Motivations of Extensive Incorporation of Uncertainty in OLE Ontologies

Vít Nováček

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
E-mail: xnovacek@fi.muni.cz

Abstract. Recently, the significance of uncertain information representation has become obvious in the Semantic Web community. This paper presents an ongoing research of uncertainty handling in automatically created ontologies. Proposal of a specific framework is provided. The research is related to OLE (Ontology LEarning), a project aimed at bottom-up generation and merging of domain specific ontologies. Formal systems that underlie the uncertainty representation are briefly introduced. We will discuss a universal internal format of uncertain conceptual structures in OLE then. The proposed format serves as a basis for inference tasks performed among an ontology. These topics are outlined as motivations of our future work.

1 Introduction

The purpose of this paper is to introduce an initial proposal of a novel representation of uncertain knowledge in domain specific ontologies. The framework presented here has been currently under research in the scope of a broader OLE project that comprises complex ontological support for automatic building of scientific portals[1].

The main objective of the OLE project is to implement a system that is able to automatically create and update a domain specific ontology for any given domain of human scientific knowledge. We emphasise an empiric approach to the ontology construction by means of bottom-up acquisition of concepts from the domain-relevant resources (documents, web pages, corpus data etc.). The acquisition process is incrementally boosted by the knowledge already stored in the ontology.

The concepts extracted from a single resource form so called miniontology that is instantly integrated into the current domain ontology[2]. The integration phase is the moment when the need of uncertainty representation arises. Even if we could obtain precise conceptual constructions from single resources (e. g.

[1] All projects mentioned here are part of the 'Information Society' program under Czech Academy of Sciences, the research grant number AV T100300419.

[2] Due to explicit orientation of this paper, the techniques of concept extraction and ontology creation methods are described in more detail in another work.

birds fly), we will experience infeasible consistency difficulties when trying to assign precise relations between the concepts in broader scope of the whole domain (as illustrated by the popular example: the fact *birds fly* collides with the statements *penguins are birds; penguins do not fly*). Besides the inconsistency handling, there are also important cognitive motivations of the utilisation of uncertainty in our empiric ontologies that led us to the proposal of a novel framework for representing uncertain knowledge. It is called $\mathcal{ANUIC}$ ($\mathcal{A}$daptive $\mathcal{N}$et of $\mathcal{U}$niversally $\mathcal{I}$nterrelated $\mathcal{C}$oncepts).

The rest of the paper is organised as follows. Section 2 summarises our overall motivations. This section also overviews important ideas from the cognitive science field that are both inspiring and relevant with respect to the topic. Formal background of uncertain information representation is briefly recalled in section 3. Sections 4 and 5 offer the framework proposal itself and basic notes on its utilisations. In section 6 there is given an illustrative example of uncertain ontology fragment generation. We conclude the paper and outline future directions of our research in section 7.

2 Motivation and Cognitive Observations

The knowledge repositories built by OLE tools must reflect the state of the respective domain empirically according to information contained in the provided resources. Such kind of knowledge is as much objective as possible, because it is not influenced by arbitrary considerations about the domain's conceptual structure, but determined by *the structure itself.*

2.1 Remedy to Emerging Inconsistencies

However, the automated empiric approach has an obvious drawback – the threat of inconsistency. As we do not generally have an infallible "oracle" to tell us how to precisely join or map newly extracted concepts to the ones that are already stored in our ontology, crisp relations between concepts are virtually impossible. We must deal with the inconsistencies somehow.

