The 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference

ISWC+ASWC 2007

Workshop 12

Uncertainty Reasoning for the Semantic Web

Workshop Organizers:

Fernando Bobillo, Paulo Costa, Claudia d'Amato, Nicola Fanizzi, Francis Fung, Thomas Lukasiewicz, Trevor Martin, Matthias Nickles, Yun Peng, Michael Pool, Pavel Smrz, Peter Vojtas

12 Nov. 2007
BEXCO, Busan KOREA

ISWC 2007 Sponsor

Emerald Sponsor

Gold Sponsor

Silver Sponsor

We would like to express our special thanks to all sponsors

ISWC 2007 Organizing Committee

General Chairs

Riichiro Mizoguchi (Osaka University, Japan)

Guus Schreiber (Free University Amsterdam, Netherlands)

Local Chair

Sung-Kook Han (Wonkwang University, Korea)

Program Chairs

Karl Aberer (EPFL, Switzerland)

Key-Sun Choi (Korea Advanced Institute of Science and Technology)

Natasha Noy (Stanford University, USA)

Workshop Chairs

Harith Alani (University of Southampton, United Kingdom)

Geert-Jan Houben (Vrije Universiteit Brussel, Belgium)

Tutorial Chairs

John Domingue (Knowledge Media Institute, The Open University)

David Martin (SRI, USA)

Semantic Web in Use Chairs

Dean Allemang (TopQuadrant, USA)

Kyung-Il Lee (Saltlux Inc., Korea)

Lyndon Nixon (Free University Berlin, Germany)

Semantic Web Challenge Chairs

Jennifer Golbeck (University of Maryland, USA)

Peter Mika (Yahoo! Research Barcelona, Spain)

Poster & Demos Chairs

Young-Tack, Park (Sonngsil University, Korea)

Mike Dean (BBN, USA)

Doctoral Consortium Chair

Diana Maynard (University of Sheffield, United Kingdom)

Sponsor Chairs

Young-Sik Jeong (Wonkwang University, Korea)

York Sure (University of Karlsruhe, German)

Exhibition Chairs

Myung-Hwan Koo (Korea Telecom, Korea)

Noboru Shimizu (Keio Research Institute, Japan)

Publicity Chair: Masahiro Hori (Kansai University, Japan)

Proceedings Chair: Philippe Cudré-Mauroux (EPFL, Switzerland

Metadata Chairs

Tom Heath (KMi, OpenUniversity, UK)

Knud Möller (DERI, National University of Ireland, Galway)

The 3rd Workshop on Uncertainty Reasoning for the Semantic Web

The Uncertainty Reasoning Workshop is an exciting opportunity for collaboration and cross-fertilization between the uncertainty reasoning community and the Semantic Web community.

Effective methods for reasoning under uncertainty are vital for realizing many aspects of the Semantic Web vision, but the ability of current-generation web technology to handle uncertainty is extremely limited. Recently, there has been a groundswell of demand for uncertainty reasoning technology among Semantic Web researchers and developers.

This surge of interest creates a unique opening to bring together two communities with a clear commonality of interest but little history of interaction. By capitalizing on this opportunity, URSW could spark dramatic progress toward realizing the Semantic Web vision

The intended audience for this workshop includes the following:

- Researchers in uncertainty reasoning technologies with interest in the Semantic Web.
- Semantic web developers and researchers.
- People in the knowledge representation community with interest in the Semantic Web.
- Ontology researchers and ontological engineers.
- Web services researchers and developers with interest in the Semantic Web.
- Developers of tools designed to support Semantic Web implementation, e.g., Jena developers, Protege and Protege-OWL developers.

We intend to have an open discussion on any topic relevant to the general subject of uncertainty in the Semantic Web (including fuzzy theory, probability theory, and other approaches). Therefore, the following list should be just an initial guide:

- Syntax and semantics for extending Semantic Web languages to include principled treatment of uncertain, incomplete information.
- Logical formalisms to support uncertainty in Semantic Web languages
- New forms to use uncertainty reasoning as a means of assessing whether similar terms in different ontologies refer to the same or similar concepts
- Architectures for applying plausible reasoning to the problem of ontology mapping
- Using fuzzy approaches to deal with imprecise concepts within ontologies
- The concept of a probabilistic ontology and its relevance to the Semantic Web
- Best practices for representing uncertain, incomplete, ambiguous, or controversial information in the Semantic Web
- The role of uncertainty as it relates to Web services
- Uncertainty-friendly interface protocols as a means to improve interoperability among Web services
- Uncertainty reasoning techniques applied to trust issues in the Semantic Web
- Existing implementations of uncertainty reasoning tools in the context of the Semantic Web
- Issues and techniques for integrating plausible inference tools
- The future of uncertainty reasoning for the Semantic Web

URSW 2007 Program Committee

Ameen Abu-Hanna - Universiteit van Amsterdam, the Netherlands.

Fernando Bobillo - University of Granada, Spain.

Paulo C. G. Costa - George Mason University, USA.

Fabio G. Cozman - Universidade de Sao Paulo, Brazil.

Claudia d'Amato - University of Bari, Italy.

Ernesto Damiani - University of Milan, Italy.

Nicola Fanizzi - University of Bari, Italy.

Francis Fung - Information Extraction & Transport, Inc., USA.

Linda van der Gaag - Universiteit Utrecht, the Netherlands.

Ivan Herman - C.W.I., the Netherlands. W3C Activity Lead for the Semantic Web.

Kathryn B. Laskey - George Mason University, USA.

Kenneth J. Laskey - MITRE Corporation, USA. Member of the W3C Advisory Board.

Thomas Lukasiewicz - Università di Roma "La Sapienza", Italy.

Anders L. Madsen - Hugin Expert A/S, Denmark.

M. Scott Marshall - Adaptive Information Disclosure, Universiteit van Amsterdam, The Netherlands.

Trevor Martin - University of Bristol, UK.

Bill McDaniel - DERI, Ireland.

Matthias Nickles - Technical University of Munich, Germany.

Leo Obrst - MITRE Corporation, USA.

Yung Peng - University of Maryland, Baltimore County, USA.

Michael Pool - Convera, Inc., USA.

Livia Predoiu - Universität Mannheim, Germany.

Dave Robertson - University of Edinburgh, UK.

Daniel Sánchez - University of Granada, Spain.

Elie Sanchez - Université de La Méditerranée Aix-Marseille II , France.

Oreste Signore - Istituto di Scienza e Tecnologie dell' Informazione "A. Faedo", Italy. Manager of the W3C Office in Italy

Nematollaah Shiri - Concordia University, Canada.

Sergej Sizov - University of Koblenz-Landau, Germany.

Pavel Smrz - Brno University of Technology, Czech Republic.

Umberto Straccia - Istituto di Scienza e Tecnologie dell' Informazione "A. Faedo", Italy.

Heiner Stuckenschmidt - Universität Mannheim, Germany.

Masami Takikawa - Cleverset, Inc., USA.

Peter Vojtas - Charles University, Czech Republic.

Technical Papers

Position Papers

Technical Papers

Probabilistic geospatial ontologies

Sumit Sen[1,2]

[1] Deptt. of Computer Science and Engg., Indian Institute of Technology Bombay, Mumbai –76 India
[2] University of Muenster, Robert Koch Str. 26, 48149 Muenster, Germany
sumitsen@uni-muenster.de

Abstract. Partial knowledge about geospatial categories is critical for knowledge modelling in the geospatial domain but is beyond the scope of conventional ontologies. Degree of overlaps between geospatial categories, especially those based on geospatial actions concepts and geospatial enitity concepts need to be specified in ontologies. We present an approach to encode probabilistic information in geospatial ontologies based on the BayesOWL approach. This paper presents a case study of using road network ontologies. Inferences within the probabilistic ontologies are discussed along with inferences across ontologies using common concepts of geospatial actions within each ontology. The results of machine-based mappings produced are verified with human generated mappings of concepts.

Keywords: geospatial ontologies, probabilistic, concept mappings, human subjects testing.

1 Introduction

Ontologies, which allow the use of probabilistic representation of categories, are under increasing focus [1]. Reasoning mechanisms using such probabilistic information, which not only allow inferring equivalent concepts but also the 'most similar' or the 'least similar' concepts are best suited for practical use of ontologies. Support for such mechanisms can also be found in cognitive sciences, which assume conceptual spaces to denote a concept [2] and distances between such spaces to explain the notion of similarity between two concepts [3]. Cognitive basis for the specification of geospatial ontologies have been favoured by many researchers [4]. However, current work in geospatial ontologies does not provide sufficient insight into the use of probabilistic knowledge in ontologies. Although mechanisms to specify such information have already been attempted, for the semantic web [5], such probabilistic ontologies have not been explored inside the geospatial domain.

This paper aims to explore this gap and illustrates the use of probabilistic ontologies in the geospatial domain. We employ the approach of BayesOWL [5] to specify probabilistic geospatial ontologies primarily related to road network entities. While we draw extensively on the ideas of BayesOWL, our work mainly concentrates on (1) extracting and using probablistic information in geospatial ontologies, (2) Inferences across geospatial ontologies based on the assumption of geospatial action concept names, and (3) its applicability to enabling semantic reference. The use of probabilistic geospatial ontologies for practical tasks of semantic translations is the main contribution of this paper.

2 Background

Existing literature in geographical information science points out the significance of geospatial ontologies as tools to represent conceptualizations in the geospatial domain. Such knowledge representation tools are mostly used to resolve semantic differences and promote interoperability between applications across information communities [6].

Agarwal [7] has discussed that a unified approach to ontology specification in the geospatial domain does not exist. Different approaches including the approaches of formal ontologies [8] and algebraic approaches [9], Rüther *et al.* [10] have evolved in parallel to the conventional approaches of Description Logic (DL) based specifications. Geospatial ontology engineering has been also proposed to enable a supportive environment for knowledge representation in the geospatial domain [11]. However the challenges for geospatial ontologies as tools of knowledge representation remain unresolved to a large extent. The primary questions that need to be answered include the following:

Gomez-Perez and Benjamins [12] have stated that the number of ontologies specified is not large enough for their use in practical and industrial scale applications. This is true for the geospatial domain and practically verified ontologies are still to be produced. In their absence it is impossible to verify their utility and hence their contributions to semantic interoperability.

With a similar point of view, it has been discussed that the tools and principles of ontologies are still viewed with skepticism even after years of research. Agarwal [7] has pointed out that geographic concepts and categories have inherent indeterminacy and vagueness; especially that emerge from human reasoning and conceptualization. It is therefore unlikely that the semantic ambiguities can be resolved without accounting for the uncertainty factor.

Geospatial ontologies have either looked at geographic space either from the point of view of the geospatial entities with it or from that of geospatial actions. A unified view, which incorporates knowledge of geospatial actions in ontologies of geospatial entities and which treats both these components of knowledge as equally important, is necessary. Kuhn [13] advocates the inclusion of actions and affordances in geospatial ontologies.

Geospatial ontologies are in need of innovative approaches to ensure their practical use. In order that geospatial conceptualizations can be encoded in ontologies, emerging techniques in ontological specifications and knowledge representation need to be adapted and experimented in the geospatial domain. These include probabilistic ontologies and inclusion of knowledge about geospatial actions and their hierarchies [14].

2.1 Need for probabilistic frameworks

We have already mentioned that uncertainties are abundant in categories of geospatial entities. Zhang and Goodchild [15] state "…and in the face of fuzziness, Boolean logic is surely less versatile in dealing with discourse that is full of heuristic metaphors, linguistic hedges and other forms of subjectivity". One of the arguments against knowledge engineering based on conventional ontologies has been against the use of rigid categories as opposed to partial, incomplete, or probablistic categories of the real world. It is also important to note that differences between such real world

categories are measurable in terms of a similarity (or a dissimilarity) score. As opposed to crisp, binary classification of instances into a certain geospatial category, it is usual to express the relative suitability of an instance to a category (such as *Road*) in comparison to others (say, *Motorway*). Note that the definition of the category itself is precise but there is only a probability, given the current knowledge about inclusions and overlaps between categories that a certain instance fits into a certain category. Although there is a tendency to associate probablistic cateogories with natural geospatial entities we need to note, that since our categories are precise, using examples of man-made entities from the transportation domain is appropriate as well.

To comprehend the notion of uncertainity or partial information, which we attempt to address it, is important to understand that there are overlaps between categories modelled within an ontology. For example, while modelling concepts of a road network ontology (shown in Fig 1), besides knowing that a class *FootPath*

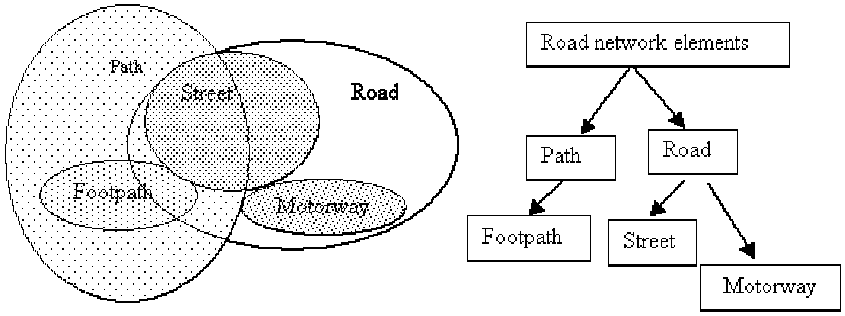


Figure 1 (a) Representation of five classes of a road network ontology. While Highway and Street are subclasses of Road, Footpath is a subclass of Path. Evidently this representation shows that Highway and Footpath are small subclasses of Road and Path respectively. Street has a major overlap with Path allthough it is not a subclass. (b) Representation of the five classes as a subsumption relation in a conventional ontology (note that in this diagram, arrows point to the subclass).

is a subclass of class *Path*, one may also know and wish to express that "*Footpath* is a small subclass of the class *Path*"; or in another case where a class *Street* and *Path* are not logically related, one may still want to say that "*Street* and *Path* are largely overlapped with each other". Users of ontologies would therefore like to know how close is a *Street* from a *Road* or a degree of similarity between *Road* and *Street*. Such tasks are beyond the scope of conventional ontologies [5], as partial knowledge is ignored as shown in the subsumption hierarchy of figure 1(a). Therefore, a mechanism to specify probabilistic ontologies and carry out reasoning tasks on them is also critical for practical use of geospatial ontologies.

