A Framework for Organization and Representation of Concept Knowledge in Autonomous Agents

Paul DAVIDSSON

Department of Computer Science, University of Lund, Box 118, S–221 00 Lund, Sweden
email: Paul.Davidsson@dna.lth.se

Abstract. In this paper the problems of organization and representation of concept knowledge are addressed from an autonomous agent perspective. The first part of the paper discusses the question of why an autonomous agent needs concepts at all and what other uses it can make of them. After the necessary and/or desirable functions of concepts have been identified, previous attempts to represent concepts are briefly surveyed to see if they support these functions. The result of the survey is negative in the sense that none of these single and simple representations are powerful enough to meet the discussed requirements. Moreover, this result indicates that, to support all the necessary and/or desirable functions, a structured composite representation is needed. Such a structure is then suggested, followed by a short discussion concerning which kinds of representations are adequate for the different parts of the structure.

1 Introduction

In the Machine Learning (ML) community there exist several different approaches for representing concepts.[1] Typical examples of such representations are logic-based descriptions [12], decision trees [19], neural networks [21], and instance-based descriptions [1]. This diversity has led to disagreement about which type of representation is the most appropriate. To settle such disputes, the different approaches are often evaluated and compared according to quantitative criteria such as accuracy, computational efficiency, and storage requirements.

On the other hand, the general opinion seems to be that a single one of these representations is sufficient to capture all the relevant aspects of a concept. The reason for this is probably that the applications using these concepts are often rather narrow, such as categorization of object-descriptions based on pre-selected features (e.g. learning from examples). In such cases there are, in fact, only one or two functions that the concepts have to serve, covering just a few aspects of concepts.

This state of affairs might be satisfying if we only wanted to use and learn concepts in such restricted domains. However, as we shall see later, it is not sufficient when dealing with *autonomous agents* acting in *dynamic environments*, since they need concepts to serve multiple functions. But which are these functions? This is a question which has not attracted enough attention within AI in general and within ML in particular. Since this issue is not discussed in the AI literature, and since humans obviously are autonomous agents and probably have

[1] In this paper "concept" refers to an agent's internal representation of a *category*. "Category", in turn, refers to a class of entities in the world.

concepts for serving similar functions, it seems appropriate to look at what functions human concepts serve or should serve (according to the cognitive sciences). In this way we may get an idea of which functions are desirable for an artificial agent. The goals of this paper are to state the requirements which these functions impose on the representation of concepts and to explore the possibility of meeting these requirements.

In the next section I will try to identify the functions of human concepts. The necessary and/or desirable functions of concepts for an artificial autonomous agent are then discussed in Section 3. Section 4 gives a brief survey of previous attempts to represent concepts to see if they support these functions. The analysis of this survey makes clear that none of these single and simple representations are powerful enough to meet the demanded requirements. Moreover, this result indicates that to support all the necessary and/or desirable functions, a structured composite representation is needed. In Section 5 such a structure is suggested, followed by a discussion in Section 6 concerning which kinds of representations are adequate for the different components of the structure. For a broader discussion of these and related topics, see [4].

2 The Functions of Human Concepts

To begin with, we can state that concepts seem to be the very stuff out of which reasoning and other cognitive processes have as their basis. However, it is possible to distinguish several functions of human concepts, some of them are (according to, for instance, [20] and [22]):

- Stability functions
- Cognitive economical functions
- Linguistic functions
- Metaphysical functions
- Epistemological functions
- Inferential functions

Concepts give our world *stability* in the sense that we can compare the present situation with similar past experiences. For instance, when confronted with a wasp,[2] we can compare this situation with a situation some years ago when we were stung by another wasp and consequently take the appropriate measures. Actually, there are two types of stability functions, *intrapersonal* and *interpersonal*. Intrapersonal stability is the basis for comparisons of cognitive states within an agent, whereas interpersonal stability is the basis for comparisons of cognitive states between agents.

By partitioning the world into categories, in contrast to always treating each individual entity separately, we decrease the amount of information we must perceive, learn, remember, communicate and reason about. In this sense we can say that categories (and thus concepts) promote *cognitive economy*. For instance, by having one representation of the category wasp instead of having a representation for every wasp we have ever experienced, we do not have to remember that the wasp we saw yesterday has a stinger or that it has a nervous system.