There are two general kinds of possible inconsistencies in an ontology (virtually any relational inconsistency can be modelled using these):

- *subsumption* inconsistency: given concepts C, D and E, the $C \subseteq D$ and $C \subseteq E$ statements may collide when we represent for example crisp *part-of* relation by the $\subseteq$ symbol (e. g.: Turkey is both part of Europe and Asia)
- *equivalence* inconsistency: given concepts C, D and E, the $C \equiv D$, $C \subset E$ and $D \equiv E$ statements are in conflict (for example when we find out in a text that '*science*', '*knowledge*' and ' *erudition*' are synonyms and at the same time we induce that '*knowledge*' is a super-concept of '*erudition*')

Such collisions are hard to be modelled in classic crisp ontology representation frameworks (see [13] or [18]). Implementation of the uncertainty into our knowledge representation is a solution for dealing with conflicts in the continuously updated ontology.

2.2 Mental Models Reflection

The second motivation lies in inspiration by the conceptual models that are characteristic for human mind. This topic is closely related to the very definition of *concept* and *meaning*. As stated for example in [12] or [4], people definitely do not represent the meaning of concepts as static crisp structures. The meanings are rather constructed as vague sets of dynamically overlapping referential associations ([12]), or so called "meaning potentials" with particular instantiation dependent on the context of concept-referring word or sequence of words ([1]).

In the rest of this paragraph, we will give a non-formal definition of a concept and its meaning in the perspective of OLE. More precise formulations related to the topic are presented in section 4. By concept we mean a representation of an entity existing in real world and/or utterable in human language. A concept is determined by its relations to another concepts in the universe then. Such "relational" definition of a concept is partly inspired by poststructuralistic philosophy (see for example [5]). Reference of a concept is then realised by instances of its relational connections. By these instances we mean especially concrete uncertainty measures assigned to each relation a concept is involved into (see section 4 for details).

Thus we can naturally represent the dynamic conceptual overlap in the meaning of [12], because the assigned relations' measures are continuously updated within new knowledge incorporation process. And by introducing a special relation of *association* we can represent the notion of meaning potentials according to [1]. Using this relation we can associate a concept with a representation of co-occurring concepts and impose another useful restriction on the meaning construction (helpful for example when resolving word-sense ambiguities).

3 Uncertainty Formalisations

The uncertain information representation frameworks are determined by three significant fields of contemporary mathematics:

1. extending the theory of measure into a more general theory of monotonous measures with respect to the classical measures of information
2. applications of (conditional) probability theory
3. extending the classical set theory into a more general fuzzy set theory

Various uncertainty extensions of the information measure theory are mentioned by Klir in [14]. However, in the computer science field there are other probabilistic theories generally accepted, mainly in the scope of:

- Bayesian networks (good overview of the topic is given in [17], specific applications are described in [18] or [13])
- non-monotonic reasoning and respective probabilistic (or possibilistic) extensions of "classical" (mainly propositional, first order or description) logics (see for example [11] or [10])

All these more or less probabilistic approaches are no doubt significant for uncertainty representation. However, we dissociate from them in our work for one main reason. As we want our ontologies to be built automatically in an empiric manner, it would be very hard to find out appropriate (conditional) probability assignments (especially in cases when the input data are sparse in some subdomains of our interest) without any background knowledge (axioms and/or inference rules) at our hand. That is why we prefer using the fuzzy sets and fuzzy logic formalisms to motivate our uncertain knowledge representation proposal.

Fuzzy sets were introduced by Zadeh in 1965 ([19]). The theory has been quite developed and widely used in many application domains so far and is quite well known. Perhaps the most important notion we will use here is a *membership function* that uniquely defines each fuzzy set, assigning a certain degree of respective set's membership to each element in a universal set X. Another crucial term is fuzzy relation (R on $X \times X$) – it is defined as a mapping $R: X \times X \rightarrow [0,1]$. Notions of reflexivity, symmetry, transitivity etc. similar to those of classical relations can be adopted even for fuzzy relations. This is very useful for example for reasoning tasks (see [9]) based on set operations. However, this intriguing topic will be discussed elaborately in another dedicated paper.