Probabilistic specifications have a strong relation in the context of using affordances and functions of geospatial entities in ontologies. The concept of categorization of manmade geospatial entities such as roads and road network components is closely associated with the functions or actions that they afford. Often, the association of such functions with certain entities is not deterministic and context sensitive. However, based on personal experience, humans are able to provide a relative value of the association between an entity and a function. Thus a *Motorway* is more strongly associated to the function of *driving* as compared to a *Street* or a *Path*. At the same time, we can argue that *Driving* is not associated to *Footpaths*. In a probabilistic ontology framework, the associations between entities and functions can be specified as probabilistic linkages. The overlaps of categories such as *Road* or

Footpath and things that afford *driving* as shown in Figure 2 below are such links and we attempt to use such overlaps in probabilistic ontologies.

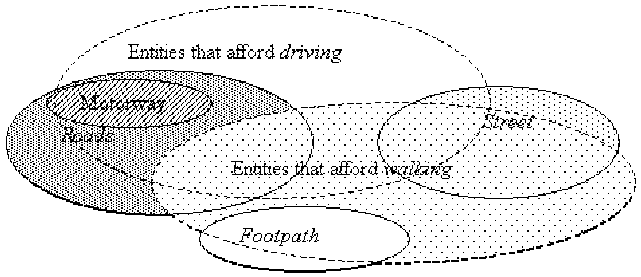


Figure 2 Sample representation of overlaps between some entity concepts and action based concepts for road networks in the UK. While ellipses with solid borders represent geospatial entities, the ellipses with dashed borders represent abstract concepts based on the entities that afford certain geospatial action.

It is important to note that translation of meanings of symbols used to represent certain entities between two agents is directly related to the affordances of the entities with respect to different geospatial actions. Affordances and functions are always in relation to a certain agent and its goals [16]. This requires that the mapping of functions and entities be updated on the basis of the context in hand. Our framework seeks to provide a mechanism for flexible translations based on reviseable probabilistic values of enitity-action linkages in a given context. Such mechanisms to specify contexts are critical for enabling pragmatics as discussed by Brodaric (2007).

2.1 Ontologies as Bayesian Networks

PR-OWL [1] and BayesOWL [5] are two approaches that use a BN based representation of ontologies. Of these, BayesOWL provides an approach for specification and reasoning.

Ding *et al* [5] developed a mechanism of expressing OWL ontologies as Bayesian networks termed as BayesOWL. The important steps to construct such ontologies are as below:

Construction of the Directed Acyclic Graph (DAG): The entity classes to be used are listed first and the topmost (most universal) concept is added to the top of the DAG as a node. Child concepts of this concept are added below the parent concept as individual nodes and the complete DAG is created by constructing the links. Each node has only 2 states (True, False)

Regular Nodes and L Nodes: The nodes created above are called Regular nodes. There are another category of nodes called L Nodes, which help in constructing Union, Intersection, Disjoint and Equivalent relationships. Since we do not use any of these relationships in our ontologies we shall ignore construction of L Nodes.

Allocating conditional probabilities: Regular nodes (other than the top node) have one conditional probability value each for its parent node. It is suggested that such conditional probability values are learnt from text classification techniques. We use the relatedness values from WordNet similarity modules to derive these values.

Applying IPFP iterations to impose P Space: Finally with given CPT values it is important for the network to learn the real values given the probability constraints to arrive at a condition where all LNodes are true. This is achieved by an Iterative

Proportional Fitting Procedure (IPFP) [17]. In case there are no L Nodes to be considered, this iterative step can be overlooked.

The principal reasoning tasks in our Bayesian network are based on computation of joint probability distributions and utilize the three methods suggested by Ding *et al*[5]. These are:

Concept Satisfiablity: if a concept based on certain states of given nodes in the network can exist. This is defined by verifying if P (e|t) = 0, where e is the given concept. For example already as discussed in § 1.2, given that a concept belongs to *Motorway* (thus P(t)=1) it cannot be a member of "Entities that afford *walking*" P (e|t) = 0 . Hence a concept of a *Motorway, which affords walking,* is not satisfied as per the representation in Figure 2.

Concept Overlap: the degree of overlap between a given concept and any other concept in the network is determined by P (e|C,t). Thus in Figure 2 we see that the overlap between *Road* and "Entities that afford *walking*" is significant whereas overlap between *Motorway* and the later is null.

Concept similarity: The advocated measure of similarity is based on Jaccard coefficient provided by Rijsbergen [18]. This measure is the ratio of the probability that an instance of the top level concept is a member of either of the two classes, with respect to the probability that the instance is a member of both the classes. The value ranges from 0 to 1. To demonstrate this if we assume that the overlap between classes as shown in figure 2, we know that the probability that an instance is a *Motorway* given that it is a *Road* is P(C|e); given that the likelihood that any instance of a road network entity (i) is a *Road* (say P(e)) (ii) is a *Motorway* (say P(C) . The similarity between the two concepts is equal to

$$MSC(e,C) = P(e \cap C) / P(e \cup C). \quad (2)$$
$$= P(e,C) / (P(e) + P(C) - P(e,C))$$

In case one of the classes is a subclass of the other, as in the case of a *Road* and a *Motorway*, the value of P(C,e) turns out to be 1 since any instance of *Motorway* is also an instance of *Road*. Thus in this case M SC (e,C) = 1 and MSC(e,C)=P(e)/P(C) which means that most similar concept among subclasses of a given class is its most specific subsumer. On the other hand, if P(C,e) = 0 for any case (and hence MSC(e,C)=0), it means that the two concepts are most disimilar. We use these equations extensively for our case studies and for further clarification of the computations the reader may refer to the explaination of BayesOWL [5].

3 Case Study: Ontologies from traffic code texts

Traffic code texts such as the Highway Code of UK[1] (HWC) and the New York Driver's Manual[2] (NYDM) are examples of formal texts, which not only mention the entities in a road network but also specify the permissible actions in the respective geographic jurisdiction. Kuhn has advocated the extraction of ontologies from such formal texts. Our case study involves the extraction of such ontologies from each of these traffic codes. We extract most frequently occurring entities and construct hierarchies of such entities. We also extract most frequently occurring actions in relation to these entities and construct hierarchies of actions as well. A further text analysis provides co-occurrence values of entity-action pairs, which are used to establish linkages between entities and their actions.

[1] www.highwaycode.gov.uk/

[2] http://www.nydmv.state.ny.us/dmanual/

In this section we discuss the extraction of probabilistic ontologies based on the text analysis. We also discuss the inferences obtained from such ontologies as opposed to conventional ontologies. It is important to note that the extraction of ontologies in this case is based on linguistic analysis and although analysis of formal texts is suggested to be a good source for building ontologies, our main purpose is to demonstrate the use of a probabilistic framework for geospatial ontologies. It is to be noted that linguistic analysis is not the cornerstone of our framework for probabilistic ontologies; rather, it serves as one of the tools, which assists in building such ontologies. Nevertheless, simplistic ontologies (as Directed Acyclic Graphs) have been developed from analysis of formal texts and we further the same methodology by using probabilistic values in the place of binary values for affordances of different road network entities.

3.1 Ontology extraction

The steps listed in § 2.1 are used to construct the BN based ontologies. The important constituents required for these are extracted from the text as follows.

1. Both texts are subjected to a Part Of Speech (POS) analysis which not only analyze the part of speech but also provides the sense of the words [19].
2. The most frequently occurring entities are used to construct a hierarchy of geospatial entities using hypernyms relations of noun terms from the WordNet lexicon [20].
3. Similarly hierarchies of geospatial action terms are used to construct the hierarchy of actions. Hypernym relations between verbs are used to construct such hierarchies.
4. WordNet-similarity modules [21] are used to extract the conditional probabilities between class and subclass relations in the two hierarchies. The CPTs thus obtained allow us to construct individual BayesOWL ontologies of entities and actions separately.
5. We go beyond this step by using the linkages between noun-verb pairs from the text analysis to link the two hierarchies together. A table of enitity concepts along with their assesed values of affordance for the given geospatial action concepts is used. The combined DAGs from the two texts are represented in figure 3 and 4 respectively. We need to clarify that the node denoting action concepts, when used in a combined DAG, represents the class of road network entities, which afford that particular action. Since the top concept for action concepts is move, we assume the top concept to be "all road network entities which afford the action *move*"

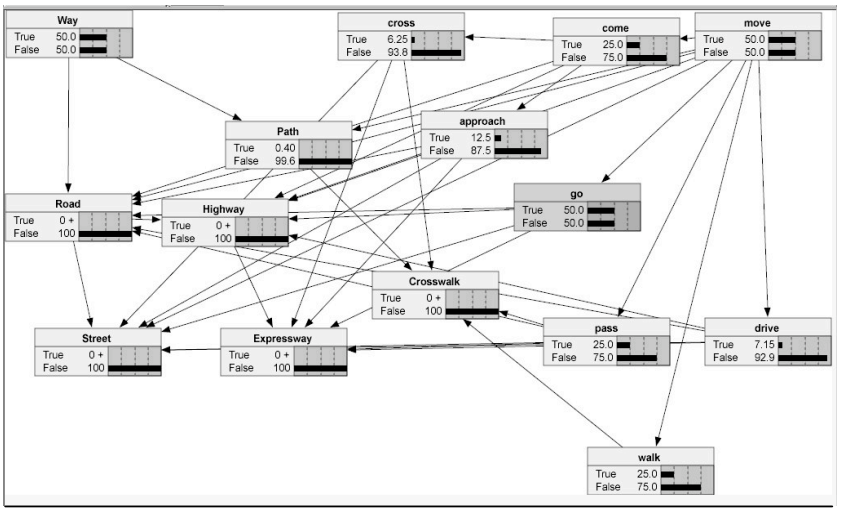


Figure 3 DAG extracted from the NYDM text, in the form of a Bayesian Network containing both geospatial enitity concepts (on the left with first letters in capitals) and action concepts (on the right). Edges within an BayesOWL ontology

3.2 Ontology reasoning and database ontologies

The main purpose of our experiments was to evaluate the utility of the developed Bayesian network based ontologies to carry out inferencing tasks for our case study.

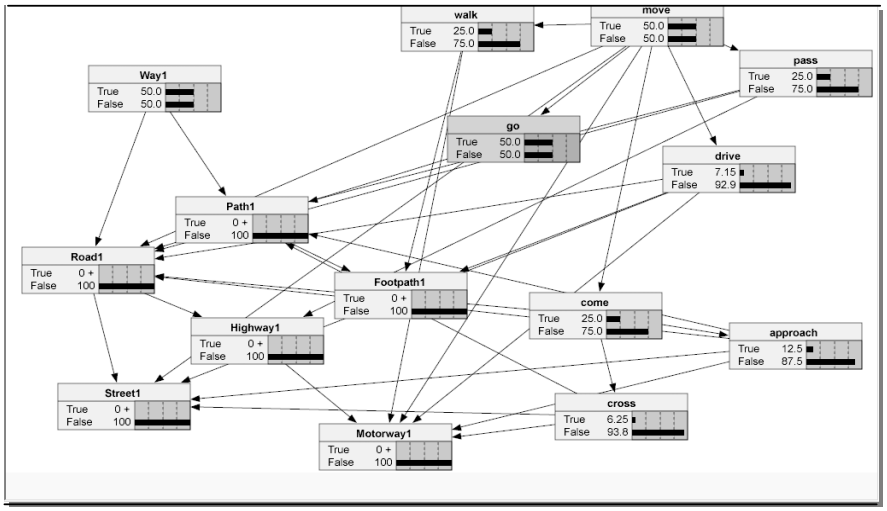


Figure 4 DAG extracted from the UK Highway code tex similar to Figure 4 above. Note that some new entity concepts (*Motorway* and *Footpath*) appear and some (*Crosswalk* and *Expressway*) are missing. The action concepts, however, remain consistent.

3.2.1 Inferences within an ontology

Given the Bayesian network ontologies as shown in figure 3 and 4, we now proceed to determine the most similar matches and most dissimilar matches within the same ontology. This is done using the notion of concept similarity described in § 2.2. We try to obtain the action concept matches in relation to the entity concepts. Table 2 depicts the results.

Table 2 Most similar and most dissimilar entity concepts of the verb concepts with in the same ontology. These are calculated on the basis of the similarity score

Entity Concept	Occurs in	Most similar action concept		Most dissimilar action concept	
Crosswalk	NYDM	cross		move,go	
Expressway	NYDM	drive		cross	
Footpath	HWC	cross		drive	
Highway	NYDM/HWC	drive	drive	walk	go,move
Motorway	HWC	drive		cross,walk	
Path	NYDM/HWC	move,go	cross	cross	move,go
Road	NYDM/HWC	drive	drive	cross,walk	cross,walk
Street	NYDM/HWC	cross,walk	cross,walk	go	Go
Way	NYDM/HWC	move,go	move,go	cross	cross,walk

3.2.2 Reasoning across ontologies with common functions

Finally we arrive at the bigger and more practical task of reasoning across ontologies. Since our two texts have differences in the list of geospatial entity concepts (the Highway code contains mention of *Footpath* and *Motorway* whereas the NY driver's manual mentions *Crosswalk* and *Expressway*, our task is to obtain the degree of overlap between these two concepts and the most similar concepts given their linkages with the common function concepts. To do this, we make an assumption that action concepts remain invariant across the ontologies such that the meanings of walk or drive remain the same (although the meaning of a *Road* and a *Highway* can differ). We create a virtual node for each node of the given ontology in the target ontology based on its conditional probabilities in respect to the action concepts (common to both ontologies). Thereafter we obtain the most similar and most dissimilar concepts based on the approach already used in § 3.2.1. Table 3 lists these top matches obtained from the two BNs.

Table 3 Most similar and dissimilar concepts of (i) the HWC in the NYDM and the NYDM in the HWC

HWC Concept	Most similar entity	Most dissimilar entity	NYDM Concept	Most similar entity	Most dissimilar entity
Footpath	Path	Expressway	Way	Way	Motorway
Highway	Way	Street	Street	Way	Street
Motorway	Road	Crosswalk	Road	Road	Street
Path	Path	Expressway	Path	Path	Motorway
Road	Road	Expressway	Highway	Path	Street
Street	Path	Street	Expressway	Road	Street
Way	Way	Expressway	Crosswalk	Path	Motorway

4 Psycholinguistic Verification

We have already stated that a simplistic evaluation of the machine based values of similarity and hence the mapping between concepts of two ontologies is not appropriate. This section explains human subjects testing based on the first case study and tries to compare the results of the machine based mappings vis-à-vis human generated ones.

