Empirical Merging of Ontologies — A Proposal of Universal Uncertainty Representation Framework

Vít Nováček[1] and Pavel Smrž[2]

[1]Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
E-mail: xnovacek@fi.muni.cz
[2]Faculty of Information Technology, Brno University of Technology
Božetěchova 2, 612 66 Brno, Czech Republic
E-mail: smrz@fit.vutbr.cz

Abstract. The significance of uncertainty representation has become obvious in the Semantic Web community recently. This paper presents our research on uncertainty handling in automatically created ontologies. A new framework for uncertain information processing is proposed. 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 discuss the universal internal format of uncertain conceptual structures in OLE then and offer a utilisation example then. The proposed format serves as a basis for empirical improvement of initial knowledge acquisition methods as well as for general explicit inference tasks.

1 Introduction

This paper introduces a novel representation of uncertain knowledge in the domain of automatic ontology acquisition. The framework presented here was designed and developed in the scope of a broader project — OLE — that comprises complex ontological support for Semantic Web applications and knowledge acquisition in general.

The main objective of the ontology acquisition platform OLE is to implement a system that is able to automatically create and update domain specific ontologies for a given domain of the scientific knowledge. We emphasise an empirical 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. The integration phase is the moment when the need of uncertainty representation arises. Even if we could obtain precise conceptual constructions from individual resources (e. g. *birds fly*), we will experience infeasible consistency difficulties when trying to establish 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 ANUIC (Adaptive Net of Universally Interrelated Concepts).

The rest of the paper is organised as follows. In Section 2 we give a concise description of the ontology acquisition process in the scope of OLE. Section 3 summarises the overall motivation for the designed uncertainty processing mechanisms. 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 4. Sections 5 and 6 define the framework itself and present basic notes on its utilisations. In Section 7, two illustrative examples of uncertain ontology fragment generation and query–processing are given. We conclude the paper and outline future directions of our research in Section 8.

2 Ontology Acquisition Process within OLE

An ontology acquisition framework is an integral part of the emerging ontology acquisition platform OLE [1, 2]. In the following subsections we give a brief overview of this tool.

As we basically process raw text data (articles, web pages' textual content, natural language corpora etc.), we can dissociate the ontology acquisition in two main phases — *text preprocessing and identification of relevant text's parts* and *creation of ontology from such parts*. These phases are described in subsections 2.1 and 2.2 here, whereas the last subsection 2.3 offers preliminary extraction results.

2.1 Text Preprocessing

OLE processes English plain-text documents and produces the respective ontology for each input resource (miniontology). To increase the efficiency, the input is preprocessed with the aim to pose at least some simple structure on the text and to reduce irrelevant data as well. Especially shallow syntactic structures (that are usually very helpful for some methods of semantic relations' acquisition) are identified in this step. Except of that, a domain dictionary is created and each of the term occurring in the dictionary is annotated by a vector that reflects its average context. This is crucial for other extraction methods, as seen below.

The preprocessing consists of *creating the domain dictionary and annotation of terms by context vectors*, *splitting of the text into sentences* (while possibly eliminating irrelevant sentences), *text tokenization*, *POS tagging and lemmatization*, and *chunking*. The steps related to processing of particular resources are based on regular expressions and performed in one pass through the input file. The promising relevance — for example the presence of a lexico-syntactic pattern — is detected and resolved (if possible) at this stage as well.

The tagging and chunking phases of preprocessing depend on task specific utilisation of NLTK natural language toolkit [3] with custom-trained Brill POS tagging algorithm [4] and fast regular expression chunking incorporated. Moreover, the usage of NLTK toolkit (which allows users to train their own POS taggers from annotated data and easily create efficient chunking rules) enables to adapt the whole OLE system even for other languages than English in future.

2.2 Taxonomy Extraction and Ontology Generation

Any extraction algorithm (such as semantic clustering, statistical co-occurrence methods or formal concept analysis) can be integrated into OLE in the form of a plug-in. Such a plug-in is responsible for the concept extraction and precise (or fuzzy) assignment of a class or a property. Then it translates the gained information into an output ontology, or passes it further to other OLE modules (like ontology-merger or reasoner).

