Why do we need probabilistic approaches to ontologies and the associated data?

Mehmet Kayaalp
Computational Model Learning Group, Lister Hill National Center for Biomedical Communications
U.S. National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894

Introduction

Today the term 'an ontology' almost always implies a knowledge representation formalism using a particular type of deterministic logic. The deterministic view of life (i.e., representing biomedical phenomena in deterministic logic) is not just simplistic but unrealistic, because science in general and modern medicine in particular are based on the principles of uncertainty.

Probability theory is the most mature and established discipline to study uncertainty; thus, it is an integral part of any scientific method that involves hypothesis testing, controlled experimentation, and analysis of observations.

Ontologies that are not capable of incorporating probabilistic relations and enabling probabilistic inference cannot truly represent everything that modern medicine, biology, and chemistry teach us.

Definitions

Knowledge is a function of

1. information (about concepts and their interrelations),
2. truth values assigned to information, and
3. justifications of information along with their truth values.

Information is an interpretation of a statement/data.

Truth values in classical epistemology are absolute (i.e., a statement is either true or false). We here adhere to the Bayesian approach, in which truth values are probabilistic. Information becomes a *belief* after assigning a truth value.

Justification of a a belief is achieved through additional information.

Ontologies are models.

Models represent what exist (in reality and/or in the minds of their designers), according to the knowledge of their designers.

Some Characteristics

Facets: Since every individual has a unique set of experiences, and backgrounds, their interpretations as well as their ontologies of the world are bound by their biases, abstractions, and beliefs. In other words, any (subset of a) conventional ontology can model only a *facet* of a part of the world as a function of a particular set of abstractions and truth assignments.

A group of ontologies of a (e.g., scientific) community involve multiple facets of the same part of the world. A *multifaceted ontological network* (**muON**) is a set of such ontologies connected through overlapping, distinctively modeled facets.

Primitives: Every model is based on a set of primitives that are not defined further with simpler constructs. The meanings of such primitives are assumed to be interpreted uniformly by all interpreters.

Formalism: A model is formal if interpreters who adhere to the same methodology that the model is based on conclude with the same interpretation. A model can be

- formal (e.g., genetic code, a computer program),
- semi-formal (e.g., a written musical composition), or
- informal (e.g., an abstract painting)

Uncertainty: Every physical phenomenon is observed within a varying degree of uncertainty. Furthermore, the inference drawn from such observations can never be absolute or certain. Since no scientific knowledge can be irrefutable, every observation or inference is only plausible with some probability p, where $0 < p < 1$. Since we can model probabilistic nature of physical phenomena, we do not have an excuse to ignore such scientific information any longer, especially when it is made readily available in research articles.

Sequentiality: As conditions change, so do the information and our knowledge about the physical phenomena. In short, every information is valid and reliable within a definite time frame, and the truth assignment on any given belief changes over time as we accumulate new knowledge.

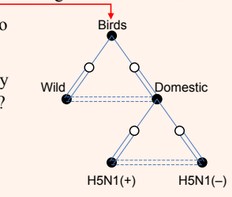
Multifaceted Ontological Networks

A multifaceted Ontological Network (muON) has been developed at the Lister Hill Center in order to map extracted information into concepts and relations of the Unified Medical Language System as well as map those concepts and relations into the information sources.

The underlying method of representation is a graphical probabilistic model called *Parameter Interdependency Networks* (**PIN**), which has been developed in order to meet the requirements of the project.

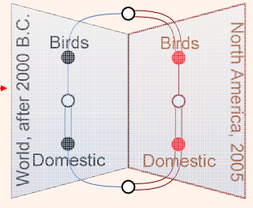
Bayesian networks and Markov random fields are other alternative methods for representing probabilistic knowledge on mathematical graphs. However, these methods have significant shortcomings. For example, their graphical notations cannot explicitly represent subsumption, union, equivalence, and disjunction between two concepts, which are essential set theoretic constructs in developing ontological relations. Furthermore, they are incapable of directly representing complex probabilistic knowledge between events.

PIN models have been developed to overcome those shortcomings.



Time: This facet is meaningful only within a particular time frame. Why? Given the domestication of birds occurred ca. 2000 B.C., and the discovery of H5N1 strain of the Avian Influenza was late 90's, such a partition of birds would be meaningless if the scope of interest is much broader than those, say, prehistoric times.

Space (as the scope of a facet) may be as important as time. For example, we are interested in the risk of an Avian Influenza outbreak in North America, say in 2005, in 2006, and later.



Prior + Data = Justified Belief

Knowledge or justified probabilistic belief is represented in muON as function of prior belief and observed data. In other words, concepts and relations in muON are directly linked to data and their probabilities are inferred/justified partly from data.

Belief updating: "Is H5N1 strain present in wild birds in North America in 2005?" To the best of our knowledge, the answer is negative. But we cannot be certain. So we may assign a p-value of 0.0001 to an affirmative answer. It is possible that we might encounter positive cases before the end of the year. Then, we should update the state of the belief accordingly.

Bibliography

Kayaalp, M. (1993) *Multifaceted Ontological Networks: Methodological Studies Toward Formal Knowledge Representation.* Master Thesis. Department of Computer Science and Engineering. Southern Methodist University, Dallas, Texas.

Kayaalp, M. (2003) *Modeling and Learning Methods*. A report to the Board of Scientific Counselors. Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine. LHNCBC-TR-2004-002, Bethesda, Maryland.

Acknowledgments

I thank to Drs. Bodenreider, Humphrey, and Rindflesch for their constructive comments.

For further information

Please contact *mehmet.kayaalp@nih.gov* or *mehmet@kayaalp.us*.