4.1 Human Subjects testing

Human subject testing was conducted for 20 participants who were native English speakers or were highly proficient speakers and long-term residents of English-speaking countries. Participants were given two sets of cards, which had names of road network entities from each ontology (the Highway Code and NY Driver's Manual). The cards bearing names of Highway Code concepts were arranged in one row. Participants were asked to arrange the cards bearing NY Driver's Manual concepts in such a way that the entities that they believed were most similar were kept closest. After this task was completed, they were asked to flip the cards and read the sections of the texts relevant to the respective entities, which occurred in the corresponding traffic code texts. These sections provided information about the different actions that were permissible on that particular road network entity. After taking as much time as they needed to read the cards, participants repeated the matching task.

The mappings generated before and after flipping the cards (and hence before and after the knowledge about entity functions was available) were recorded and analyzed. The tests took not more than 20 minutes and were administered with no interference once the initial instructions were given. All 20 participants volunteered willingly and were debriefed at the end of the tests.

4.2 Analysis

Table 4 below summarizes some of the mappings generated from the human subject tests. We note that most dissimilar mappings are not reported here for sake of simplicity. We note that other than the cases of *Street* and *Motorway* most similar mappings also appear in machine-based mappings.

It is also important to note that the covariance of the mapping values in respect to age and gender was found to be insignificant (0.06). The variance of mappings produced

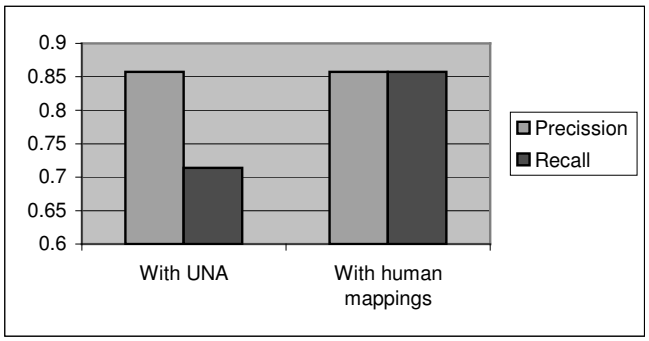
Table 4: Human generated mappings. Most similar and dissimilar concepts of the HWC in the NYDM

before reading the texts about entity functions			*after reading the texts about entity functions*		
NYDM Concept	*Most similar entity*	*Values (0 to 3)*	*NYDM Concept*	*Most similar entity*	*Value (0 to 3)*
Way	Way	2.7	Way	Way	1.8
Street	Street	2.7	Street	Street	2.85
Road	Road	2.65	Road	Road	2.95
Path	Path	2.5	Path	Path	1.2
Highway	Highway	0.875	Highway	Road	1.25
Expressway	Motorway	2.55	Expressway	Motorway	2.5
Crosswalk	Footpath	0.95	Crosswalk	Path	1.0

by subjects who have driven in both countries was found to be slightly lower than those who have driven only in one but this was fairly insignificant (0.09).

We have already discussed that there is a close resemblance in the machine based mappings and the human based mappings although they are not identical. It is possible to report precision and recall of the mappings in terms of false positives (when a true match is overlooked) and false negatives (when a incorrect match is reported), using a unique name assumption (assuming that entities which have same names in both ontologies are the same entities). This is not a good evaluation of the performance of the machine based mapping because naming heterogeneity is abundant in most cases. For example, the term *Highway* is used differently in the HWC and the NYDM and this is concurrent with the use of the word in the two countries as well. This is also evident from the results of our human subject tests. Thus evaluation of machine-based mappings warrants the use of human subjects testing to ascertain the goodness of the results.

The Graph 2 (below) compares the precision and recall values based on the unique name assumption and on the mappings produced by the human subject tests. The recall value remains the same (mainly due to the mismatch of the entity *Street* in the machine-based mappings). However recall has been shown to improve.



Graph 3 Comparing evaluations of machine-based mapping in the (i) absence or with Unique Name Assumption) and (ii) presence of human mapping values

5 Conclusions and Future Work

We have reported on a mechanism to design probabilistic ontologies in the geospatial domain. The use of text analysis to obtain information to construct such ontologies was discussed. Inferences based on such probabilistic geospatial ontologies provided results such as most similar and most dissimilar concepts within and across ontologies. Such results are comparable to human generated mappings. The precision and recall of ontology mapping exercises was found to be good with unique name assumptions of entities. The performance improved when human generated mappings were used as benchmarks. We summarize our conclusions from these case studies as follows:

1) Ontologies of geospatial entities need to be extended with probabilistic frameworks in order to enable rich and practical inferences such as concept similarity and concept overlaps.
2) It is possible to use both hierarchies of geospatial entities as well as geospatial actions and link them with probabilistic knowledge about affordances of geospatial entities.

3) The use of probabilistic geospatial ontologies for mappings between most similar entities mimics, to a large extent, the human mechanism of semantic translations of entity names. Our results provide support to the hypothesis that knowledge about geospatial actions and affordances to such actions are a critical part of geospatial knowledge.

This is only a first step in our experimental validation and our experience has shown that there exist many themes for future work. These include
(1) Inclusion of Disjoint, Equivalent, Intersection and Union relations: For simplification of our case study these relations were avoided although these relations can be easily determined from WordNet during text analysis. Using such relations in future will require use of some iterative algorithm such as Decomposed IPFP in order to enforce truth conditions of the LNodes in BayesOWL [17].
(2) Testing on industrial scale: this experiment, although at a prototype scale aims, in the end, to solve semantic problems, which occur at industrial scale.
(3) Machine based learning: The human mappings, especially that of the experts, are considered as the ideal mappings. Human interactions and judgments for most similar concepts can be used to improve heuristics involved in specification of entity-action linkages.

Acknowledgment:
The work presented in this paper was funded by Ordnance Survey, UK. Coments from three anynomus reviewers helped to improve the paper to its present form. The author is also thankful to members of MUSIL for their inputs.

References

1. Costa, P.C.G. and K.B. Laskey. *PR-OWL: A Framework for Probabilistic Ontologies*. in I*nternational Conference on Formal Ontology in Information Systems (FOIS 2006)*. 2006. Baltimore, MD: IOS Press.
2. Gärdenfors, P., *Conceptual Spaces - The Geometry of Thought*. 2000, Cambridge, MA: Bradford Books, MIT Press.
3. Raubal, M., *Formalizing Conceptual Spaces*, in *Formal Ontology in Information Systems, Proceedings of the Third International Conference (FOIS 2004)*, A. Varzi, et al., Editors. 2004, IOS Press: Amsterdam, NL. p. 153-164.
4. Kuhn, W., *Semantic Reference Systems.* International Journal of Geographical Information Science, 2003. **17**(5): p. 405-409.
5. Ding, Z., Y. Peng, and R. Pan, *BayesOWL: Uncertainty Modelling in Semantic Web Ontologies*, in *Soft Computing in Ontologies and Semantic Web*. 2005, Springer-Verlag.
6. Kuhn, W., *Geospatial Semantics: Why, of What, and How?* Journal on Data Semantics, 2005. **III**(2005).
7. Agarwal, P., *Ontological considerations in GIScience*. Int. Journal of Geographical Information Science, 2005. **19**(5): p. 501-536.
8. Bittner, T. and A. Frank, *On the design of formal theories of geographic space*. Journal of Geographical Systems, 1999. **1**(3): p. 237-275.
9. Raubal, M. and W. Kuhn, *Ontology-Based Task Simulation.* Spatial Cognition and Computation, 2004. **4**(1): p. 15-37.

10. Rüther, C., W. Kuhn, and Y. Bishr. *An algebraic description of a common ontology for ATKIS and GDF*. in *3rd AGILE Conference on Geographic Information Science*. 2000. Helsinki/Espoo, Finland.
11. Klien, E. and F. Probst. *Requirements for Geospatial Ontology Engineering*. in *Conference on Geographic Information Science (AGILE 2005)*. 2005. Estoril, Portugal.
12. Gomez-Perez, A. and V.R. Benjamins, *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web - 13th International Conference, EKAW 2002 (LNAI 2473)*. 2002: Springer-Verlag Heidelberg.
13. Kuhn, W., *Ontologies in Support of Activities in Geographic Space.* International Journal of Geographical Information Science, 2001. **15**(7): p. 613-631.
14. Sen, S. and K. Janowicz. *Semantics of Motion verbs*. in *Workshop on Spatial Language and Dialogue (5th Workshop on Language and Space)*. 2005. Delmenhorst, Germany.
15. Zhang, J.X. and M. Goodchild, *Uncertainty in Geographical Information*. 2002, Taylor and Francis: New York.
16. Norman, D., *Affordance, Conventions, and Design.* Interactions, 1999(May + June): p. 38-42.
17. Peng, Y. and Z. Ding. *Modifying Bayesian Networks by Probability Constraints*. in *21st Conference on Uncertainty in Artificial Intelligence (UAI-2005)*. 2005. Edinburgh, Scotland.
18. Rijsbergen, V., *Information Retrieval*. 2nd ed. 1979, London: Butterworths.
19. Decadt, B., et al. *GAMBL, Genetic Algorithm Optimization of Memory-Based WSD.* in *Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*. 2004. Barcelona, Spain.
20. Miller, G., *Wordnet: An Online Lexical Database.* Int. Journal of Lexicography, 1990. **3**(4): p. 235-312.
21. Patwardhan, S. and T. Pedersen. *Using WordNet Based Context Vectors to Estimate the Semantic Relatedness of Concepts*. in *EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*. 2006. Trento, Italy.

A Framework for Representing Ontology Mappings under Probabilities and Inconsistency

Andrea Calì [1], Thomas Lukasiewicz [2, 3], Livia Predoiu [4], and Heiner Stuckenschmidt [4]

[1] Computing Laboratory, University of Oxford, UK
andrea.cali@comlab.ox.ac.uk
[2] Dipartimento di Informatica e Sistemistica, Sapienza Università di Roma, Italy
lukasiewicz@dis.uniroma1.it
[3] Institut für Informationssysteme, Technische Universität Wien, Austria
lukasiewicz@kr.tuwien.ac.at
[4] Institut für Informatik, Universität Mannheim, Germany
{heiner,livia}@informatik.uni-mannheim.de

Abstract. Creating mappings between ontologies is a common way of approaching the semantic heterogeneity problem on the Semantic Web. To fit into the landscape of semantic web languages, a suitable, logic-based representation formalism for mappings is needed. We argue that such a formalism has to be able to deal with uncertainty and inconsistencies in automatically created mappings. We analyze the requirements for such a mapping language and present a formalism that combines tightly integrated description logic programs with independent choice logic for representing probabilistic information. We define the language, show that it can be used to resolve inconsistencies and merge mappings from different matchers based on the level of confidence assigned to different rules. We also analyze the computational aspects of consistency checking and query processing in tightly integrated probabilistic description logic programs.

1 Introduction

The problem of aligning heterogeneous ontologies via semantic mappings has been identified as one of the major challenges of semantic web technologies. In order to address this problem, a number of languages for representing semantic relations between elements in different ontologies as a basis for reasoning and query answering across multiple ontologies have been proposed [21]. In the presence of real world ontologies, it is unrealistic to assume that mappings between ontologies are created manually by domain experts, since existing ontologies, e.g., in the area of medicine contain thousands of concepts and hundreds of relations. Recently, a number of heuristic methods for matching elements from different ontologies have been proposed that support the creation of mappings between different languages by suggesting candidate mappings (e.g., [7]). These methods rely on linguistic and structural criteria. Evaluation studies have shown that existing methods often trade off precision and recall. The resulting mapping either contains a fair amount of errors or only covers a small part of the ontologies involved [6,8]. To leverage the weaknesses of the individual methods, it is common practice to combine the results of a number of matching components or even the results of different matching systems to achieve a better coverage of the problem [7].

This means that automatically created mappings often contain uncertain hypotheses and errors that need to be dealt with, as briefly summarized as follows:

- mapping hypotheses are often oversimplifying, since most matchers only support very simple semantic relations (mostly equivalence between individual elements);
- there may be conflicts between different hypotheses for semantic relations from different matching components and often even from the same matcher;
- semantic relations are only given with a degree of confidence in their correctness.

If we want to use the resulting mapping, we have to find a way to deal with these uncertainties and errors in a suitable way. We argue that the most suitable way of dealing with uncertainties in mappings is to provide means to explicitly represent uncertainties in the target language that encodes the mappings. In this paper, we address the problem of designing a mapping representation language that is capable of representing the kinds of uncertainty mentioned above. We propose an approach to such a language, which is based on an integration of ontologies and rules under probabilistic uncertainty.

There is a large body of work on integrating ontologies and rules, which is a promising way of representing mappings between ontologies. One type of integration is to build rules on top of ontologies, that is, rule-based systems that use vocabulary from ontology knowledge bases. Another form of integration is to build ontologies on top of rules, where ontological definitions are supplemented by rules or imported from rules. Both types of integration have been realized in recent hybrid integrations of rules and ontologies, called *description logic programs* (or *dl-programs*), which have the form $KB = (L, P)$, where L is a description logic knowledge base, and P is a finite set of rules involving either queries to L in a loose integration [5] or concepts and roles from L as unary resp. binary predicates in a tight integration [16] (see especially [5,18,16] for detailed overviews on the different types of description logic programs).

Other works explore formalisms for *uncertainty reasoning in the Semantic Web* (an important recent forum for approaches to uncertainty in the Semantic Web is the annual *Workshop on Uncertainty Reasoning for the Semantic Web (URSW)*; there also exists a W3C Incubator Group on *Uncertainty Reasoning for the World Wide Web*). There are especially probabilistic extensions of description logics [12], web ontology languages [2,3], and description logic programs [15] (to encode ambiguous information, such as "John is a student with the probability 0.7 and a teacher with the probability 0.3", which is very different from vague/fuzzy information, such as "John is tall with degree of truth 0.7"). In particular, [15] extends the loosely integrated description logic programs of [5] by probabilistic uncertainty as in Poole's independent choice logic (ICL) [20]. The ICL is a powerful representation and reasoning formalism for single- and also multi-agent systems, which combines logic and probability, and which can represent a number of important uncertainty formalisms, in particular, influence diagrams, Bayesian networks, Markov decision processes, normal form games, and Pearl's causal models [10].