[2] Here, and in the following, "wasp" refers to the black- and yellow-striped wasp (yellow-jacket).

The *linguistic function* is mainly providing semantics for linguistic entities (words), so that they can be translated and synonymy relations be revealed. For instance, the fact that the English word "wasp" and the Swedish word "geting" have the same meaning enables us to translate "wasp" into "geting" and vice versa. Furthermore, it seems that it is the linguistic function together with the interpersonal stability function that make it possible for us to communicate (by using a language).

In philosophy, metaphysics deals with issues concerning how the world is, while epistemology deals with issues concerning *how we know* how the world is. Thus, we might say that the *metaphysical functions* of a concept are those that determine what makes an entity an instance of a particular category. For example, we say that something actually is a wasp if it has a particular genetic code or something like that.[3] The *epistemological functions* then, are those that determine how we decide whether the entity is an instance of a particular category. For instance, we recognize a wasp by colour, bodyshape and so on.[4]

Finally, concepts allow us to *infer* non-perceptual information from the perceptual information we get from an entity, and to make predictions concerning it. Thus, we can say that concepts enable us to go beyond the information given. For instance, by perceptually recognizing a wasp we can infer that it is able to hurt us. We know that all wasps have a stinger and that a stinger can be used for hurting other animals.

3 Functions of Concepts in Artificial Autonomous Agents

As pointed out in the introduction, the functions of concepts have never really been subject for discussion in the AI-literature. The main reason for this is probably that AI researchers often do not study problems from an autonomous agent perspective. Instead, they make the assumption that the concepts acquired are to be used for some classification task under human supervision. Typically, the task can be described as: finding a concept description such that the system correctly can classify the given training instances (described by a number of observable features). Thus, in my terms, the function of the acquired concept is mainly of an epistemological nature. The other functions are to a great extent taken cared of by the human supervisor. In my opinion, this leads to a narrow view of the problem where several difficulties are ignored. To get a broader view, I will base the discussion of functions of concepts on my own reflections on the previous section.

The functions of intrapersonal stability and cognitive economy are of course important, but they are trivial in the sense that they emerge more or less automatically for the agent just by having concepts, independently of the choice of representation. By analogy with the stability functions, we can say that an agent can have both intrapersonal and interpersonal linguistic functions. Where the intrapersonal function is a rather weak one, implied only by the fact that the concepts have names internal to the agent. This function is, of course, also trivial in the same sense as above. But what about the interpersonal stability and linguistic

[3] I use the word "metaphysic" in a more pragmatic way than in philosophy. In my notion that which makes an entity an instance of a particular category is decided by some kind of consensus amongst the agents in the domain. The example with the wasp is unfortunate though, since there exist several competing views of biological taxonomy. For instance, the cladistic view, which categorizes species according to shared derived features, and the phenetic view, which categorizes them on the basis of overall similarity.

[4] For human concepts this distinction is maybe not as clean-cut and unproblematic as described here, see [9], but nevertheless it suites my purposes very well.

functions? They are clearly not necessary in a one-agent scenario. However, if we are interested in a multi-agent scenario with communicating agents, the concepts must support also these functions.

However, it is the remaining three functions, the metaphysical, the epistemological and the inferential, that are the most interesting, and the ones I will concentrate on in the remaining part of this paper. Since an autonomous agent obviously should be able to classify objects in ordinary situations, the epistemological function is necessary. The metaphysical functions can of course be useful for an agent to have, but in most cases it seems that it can manage without them. Finally, if the agent is to be able to reason and plan about objects it is necessary that it have at least some inferential functions.

4 Representation of Concepts

Traditionally in AI, categories are treated as equivalence classes that can be characterized in terms of necessary and sufficient conditions. This is a rather strong version of what in cognitive psychology is called the *classical view* [23].