The taxonomy (*is-a*) relation is crucial for ontology development. Therefore we have implemented methods for its acquisition first so that we could experiment with practical application of our proposal of novel uncertainty representation framework. In order to build taxonomic skeleton for our ontologies we have implemented a basic pattern-driven *is-a* relation extraction plug-in with relatively high precision but low recall. The pattern-based method gains classes (intensions) and individuals (extensions) that are directly lexicalised in the resources. To increase the overall precision of our system, we have also devised and implemented a novel method that utilises hierarchical clustering of domain terms and consequent autonomous class annotation. This method considers the terms in processed resources as extensions and tries to annotate their groups by appropriate intensional identifiers using the WordNet lexical database [5]. See [2] for detailed description of these methods and their implementation.

The extracted information is stored in the universal format proposed here in Section 5, no matter which extraction technique has been used. The output ontology file can be produced by applying respective translation rules. These rules are implemented as another independent plug-in (likewise the extraction algorithms) responsible for producing the output file in a desired format. Currently, the OWL DL format with our own basic fuzzy extensions is supported, but OLE is able to produce any other format by the same mechanism.

2.3 Preliminary Results of Taxonomy Extraction

Due to problems with evaluation of automatic ontology acquisition (as articulated for example in [6]) we have performed only orientational measures. For the pattern based method, we tested the system with patterns given in Table 1 below[1]. The patterns are presented in common regular expression–like syntax.

Id	The pattern
1	NP such as (NPList | NP)
2	such NP as (NPList | NP)
3†	(NPList | NP) (and | or)other NP
4	NP (including | especially) (NPList | NP)
5‡	(NPList | NP) (is | was)an? NP
6†	(NPList | NP) is the NP
7†	(NPList | NP) and similar NP
8†	NP like (NPList | NP)

Table 1. Patterns for *is-a* relation

Other patterns can be added easily, but the patterns presented in the table were found to be sufficient for basic evaluation.

For the approximate manual evaluation we randomly chose ten resources from the whole document set ($12,969$ automatically downloaded articles from computer science domain in this case). For each miniontology created by OLE system, we computed precision as the ratio of "reasonable" relations compared to all extracted relations. The recall was computed as the ratio of number of extracted terms (nouns) to all terms present in the resource. For all the measures of informal precision (*Pr.*) and recall (*Rec.*), an average value was computed. We present these results in Table 2, provided with respective average original resource size and number of all concepts extracted (in the *M1* row).

In the same table, there are also similar results of clustering–based technique (in the *M2* rows). Due to the strenuousness of manual evaluation of large ontologies we used only a set of 131 concepts (non–unique individuals) from a coherent computer science domain resource. 60 unique individuals and 47 classes were induced. We distinguished between *class–class* and *class–individual* relationships when analysing the precision. The method's approximate recall is 100%, because it processes all the terms within the input data.

Precision values for both methods are quite high when we look at the I column in the table. The I values present an improvement in precision over a base–line, which is computed as $\frac{R_R}{N(N-1)}$, where R_R stands for number of reasonable relations and N is the number of concepts in an ontology[2]. Moreover, it is only a "crisp" precision of the extraction phase.

When we incorporate empirical merging of the miniontologies by means of our uncertainty representation framework proposed in Section 5, we can significantly

[1] † — introduced by author, ‡ — modified by author, others adopted according to [7] and [8]; however, the devision of simple patterns is quite easy, therefore similar patterns can be found even in other works.

[2] The $N(N-1)$ is number of all *is-a* relations that can be assigned among all concepts.

Method	Res. sz. (wrd.)	No. of conc.	No. of rel.	Pr. (%)	Rec. (%)	I (%)
M1 avg.	4093	22.6	14.5	61.16	1.57	3399.17
M2 cl.–cl.	-	47	99	38.38	100	2183.62
M2 cl.–indiv.	-	60	62	51.61	100	5691.05
M2 sum–up	486	107	161	44.99 (avg.)	100 (avg.)	3937.34 (avg.)

Table 2. Selected results of OLE's taxonomy extraction tools

improve the values of precision (among other things) in a certain sense, as shown in Section 7 in more detail.

3 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 objective as possible, because it is not influenced by arbitrary considerations about the domain's conceptual structure, but determined by the structure itself.