In this paper, we propose a language for representing and reasoning with uncertain and possibly inconsistent mappings, where the tight integration between ontology and rule languages (namely, the tightly integrated disjunctive description logic programs of [16]) is combined with probabilistic uncertainty (as in the ICL). The resulting language has the following useful features, which will be explained in more detail later:

- The semantics is based on a tight integration of the rule and the ontology language. This enables us to have description logic concepts and roles in both rule bodies and rule heads. This is necessary if we want to use rules to combine ontologies.

– The rule language is quite expressive. In particular, we can have disjunctions in rule heads and nonmonotonic negations in rule bodies. This gives a rich basis for refining and rewriting automatically created mappings for resolving inconsistencies.
– The integration with probability theory provides us with a sound formal framework for representing and reasoning with confidence values. In particular, we can interpret the confidence values as error probabilities and use standard techniques for combining them. We can also resolve inconsistencies by using trust probabilities.
– In [1], we show that consistency checking and query processing in the new rule language are decidable resp. computable, and can be reduced to their classical counterparts in tightly integrated disjunctive description logic programs. We also analyze the complexity of consistency checking and query processing in special cases.
– In [1], we show that there are tractable subsets of the language that are of practical relevance. In particular, we show that when ontologies are represented in *DL-Lite*, reasoning in the language can be done in polynomial time in the data complexity.

2 Representation Requirements

The problem of ontology matching can be defined as follows [7]. Ontologies are theories encoded in a certain language L. In this work, we assume that ontologies are encoded in OWL DL or OWL Lite. For each ontology O in language L, we denote by $Q(O)$ the matchable elements of the ontology O. Given two ontologies O and O', the task of matching is now to determine correspondences between the matchable elements in the two ontologies. Correspondences are 5-tuples (id, e, e', r, n) such that

– id is a unique identifier for referring to the correspondence;
– $e \in Q(O)$ and $e' \in Q(O')$ are matchable elements from the two ontologies;
– $r \in R$ is a semantic relation (in this work, we consider the case where the semantic relation can be interpreted as an implication);
– n is a degree of confidence in the correctness of the correspondence.

From this general description of automatically generated correspondences between ontologies, we can derive a number of requirements for a formal language for representing the results of multiple matchers as well as the contained uncertainties:

– *Tight integration of mapping and ontology language:* The semantics of the language used to represent the correspondences between elements in different ontologies has to be tightly integrated with the semantics of the ontology language used (in this case OWL). This is important if we want to use the correspondences to reason across different ontologies in a semantically coherent way. In particular, this means that the interpretation of the mapped elements depends on the definitions in the ontologies.

– *Support for mappings refinement:* The language should be expressive enough to allow the user to refine oversimplifying correspondences suggested by the matching system. This is important to be able to provide a precise account of the true semantic relation between elements in the mapped ontologies. In particular, this requires the ability to describe correspondences that include several elements from the two ontologies.

– *Support for repairing inconsistencies:* Inconsistent mappings are a major problem for the combined use of ontologies because they can cause inconsistencies in the mapped ontologies. These inconsistencies can make logical reasoning impossible, since everything can be derived from an inconsistent ontology. The mapping language should be able to represent and reason about inconsistent mappings in an approximate fashion.

– *Representation and combination of confidence:* The confidence values provided by matching systems is an important indicator for the uncertainty that has to be taken into account. The mapping representation language should be able to use these confidence values when reasoning with mappings. In particular, it should be able to represent the confidence in a mapping rule and to combine confidence values on a sound formal basis.

– *Decidability and efficiency of instance reasoning:* An important use of ontology mappings is the exchange of data across different ontologies. In particular, we normally want to be able to ask queries using the vocabulary of one ontology and receive answers that do not only consist of instances of this ontology but also of ontologies connected through ontology mappings. To support this, query answering in the combined formalism consisting of ontology language and mapping language has to be decidable and there should be efficient algorithms for answering queries at least for relevant cases.

Throughout the paper, we use real data form the Ontology Alignment Evaluation Initiative[1] to illustrate the different aspects of mapping representation. In particular, we use examples from the benchmark and the conference data set. The benchmark dataset consists of five OWL ontologies (tests 101 and 301 to 304) describing scientific publications and related information. The conference dataset consists of about 10 OWL ontologies describing concepts related to conference organization and management. In both cases, we give examples of mappings that have been created by the participants of the 2006 evaluation campaign. In particular, we use mappings created by state-of-the-art ontology matching systems like falcon, hmatch, and coma++.

3 Description Logics

In this section, we recall the expressive description logics $\mathcal{SHIF}(\mathbf{D})$ and $\mathcal{SHOIN}(\mathbf{D})$, which stand behind the web ontology languages OWL Lite and OWL DL [13], respectively. Intuitively, description logics model a domain of interest in terms of concepts and roles, which represent classes of individuals and binary relations between classes of individuals, respectively. A description logic knowledge base encodes especially subset relationships between concepts, subset relationships between roles, the membership of individuals to concepts, and the membership of pairs of individuals to roles.

3.1 Syntax. We first describe the syntax of $\mathcal{SHOIN}(\mathbf{D})$. We assume a set of *elementary datatypes* and a set of *data values*. A *datatype* is either an elementary datatype or a set of data values (*datatype oneOf*). A *datatype theory* $\mathbf{D} = (\Delta^{\mathbf{D}}, \cdot^{\mathbf{D}})$ consists of a *datatype domain* $\Delta^{\mathbf{D}}$ and a mapping $\cdot^{\mathbf{D}}$ that assigns to each elementary datatype a subset of $\Delta^{\mathbf{D}}$ and to each data value an element of $\Delta^{\mathbf{D}}$. The mapping $\cdot^{\mathbf{D}}$ is extended to all datatypes by $\{v_1, \ldots\}^{\mathbf{D}} = \{v_1^{\mathbf{D}}, \ldots\}$. Let $\mathbf{A}$, $\mathbf{R}_A$, $\mathbf{R}_D$, and $\mathbf{I}$ be pairwise disjoint (denumerable) sets of *atomic concepts*, *abstract roles*, *datatype roles*, and *individuals*, respectively. We denote by $\mathbf{R}_A^-$ the set of *inverses* R^- of all $R \in \mathbf{R}_A$.

A *role* is any element of $\mathbf{R}_A \cup \mathbf{R}_A^- \cup \mathbf{R}_D$. *Concepts* are inductively defined as follows. Every $\phi \in \mathbf{A}$ is a concept, and if $o_1, \ldots, o_n \in \mathbf{I}$, then $\{o_1, \ldots, o_n\}$ is a concept (*oneOf*). If ϕ, ϕ_1, and ϕ_2 are concepts and if $R \in \mathbf{R}_A \cup \mathbf{R}_A^-$, then also $(\phi_1 \sqcap \phi_2)$, $(\phi_1 \sqcup \phi_2)$, and $\neg\phi$ are concepts (*conjunction*, *disjunction*, and *negation*, respectively), as well as $\exists R.\phi$, $\forall R.\phi$, $\geqslant nR$, and $\leqslant nR$ (*exists*, *value*, *atleast*, and *atmost restriction*,

[1] http://oaei.ontologymatching.org/2006/

respectively) for an integer $n \geqslant 0$. If D is a datatype and $U \in \mathbf{R}_D$, then $\exists U.D$, $\forall U.D$, $\geqslant nU$, and $\leqslant nU$ are concepts (*datatype exists*, *value*, *atleast*, and *atmost restriction*, respectively) for an integer $n \geqslant 0$. We write $\top$ and $\bot$ to abbreviate the concepts $\phi \sqcup \neg\phi$ and $\phi \sqcap \neg\phi$, respectively, and we eliminate parentheses as usual.

An *axiom* has one of the following forms: (1) $\phi \sqsubseteq \psi$ (*concept inclusion axiom*), where ϕ and ψ are concepts; (2) $R \sqsubseteq S$ (*role inclusion axiom*), where either $R, S \in \mathbf{R}_A \cup \mathbf{R}_A^-$ or $R, S \in \mathbf{R}_D$; (3) $\mathrm{Trans}(R)$ (*transitivity axiom*), where $R \in \mathbf{R}_A$; (4) $\phi(a)$ (*concept membership axiom*), where ϕ is a concept and $a \in \mathbf{I}$; (5) $R(a, b)$ (resp., $U(a, v)$) (*role membership axiom*), where $R \in \mathbf{R}_A$ (resp., $U \in \mathbf{R}_D$) and $a, b \in \mathbf{I}$ (resp., $a \in \mathbf{I}$ and v is a data value); and (6) $a = b$ (resp., $a \neq b$) (*equality* (resp., *inequality*) *axiom*), where $a, b \in \mathbf{I}$. A *(description logic) knowledge base* L is a finite set of axioms. For decidability, number restrictions in L are restricted to simple abstract roles [14].

The syntax of $\mathcal{SHIF}(\mathbf{D})$ is as the above syntax of $\mathcal{SHOIN}(\mathbf{D})$, but without the oneOf constructor and with the atleast and atmost constructors limited to 0 and 1.

3.2 Semantics. An *interpretation* $\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})$ relative to a datatype theory $\mathbf{D} = (\Delta^\mathbf{D}, \cdot^\mathbf{D})$ consists of a nonempty (*abstract*) *domain* $\Delta^\mathcal{I}$ disjoint from $\Delta^\mathbf{D}$, and a mapping $\cdot^\mathcal{I}$ that assigns to each atomic concept $\phi \in \mathbf{A}$ a subset of $\Delta^\mathcal{I}$, to each individual $o \in \mathbf{I}$ an element of $\Delta^\mathcal{I}$, to each abstract role $R \in \mathbf{R}_A$ a subset of $\Delta^\mathcal{I} \times \Delta^\mathcal{I}$, and to each datatype role $U \in \mathbf{R}_D$ a subset of $\Delta^\mathcal{I} \times \Delta^\mathbf{D}$. We extend $\cdot^\mathcal{I}$ to all concepts and roles, and we define the *satisfaction* of an axiom F in an interpretation $\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})$, denoted $\mathcal{I} \models F$, as usual [13]. We say $\mathcal{I}$ *satisfies* the axiom F, or $\mathcal{I}$ is a *model* of F, iff $\mathcal{I} \models F$. We say $\mathcal{I}$ *satisfies* a knowledge base L, or $\mathcal{I}$ is a *model* of L, denoted $\mathcal{I} \models L$, iff $\mathcal{I} \models F$ for all $F \in L$. We say L is *satisfiable* iff L has a model. An axiom F is a *logical consequence* of L, denoted $L \models F$, iff every model of L satisfies F.

4 Description Logic Programs

In this section, we recall the novel approach to *description logic programs* (or *dl-programs*) $KB = (L, P)$ from [16], where KB consists of a description logic knowledge base L and a disjunctive logic program P. Their semantics is defined in a modular way as in [5], but it allows for a much tighter integration of L and P. Note that we do not assume any structural separation between the vocabularies of L and P. The main idea behind their semantics is to interpret P relative to Herbrand interpretations that are compatible with L, while L is interpreted relative to general interpretations over a first-order domain. Thus, we modularly combine the standard semantics of logic programs and of description logics, which allows for building on the standard techniques and results of both areas. As another advantage, the novel dl-programs are decidable, even when their components of logic programs and description logic knowledge bases are both very expressive. See especially [16] for further details on the new approach to dl-programs and for a detailed comparison to related works.

4.1 Syntax. We assume a first-order vocabulary Φ with finite nonempty sets of constant and predicate symbols, but no function symbols. We use Φ_c to denote the set of all constant symbols in Φ. We also assume a set of data values $\mathbf{V}$ (relative to a datatype theory $\mathbf{D} = (\Delta^\mathbf{D}, \cdot^\mathbf{D})$) and pairwise disjoint (denumerable) sets $\mathbf{A}$, $\mathbf{R}_A$, $\mathbf{R}_D$, and $\mathbf{I}$ of atomic concepts, abstract roles, datatype roles, and individuals, respectively, as in

Section 3. We assume that (i) Φ_c is a subset of $\mathbf{I} \cup \mathbf{V}$, and that (ii) Φ and $\mathbf{A}$ (resp., $\mathbf{R}_A \cup \mathbf{R}_D$) may have unary (resp., binary) predicate symbols in common.

Let $\mathcal{X}$ be a set of variables. A *term* is either a variable from $\mathcal{X}$ or a constant symbol from Φ. An *atom* is of the form $p(t_1, \ldots, t_n)$, where p is a predicate symbol of arity $n \geqslant 0$ from Φ, and $t_1, \ldots, t_n$ are terms. A *literal* l is an atom p or a default-negated atom $not\, p$. A *disjunctive rule* (or simply *rule*) r is an expression of the form

$$\alpha_1 \vee \cdots \vee \alpha_k \leftarrow \beta_1, \ldots, \beta_n, not\, \beta_{n+1}, \ldots, not\, \beta_{n+m}\,, \tag{1}$$

where $\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_{n+m}$ are atoms and $k, m, n \geqslant 0$. We call $\alpha_1 \vee \cdots \vee \alpha_k$ the *head* of r, while the conjunction $\beta_1, \ldots, \beta_n, not\, \beta_{n+1}, \ldots, not\, \beta_{n+m}$ is its *body*. We define $H(r) = \{\alpha_1, \ldots, \alpha_k\}$ and $B(r) = B^+(r) \cup B^-(r)$, where $B^+(r) = \{\beta_1, \ldots, \beta_n\}$ and $B^-(r) = \{\beta_{n+1}, \ldots, \beta_{n+m}\}$. A *disjunctive program* P is a finite set of disjunctive rules of the form (1). We say P is *positive* iff $m = 0$ for all disjunctive rules (1) in P. We say P is a *normal program* iff $k \leqslant 1$ for all disjunctive rules (1) in P.

A *disjunctive description logic program* (or *disjunctive dl-program*) $KB = (L, P)$ consists of a description logic knowledge base L and a disjunctive program P. We say KB is *positive* iff P is positive. It is a *normal dl-program* iff P is a normal program.

4.2 Semantics. We now define the answer set semantics of disjunctive dl-programs as a generalization of the answer set semantics of ordinary disjunctive logic programs. In the sequel, let $KB = (L, P)$ be a disjunctive dl-program.

A *ground instance* of a rule $r \in P$ is obtained from r by replacing every variable that occurs in r by a constant symbol from Φ_c. We denote by $ground(P)$ the set of all ground instances of rules in P. The *Herbrand base* relative to Φ, denoted HB_Φ, is the set of all ground atoms constructed with constant and predicate symbols from Φ. We use DL_Φ to denote the set of all ground atoms in HB_Φ that are constructed from atomic concepts in $\mathbf{A}$, abstract roles in $\mathbf{R}_A$, and datatype roles in $\mathbf{R}_D$.