4.1 The Classical View

According to the classical view, all instances of a category share common features that are singly necessary and jointly sufficient for determining category membership. Thus, it would be possible to represent a concept by these features. Categorization would then be a matter of straightforward application of this "definition". Some of the representation languages that have been used for such definitions are: *logic-based notation* (attribute-value pairs) [12, 15], *decision trees* [19] and *semantic networks* [25].

However, as often noted in recent cognitive science literature (see for instance [23] and [22]), there are some problems with the classical view. The most serious problem is probably that it is often not possible to find necessary and sufficient features for natural categories, in contrast to artificial categories.[5] This problem is sometimes called the ontological problem [2]. Moreover, there are unclear cases of category membership. For instance, it is hard to decide for some objects whether they are a bowl or a cup.

Furthermore, assuming that a classical definition exists for a category, it is interesting to notice that instead of using the classical definition we often use non-necessary features to categorize objects of the category. Thus, it seems that, at least humans, do not use classical definitions to implement the epistemological function.

Finally, it is generally believed that some exemplars of a category are more typical than others. *Prototype* usually refers to the best representative(s) or most typical instance(s) of a category as opposed to the treatment of categories as equivalence classes. For instance, it has been shown that (at least for the experiment subjects) robins are prototypical birds whereas penguins are not.

Thus, it seems clear that the classical view cannot explain all aspects of natural concepts. In response to this, the *probabilistic* and the *exemplar view* [23] have been presented by

[5] Artificial categories are typically categories that are constructed for a particular experiment, whereas natural categories are those that have evolved in a natural way through everyday use. Artificial categories are often constructed to be specified by a short and simple definition in terms of necessary and sufficient conditions.

cognitive scientists as views being more realistic and consistent with empirical findings. These views have also been adopted by some scientists working in the ML field.

4.2 Non-classical Views

According to the probabilistic view, concepts are represented by a summary representation in terms of features that may be only probable or characteristic of category members. Membership in a category is graded rather than all-or-none. Better members have more characteristic properties than the poorer ones. An object will then be categorized as an instance of some category if, for example, it possesses some critical number of properties, or sum of weighted properties, included in the summary representation of that category.

Followers of the probabilistic view in AI are, for instance, de la Maza [5] who calls his type of representation *augmented prototypes*. Fisher's [6] *probabilistic concept tree* represents a taxonomy of probabilistic concepts.

Those in favor of the exemplar view argue that categories may be represented by some of their individual exemplars, and that concepts thus are represented by representations of these exemplars. A new instance is then categorized as a member of a category if it is sufficiently similar to one or more of the category's known exemplars. There are several models consistent with the exemplar view. One such model is the *proximity* model that simply stores all instances. An instance is categorized as a member of the category that contains its most similar stored exemplar. Another model is the *best examples* model. It only stores selected, typical instances. This model assumes that a prototype exists for each category and that it is represented as a subset of the exemplars of the category. Another possible alternative is that the prototype is a non-existing "average" instance that is derived from the known instances.

In AI, Kibler and Aha [7] have experimented with both the proximity model and selected examples models where a subset of the instances are stored. Systems that use this kind of representation often use some version of the nearest neighbor algorithm to classify unknown instances. That is, a novel instance is classified according to its most similar known instance. Musgrove and Phelps [16] have chosen to have a singular representation of the average member (not necessarily an actual instance) of the category, which they call the prototype. Nagel [17] presents a best examples model that, in addition to the prototype(s), stores transformations that transform less typical instances to a prototype. Learning systems that use specific instances rather than abstractions to represent concepts have been labeled *instance-based* [1].

A quite different approach to non-traditional concept representation is taken by Michalski and his colleagues [13, 3]. Their representation has two components, the *base concept representation* (BCR) and the *inferential concept interpretation* (ICI). The BCR is a classical representation that is supposed to capture typical and relevant aspects of the category, whereas the ICI should handle exceptional or borderline cases. When categorizing an unknown object, the object is first matched against the BCR. Then, depending on the outcome, the ICI either extends or specializes the base concept representation to see if the object really belongs to the category.

These non-traditional representations are sometimes commonly called prototype-based representations. In addition to what is normally called prototype-based representations, there has in the last few years been a growing optimism about the capability of *neural networks* for dealing with concepts. For instance, *backpropagation networks* [21] have been suggested for the learning and representation of concepts.