Many variations branching from the original Zadeh's idea have been developed until now. Some of them (besides the original fuzzy sets) are quite significant with respect to our research topic. The **fuzzy rough sets** and **rough fuzzy sets** (as introduced by Dubois and Prade in [6]) are based on approximations of (fuzzy) sets using a fuzzy similarity relation or crisp equivalence classes. Using these theories we can structure our conceptual universe as such approximation space in various perspectives (according to the relation used). **Intuitionistic fuzzy sets** (see [3]) based on combination of membership and non-membership degree can be used when dealing with negative knowledge in our ontologies.

4 $\mathcal{ANUIC}$ Proposal

$\mathcal{ANUIC}$ ($\mathcal{A}$daptive $\mathcal{N}$et of $\mathcal{U}$niversally $\mathcal{I}$nterrelated $\mathcal{C}$oncepts) forms a backbone of the uncertainty representation in OLE. The formal definition of $\mathcal{ANUIC}$ and issues regarding possible problems, modifications of basic empiric approach as well as reasoning perspectives are mentioned in this section.

4.1 Formal Definition

The concepts are stored in a special fuzzy network structure. The network is an oriented graph $G = (V, E)$, where V is a set of stored concepts and E is a set of ordered tuples (u, v), where $u \in V, v \in V$. The edges are induced by imprecise concept relations. Multiple edges are allowed as there can exist multiple relations between concepts. A node is a tuple in the form of (c, S, R, A), where:

- c is a **core** word of the concept. It serves as a master reference index and is computed as the most frequently occurring word in the scope of the hyperohyponymy relation instance with the highest associated μ-measure (see

below what the μ-measure is). The hyperohyponymy relation was chosen because it is commonly considered as a basic relation when forming a knowledge basis.

- S is a **synonymic** set of tuples in the form of $(s, \mu(s))$, where $s \in V$ is a concept, that was found to be synonymous with this one (in the meaning of the synonymy relation of respective core words). $\mu(s) \in [0, 1]$ is a μ-measure assigned to this observation.
- R is a **relational** set of tuples in the form of $(r, c_r, \mu(r))$, where $r \in N$ is an identifier of a relation from a given set N (its members can be usual lexico-semantic relations, such as hyperohyponymy, holonymy, meronymy, or domain-specific relations like *used_for*, *appears_in*, *method_of* and so forth). The $c_r \in V$ is again a concept, which is related with the current one by r, and $\mu(r) \in [0, 1]$ is the μ-measure assigned to this observation.
- A is an **associative** set of tuples in the form of (r, W), where $r \in N$ is identifier of a relation. W is a limited set of tuples (w, f_w) of top-most superconcept identifiers of word classes (w) and respective absolute frequencies (f_w) of their appearances in the context that led to induction of relation r (with respect to the core word c). Only the most frequent contexts are recorded. This set supports the meaning potentials remark from section 2.

4.2 Conviction Function

For computation of the μ-measure (that is, the membership/appropriateness function value) $\mu(r)$ for a relation r that is corresponding to a (c_1, c_2) edge we devise the following heuristic "conviction" formula (derived from the standard sigmoid function):

$$\mu(r) = \frac{1}{1 + e^{-s(f_r - \alpha)}}$$

where $f_r = \frac{f(r(c1,c2))}{\sum_{c \in V} f(r(c1,c))}$ is the relative frequency of relation observations in input data, s is a parameter regulating the "steepness" of the function and α influences the placement of the inflexion point. The domain of the function is real interval $(0, 1\rangle$ (but only rational numbers obviously appear as an input). The range is real interval $(0, 1)$.