An *interpretation* I is any subset of HB_Φ. Informally, every such I represents the Herbrand interpretation in which all $a \in I$ (resp., $a \in HB_\Phi - I$) are true (resp., false). We say an interpretation I is a *model* of a description logic knowledge base L, denoted $I \models L$, iff $L \cup I \cup \{\neg a \mid a \in HB_\Phi - I\}$ is satisfiable. We say I is a *model* of a ground atom $a \in HB_\Phi$, or I *satisfies* a, denoted $I \models a$, iff $a \in I$. We say I is a *model* of a ground rule r, denoted $I \models r$, iff $I \models \alpha$ for some $\alpha \in H(r)$ whenever $I \models B(r)$, that is, $I \models \beta$ for all $\beta \in B^+(r)$ and $I \not\models \beta$ for all $\beta \in B^-(r)$. We say I is a *model* of a set of rules P iff $I \models r$ for every $r \in ground(P)$. We say I is a *model* of a disjunctive dl-program $KB = (L, P)$, denoted $I \models KB$, iff I is a model of both L and P.

We now define the answer set semantics of disjunctive dl-programs by generalizing the ordinary answer set semantics of disjunctive logic programs. We generalize the definition via the FLP-reduct [9] (which coincides with the answer set semantics defined via the Gelfond-Lifschitz reduct [11]). Given a dl-program $KB = (L, P)$, the *FLP-reduct* of KB relative to an interpretation $I \subseteq HB_\Phi$, denoted KB^I, is the dl-program (L, P^I), where P^I is the set of all $r \in ground(P)$ such that $I \models B(r)$. An interpretation $I \subseteq HB_\Phi$ is an *answer set* of KB iff I is a minimal model of KB^I. A dl-program KB is *consistent* (resp., *inconsistent*) iff it has an (resp., no) answer set.

We finally define the notions of *cautious* (resp., *brave*) *reasoning* from disjunctive dl-programs under the answer set semantics as follows. A ground atom $a \in HB_\Phi$ is a *cautious* (resp., *brave*) *consequence* of a disjunctive dl-program KB under the answer set semantics iff every (resp., some) answer set of KB satisfies a.

4.3 Semantic Properties. We now summarize some important semantic properties of disjunctive dl-programs under the above answer set semantics. In the ordinary case, every answer set of a disjunctive program P is also a minimal model of P, and the converse holds when P is positive. This result holds also for disjunctive dl-programs.

The following theorem shows that the answer set semantics of disjunctive dl-programs faithfully extends its ordinary counterpart. That is, the answer set semantics of a disjunctive dl-program with empty description logic knowledge base coincides with the ordinary answer set semantics of its disjunctive program.

Theorem 4.1 (see [16]). *Let $KB{=}(L, P)$ be a disjunctive dl-program with $L{=}\emptyset$. Then, the set of all answer sets of KB coincides with the set of all ordinary answer sets of P.*

The next theorem shows that the answer set semantics of disjunctive dl-programs also faithfully extends (from the perspective of answer set programming) the first-order semantics of description logic knowledge bases. That is, $\alpha \in HB_\Phi$ is true in all answer sets of a positive disjunctive dl-program $KB = (L, P)$ iff α is true in all first-order models of $L \cup ground(P)$. In particular, $\alpha \in HB_\Phi$ is true in all answer sets of $KB = (L, \emptyset)$ iff α is true in all first-order models of L. Note that the theorem holds also when α is a ground formula constructed from HB_Φ using the operators $\wedge$ and $\vee$.

Theorem 4.2 (see [16]). *Let $KB = (L, P)$ be a positive disjunctive dl-program, and let α be a ground atom from HB_Φ. Then, α is true in all answer sets of KB iff α is true in all first-order models of $L \cup ground(P)$.*

4.4 Representing Mappings. Tightly integrated disjunctive dl-programs $KB = (L, P)$ provide a natural way for representing mappings between two heterogeneous ontologies O_1 and O_2 as follows. The description logic knowledge base L is the union of two independent description logic knowledge bases L_1 and L_2 (representing O_1 resp. O_2) with signatures $\mathbf{A}_1$, $\mathbf{R}_{A,1}$, $\mathbf{R}_{D,1}$, $\mathbf{I}_1$ and $\mathbf{A}_2$, $\mathbf{R}_{A,2}$, $\mathbf{R}_{D,2}$, $\mathbf{I}_2$, respectively, such that $\mathbf{A}_1 \cap \mathbf{A}_2 = \emptyset$, $\mathbf{R}_{A,1} \cap \mathbf{R}_{A,2} = \emptyset$, $\mathbf{R}_{D,1} \cap \mathbf{R}_{D,2} = \emptyset$, and $\mathbf{I}_1 \cap \mathbf{I}_2 = \emptyset$. Note that this can easily be achieved for any pair of ontologies by a suitable renaming. A mapping between elements e and e' from L_1 and L_2, respectively, is then represented by a simple rule $e'(\overrightarrow{x}) \leftarrow e(\overrightarrow{x})$ in P, where $e \in \mathbf{A}_1 \cup \mathbf{R}_{A,1} \cup \mathbf{R}_{D,1}$, $e' \in \mathbf{A}_2 \cup \mathbf{R}_{A,2} \cup \mathbf{R}_{D,2}$, and $\overrightarrow{x}$ is a suitable variable vector. Note that the fact that we demand that the signatures of L_1 and L_2 are disjoint guarantees that the rule base that represents mappings between different ontologies is stratified as long as there are no cyclic mapping relations.

Taking some examples from the conference data set of the OAEI challenge 2006, we find e.g. the following mappings that were created by automatic matching systems:[2]

$$\begin{aligned} &\mathit{NegativeReview}(X) \leftarrow \mathit{Review}(X)\,; \\ &\mathit{NeutralReview}(X) \leftarrow \mathit{Review}(X)\,; \\ &\mathit{PositiveReview}(X) \leftarrow \mathit{Review}(X)\,. \end{aligned}$$

Another example of created mapping relations are the following:[3]

$$\begin{aligned} &\mathit{EarlyRegisteredParticipant}(X) \leftarrow \mathit{participant}(X)\,; \\ &\mathit{LateRegisteredParticipant}(X) \leftarrow \mathit{participant}(X)\,. \end{aligned}$$

[2] Results of the hmatch system for mapping the SIGKDD on the EKAW Ontology.

[3] Results of the hmatch system for mapping the CRS on the EKAW Ontology.

Both of these sets of correspondences are examples of mappings that introduce inconsistency in the target ontology. The reason is that the three concepts $NegativeReview$, $NeutralReview$, and $PositiveReview$, as well as the two concepts $EarlyRegistered$-$Participant$ and $LateregisteredParticipant$ are defined to be disjoint in the corresponding ontologies. Using the rules as shown above will make an instance of the concept $Review$ (resp., $participant$) a member of disjoint classes. In [17], we have presented a method for detecting such inconsistent mappings. There are different approaches for resolving this inconsistency. The most straightforward one is to drop mappings until no inconsistency is present anymore. Peng and Xu [19] have proposed a more suitable method for dealing with inconsistencies in terms of a relaxation of the mappings. In particular, they propose to replace a number of conflicting mappings by a single mapping that includes a disjunction of the conflicting concepts. In the first example above, we would replace the three rules by the following one:

$$NegativeReview(X) \vee NeutralReview(X) \vee PositiveReview(X) \leftarrow Review(X)\,.$$

This new mapping rule can be represented in our framework and resolves the inconsistency. In this particular case, it also correctly captures the meaning of the concepts.

In principle, the second example can be solved using the same approach. In this case, however, the actual semantics of the concepts can be captured more accurately by refining the rules and making use of the full expressiveness of the mapping language. In particular, we can resolve the inconsistency by extending the body of the mapping rules with additional requirements:

$$EarlyRegisteredParticipant(X) \leftarrow participant(X) \wedge RegisterdbeforeDeadline(X)\,;$$
$$LateRegisteredParticipant(X) \leftarrow participant(X) \wedge not\ RegisteredbeforeDeadline(X)\,.$$

This refinement of the mapping rules resolves the inconsistency and also provides a more correct mapping. A drawback of this approach is the fact that it requires manual post-processing of mappings. In the next section, we present a probabilistic extension of tightly integrated disjunctive dl-programs that allows us to directly use confidence estimations of matching engines to resolve inconsistencies and to combine the results of different matchers.

5 Probabilistic Description Logic Programs

In this section, we present a *tightly integrated* approach to *probabilistic disjunctive description logic programs* (or simply *probabilistic dl-programs*) *under the answer set semantics*. Differently from [15] (in addition to being a tightly integrated approach), the probabilistic dl-programs here also allow for disjunctions in rule heads. Similarly to the probabilistic dl-programs in [15], they are defined as a combination of dl-programs with Poole's ICL [20], but using the tightly integrated disjunctive dl-programs of [16] (see Section 4), rather than the loosely integrated dl-programs of [5]. Poole's ICL is based on ordinary acyclic logic programs P under different "choices", where every choice along with P produces a first-order model, and one then obtains a probability distribution over the set of all first-order models by placing a probability distribution over the different choices. We use the tightly integrated disjunctive dl-programs under the answer set semantics of [16], instead of ordinary acyclic logic programs under

their canonical semantics (which coincides with their answer set semantics). We first introduce the syntax of probabilistic dl-programs and then their answer set semantics.

5.1 Syntax. We now define the syntax of probabilistic dl-programs and probabilistic queries to them. We first introduce choice spaces and probabilities on choice spaces.

A *choice space* C is a set of pairwise disjoint and nonempty sets $A \subseteq HB_{\Phi} - DL_{\Phi}$. Any $A \in C$ is an *alternative* of C and any element $a \in A$ an *atomic choice* of C. Intuitively, every alternative $A \in C$ represents a random variable and every atomic choice $a \in A$ one of its possible values. A *total choice* of C is a set $B \subseteq HB_{\Phi}$ such that $|B \cap A| = 1$ for all $A \in C$ (and thus $|B| = |C|$). Intuitively, every total choice B of C represents an assignment of values to all the random variables. A *probability* μ on a choice space C is a probability function on the set of all total choices of C. Intuitively, every probability μ is a probability distribution over the set of all variable assignments. Since C and all its alternatives are finite, μ can be defined by (i) a mapping $\mu\colon \bigcup C \rightarrow [0,1]$ such that $\sum_{a \in A} \mu(a) = 1$ for all $A \in C$, and (ii) $\mu(B) = \Pi_{b \in B} \mu(b)$ for all total choices B of C. Intuitively, (i) defines a probability over the values of each random variable of C, and (ii) assumes independence between the random variables.

A *probabilistic dl-program* $KB = (L, P, C, \mu)$ consists of a disjunctive dl-program (L, P), a choice space C such that no atomic choice in C coincides with the head of any rule in $ground(P)$, and a probability μ on C. Intuitively, since the total choices of C select subsets of P, and μ is a probability distribution on the total choices of C, every probabilistic dl-program is the compact representation of a probability distribution on a finite set of disjunctive dl-programs. Observe here that P is fully general and not necessarily stratified or acyclic. We say KB is *normal* iff P is normal. A *probabilistic query* to KB has the form $\exists(c_1(\boldsymbol{x}) \vee \cdots \vee c_n(\boldsymbol{x}))[r, s]$, where $\boldsymbol{x}, r, s$ is a tuple of variables, $n \geqslant 1$, and each $c_i(\boldsymbol{x})$ is a conjunction of atoms constructed from predicate and constant symbols in Φ and variables in $\boldsymbol{x}$. Note that the above probabilistic queries can also be easily extended to conditional expressions as in [15].

5.2 Semantics. We now define an answer set semantics of probabilistic dl-programs, and we introduce the notions of consistency, consequence, tight consequence, and correct and tight answers for probabilistic queries to probabilistic dl-programs.

Given a probabilistic dl-program $KB = (L, P, C, \mu)$, a *probabilistic interpretation* Pr is a probability function on the set of all $I \subseteq HB_{\Phi}$. We say Pr is an *answer set* of KB iff (i) every interpretation $I \subseteq HB_{\Phi}$ with $Pr(I) > 0$ is an answer set of $(L, P \cup \{p \leftarrow \mid p \in B\})$ for some total choice B of C, and (ii) $Pr(\bigwedge_{p \in B} p) = \sum_{I \subseteq HB_{\Phi},\, B \subseteq I} Pr(I) = \mu(B)$ for every total choice B of C. Informally, Pr is an answer set of $KB = (L, P, C, \mu)$ iff (i) every interpretation $I \subseteq HB_{\Phi}$ of positive probability under Pr is an answer set of the dl-program (L, P) under some total choice B of C, and (ii) Pr coincides with μ on the total choices B of C. We say KB is *consistent* iff it has an answer set Pr.

We define the notions of consequence and tight consequence as follows. Given a probabilistic query $\exists(q(\boldsymbol{x}))[r, s]$, the *probability* of $q(\boldsymbol{x})$ in a probabilistic interpretation Pr under a variable assignment σ, denoted $Pr_{\sigma}(q(\boldsymbol{x}))$ is defined as the sum of all $Pr(I)$ such that $I \subseteq HB_{\Phi}$ and $I \models_{\sigma} q(\boldsymbol{x})$. We say $(q(\boldsymbol{x}))[l, u]$ (where $l, u \in [0,1]$) is a *consequence* of KB, denoted $KB \mid\!\sim (q(\boldsymbol{x}))[l, u]$, iff $Pr_{\sigma}(q(\boldsymbol{x})) \in [l, u]$ for every answer set Pr of KB and every variable assignment σ. We say $(q(\boldsymbol{x}))[l, u]$ (where $l, u \in [0,1]$) is a *tight consequence* of KB, denoted $KB \mid\!\sim_{tight} (q(\boldsymbol{x}))[l, u]$, iff l (resp., u) is the

infimum (resp., supremum) of $Pr_\sigma(q(\boldsymbol{x}))$ subject to all answer sets Pr of KB and all σ. A *correct* (resp., *tight*) *answer* to a probabilistic query $\exists(c_1(\boldsymbol{x}) \vee \cdots \vee c_n(\boldsymbol{x}))[r,s]$ is a ground substitution θ (for the variables $\boldsymbol{x}, r, s$) such that $(c_1(\boldsymbol{x}) \vee \cdots \vee c_n(\boldsymbol{x}))[r,s]\,\theta$ is a consequence (resp., tight consequence) of KB.

