

Improving the efficiency of ontology engineering by introducing prototypicality

Xavier Aimé¹ and Frédéric Fürst² and Pascale Kuntz³ and Francky Trichet⁴

1 Introduction

Currently, a lot of Information Systems exploit ontologies which are defined as conceptual representations of knowledge related to a given field and are built according to a consensus shared by the members of an ingroup. Classically, an ontology is composed of ordered sets of concepts and properties. However, this model loses a part of knowledge the members of the ingroup have on the domain. In particular, it does not specify how a concept is representative of its upper concept. Nevertheless, this notion, called *prototypicality* in cognitive psychology, plays a major role in the construction of any conceptual categorization[4]. We propose to introduce this notion in ontology engineering to complete the classical models.

2 Introducing prototypicality in ontology design

We introduce the prototypicality by weighting the different kinds of links in the ontology. These weights form numerical gradients which express three kinds of prototypicality:

- **Semiotic-Based Conceptual Prototypicality Gradient (SPG)** defined on isa links between two concepts: two concepts, so that one inherits from the other, can be more or less close together. More precisely, in a sibship of concepts, some of them are more prototypical of their father than the others. For example, among all plane types, the most representative in our mind of the *plane* concept is rather a modern and commercial plane, and not the first biplane or a plane propelled by muscular strength.
- **Lexical Prototypicality Gradient (LPG)** defined on links between a concept (or property) and each term that expresses the concept (or property): for a given concept (or a property) with several terms used to name it, some terms are most often used than the others. For example, *plane* is rather used than *aeroplane* or *heavier-than-air*.
- **Instance Prototypicality Gradient (IPG)** defined on links between a concept and each instance of the concept: among the instances of a given concept, some instances are more representative than the others. For example, when thinking to the concept of *plane-which-flew-over-the-Atlantic*, we think rather to the *Spirit-of-St.-Louis* than to the current planes that fly across the Atlantic.

From a semiotic point of view [3], SPG is composed of three components: (1) an **intensional component**, based on the comparison between the concept intensions (*i.e* the properties of concepts), (2) an

extensional component, based on the comparison between concept instances and (3) an **expressional component**, based on the comparison between the concept terms which appear in a corpus. The computation of the three components is based on specific requirements: (i) a set of concept instances, which must be representative of the cognitive universe of the ingroup, (ii) a corpus of documents representative of the cognitive universe of the ingroup.

Moreover, the intensional component can be refined by weighting links between properties and concepts: these weights express, for each concept and each property, the weights quantify the degree of significance of the property. They are currently specified manually. For each property, all the associated concepts have to be ordered on the interval [0,1]. For example, for the property *has_author*, the concept *Scientific article* will be placed first (close to 1), the concept *Press article* will be placed after (because to have an author is less important for this concept) and the concept *Instructions leaflet* will be placed close to 0. Obviously in real life applications, all these resources are not necessarily available. However, our approach prototypicalities remains valid even for limited resources.

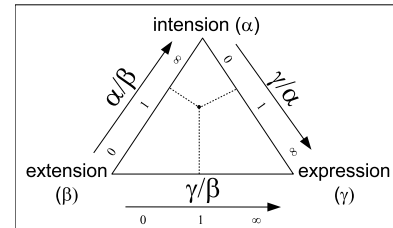


Figure 1. The semiotic triangle with coefficients. γ/α close to 0 means that the conceptualization of the domain is more intensional than expressional; this ratio close to ∞ means the opposite and this ratio equal to 1 means that the conceptualization is as intensional as expressional.

Moreover, our model allows to modulate the influence of each component of the SPG, in order to take the domination of the intensional, extensional and/or expressional aspects in the conceptualisation of the domain into account. These dominations are conditioned by the domain, the ingroup and the applicative context. For example, in the field of mathematics, the intensional aspect prevails. In the domain of biological species, a biologist builds its conceptualisation on biological properties (the intensional aspect prevails), but most of people usually use extensional conceptualisations, based on the animals encountered during their lives. α , β and γ are positive coefficients which weight each component of the SPG. To normalize the values, we impose the SPG and each of its component to be between 0 (minimum representativeness) and 1 (maximum representativeness), and $\alpha + \beta + \gamma = 1$. The values of these 3 coefficients

¹ LINA, Université de Nantes & Tennaxia, France, xaime@tennaxia.fr

² MIS, Université de Picardie, France, frederic.furst@u-picardie.fr

³ LINA, Université de Nantes, France, pascale.kuntz@univ-nantes.fr

⁴ LINA, Université de Nantes, France, francky.trichet@univ-nantes.fr

can be set arbitrarily or experimentally; we also provide a method to set them automatically, by calculating average ratio between instances and concepts in the corpus and average ratio between number of properties and number of instances of concept.

Works in cognitive psychology show that the emotional state of a person influences his perception of categories: the more a person is stressed, the more his mind is concentrated on objects cognitively close to those he is directly concerned by[2]. For that reason, we introduce an **emotional factor**, represented by a coefficient δ , that can vary from 0 to 1, for an open state of mind, and from 1 to ∞ , for a closed state of mind. SPG, LPG and IPG are powered by $1/\delta$, in order to modulate the gradients according to a state of mind: for an open state of mind, weak values of prototypicality are increasing and for a closed state of mind, they are decreasing.