3.1 Remedy to Emerging Inconsistencies

Nevertheless, the automated empiric approach has an obvious drawback – the threat of inconsistency and possible errors. 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[3]):

- *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 (supposing Europe and Asia are disjunct, the '*Turkey is both part of Europe and Asia*' statement is inconsistent);
- *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 [9] or [10]). Implementation of the uncertainty into our

[3] As a matter of fact, even the equivalence inconsistency can be modelled by the subsumption one, but we give both of them in order to show clear examples.

knowledge representation is a solution for dealing with conflicts in the continuously updated ontology.

3.2 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 [11] or [12], 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 [11], or so called “meaning potentials” with particular instantiation dependent on the context of concept-referring word or sequence of words [13]. These overlapping structures can also be viewed as interconnected in an associative network presented in [14]. We address all these issues in the framework proposal.

In the rest of this section, we will give an informal definition of a concept and its meaning in the perspective of OLE. More precise formulations related to the topic are presented in Section 5. 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 [15]). 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 5 for details).

Thus we can naturally represent the dynamic conceptual overlap in the meaning of [11], 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 [13]. Using this relation we can associate a concept with a representation of selected co-occurring concepts and impose another useful restriction on the meaning construction (helpful for example when resolving word-sense ambiguities).

4 Uncertainty Formalisations

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

1. extensions of 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. extensions of the classical set theory into a more general fuzzy set theory.

Various uncertainty extensions of the information measure theory are mentioned by Klir in [16]. 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 [10] or [9]);
- non-monotonic reasoning and respective probabilistic (or possibilistic) extensions of "classical" (mainly propositional, first order or description) logics (see for example [18] or [19]).

All these more or less probabilistic approaches are no doubt significant for uncertainty representation. However, we dissociate from them in our work for a few important reasons. As we want our ontologies to be built automatically in an empirical manner, it would be very hard to find out appropriate (conditional) probability assignments without any background knowledge (axioms and/or inference rules) at our hand except of the knowledge given by frequencies of particular evidences. Moreover, we would like to assign similar and quite high "belief" measures to certain instances of some relations. Imagine we would like to make our system quite strongly believe that *dog* is very likely a *canine* as well as a *pet*. The strong believe can be intuitively represented for instance as 0.8 value and higher within the $\langle 0,1\rangle$ scale. Suppose we induce this belief–measure from data on a probabilistic basis — then we can assign values equal to at most 0.5 to each of the relation instances if we want to have them as similar as possible and reasonably high at the same time. Moreover, the probabilities can limitary decrease to 0 for very large amounts of data with uniform distribution of instances of particular relations.

Coping with these facts would obviously break axioms of usual probability or information measure theory. But with a relatively little effort, we can quite naturally avoid these problems using the notion of fuzzy measure. 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 [20]. The theory has been quite developed and widely used in many application domains so far and is quite well known. 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 $R:\ X \times X \ \rightarrow\ \langle 0,1\rangle$. 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 explicit reasoning tasks (see [21]) based on set operations. However, this intriguing topic will be discussed more elaborately in another dedicated paper.

5 ANUIC Proposal

ANUIC (Adaptive Net of Universally Interrelated Concepts) forms a backbone of the uncertainty representation in OLE. The formal definition of ANUIC and a few comments on the topic are mentioned in this section.

5.1 Formal Definition

The concepts are stored in a special fuzzy network structure. The network is an oriented multigraph $G = (V, E)$, 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, R, A), where:

- c is a reference word or collocation (a term in general) of the concept. It serves as a master reference index for the node in the network;
- 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 (*is-a*), synonymy, 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 $\mu(r) \in \langle 0, 1\rangle$ is the fuzzy μ–measure assigned to this observation — see below what exactly this measure represents;
- A is an **associative** set of numeric centroid vectors that are representing the terms occurring near the reference term in average (either throughout the whole domain or specific subdomains); numeric elements of the vectors are gained through mapping of domain terms to integers using a domain dictionary. This supports the meaning potentials remark from Section 3, among other things like induction of vector space on the domain texts (useful for example for concept clustering).

5.2 Conviction Function

Fuzzy appropriateness (μ–measure) of a relation r (for example the *is-a* relation) between concepts (c_1, c_2) is given by a special conviction function (derived from standard sigmoid):

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

where $f_r = \frac{f(r(c1,c2))}{\sum_{c \in V} f(r(c1,c))}$ is the relative frequency of relation instance observations in input data, s is a parameter regulating the "steepness" of the function and β 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 $(0, 1)$.