5.3 Representing and Combining Confidence Values. The probabilistic extension of disjunctive dl-programs $KB = (L, P)$ to probabilistic dl-programs $KB' = (L, P, C, \mu)$ provides us with a means to explicitly represent and use the confidence values provided by matching systems. In particular, we can interpret the confidence value as an *error probability* and state that the probability that a mapping introduces an error is $1 - n$. Conversely, the probability that a mapping correctly describes the semantic relation between elements of the different ontologies is $1 - (1 - n) = n$. This means that we can use the confidence value n as a probability for the correctness of a mapping. The indirect formulation is chosen, because it allows us to combine the results of different matchers in a meaningful way. In particular, if we assume that the error probabilities of two matchers are independent, we can calculate the joint error probability of two matchers that have found the same mapping rule as $(1 - n_1) \cdot (1 - n_2)$. This means that we can get a new probability for the correctness of the rule found by two matchers which is $1 - (1 - n_1) \cdot (1 - n_2)$. This way of calculating the joint probability meets the intuition that a mapping is more likely to be correct if it has been discovered by more than one matcher because $1 - (1 - n_1) \cdot (1 - n_2) \geqslant n_1$ and $1 - (1 - n_1) \cdot (1 - n_2) \geqslant n_2$.

In addition, when merging inconsistent results of different matching systems, we weigh each matching system and its result with a (user-defined) *trust probability*, which describes our confidence in its quality. All these trust probabilities sum up to 1. For example, the trust probabilities of the matching systems m_1, m_2, and m_3 may be 0.6, 0.3, and 0.1, respectively. That is, we trust most in m_1, medium in m_2, and less in m_3. Note that similarly one can associate trust probabilities with single mapping rules.

We illustrate this approach using an example from the benchmark data set of the OAEI 2006 campaign. In particular, we consider the case where the publication ontology in test 101 (O_1) is mapped on the ontology of test 302 (O_2). Below we show some mappings that have been detected by the matching system hmatch that participated in the challenge. The mappings are described as rules in P, which contain a conjunct indicating the matching system that has created it and a number for identifying the mapping. These additional conjuncts are atomic choices of the choice space C and link probabilities (which are specified in the probability μ on the choice space C) to the rules (where the common concept $Proceedings$ of both ontologies O_1 and O_2 is renamed to the concepts $Proceedings_1$ and $Proceedings_2$, respectively):

$$Book(X) \leftarrow Collection(X) \wedge hmatch_1\,;$$
$$Proceedings_2(X) \leftarrow Proceedings_1(X) \wedge hmatch_2\,.$$

We define the choice space according to the interpretation of confidence described above. The resulting choice space is $C = \{\{hmatch_i, not\_hmatch_i\} \mid i \in \{1, 2\}\}$. It comes along with the probability μ on C, which assigns the corresponding confidence value n to each atomic choice $hmatch_i$ and the complement $1 - n$ to the atomic choice not_hmatch_i. In our case, we have $\mu(hmatch_1) = 0.62$, $\mu(not\_hmatch_1) = 0.38$, $\mu(hmatch_2) = 0.73$, and $\mu(not\_hmatch_2) = 0.27$.

The benefits of this explicit treatment of the uncertainty becomes clear when we now try to merge this mapping with the result of another matching system. Below are two examples of rules that describe correspondences for the same ontologies that have been found by the falcon system:

$$InCollection(X) \leftarrow Collection(X) \wedge falcon_1 ;$$
$$Proceedings_2(X) \leftarrow Proceedings_1(X) \wedge falcon_2 .$$

Here, the confidence encoding yields the choice space $C' = \{\{falcon_i, not\_falcon_i\} \mid i \in \{1,2\}\}$ along with the probabilities $\mu'(falcon_1) = 0.94$ and $\mu'(falcon_2) = 0.96$.

Note that just putting together the rules without considering the choice space would lead to the same inconsistency problems shown in the last section, because the concepts $Book$ and $InCollection$ are disjoint. Further, the fact that the mapping between the concepts $Proceeding_1$ and $Proceeding_2$ has been found by both matchers is not considered and this mapping rule would have the same status as any other rule in the mapping.

Suppose we associate with hmatch and falcon the trust probabilities 0.55 and 0.45, respectively. Based on the interpretation of confidence values as error probabilities, and on the use of trust probabilities when resolving inconsistencies between rules, we can now define a merged mapping set that consists of the following rules:

$$Book(X) \leftarrow Collection(X) \wedge hmatch_1 \wedge sel\_hmatch_1 ;$$
$$InCollection(X) \leftarrow Collection(X) \wedge falcon_1 \wedge sel\_falcon_1 ;$$
$$Proceedings_2(X) \leftarrow Proceedings_1(X) \wedge hmatch_2 ;$$
$$Proceedings_2(X) \leftarrow Proceedings_1(X) \wedge falcon_2 .$$

The new choice space C'' and the new probability μ'' on C'' are obtained from $C \cup C'$ and $\mu \cdot \mu'$ (which is the product of μ and μ', that is, $(\mu \cdot \mu')(B \cup B') = \mu(B) \cdot \mu'(B')$ for all total choices B of C and B' of C'), respectively, by adding the alternative $\{sel\_hmatch_1, sel\_falcon_1\}$ and the probabilities $\mu''(sel\_hmatch_1) = 0.55$ and $\mu''(sel\_falcon_1) = 0.45$ for resolving the inconsistency between the first two rules.

It is not difficult to verify that, due to the independent combination of alternatives, the last two rules encode that the rule $Proceedings_2(X) \leftarrow Proceedings_1(X)$ holds with the probability $1 - (1 - \mu''(hmatch_2)) \cdot (1 - \mu''(falcon_2)) = 0.9892$, as desired.

6 Summary and Outlook

We have presented a rule-based framework for representing ontology mappings that supports the resolution of inconsistencies on a symbolic and a numeric level. While the use of disjunction and nonmonotonic negation allows the rewriting of inconsistent rules, the probabilistic extension of the language allows us to explicitly represent numeric confidence values as error probabilities, to resolve inconsistencies by using trust probabilities, and to reason about these on a numeric level. While being expressive and well-integrated with description logic ontologies, the language is still decidable and has data-tractable subsets that make it particularly interesting for practical applications.

We leave for future work the implementation of the language and the performing of experiments on the basis of large data sets, to further substantiate our claims that this formal framework is suited for realistic applications of ontology mappings.

Acknowledgments. Andrea Calì is supported by the EU STREP FET project TONES (FP6-7603). Thomas Lukasiewicz is supported by the German Research Foundation

(DFG) under the Heisenberg Programme and by the Austrian Science Fund (FWF) under the project P18146-N04. Heiner Stuckenschmidt and Livia Predoiu are supported by an Emmy-Noether Grant of the German Research Foundation (DFG).

References

1. A. Calì and T. Lukasiewicz. Tightly integrated probabilistic description logic programs for the Semantic Web. In *Proc. ICLP-2007*, pp. 428-429. *LNCS* 4670, Springer, 2007. (See also Report RR-1843-07-05, Institut für Informationssysteme, TU Wien, 2007.)
2. P. C. G. da Costa. *Bayesian Semantics for the Semantic Web*. Doctoral Dissertation, George Mason University, Fairfax, VA, USA, 2005.
3. P. C. G. da Costa and K. B. Laskey. PR-OWL: A framework for probabilistic ontologies. In *Proc. FOIS-2006*, pp. 237–249. IOS Press, 2006.
4. D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. *DL-Lite*: Tractable description logics for ontologies. In *Proc. AAAI-2005*, pp. 602–607. AAAI Press, 2005.
5. T. Eiter, T. Lukasiewicz, R. Schindlauer, H. Tompits. Combining answer set programming with description logics for the Semantic Web. In *Proc. KR-2004*, pp. 141–151. AAAI Press, 2004. (See also Report RR-1843-07-04, Institut für Informationssysteme, TU Wien, 2007.)
6. J. Euzenat, M. Mochol, P. Shvaiko, H. Stuckenschmidt, O. Svab, V. Svatek, W. R. van Hage, and M. Yatskevich. First results of the ontology alignment evaluation initiative 2006. In *Proc. ISWC-2006 Workshop on Ontology Matching*.
7. J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer, Heidelberg, Germany, 2007.
8. J. Euzenat, H. Stuckenschmidt, and M. Yatskevich. Introduction to the ontology alignment evaluation 2005. In *Proc. K-CAP-2005 Workshop on Integrating Ontologies*.
9. W. Faber, N. Leone, and G. Pfeifer. Recursive aggregates in disjunctive logic programs: Semantics and complexity. In *Proc. JELIA-2004*, pp. 200–212. *LNCS* 3229, Springer, 2004.
10. A. Finzi and T. Lukasiewicz. Structure-based causes and explanations in the independent choice logic. In *Proc. UAI-2003*, pp. 225–232. Morgan Kaufmann, 2003.
11. M. Gelfond and V. Lifschitz. Classical negation in logic programs and disjunctive databases. *New Generation Comput.*, 9(3/4):365–386, 1991.
12. R. Giugno and T. Lukasiewicz. P-$\mathcal{SHOQ}(\mathbf{D})$: A probabilistic extension of $\mathcal{SHOQ}(\mathbf{D})$ for probabilistic ontologies in the Semantic Web. In *Proc. JELIA-2002*, pp. 86–97. *LNCS* 2424, Springer, 2002.
13. I. Horrocks and P. F. Patel-Schneider. Reducing OWL entailment to description logic satisfiability. In *Proc. ISWC-2003*, pp. 17–29. *LNCS* 2870, Springer, 2003.
14. I. Horrocks, U. Sattler, and S. Tobies. Practical reasoning for expressive description logics. In *Proc. LPAR-1999*, pp. 161–180. *LNCS* 1705, Springer, 1999.
15. T. Lukasiewicz. Probabilistic description logic programs. *Int. J. Approx. Reason.*, 45(2):288–307, 2007.
16. T. Lukasiewicz. A novel combination of answer set programming with description logics for the Semantic Web. In *Proc. ESWC-2007*, pp. 384-398. *LNCS* 4519, Springer, 2007.
17. C. Meilicke, H. Stuckenschmidt, and A. Tamilin. Repairing ontology mappings. In *Proc. AAAI-2007*, pp. 1408–1413. AAAI Press, 2007.
18. B. Motik, I. Horrocks, R. Rosati, and U. Sattler. Can OWL and logic programming live together happily ever after? In *Proc. ISWC-2006*, pp. 501–514. *LNCS* 4273, Springer, 2006.
19. P. Wang and B. Xu. Debugging ontology mapping: A static method. *Computation and Intelligence*, 2007. To appear.
20. D. Poole. The independent choice logic for modelling multiple agents under uncertainty. *Artif. Intell.*, 94(1/2):7–56, 1997.
21. L. Serafini, H. Stuckenschmidt, and H. Wache. A formal investigation of mapping languages for terminological knowledge. In *Proc. IJCAI-2005*, pp. 576–581, 2005.

A Mass Assignment Approach to Granular Association Rules for Multiple Taxonomies

Trevor Martin[1,2], Yun Shen[1] and Ben Azvine[2]

[1] AI Group, University of Bristol, BS8 1TR UK
[2] Intelligent Systems Lab, BT, Adastral Park, Ipswich IP5 3RE, UK
{Trevor.Martin, Yun.Shen}@bristol.ac.uk, Ben.Azvine@bt.com

Abstract. The use of hierarchical taxonomies to organise information (or sets of objects) is a common approach for the semantic web and elsewhere, and is based on progressively finer granulations of objects. In many cases, seemingly crisp granulation disguises the fact that categories are based on loosely defined concepts which are better modelled by allowing graded membership. A related problem arises when different taxonomies are used, with different structures, as the integration process may also lead to fuzzy categories. Care is needed when information systems use fuzzy sets to model graded membership in categories - the fuzzy sets are not disjunctive possibility distributions, but must be interpreted conjunctively. We clarify this distinction and show how an extended mass assignment framework can be used to extract relations between fuzzy categories. These relations are association rules and are useful when integrating multiple information sources categorised according to different hierarchies. Our association rules do not suffer from problems associated with use of fuzzy cardinalities. An example of discovering associated film genres is given.

Keywords: fuzzy, granules, association rules, hierarchies, mass assignments, semantic web, iPHI

1 1 Introduction

The use of taxonomic hierarchies to organise information and sets of objects into manageable chunks (granules) is widespread. Granules were informally defined by Zadeh [1] as a way of decomposing a whole into parts, generally in a hierarchical way. We can regard a hierarchical categorisation as a series of progressively finer granulations, allowing us to represent problems at the appropriate level of granularity.

The idea of a taxonomy serves as an organisational principle for libraries, for document repositories, for corporate structure, for the grouping of species and very many other applications. It is therefore no surprise to note that the semantic web adopts hierarchical taxonomies as a fundamental structure, using the *subClassOf* construct. Although in principle the idea of a taxonomic hierarchy is crisply defined, in practice there is often a degree of arbitrariness in its definition. For example, we might divide the countries of the world by continent at the top level of a taxonomic hierarchy. However, continents do not have crisp definitions - Europe contains some

definite members (e.g. France, Germany) but at the Eastern and South-Eastern border, the question of which countries belong / do not belong is less clear. Iceland is generally included in Europe despite being physically closer to Greenland (part of North America). Thus although the word "Europe" denotes a set of countries (i.e. it is a granule) and can be used as the basis for communication between humans, it does not have an unambiguous definition in terms of the elements that belong to the set. Different "authorities" adopt different definitions - the set of countries eligible to enter European football competitions differs from the set of countries eligible to enter the Eurovision song contest, for example.

Of course, mathematical and some legal taxonomic structures are generally very precisely defined - the class of polyhedra further subdivides into triangles, quadrilaterals, etc and triangles may be subdivided into equilateral, isosceles etc. Such definitions admit no uncertainty. Most information systems model the world in some way, and need to represent categories which correspond to the loosely defined classes used by humans in natural language. For example, a company may wish to divide adults into customers and non-customers, and then sub-divide these into high-value customers, dissatisfied customers, potential customers, etc. Such categories are not necessarily distinct (i.e. they may be a covering rather than a partition) but more importantly, membership in these categories is graded - customer X may be highly dissatisfied and about to find a new supplier whilst customer Y is only mildly dissatisfied. We argue that most hierarchical taxonomies involve graded or loosely defined categories, but the nature of computerised information systems means that a more-or-less arbitrary decision has to be made on borderline cases, giving the taxonomy the appearance of a crisp, well-defined hierarchy. This may not be a problem as long as a rigorous and consistent criterion for membership is used (e.g. a dissatisfied customer is defined as one who has made at least two calls complaining about service), but the lack of subjectivity in a definition is rare. The use of graded membership (fuzziness) in categories enhances their expressive power and usefulness.