So, how should autonomous agents represent concepts? Should we use logic-based representations, decision trees, instance-based or probabilistic representations or maybe neural networks? Let us analyze these questions in terms of the functions that the concepts should be able to serve, starting with the epistemological function.

The epistemological function of concepts is what makes an agent able to classify objects on the basis of the perceptual input that it receives from the environment. It was mentioned earlier that it is not possible to find a definition based on necessary and sufficient conditions for all natural categories (the ontological problem). But even if such a definition exists, as is true for many categories, it is often based on features that under normal circumstances are not detectable by an agent's perceptual system. Examples of such features are; atomic structure, genetic code and functionality (of the instances of the category). Thus, classical definitions are not adequate for perceptual classification, and consequently not appropriate for supporting the epistemological function. Contrary to this conclusion, it is very common in ML to try to make a classical definition of a category based directly on the perceptual data.

The metaphysical function of concepts, on the other hand, is what determines if an object actually is an instance of a particular category. Such crucial conditions for category membership might not exist for all categories (an obvious consequence of the ontological problem). Also, these conditions must hold for every instance of the category and must be explicitly stated. This implies that prototype-based representations are not adequate for supporting the metaphysical function.

Finally, to implement the inferential function we must have some "encyclopedic" knowledge about the category and its members. Kirsh [8] has called this collection of knowledge "a package of associated glop". It includes mainly declarative knowledge, in contrast to the more procedural classification knowledge. Concerning the representation of this knowledge, it is obviously not adequate to use classical definitions or prototype-based representations. They should, of course, be used for classification purposes only (representing classification knowledge).

5 The Idea of a Composite Structure

The discussion above makes clear that it is not possible for a single and simple structure could capture all the relevant aspects of a concept. We need a richer composite representation that, in some way, is structured according to the functions of the concept to be represented. Therefore, I propose the structure illustrated in Figure 1 as a reasonable candidate for the representation of concepts by autonomous agents. The dashed boxes in the figure indicate optional components. All parts of the representation are not always necessary or even adequate. Metaphysical representation only exists for some concepts and might, besides, be irrelevant for an autonomous agent. External designators are only necessary for communicating agents in a multi-agent scenario. On the other hand, it might be convenient to have more than one epistemological representation, since the perceptual classification often is dependent of the situation (context). For instance, in daylight wasps can be recognized by their look, whereas they must be recognized by their sound when it is dark.

Let us illustrate the idea of composite representation using the category "wasp". The kind of information that the epistemological representation may include are that; wasps are

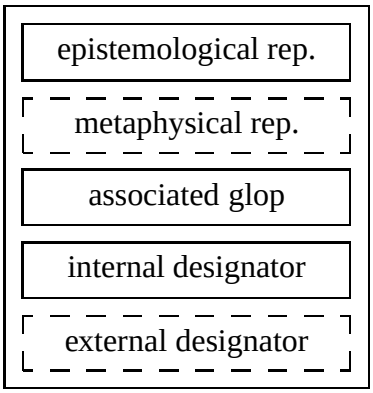

Figure 1: Composite Concept Representation

black and yellow striped, cylinder-shaped, approximately two centimeters long and half a centimeter in diameter, hum, have two wings, and so on. The metaphysical representation on the other hand, may include information of the genetic code of wasps. The kinds of encyclopedic knowledge that the associated glop would include are for instance, that they can hurt other animates with their stings and that they live in collectives. The internal designator could be something like "organism.animate.xxx"[6], whereas the external designator would be "wasp" (in an environment where communication is based on the English language, that is).

There are of course no sharp distinction between what information that is included in these representations in the sense that they may contain redundant information. For example, besides being an essential part of the epistemological representation, the fact that wasps have wings is a rather natural part of the encyclopedic knowledge represented in the associated glop. However, the fact is probably not represented in the same way in these representations. For instance, it may be rather implicitly represented in a prototype-based representation for the epistemological representation and explicitly represented in a logic-based notion for the associated glop.