Proper adjustment of the parameters defines the reflection of the impact of frequency on the fuzzy appropriateness μ-measure of the spotted relation. Thus we can regulate for example the "conservativeness" of the system (in the meaning of influence of major or minor observations to the overall conviction). The function is continuous and thus can be implemented in a very straightforward way. However, it can easily imitate discontinuous jumps in the shape of the curve, which is also very useful. Examples showing shapings of the conviction function are displayed in Figure 1. In Table 1 there are given the parameter values corresponding to the respective shapes of the conviction function. As we can see from these examples, the proposed conviction function allow us to naturally simulate the relative influence the observation frequency has on the relevancy of the spotted relation instance. To be more specific, consider the following overview:

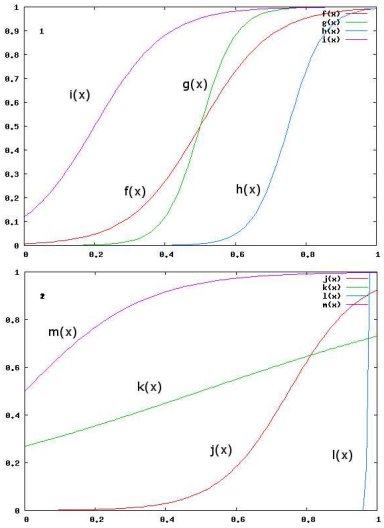

Fig. 1. Examples of various shapes of the conviction function

Plot number	Curve name	s	α
1	f(x)	10	0.5
1	g(x)	20	0.5
1	h(x)	20	0.75
1	i(x)	10	0.2
2	j(x)	10	0.75
2	k(x)	2	0.5
2	l(x)	5000	0.97
2	m(x)	6	0.0

Table 1. The conviction function parameter settings

– On the plot with label 1 there are given only slightly deformed curves that are quite similar to the standard sigmoid shape. The functions with α set to 0.5 and sufficiently high s reflect symmetric impact of the frequency on the μ-measure, raising from 0 to 1.
– The shapes presented on the plot 2 show us the flexibility of the proposed conviction function more illustratively:
  • Shape labelled as $m(x)$ presents quite "hesitating" function that assigns relatively high μ-measures (greater than 0.5) even to small frequencies, thus making the system partially believe in almost every evidence, yet preferring the higher frequencies significantly.
  • We can also acquire an almost "linear" curve shape (the $k(x)$ label), however it is more convenient to take directly the frequency as the μ-measure if we want a reasonable linear formula.
  • The $j(x)$ function presents a shape assigning relatively low values (in the meaning that they are quite far from 1) even for frequency near or equal

to 1. It reflects an "opinion" of the system that even a provisionally sure fact can never be absolutely valid if we consider future observations.
- The shape given by $l(x)$ presents a very "conservative" settings – only very high frequency will get a μ-measure significantly higher than 0, observations with minor frequencies are ignored. The α parameter presents a threshold of these ignored frequencies here.

5 Notes on the μ-measures Interpretation and Processing

In the following few paragraphs we present basic ideas related to utilisations of the notions described in the previous section.

5.1 Learning and Propagation of the Conviction Function Parameters

Given a reasonable and comprehensive portion of annotated concept data from an ontology domain, we can learn specific conviction function parameter settings for particular concepts (besides of selected parameters valid for the rest of an ontology). Thus we can reflect for instance whether a concept tends to have more instances of a relation at a time – the conviction function should assign almost same high values to almost equal (but relatively low) frequencies. The parameter settings can then spread over transitive strings in ontology within reasoning operations performed on $\mathcal{ANUIC}$ format. However, the annotation of data (and/or perhaps even implementation of unsupervised learning methods) as well as concrete implementation are still mostly subjects of future research.

5.2 Data View Perspectives

The μ-measures of relations in ontology allow us to impose various perspectives upon the stored data. The primary perspectives are *set-oriented* perspective (e. g. fuzzy set constructed by the μ-measures of subconcepts related to a concept or crisp approximations of the ontology structure given by some specific α-cuts) and *relation-oriented* perspective (e. g. the fuzzy synonymy relation). These perspectives allow us to develop reasoning procedures using the results of the existing theory of uncertain fuzzy inference (see for example [2], [9], [7] or [11]).