This function maps relative frequencies of respective observations in input data to the fuzzy appropriateness measure of the relation. It can model various natural characteristics of human mind like conservativeness, open–mindness (in the meaning of influence of major or minor observations to the overall conviction) and so forth[4].

[4] Thus we can for example fix the meaning of a specific group of concepts and allow meaning variations for another one.

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 shapes of the conviction function are displayed in Figure 1[5]. As we can see on examples, the proposed conviction function allow us to naturally simulate the relative influence the observation frequency has on the relevancy of the observed relation instance. To be more specific, consider the following overview:

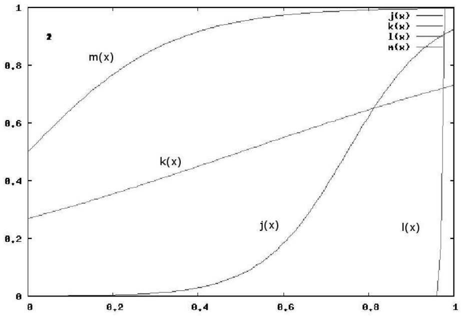

Fig. 1. Examples of various shapes of the conviction function

- 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.
- 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 β parameter presents a threshold of these ignored frequencies here.

6 Notes on the μ-measures Interpretation and Processing

In the following subsections we present basic ideas related to utilisations of the principles described in the previous section. We introduce notions of *implicit* and *explicit* reasoning with respect to the automatic empirical ontology acquisition and merging. The notions are also supported by preliminary examples given in Section 7.

[5] With the relative frequency and μ-measure on the horizontal and vertical axes respectively.

6.1 Implicit Reasoning

The implicit reasoning plays mayor role in learning of new knowledge by integration of various examples of empirical evidence for a relation between concepts in an ontology. We induce knowledge by a kind of implicit inference based on comparing the stored information and new sources of evidence in a well–defined manner.

The process of integration of newly coming facts is similar to the process of how people construct their conceptual representations — first they have an almost crisp evidence of a relation between objects (for example that *dogs have four legs*). This opinion is strengthened by further observations of four–legged dogs, but one day they see a cripple dog having only three legs. So they have to add another instance of the "have–number–of–legs" relation, but with much more decreased relevancy (unless they keep seeing other and other different three–legged dogs).

This is analogous to the ontology merging task — when we have a large amount of miniontologies gained from a vast number of domain resources, we can join them simply using their mutual insertion into one complex ANUIC structure. After proper configuration of the conviction function parameters we have qualitatively different representation of the domain — many formerly incorrect relations are mostly marginalised, whereas the empirically more valid relations obtain high μ-measures, signalising strong belief in their appropriateness.

We have found that very good heuristic for configuration of the conviction function parameters presented in Section 5 is dynamic setting of the β inflexion point value. The steepness parameter s can be set arbitrarily (however higher values are generally better for they cause better discrimination). The β for a concept c and relation R is set as:

$$\beta = \frac{1}{|\{\hat{c}|(c,\hat{c}) \in R\}|} \;.$$

Moreover, any relative frequency f higher than 0.5 is modified by weighing the β parameter with $1-(f-0.5)$ expression. Only then we obtain for example natural conviction of (almost) 1 when we deal with a single relation instance. Thus we can discriminate very well between the relation instances with significant and insignificant frequencies due to the shape of the conviction function[6]. Concrete example of such an ontology merge is given in Section 7.1.

6.2 Explicit Reasoning

Explicit reasoning conforms to classical definition of reasoning — it stems from explicit inference of new facts based on the facts already stored in an ontology and corresponding rules tailored to our uncertain knowledge representation. It

[6] Supposing that the higher the relation frequency is with respect to the average relative frequency for relation edges coming from the c concept, the more is the relation significant and vice versa.

can always be reduced on query–answering. The mechanisms underlying the query processing proposal are rather fuzzy set–based then logic–founded. Thus we can answer also queries difficult or even infeasible when using a classical logical formalism (see Section 7.2 for an example of the query–processing and possible utilisation sketches).