This composite structure enables concepts to serve all the functions listed before. The epistemological and metaphysical representations support the epistemological and metaphysical functions respectively. The associated glop supports the inferential function. The internal designator supports the intrapersonal stability, whereas the external designator supports both the interpersonal stability and the linguistic function.

An issue that is not yet discussed is how these concept structures should be organized and how they relate to each other. We know that categories can be hierarchically organized in taxonomies. For instance, "fruit" is a superordinate category to "apple", whereas "Red Delicious" is a subordinate category. Taxonomies serve an important function by promoting cognitive economy. Since categories *inherit* features from their superordinate categories, it is possible to reduce the amount of information that have to be stored at each level in the hierarchy. For instance, if we know that all fruits are sweet, we do not have to remember that apples are sweet (if we know that apples are fruits).[7] Thus, we need to complete the

[6]The choice of the internal designator is entirely up to the system, it should be as convenient and effective as possible for the system.

[7]However, it seems that it is mainly encyclopedic knowledge that can be inherited in this manner. It is not clear how this could be done with classification knowledge (epistemological and metaphysical). If it is possible,

composite representation suggested above with taxonomical information, so that the concepts together form a tree-structure.

Depending on the situation the composite concept representation is *accessed* (or retrieved) in different ways. External "stimuli" in the form of direct perception of objects access the concept via the epistemological representation. Thus, the epistemological representation reminds of *percepts* as described by Sowa [24] in the context of his theory of conceptual graphs. If, on the other hand, the external stimuli is on the linguistic level, as when communicating with other agents, the concept is accessed via the external designator. Finally, if the stimulus is internal, like in the case of reasoning, the concept is accessed via the internal designator.

6 How Should the Components be Represented?

We now know *what* information that should be included in the different parts of the composite structure, but not *how* it should be represented. In Section 4.3 it was pointed out which types of representations that are not appropriate for the different parts. In this section I will briefly discuss which kinds of representations that might be appropriate.

It was concluded above that classical definitions are not adequate for supporting the epistemological function. Moreover, the bulk of research on this matter in cognitive science suggests that humans probably use some kind of prototype-based representations for this purpose. Thus, it seems that a prototype-based representation would be a good choice for the epistemological representation. Since there exist several such representations - probabilistic, instance-based, and different types of neural networks - it is a subject for future research to find the best choice of these or, perhaps, invent better ones.

The reasons for not using prototype-based representations to support the metaphysical function were that the crucial condition for category membership must hold for every instance and be explicitly stated. Thus, the implementation of the metaphysical function demands, almost by definition, a classical definition. The most common ways to express such a definition are either in a logic-based notation, such as predicate logic, or by a decision tree.

To support the inferential function we need some encyclopedic knowledge about the category and its members. This knowledge might be seen as a collection of universal or probabilistic rules. Seen from this perspective, it seems natural to express it in some logic-based notation. But alternative approaches exist, for instance, *semantic networks* [18] and *frames* [14]. (In fact, one might see the composite structure suggested above as a "meta-frame" where the five parts correspond to slots to be filled in.) However, results from the research of the Cyc project [11] indicate that it is necessary to combine several of these representation languages to capture all the "associated glop" that might be relevant for an autonomous agent.[8]

---

it could significantly increase the effectiveness of the classification process (as is done in, for instance, UNIMEM [10]).

[8] Using the terminology used in this article, the long-term goal of the Cyc project can be described as capturing all the associated glop of every category known to man.

7 Conclusions

The main lesson to learn from this paper is that there are different kinds of knowledge about concepts that are used for different functions. We have concluded that a single and simple structure does not suffice to account for all the functions that we want concepts to serve. Instead, an autonomous agent must have a complex (composite) concept representation. A suggestion for such a structure which supports the most important functions has been presented. It has an epistemological representation for perceptual (normal) categorization and an optional metaphysical representation for more "scientific" categorization. As we have seen, it seems that some kind of prototype-based representation is the best alternative for the epistemological representation, whereas a classical representation is the most appropriate for the metaphysical. To be able to reason and make predictions about the category and its members, the agent needs a large amount of encyclopedic knowledge. This is stored in the "associated glop". Moreover, to support stability and linguistic functions, the structure also includes an internal and an external designator.