A related problem arises when trying to combine multiple sources of information that have been categorised in some way (often hierarchically). For example, the category of "vintage wine" has a different (but objective) definition, depending on the country of origin. To a purist, vintage wines are made from grapes harvested in a single year – however, the European Union allows up to 5% of the grapes to be harvested in a different year, the USA allows 15% in some cases and 5% in others, while other countries such as Chile and South Africa may allow up to 25%. Thus even taking a simple (crisp) granulation of wines into vintage and non-vintage categories can lead to problems if we try to integrate different sources.

In this paper we describe a new method for calculating association rules to find correspondences between fuzzy granules in different hierarchies (with the same underlying universe). We discuss the semantics of fuzzy sets when used to describe granules, and introduce a mass assignment-based method to rank association rules and show that the new method gives more satisfactory results than approaches based on fuzzy cardinalities. Ongoing work is focused on comparison of this approach to others (e.g. on ontology merging benchmarks), and with application to merging classified directory content.

2 Background

This work take place in the context of the iPHI system (intelligent Personal Hierarchies for Information) [2] which aims to combine and integrate multiple sources of information and to configure access to the information based on an individual's personal categories. We assume here that the underlying entities (instances) that are being categorised are known unambiguously - when integrating multiple sources, this is often not the case. We have outlined SOFT (the Structured Object Fusion Toolkit) elsewhere [3] as one solution to this problem.

2.1 Fuzzy Sets in Information Systems

Many authors (e.g. [4]) have proposed the use of fuzzy sets to model uncertain values in databases and other knowledge based applications . The standard interpretation of a fuzzy set in this context is as a *possibility distribution* - that is to say it represents a single valued attribute which is not known exactly. For example we might use the fuzzy set *tall* to represent the height of a specific person or *low* to represent the value shown on a dice. The fuzzy sets *tall* and *low* admit a range of values, to a greater or lesser degree; the actual value is taken from the range. Knowing that a dice value *val* is *even* restricts the possible values to *val=2 XOR val=4 XOR val=6* (where *XOR* is an exclusive or). If a fuzzy set on the same universe is defined as *low = {1/1, 2/1, 3/0.4}* then knowing the value *val* is *low* restricts the possible values to *val=1 XOR val=2 XOR val=3* with corresponding memberships.

The conjunctive interpretation of a fuzzy set occurs when the attribute can have multiple values. For example, a person may be able to speak several languages; we could model this as a fuzzy set of languages, where membership would depend on the degree of fluency. This is formally a relation rather than a function on the underlying sets. Our position is to make a distinction between the conjunctive interpretation - modelled by a fuzzy relation – and the disjunctive interpretation – modelled by a possibility distribution. To emphasise the distinction, we use the notation

$$F(a) = \{x/\mu(x) \mid x \in U\}$$

to denote a single valued attribute F of some object a (i.e. a possibility distribution over a universe U) and

$$R(a) = [x/\chi(x) \mid x \in U]$$

to denote a multi-valued attribute (relation). Granules represent the latter case, since we have multiple values that satisfy the predicate to a greater or lesser degree.

2.2 Association Rules

In creating association rules within transaction databases (e.g. [5], see also [7] for a clear overview), the standard approach is to consider a table in which columns correspond to items and each row is a transaction. A column contains 1 if the item was bought, and 0 otherwise. The aim of association rule mining is to determine whether or not there are links between two disjoint subsets of items – for example, do customers generally buy biscuits and cheese when beer, lager and wine are bought?

Let X denote the set of items, so that any transaction can be represented as $tr \subseteq X$ and we have a multiset Tr of transactions. We must also specify two non-overlapping subsets of X, s and t. An association rule is of the form $S => T$ where S (resp T) is the set of transactions containing the items s (resp t). The rule is interpreted as stating that when the items in s appear in a transaction, it is likely that the items in t will also appear i.e. it is not an implication in the formal logical sense.

Most authors use two measures to assess the significance of association rules, although these measures can be misleading in some circumstances. The support of a rule is the fraction of transactions in which both S and T appear, and the confidence of a rule is an estimate (based on the samples) of the conditional probability of T given S

$$Support(S,T) = |S \cap T|$$

and

$$Conf(S,T) = \frac{|S \cap T|}{|S|}$$

where we operate on multisets rather than sets. Typically a threshold is chosen for the support, so that only frequently occurring sets of items s and t are considered; a second threshold filters out rules of low confidence.

Various approaches to fuzzifying association rules have been proposed e.g. [6-8]. The standard extension to the fuzzy case is to treat the (multi-) sets S, T as fuzzy and find the intersection and cardinality using a t-norm and sigma-count respectively.

$$Conf(S,T) = \frac{\sum_{x \in X} \mu_{S \cap T}(x)}{\sum_{x \in X} \mu_S(x)}$$

Note that many authors just refer to fuzzy sets, rather than multisets.

As pointed out by [7], using min and the sigma count for cardinality can be unsatisfactory because it does not distinguish between several tuples with low memberships and few tuples with high memberships - for example,

$$S = [x_1/1, x_2/0.01, x_3/0.01, \ldots, x_{1000}/0.01]$$
$$T = [x_1/0.01, x_2/1, x_3/0.01, \ldots, x_{1000}/0.01]$$

leads to

$$Conf(S,T) = \frac{1000 \times 0.01}{1 + 999 \times 0.01} \approx 0.91$$

which is extremely high for two almost disjoint sets (this example originally appeared in [9]). Using a fuzzy cardinality (i.e. a fuzzy set over the possible cardinality values) is also potentially problematic.

For these reasons, we propose the use of mass assignment theory in calculating the support and confidence of association rules between fuzzy categories.

The fuzziness in our approach arises because we allow partial membership in categories – for example, instead of looking for an association between biscuits and beer, we might look for an association between *alcoholic drinks* and *snack foods*. It is important to note that we are dealing with conjunctive fuzzy sets (monadic fuzzy relations) here. Mass assignment theory is normally applied to fuzzy sets representing possibility distributions and the operation of finding the conditional probability of one fuzzy sets given another is known as semantic unification [10]. This rests on the underlying assumption of a single valued attribute – a different approach is required to find the conditional probability when we are dealing with set-valued attributes.

2.3 Mass Assignments

A mass assignment [11] (see also [12]) is a distribution over a power set, representing disjunctive uncertainty about a value. For a universe U

$$m : P(U) \rightarrow [0,1]$$
$$\sum_{X \subseteq U} m(X) = 1 \qquad (1)$$

The mass assignment is related to a fuzzy set (possibility distribution) A as follows:
Let μ_A be the membership function of A with range

$R(\mu_A) = \{\mu_A^1, \mu_A^2, \ldots, \mu_A^m\}$

such that $\mu_A^1 > \mu_A^2 > \ldots > \mu_A^m$

and Ai be the alpha-cuts at these values i.e.

$A_i = \{x | \mu_A(x) \geq \mu_A^i\}$

(also known as the focal elements)
Then

$$m_A(A_i) = \mu_A^i - \mu_A^{i+1} \qquad (2)$$

Given a fuzzy set A, the corresponding mass assignment can be written as

$M(A) = \{A_i : m_A(A_i) \mid A_i \subseteq A\}$

where conventionally only the focal elements (non-zero masses) are listed in the mass assignment. The mass assignment represents a family of probability distributions on U, with the restrictions

$$p : U \rightarrow [0,1]$$
$$\sum_{x \in U} p(x) = 1 \qquad (3)$$
$$m(\{x\}) \leq p(x) \leq \sum_{x \in X} m(X)$$

For example, if $X = \{a, b, c, d\}$ and A is the fuzzy set

$\{a/1, b/0.8, c/0.3, d/0.2\}$

then

$M(A) = \{\{a\} : 0.2, \{a, b\} : 0.5, \{a, b, c\} : 0.1, \{a, b, c, d\} : 0.2\}$

In the example above, $p(a) = 0.4$, $p(b) = 0.3$, $p(c) = 0.1$, $p(d) = 0.2$ is a possible distribution, obtained by allocating the mass of 0.5 on the set $\{a, b\}$ to a (0.2) and b (0.3), and so on. We can also give a mass assignment definition of the cardinality of a fuzzy set as a distribution over integers

$$p(|A| = n) = \sum_{\substack{A_i \subseteq A \\ |A_i| = n}} m_A(A_i)$$

for $0 \leq n \leq |U|$

In the example above, $p(|A| = 1) = 0.2$, $p(|A| = 2) = 0.5$, etc. Clearly in this framework, the cardinality of a fuzzy set can be left as a distribution over integer values, or an expected value can be produced from this distribution in the usual way. A similar definition of fuzzy cardinality was proposed by [13], also motivated by the problem of fuzzy association rules.

Baldwin introduced the least prejudiced distribution (lpd) which is a specific distribution satisfying (3) above but also obeying

$$lpd_A(x) = \sum_{x \in A_i} \frac{m(A_i)}{|A_i|} \qquad (4)$$

where $|A|$ indicates the cardinality of the set A and the summation is over all focal elements containing x.

Informally, wherever mass is associated with a non-singleton focal element, it is shared equally between the members of the set. Clearly a least prejudiced distribution is a restriction of the original assignment.

The steps from lpd to mass assignment and then to fuzzy set can be reversed, so that we can derive a unique fuzzy set for any frequency distribution on a finite universe, by assuming the relative frequencies are the least prejudiced distribution (proof in [14]).

If the relative frequencies are written

$$L_A = \{L_A(x_1), L_A(x_2), \ldots, L_A(x_n)\}$$

such that

$$L_A(x_1) > L_A(x_2) > \ldots > L_A(x_n)$$

then we can define

$$A_i = \{x | x \in U \wedge L_A(x) \geq L_A(x_i)\}$$

and the fuzzy set memberships are given by

$$\mu_A(x_i) = |A_i| \times L_A(x_i) + \sum_{j=i+1}^{n} \left(|A_j| - |A_{j-1}|\right) \times L_A(x_j)$$

2.4 Fuzzy relations and mass assignments

A relation is a conjunctive set of ordered n-tuples i.e. it represents a conjunction of n ground clauses. For example, if U is the set of dice scores then we could define a predicate *differBy4or5* on $U \times U$ as the set of pairs

[(1, 6), (1, 5), (2, 6), (5, 1), (6, 1), (6, 2)]

This is a conjunctive set in that each pair satisfies the predicate. In a similar way, a fuzzy relation represents a set of n-tuples that satisfy a predicate to a specified degree. Thus *differByLargeAmount* could be represented by

[(1, 6)/1, (1, 5)/0.6, (2, 6)/0.6, (5, 1)/0.6, (6, 1)/1, (6, 2)/0.6]

2.5 Mass-based association rules

We consider two granules, represented as monadic fuzzy relations S and T on the same domain, and wish to calculate the degree of association between them. For example, consider a database of sales employees, salaries and sales figures. We can categorise employees according to whether their salaries are *high*, *medium* or *low* and also according to whether their sales figures are *good*, *moderate* or *poor*. A mining task might be to find out whether the *good* sales figures are achieved by the *highly paid* employees. For example, given the table

name	sales	salary
a	100	1000
b	80	400
c	50	800
d	20	700

we might define the monadic fuzzy relations

$S = goodSales = [a/1, b/0.8, c/0.5, d/0.2]$

and

$T = highSalary = [a/1, b/0.4, c/0.8, d/0.7]$

These represent sets of values (1-tuples) that all satisfy the related predicate to a degree. The confidence in an association rule can be calculated as follows:

For a source granule

$$S = \left[x_1 / \chi_S(x_1),\ x_2 / \chi_S(x_2), \ldots,\ x_{|S|} / \chi_S\left(x_{|S|}\right) \right]$$

and a target granule

$$T = \left[x_1 / \chi_T(x_1),\ x_2 / \chi_T(x_2), \ldots,\ x_{|T|} / \chi_T\left(x_{|T|}\right) \right]$$

we can define the corresponding mass assignments as follows. Let the set of distinct memberships in S be

$$\left\{ \chi_S^{(1)}, \chi_S^{(2)}, \ldots, \chi_S^{(n_S)} \right\}$$

where

$$\chi_S^{(1)} > \chi_S^{(2)} > \ldots > \chi_S^{(n_S)}$$

and $n_S \le |S|$

Let

$$S_1 = \left\{ \left[x \,\middle|\, \chi_S(x) = \chi_S^{(1)} \right] \right\}$$

$$S_i = \left\{ \left[x \,\middle|\, \chi_S(x) \ge \chi_S^{(i)} \right] \right\} \cup S_{i-1} \qquad 1 < i \le n_S$$

Then the mass assignment corresponding to S is

$\left\{ S_i : m_S(S_i) \right\}, \quad 1 \le i \le n_S$

where $m_S(S_k) = \chi_S^{(k)} - \chi_S^{(k+1)}$

and we define

$$\chi_S^{(i)} = 0 \quad if \quad i > n_S$$

For example, the fuzzy relation

$S = [a/1, b/0.8, c/0.5, d/0.2]$

has the corresponding mass assignment

$$M_S = \left\{ \{[a]\} : 0.2, \{[a],[a,b]\} : 0.3, \{[a],[a,b],[a,b,c]\} : 0.3, \{[a],[a,b],[a,b,c],[a,b,c,d]\} : 0.2 \right\}$$

The mass assignment corresponds to a distribution on the power set of relations, and we can define the least prejudiced distribution in the same way as for the standard mass assignment. In the example above

$$L_S = \left\{ [a] : 0.5, [a,b] : 0.3, [a,b,c] : 0.15, [a,b,c,d] : 0.05 \right\}$$

We can now calculate the confidence in the association between the granules S and T using mass assignment theory. In general, this will be an interval as we are free to move mass (consistently) between elements of each S_i .and T_j

For two mass assignments

$M_S = \left\{ \left\{ S_{p_i} \right\} : m_S(S_i) \right\}, \quad 1 \le p_i \le i \le n_S$