Despite of this, we can always reduce our knowledge repository to the OWL DL format [22]. We can gain a crisp Description Logics approximation by performing an α–reduction using respective α–cuts[7] on fuzzy constructs contained in the ontology and by elimination of possible relations that are restricted in OWL DL. Then we can use widely–adopted Description Logics[8] reasoning on such an approximation in order to learn less–expressive but crisp facts from our knowledge base.

7 Examples of the ANUIC Framework Utilisation

We give an example on practical utilisation of the representation properties of ANUIC for real world data in the first subsection. The second subsection offers an example of how a query could be processed by the ANUIC–based empirical inference engine. We also mention possible related utilisations of the framework within our another project.

7.1 Ontology Merging

We tested the ontology merging on a set of $3,272$ automatically downloaded articles from the computer science domain. The overall size of the resources was $20,405,014$ words. We produced the respective miniontologies by pattern–based OLE module and merged them into one ANUIC structure. Thus we gained a taxonomy with $5,538$ classes, $9,842$ individuals[9] and $61,725$ mutual *is-a* relations. A sample from this ontology is given on Figure 2 — the ovals represent classes, squares individuals and arrows go from sub-concept to its super-concept, labelled by respective μ–measures. For the μ–measures computation we used the dynamic β assignment heuristics described in Section 6.1 and s parameter set to 100, which performed best among various other settings.

It is very hard to formally decide what is the representation's exact improvement when compared to the knowledge stored in the former crisp miniontologies. But we can again give at least an informal statistics — when we consider only the relations with highest μ–measure(s) relevant for a particular concept[10], we can compute an approximate ratio of "reasonable" relations similar to the one

[7] An α–cut of a fuzzy set A is a classical crisp set of objects that have a membership value higher than $\alpha \in \langle 0, 1 \rangle$ with respect to A.

[8] Currently the $\mathcal{SROIQ}$ Description Logic is implemented in OWL DL, version 1.1. proposal, see [23].

[9] We empirically assume that a concept is an individual as long as it has no hyponyms.

[10] Which is by the way a very strong restriction, the range of possible interpretations of the concrete conviction values is much higher.

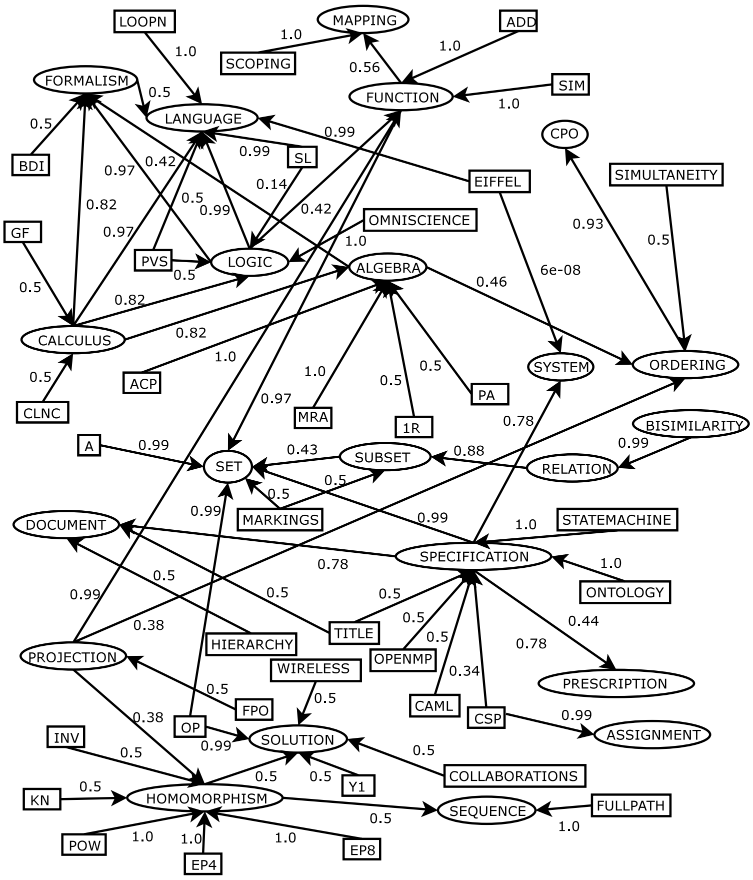

Fig. 2. Sample of the merged ontology

presented in Section 2.3. We computed the ratio on a random sample of 50 relations from the whole merged ontology and obtained the value 86 %. We cannot formally compare this ratio even to the informal measures given in Section 2.3, but we clearly see that this truly means a kind of improvement under a certain perspective.