5.3 Coping with Sparse Input Data

The network constructed this way squares with the ideas presented in section 2 and conforms with the very intuitive notion of how people natively represent concepts in their minds. The dynamics of the system rests on continuous updating of all the μ-measures from the spotted data. However, the real world data are not homogeneous in the frequency distribution of particular concepts. For some rarely occurring but important words the empirically measure could easily be unsuitable for further utilisation of such knowledge. Therefore additional

"referees" must be incorporated especially for terms with low frequency (and even for the other ones). Existing lexical databases and electronic thesauri are good for correcting the possibly invalid uncertain measures gained by empiric evaluation of sparse data. In order to combine more resources of such external judgement, usage of WordNet ([8]) lexical database with Bonito2 word sketches ([16]) and Roget's electronic thesauri services ([15]) is appropriate.

5.4 Conscious and Unconscious Operations

In an analogy with human mind, two kind of operations within the $\mathcal{ANUIC}$ knowledge base are possible. We call them *conscious* and *unconscious* operations. The former are triggered by external incentives – mainly observations from the input data, user queries, or administrator commands (for example dumping the knowledge base in order to examine concept shifts over time later or learning the conviction function parameters).

The *unconscious* operations are run by the knowledge base itself and are of the same importance as the conscious ones. These operations are mainly reasoning tasks like merging of concepts that have a reasonable high measure of synonymy or inverse operation of splitting concepts. They should be run when there are no extensive computational demands on the knowledge base. Proper implementation of such operations helps to improve the consistency and manage the redundancies in the stored data.

6 An Example of Uncertain Ontology Fragment Creation

We offer a very simple example of uncertain ontology integration within $\mathcal{ANUIC}$ below. In the Figure 2 there is given a sample text as an input for miniontology extraction. The words referring to respective concepts in our fragment are marked with a index that represents them in the Table 2.

... In the fairy book there is a lot of information on tritons, **mermaids**$_{\mathbf{c_1}}$, sea snakes and other **mythical creatures**$_{\mathbf{c_2}}$...
... **Mermaids**$_{\mathbf{c_1}}$ are considered as **females**$_{\mathbf{c_3}}$...
... for **sylphs**$_{\mathbf{c_4}}$, especially **mermaids**$_{\mathbf{c_1}}$, are banned to interfere with humans...

Fig. 2. The text with concepts to be merged

The Table 2 shows relevant *is-a* taxonomic relations (noted as r) before and after the merging of the miniontology with a domain ontology[3]. Both halfs of the table present a matrix with indices from concept set and values from the

[3] The μ-measure function with parameters $s = 10$ and $\alpha = 0.2$ is used. The initial relative frequencies for observations $r(c_1, c_2)$, $r(c_1, c_4)$, $r(c_1, c_3)$ are $\frac{4}{7}$, 0, $\frac{3}{7}$ respectively.

$\mu(r)$ range (for states before and after the integration). Semantics of a matrix element e_{c_1,c_2} is: "the concept c_1 (the row index) *is a* concept c_2 (the column index) with conviction given by the appropriateness measure e_{c_1,c_2}".

$\mu(r)$	c_1	c_2	c_3	c_4	c_1	c_2	c_3	c_4
c_1	1.0	0.976	0.908	0.0	1.0	0.953	0.881	0.296
c_2	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0
c_3	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
c_4	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0

Table 2. The *is-a* relation μ-measure values before and after integration

7 Conclusion and Future Work

We presented an $\mathcal{ANUIC}$ framework for natural dealing with uncertain knowledge in ontologies. The framework is motivated by intuitive, yet valuable notion of representation of uncertainty in human mind. The theoretical background of fuzzy sets methodology allows to develop an appropriate calculus and consecutively build inference tools to reason among the concepts stored in $\mathcal{ANUIC}$.