The ideas presented in this paper stem from a fresh view on concepts. Concepts should not only be used for some limited classification task. Instead, they should provide the basis for most of an agent's cognitive tasks. The ideas presented here are clearly not fully developed; rather, they suggest possible starting points for future research. Moreover, this paper is written with categories of physical objects in mind. Thus, another subject for future research is to investigate whether this approach can be modified to represent categories of abstract entities such as events and actions.

Acknowledgements

I wish to thank Eric Astor and Peter Gärdenfors for helpful comments and suggestions.

References

[1] D.W. Aha, D. Kibler, and M.K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.

[2] J. Amsterdam. Some philosophical problems with formal learning theory. In *AAAI-88*, pages 580–584, 1988.

[3] F. Bergadano, S. Matwin, R.S. Michalski, and J. Zhang. Learning two-tiered descriptions of flexible concepts: The POSEIDON system. *Machine Learning*, 8(1):5–43, 1992.

[4] P. Davidsson. Concept acquisition by autonomous agents: Cognitive modeling versus the engineering approach. LUCS 12, ISSN 1101-8453, Cognitive Science, Lund University, Sweden, 1992.

[5] M. de la Maza. A prototype based symbolic concept learning system. In *Eighth International Workshop on Machine Learning*, pages 41–45, 1991.

[6] D.H. Fisher. A computational account of basic level and typicality effects. In *AAAI-88*, pages 233–238, 1988.

[7] D. Kibler and D. Aha. Learning representative exemplars of concepts. In *Fourth International Workshop on Machine Learning*, pages 24–30, Irvine, CA, 1987.

[8] D. Kirsh. Second-generation AI theories of learning. *Behavioral and Brain Sciences*, 9:658–659, 1986.

[9] G. Lakoff. *Women, Fire, and Dangerous Things: What categories reveal about the mind*. The University of Chicago Press, 1987.

[10] M. Lebowitz. Experiments with incremental concept formation: UNIMEM. *Machine Learning*, 2(2):103–138, 1987.

[11] D.B. Lenat and R.V. Guha. *Buildig Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley, 1990.

[12] R.S. Michalski. Pattern recognition as rule-guided inductive inference. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2(4):349–361, 1980.

[13] R.S. Michalski. How to learn imprecise concepts: A method for employing a two-tiered knowledge representation in learning. In *Fourth International Workshop on Machine Learning*, pages 50–58, Irvine, CA, 1987.

[14] M. Minsky. A framework for representing knowledge. In P.H. Winston, editor, *The Psychology of Computer Vision*, pages 211–277. McGraw-Hill, 1975.

[15] T.M. Mitchell. Version spaces: A candidate elimination approach to rule learning. In *IJCAI-77*, pages 305–310, Cambridge, MA, 1977.

[16] P.B. Musgrove and R.I. Phelps. An automatic system for acquisition of natural concepts. In *ECAI-90*, pages 455–460, Stockholm, Sweden, 1990.

[17] D.J. Nagel. *Learning Concepts with a Prototype-based Model for Concept Representation*. PhD thesis, Rutgers, The State University of New Jersey, 1987.

[18] M.R. Quillian. Semantic memory. In M. Minsky, editor, *Semantic Information Processing*, pages 227–270. MIT Press, 1968.

[19] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[20] G. Rey. Concepts and stereotypes. *Cognition*, 15:237–262, 1983.

[21] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol.1: Foundations*. MIT Press, 1986.

[22] E.E. Smith. Concepts and thought. In R.J. Sternberg and E.E. Smith, editors, *The Psychology of Human Thought*. Cambridge University Press, 1988.

[23] E.E. Smith and D.L. Medin. *Categories and Concepts*. Harvard University Press, 1981.

[24] J.F. Sowa. *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, 1984.

[25] P. Winston. Learning structural descriptions from examples. In *The Psychology of Computer Vision*, pages 157–209. McGraw-Hill, 1975. Also in Readings In Knowledge Representation, ed. R. Brachman and H. Levesque, Morgan Kaufmann, 1985.