7.2 Query Processing and Possible Utilisations

In the following we show how can a vague but very useful query be processed using ANUIC–based explicit reasoning. Suppose we have the query:

Are the *network* and the *graph* concepts similar?

Such a query can hardly be modelled in any classical logic. Nevertheless, it can be very useful — let us give one example for all. The answer to such a query is very significant when we consider different domains. In the computer science domain, for instance, the *network* and *graph* concepts are quite similar (network can be viewed as a kind of graph). On the other hand, in the sociology domain there is no observable similarity between these concepts, albeit the *network* term is widely used (*social network* etc.). Thus we can efficiently use such kind of questions for example in the task of discourse identification.

Now how do we process the above query? Suppose we have the following four kinds of relations stored in our ANUIC structure:

1. *synonymy* (s identifier) — usual lexico–semantic relation of meaning similarity; however, this relation does not have to be sufficient when processing vague queries among our empirical knowledge repository;
2. *hyperohyponymy* (h identifier) — super/sub–concept lexico–semantic relation;
3. *association* (a identifier) — arbitrary co–occurrence relation, its μ–measure shows how often the concepts appear in the vicinity of each other;
4. *antonymy* (t identifier) — lexico–semantic relation of meaning dissimilarity.

Let us encode *network* and *graph* concepts as n, g respectively. Let $\mu_r(n, g)$ be the value of μ–measure of relation r between n and g. Then we can express empirical similarity (ψ) as:

$$\psi(n, g) = \gamma_1(\mu_s(n, g) - \mu_t(n, g)) + \gamma_2(\mu_h(n, g) + \mu_h(g, n)) + \gamma_3(\mu_a(n, g) + \mu_a(g, n)),$$

where $\gamma_1, \ldots, \gamma_3$ are real coefficients such that $\gamma_1, \gamma_2, \gamma_3 > 0$ and $\gamma_1 > \gamma_2 > \gamma_3$.

After selecting the γ_i, $i \in \{1, \ldots, 3\}$ coefficients appropriately, we can define a non–decreasing scale of possible similarity values and map their consequent intervals to the respective scale of linguistic fuzzy labels (for example *distinct*, *almost dissimilar*, *little similar*, *moderately similar*, *very similar*, *almost same*, *same*). Thus we can straightforwardly answer the query and/or pass the numeric value for further processing. Supposing we have inserted sufficient amount of data in our knowledge base, answers like this are useful even for rarely occurring concepts and relations. Moreover, the time complexity of the query processing itself is constant — we only need to get the 6 μ–measure values and add them up[11].

The inference engine based on ANUIC format can be directly used in the scope of general knowledge acquisition as well as within more specific Semantic Web tasks. It can be very useful for example for another project we are involved in — PortaGe — that is aimed on automatic generation and personalisation of scientific Semantic Web portals [1]. We can employ the uncertainty representation for example in the automatic extraction of metainformation from the scientific documents, citation analysis, metasearch in digital libraries, analysis of various web pages, meta-data annotation of web resources and source-change analysis. The ontology support would be useful even for general semantics–enhanced search and retrieval tasks among the particular portal's domain.

[11] The lookup for values is performed on efficient hash–like structures.

8 Conclusions and Future Work

We presented the ANUIC framework that deals 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 novel inference tools to reason among the concepts stored in expressive ANUIC format very efficiently.

Our future work will focus on incorporation of results of another extraction methods (mainly our clustering–based technique) into the ANUIC ontologies in order to increase the recall. A formal development and validation of a specific calculus for ANUIC explicit reasoning is needed then. We will also devise formal evaluation methods and test the framework properly using various data from other distinct domains of available resources. Finally, the mutual correspondence and transformation possibilities between ontologies in ANUIC format and formats like OWL extended by possible fuzzy modifications must be examined. All of the mentioned tasks are no doubt hard, but we demand it will be very challenging to pursue them and refine the ideas behind to gain a sustainable, expressive and efficient universal model of representation of uncertain knowledge.