The research results presented here are mostly in the phase of proposal, so a lot of work still has to be done. First objective for future work is to find efficient implementation methods as a proof of concept for the proposed structure. Additional psycholinguistic experiments should help with proper setting of the parameters for the μ-measure function presented in section 4 then. Invention and formal validation of a specific calculus for $\mathcal{ANUIC}$ is also needed. Then we can evaluate the framework using real world data from distinct domains of OLE project. Finally, the mutual correspondence and lossless transformation possibilities between ontologies represented in $\mathcal{ANUIC}$ and in current formats like OWL must be examined. All of the mentioned tasks are no doubt hard, but we demand it would be very challenging to pursue them and refine the ideas behind to gain a sustainable and efficient universal model of representation of uncertain knowledge.

References

1. J. Allwood. Meaning potentials and context: Some consequences for the analysis of variation and meaning. In *Cognitive Approaches to Lexical Semantics*, pages 29–66. Mouton de Gruyter, Berlin, 2003.
2. X. Li B. Wang, W. Liu, and Y. Shi. A new sparse rule-based fuzzy reasoning method. In *Proceedings of the Fourth International Conference on Hybrid Intelligent Systems (HIS'04)*, pages 462–467. IEEE Computer Society, 2004.
3. C. Cornelis and E. Kerre. Inclusion measures in intuitionistic fuzzy set theory. In *ECSQARU '03: Proceedings of the 7th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 345–356, London, UK, 2003. Springer-Verlag.

4. H. Cuyckens, R. Dirven, and J. R. Taylor, editors. *Cognitive Approaches to Lexical Semantics*, volume 23. Mouton de Gruyter, Berlin, cognitive linguistics research edition, 2003.
5. J. Derrida. *A Derrida Reader: between the Blinds*. Harvester Wheatsheaf, New York, 1991.
6. D. Dubois and H. Prade. Rough fuzzy sets and fuzzy rough sets. *International Journal of General Systems*, 17:191–209, 1990.
7. I. Düntsch. A logic for rough sets. *Theoretical Computer Science*, 179(1-2):427–436, 1997.
8. C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
9. L. Garmendia and A. Salvador. Computing a transitive opening of a reflexive and symmetric fuzzy relation. In *ECSQARU '05: Proceedings of the 7th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 587–599, London, UK, 2005. Springer-Verlag.
10. R. Giugno and T. Lukasiewicz. P-$\mathcal{SHOQ}$(D): A probabilistic extension of $\mathcal{SHOQ}$(D) for probabilistic ontologies in the semantic web. In *JELIA '02: Proceedings of the European Conference on Logics in Artificial Intelligence*, pages 86–97, London, UK, 2002. Springer-Verlag.
11. C. M. Teng H. E. Kyburg, J. Kyburg. *Uncertain Inference*. Cambridge University Press, Cambridge, 2001.
12. D. Hofstadter. *Fluid Concepts & Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. Basic Books, New York, 1995.
13. M. Holi and E. Hyvönen. A method for modeling uncertainty in semantic web taxonomies. In *WWW Alt. '04: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 296–297, New York, NY, USA, 2004. ACM Press.
14. G. J. Klir and M. J. Wierman. *Uncertainty-Based Information: Elements of Generalized Information Theory*. Physica-Verlag/Springer-Verlag, Heidelberg and New York, 1999.
15. J. L. Old. The semantic structure of roget's, a whole-language thesaurus, 2003.
16. P. Rychlý and P. Smrž. Manatee, bonito and word sketches for czech. In *Proceedings of the Second International Conference on Corpus Linguisitcs*, pages 124–132. Saint-Petersburg State University Press, 2004.
17. Y. Xiang. *Probabilistic Reasoning in Multi-agent Systems: A Graphical Models Approach*. Cambridge University Press, Cambridge, 2002.
18. R. Pan Y. Peng, Z. Ding. Bayesowl: A probabilistic framework for uncertainty in semantic web. In *Proceedings of Nineteenth International Joint Conference on Artificial Intelligence (IJCAI05)*, 2005.
19. L. A. Zadeh. Fuzzy sets. *Journal of Information and Control*, 8:338–353, 1965.