to the ontology and the data.

In First Order Logic (of which DLs are a fragment), we define this meaning in terms of models that are supposed to be an analogue of the real world situation being modelled. These models have a very simple structure, consisting of sets of elements representing objects of a particular type (e.g., there may be a set of elements corresponding to people and another set corresponding to vehicles) and relationships between such elements (e.g., a "drives" relationship may exist between people and vehicles). Exactly the same kind of model is used in databases: objects in the world are modeled as values (elements) and relationships as tables.

One benefit of ontologies is that they provide a coherent and user-centric view of the domain being modelled. This can help to identify possible misunderstandings and disagreements, ensuring that the understanding of meaning is shared across users and applications. Ontology based information systems can exploit the structural knowledge in an ontology to provide numerous benefits when compared to standard database systems. For example, the ontology provides a view of the data that is familiar to users, and is independent of the logical of physical arrangement of the data in the storage system. Queries can use these familiar terms, and answers to queries can reflect not only the data, but also background knowledge from the ontology. The ontology can also be used to support the integration of multiple information sources, with the ontology terms acting as anchor points for data from the different sources, and it allows for some or all of these sources to be incomplete and semi-structured, with the ontology "filling in" the missing data using structural knowledge.

The availability of tools and reasoning systems has contributed to the increasingly widespread use of OWL, and it has become the de facto standard for ontology development in fields as diverse as biology, medicine, geography, geology, agriculture and defence. Applications of OWL are particularly prevalent in the life sciences,m where it has been used by the developers of several large biomedical ontologies, including the SNOMED, GO and BioPAX ontologies, the Foundational Model of Anatomy (FMA) and the National Cancer Institute thesaurus.