Acknowledgements

This work has been supported by the Academy of Sciences of the Czech Republic, 'Information Society' national research program, the grant AV 1ET100300419, and partially by the Czech Republic Ministry of Education, Research Plan MSM 6383917201.

References

1. Nováček, V., Smrž, P.: Ontology acquisition for automatic building of scientific portals. In: LNCS. Volume 3831., Springer-Verlag Berlin Heidelberg (2006)
2. Nováček, V.: Ontology learning. Master's thesis, Faculty of Informatics, Masaryk University, Czech Republic (2006) Available at (February 2006): `http://nlp.fi.muni.cz/~xnovacek/files/dp.pdf`.
3. NLTK: NLTK: Natural Language Toolkit – Technical Reports. (2005) Available at (February 2006): `http://nltk.sourceforge.net/tech/index.html`.
4. Brill, E.: A report of recent progress in transformation-based error-driven learning. In: Proc. ARPA Human Language Technology Workshop '94, Princeton, NJ (1994)
5. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press (1998)
6. Brewster, C., Alani, H., Dasmahapatra, S., Wilks, Y.: Data driven ontology evaluation. In: Proceedings of LREC 2004. (2004)
7. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th conference on Computational linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1992) 539–545

8. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Web-scale information extraction in KnowItAll: (preliminary results). In: Proceedings of WWW '04, New York, NY, USA, ACM Press (2004) 100–110
9. Holi, M., Hyvönen, E.: A method for modeling uncertainty in semantic web taxonomies. In: Proceedings of WWW Alt. '04, New York, NY, USA, ACM Press (2004) 296–297
10. Y. Peng, Z. Ding, R.P.: BayesOWL: A probabilistic framework for uncertainty in semantic web. In: Proceedings of Nineteenth International Joint Conference on Artificial Intelligence (IJCAI05). (2005)
11. Hofstadter, D.: Fluid Concepts & Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought. Basic Books, New York (1995)
12. Cuyckens, H., Dirven, R., Taylor, J.R., eds.: Cognitive Approaches to Lexical Semantics. Cognitive linguistics research edn. Volume 23. Mouton de Gruyter, Berlin (2003)
13. Allwood, J.: Meaning potentials and context: Some consequences for the analysis of variation and meaning. In: Cognitive Approaches to Lexical Semantics. Mouton de Gruyter, Berlin (2003) 29–66
14. Ruge, G.: Combining corpus linguistics and human memory models for automatic term association. In Strzalkowski, T., ed.: Natural Language Information Retrieval. Kluwer Academic Publishers (1999) 75–98
15. Derrida, J.: A Derrida Reader: between the Blinds. Harvester Wheatsheaf, New York (1991)
16. Klir, G.J., Wierman, M.J.: Uncertainty-Based Information: Elements of Generalized Information Theory. Physica-Verlag/Springer-Verlag, Heidelberg and New York (1999)
17. Xiang, Y.: Probabilistic Reasoning in Multi-agent Systems: A Graphical Models Approach. Cambridge University Press, Cambridge (2002)
18. Kyburg, H.E., Kyburg, J., Teng, C.M.: Uncertain Inference. Cambridge University Press, Cambridge (2001)
19. Giugno, R., Lukasiewicz, T.: P-$\mathcal{SHOQ}$(D): A probabilistic extension of $\mathcal{SHOQ}$(D) for probabilistic ontologies in the semantic web. In: Proceedings of JELIA '02, London, UK, Springer-Verlag (2002) 86–97
20. Zadeh, L.A.: Fuzzy sets. Journal of Information and Control **8** (1965) 338–353
21. Garmendia, L., Salvador, A.: Computing a transitive opening of a reflexive and symmetric fuzzy relation. In: Proceedings of ECSQARU '05, London, UK, Springer-Verlag (2005) 587–599
22. Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., Stein, L.A.: OWL Web Ontology Language Reference. (2004) Available at (February 2006): `http://www.w3.org/TR/owl-ref/`.
23. Horrocks, I., Kutz, O., Sattler, U.: The even more irresistible $\mathcal{SROIQ}$. Technical report, University of Manchester (2005)