2.1 Intensional component

The intensional component of the SPG measures the representativity of a concept c for its father c_f by comparing properties of the two concepts. Like in [1], concepts are represented by vectors in the vectorial space of properties. But comparing with Au Leng and Leng our vector coordinates are not fuzzy truth values but values which estimate the significance of the properties in the concept definitions. The prototype of the upper concept c_f is calculated as a vector $\vec{p}_{c_f}$, and the intensional prototypicality between c and c_f is the euclidian distance between $\vec{p}_{c_f}$ and $\vec{v}_c$, the vector that represents c . The prototype of a concept c_f has been introduced in [1] as the average of the vectors associated to the children of c_f . But [1] only considers the concepts which directly inherit from c_f , whereas we extend the calculation to all the descendants of c_f , because properties which appear only in an indirect descendant of c_f can nevertheless appear in the prototype of c_f . For example, to wear a white coat is not a property of the concept *Researcher*, but the prototype of *Researcher* can have this property.

2.2 Extensional component

The extensional component of the SPG measures the representativity of a concept c for its father c_f by comparing the sets of instances of the two concepts. The more instances of c_f are instances of c , the more c is prototypical of c_f . For example, someone who owns a dozen of cats considers a cat more prototypical with respect to the concepts *Animal* than someone who owns a goldfish. Calculating this component supposes that instances are fixed in the ontology. We use a logarithmic function in order to obtain a behavior of the extensional component close to human evaluation (prototypicalities of concepts with very few instances are not too close to 0).

2.3 Expressional component

The expressional component of the SPG measures the representativity of a concept c for its father c_f by comparing the way they are expressed. More c is expressed, more it is representative of c_f . A first estimation of the expressional prototypicality of c is based on the number of terms: more c has terms, more the place that c occupies in the cognitive universe of the ingroup is wide. For example, the concept *Horse*, that has a lot of synonyms (e.g. *cob*, *nag*, *gee-gee* or *jade*) is more prototypical of the concept *Animal* than the concept *Raccoon* that has only one term. If a corpus considered as representative of the domain is available, it can be used to refine this first estimation: the more the terms of c or its descendants are present

in the corpus, the more c is expressed in the cognitive universe of the ingroup, and the more c is prototypical of c_f . The occurrences of terms are weighted according to the structures of the documents. For example, an occurrence that appear in a title or in a keyword list counts more than an occurrence that appears in a paragraph. We also take into account the distribution of occurrences between documents, because a term that appears often but in few documents should have smaller weight than one which appears not very often but in most documents.

3 Applications of prototypicality

Prototypicality can be used for **ontology validation** and **ontology evolution**. If some values of SPG are very low, the corresponding *isa* links should be considered as inadequate.

Prototypicality seems to be useful for **ontology-based information retrieval**. In particular, request can be expanded straight down to representative concepts. For example, a request about *Bird* can be expanded to *Sparrow*, *Partridge* and *Duck*, but not to *Ostrich*. An horizontal expansion can also be considered. For example, a request about *Cat* can be expanded to *Dog*, which is as representative of *Domestic animal* as *Cat*, but not expanded to *Crocodile*. Moreover, prototypicality can be used to sort results of a expanded request from high to low prototypicality.

Prototypicality can also be used for **ontology personalization**. Even if an ontology is a shared conceptualisation within an ingroup, the instances of concepts are not necessarily the same for all the ingroup members. For example, in the case of an ontology about trading activity, two marketing men can agree with the concepts in the ontology, but not with the instances: they do not meet the same customers, they do not sell the same goods and so on. In a similar idea, documents representative of the knowledge of the ingroup members on the domain may be different at the individual scale. Then, for a given Domain Ontology (DO), gradients of conceptual, lexical and/or instance prototypicality are calculated from instances, corpus and property weights peculiar to a user (or to a group of users). DO enriched with prototypicality is called Personalized Domain Ontology (PDO) and several PDO can be derived from one DO.

Our method is implemented in TOOPRAG (*a Tool for Pragmatics in ontology*), and is applied in REDENE, a project funded by Tennaxia, a company which a compagny which develops services and software for legal intelligence in the domain of environment (www.tennaxia.com). Our approach has been experimentally tested on two real-life situations main experiments: an experiment with a wide ontology developed for REDENE (10000 concepts) and 1100 french and european legal texts, and an experiment with the Eurovoc thesaurus. Prototypicalities calculated with our method have been validated by experts of the considered domains.

REFERENCES

- [1] C. M. Au Yeung and H. F. Leung, 'Ontology with likeliness and typicality of objects in concepts', in *Proceedings of the 25th International Conference on Conceptual Modeling (ER'2006)*, volume 4215, pp. 98–111. Springer - LNCS, (2006).
- [2] M. Mikulincer, P. Kedem, and D. Paz, 'The impact of trait anxiety and situational stress on the categorization of natural objects', *Anxiety Research*, 2, 85–101, (1990).
- [3] C.W. Morris, *Foundations of the Theory of Signs*, Chicago University Press, 1938.
- [4] E. Rosch, 'Cognitive reference points', *Cognitive Psychology*, (7), 532–547, (1975).