$M_T = \left\{ \left\{ T_{q_j} \right\} : m_T(S_j) \right\}, \quad 1 \le q_j \le j \le n_T$

the composite mass assignment is

$M_C = M_S \oplus M_T$

$= \left\{ X : m_C(X) \right\}$

where m_C is specified by the composite mass allocation function $C\left(i, j, S_{p_i}, T_{q_j}\right)$ subject to

$$\sum_{j=1}^{n_T} \sum_{\substack{1 \le q_j \le j \\ 1 \le p_i \le i}} C\left(i, j, S_{p_i}, T_{q_j}\right) = m_S(S_i)$$

$$\sum_{i=1}^{n_S} \sum_{\substack{1 \le p_i \le i \\ 1 \le q_j \le j}} C\left(i, j, S_{p_i}, T_{q_j}\right) = m_T(T_j)$$

This can be visualised using a mass tableau (see [11]) Each row (column) represents a focal element of the mass assignment, and is split into sub-rows (sub-columns). The mass associated with a row (column) is shown at the far left (top) and can be distributed amongst the sub-rows (sub-columns). For example consider the granules

$S = [a/1, b/0.8, c/0.5, d/0.2]$ and

$T = [a/1, b/0.4, c/0.8, d/0.7]$

The rule confidence is given by equation (5)

$$Conf(S \to T) = \left(\frac{\sum_{i=1}^{n_S} \sum_{j=1}^{n_T} \sum_{\substack{1 \le q_j \le j \\ 1 \le p_i \le i}} C\left(i, j, S_{p_i}, T_{q_j}\right) \times \left| S_{p_i} \cap T_{q_j} \right|}{\sum_{i=1}^{n_S} \sum_{j=1}^{n_T} \sum_{\substack{1 \le q_j \le j \\ 1 \le p_i \le i}} C\left(i, j, S_{p_i}, T_{q_j}\right) \times \left| S_{p_i} \right|} \right) \qquad (5)$$

		0.2	0.1		0.3			0.4			
		a	*a*	*ac*	*a*	*ac*	*acd*	*a*	*ac*	*acd*	*abcd*
0.2	*a*	0.2									
0.3	*a*		0.1								
	ab										0.2
0.3	*a*				0.3						
	ab										
	abc										
0.2	*a*							0.2			
	ab										
	abc										
	abcd										

(a) $$Conf(S \rightarrow T) = \frac{0.2\times1 + 0.1\times1 + 0.2\times2 + 0.3\times1 + 0.2\times1}{0.2\times1 + 0.1\times1 + 0.2\times2 + 0.3\times1 + 0.2\times1} = 1$$

		0.2	0.1		0.3			0.4			
		a	*a*	*ac*	*a*	*ac*	*acd*	*a*	*ac*	*acd*	*abcd*
0.2	*a*	0.2									
0.3	*a*										
	ab		0.1					0.2			
0.3	*a*										
	ab										
	abc				0.3						
0.2	*a*										
	ab										
	abc										
	abcd							0.2			

(b) $$Conf(S \rightarrow T) = \frac{0.2\times1 + 0.1\times1 + 0.2\times1 + 0.3\times1 + 0.2\times1}{0.2\times1 + 0.1\times2 + 0.2\times2 + 0.3\times3 + 0.2\times4} = 0.4$$

Fig 1 - Composite mass allocation (a) maximising and (b) minimising association rule confidence

Clearly the mass can be allocated in many ways, subject to the column constraints and it is not always straightforward to find the minimum and maximum confidences arising from different composite mass allocations. Two extreme examples are shown in Fig 1, so that the confidence in the association rule between the two granules lies in the interval [0.4, 1]. In general there can be considerable computation involved in finding the maximum and minimum confidences for a rule. When ranking association rules it is preferable to have a single figure for confidence, rather than an interval which can lead to ambiguity in the ordering.

We can redistribute the mass according to the least prejudiced distribution i.e. split the mass in each row (column) equally between its sub-rows (sub-columns) and taking the product as the mass in each cell. In this case, the calculation is simplified by (a) combining rows (columns) with the same label and (b) re-ordering the summations. This enables us to calculate association confidences with roughly *O(n)* complexity, rather than $O(n^4)$ where n is the number of focal elements in the source granule S. The confidence is then given by

$$Conf_{LPD}(S,T) = \frac{\sum_{i=1}^{n_S}\sum_{j=1}^{n_T} LPD_S(S_i) \times LPD_T(T_j) \times |S_i \cap T_j|}{\sum_{i=1}^{n_S} LPD_S(S_i) \times |S_i|} \qquad (6)$$

(due to the nested structure of the sets, the numerator does not require a double summation but can be calculated by stepping through the cells on the leading diagonal). If we choose the least prejudiced distribution and re-arrange sub-rows into single rows with the same label (also columns) we obtain the following intersections

		0.45 a	0.25 ac	0.2 acd	0.1 abcd
0.5	a	a	a	a	a
0.3	ab	a	a	a	ab
0.15	abc	a	ac	ac	abc
0.05	abcd	a	ac	acd	abcd

and the numerator for the rule confidence is

$0.5 \times (0.45+0.25+0.2+0.1) \times 1$
$+ 0.3 \times (0.45+0.25+0.2) \times 1 + 0.3 \times 0.1 \times 2$
$+ 0.15 \times 0.45 \times 1 + 0.15 \times (0.25+0.2) \times 2 + 0.15 \times 0.1 \times 3$
$+ 0.05 \times 0.45 \times 1 + 0.05 \times 0.25 \times 2 + 0.05 \times 0.2 \times 3 + 0.05 \times 0.1 \times 4$

giving a confidence of 0.67 - lying in the interval shown in Fig 1 (obviously). Using the LPD allows us to replace the calculation in eq 5 with straightforward calculations of the expected values of the cardinality of the source set and the intersection.

The example above gives a similar result to the cardinality-based method, but this is not always the case. For example if

$$S = [x_1/1, x_2/0.01, x_3/0.01, \ldots, x_{1000}/0.01]$$
$$T = [x_1/0.01, x_2/1, x_3/0.01, \ldots, x_{1000}/0.01]$$

then a fuzzy cardinality based approach gives a confidence of $10/10.99 \approx 0.91$ whereas our approach gives approximately 10^{-5}. Clearly this is a far more reasonable answer, as there are no elements with strong membership in both granules.

3 Experiment

We have carried out preliminary tests on the approach by finding associations between movie genres from different online sources. Ongoing work is focusing on finding associations between music genres, categories in different classified business directories and also on comparative studies using the ontology matching benchmarks, where suitable instance data is available.

The two online movie databases IMDB and Rotten Tomatoes have been used in previous work [15] to test instance matching methods. We have used the SOFT method to establish correspondence between the (roughly) 95000 movies in the databases. Within these two sources, movies are assigned to one or more genres and our task is to find strong associations between genres. The genres form a fairly flat hierarchy, although in principle one would expect genres to form a deeper hierarchical structure (e.g. comedy could be sub-divided into slapstick, satire, situation comedy, etc). At this stage, there is no benchmark for comparison but the results are intuitively reasonable as shown in Fig 2.

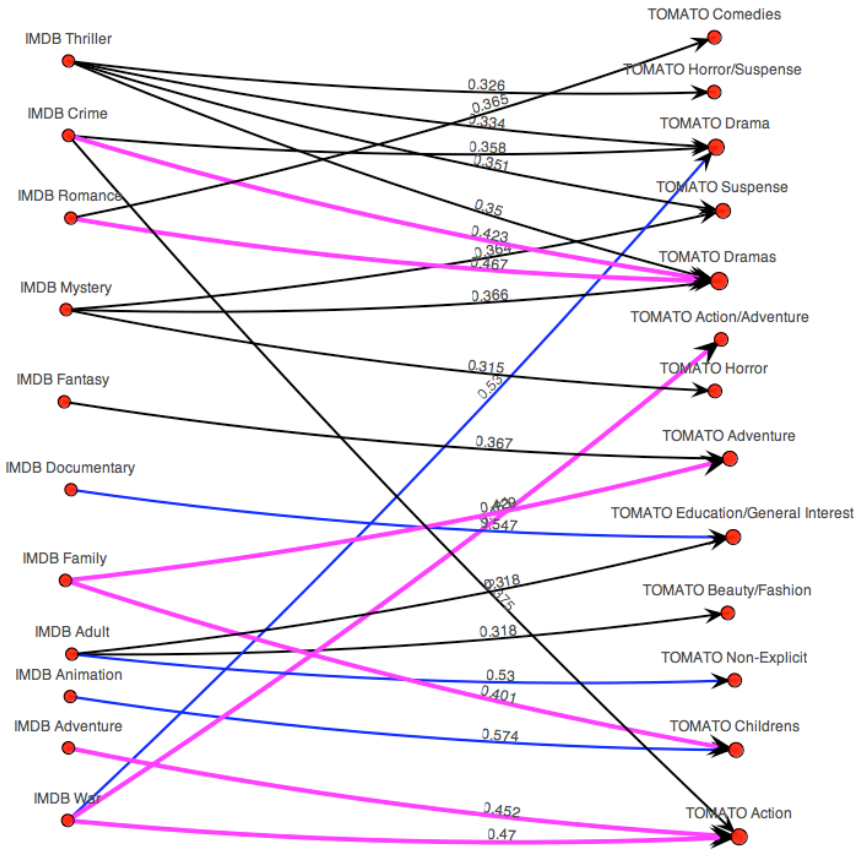


Fig 2 - strong associations from source IMDB genres (left) to target Rotten Tomato genres (right). Edge labels denotes the association strength.

4 Summary

We have described a new method for generating association rules between granules in different information hierarchies. These rules enable us to find related categories without leading to spurious relations suggested by association rules based on fuzzy cardinalities. Results were presented for discovery of links between film genres in different classification hierarchies, giving intuitively reasonable associations. The new method is currently undergoing further tests, looking at benchmark instance-matching problems, finding associations between music genres and finding links between categories in different classified business directories.

Acknowledgement : this work was partly funded by BT and the Defence Technology Centre Data and Information Fusion. We would like to thank the referees for their careful reviews and helpful suggestions.

5 References

[1] Zadeh, L. A., "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets and Systems*, vol. 90, pp. 111-127, 1997.

[2] Martin, T. P. and B. Azvine, "Acquisition of Soft Taxonomies for Intelligent Personal Hierarchies and the Soft Semantic Web," *BT Technology Journal*, vol. 21, pp. 113-122, 2003.

[3] Martin, T. P. and B. Azvine, "Soft Integration of Information with Semantic Gaps," in *Fuzzy Logic and the Semantic Web*, E. Sanchez, Ed.: Elsevier, 2005.

[4] Bosc, P. and B. Bouchon-Meunier, "Databases and Fuzziness - Introduction," *International Journal of Intelligent Systems*, vol. 9, pp. 419, 1994.

[5] Agrawal, R. and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," presented at Very large data bases, Santiago, 1994.

[6] Bosc, P. and O. Pivert, "On Some Fuzzy Extensions of Association Rules," presented at IFSA world congress, Vancouver, Canada, 2001.

[7] Dubois, D., E. Hullermeier, and H. Prade, "A systematic approach to the assessment of fuzzy association rules," *Data Mining and Knowledge Discovery*, vol. 13, pp. 167-192, 2006.

[8] Kacprzyk, J. and S. Zadrozny, "Linguistic Summarization of Data Sets Using Association Rules," presented at Fuzzy systems; Exploring new frontiers, St Louis, MO, 2003.

[9] Martin-Bautista, M. J., M. A. Vila, H. L. Larsen, and D. Sanchez, "Measuring Effectiveness in Fuzzy Information Retrieval," presented at Flexible Query Answering Systems (FQAS), 2000.

[10] Baldwin, J. F., J. Lawry, and T. P. Martin, "Efficient Algorithms for Semantic Unification," presented at Information Processing and the Management of Uncertainty, Spain, 1996.

[11] Baldwin, J. F., "The Management of Fuzzy and Probabilistic Uncertainties for Knowledge Based Systems," in *Encyclopedia of AI*, S. A. Shapiro, Ed., 2nd ed: John Wiley, 1992, pp. 528-537.

[12] Dubois, D. and H. Prade, "On Several Representations of an Uncertain Body of Evidence," in *Fuzzy Information and Decision Processes*, M. M. Gupta and E. Sanchez, Eds.: North Holland, 1982.

[13] Delgado, M., D. Sanchez, M. J. Martin-Bautista, and M. A. Vila, "A probabilistic definition of a nonconvex fuzzy cardinality," *Fuzzy Sets and Systems*, vol. 126, pp. 177-190, 2002.

[14] Baldwin, J. F., J. Lawry, and T. P. Martin, "A Mass Assignment Theory of the Probability of Fuzzy Events," *Fuzzy Sets and Systems*, vol. 83, pp. 353-367, 1996.

[15] Martin, T. P. and Y. Shen, "Improving access to multimedia using multi-source hierarchical meta-data," in *Adaptive Multimedia Retrieval: User, Context, and Feedback*, vol. LNCS vol 3877, *LNCS*: Springer, 2006, pp. 266 - 278.

Rough Description Logics for Modeling Uncertainty in Instance Unification

Michel C.A. Klein[1], Peter Mika[2], and Stefan Schlobach[1]

[1] Vrije Universiteit Amsterdam, {michel.klein|schlobac}@few.vu.nl
[2] Yahoo! Research Barcelona, pmika@yahoo-inc.com

Abstract. Instance-unification is a prime example for uncertainty on the Semantic Web, as it is not always possible to automatically determine with absolute certainty whether two references denote the same object or not. In this paper, we present openacademia, a semantics-based system for the management of distributed bibliographic information collected from the Web, in which the Instance Unification problem is ubiquitous. Our tentative solution is Rough DL, a simple extension of classical Description Logics, which allows for approximations of vague concept. This shows that already a simple formalism for dealing with uncertain information in a qualitative way can provide an elegant solution to practical problems on the Semantic Web.

1 Introduction

When information is gathered from the Web it often occurs that multiple descriptions of the same resource are found. In that case the duplicate resources should be identified and their descriptions have to be combined.

Failing to effectively deal with duplicate results negatively affects the workings of all search engines. In a search system for publications instance unification is important for at least two types of information: persons and publications. If coreferences of persons [3, 4] are not resolved one will obtain an incomplete list of publications when querying for publications of a specific person (e.g. because the system does not recognize that the authors 'John Smith' and 'John J.B. Smith' are the same person. One may face the opposite situation of receiving irrelevant results, such as when a system assumes that authors with the same name are the same person, as is common in most existing NLP based publication search engines such as Google Scholar and CiteSeer. On the other hand, if publications are not unified [8, 6] the result will contain duplicates, which makes browsing the results difficult and obscures publication counts, an important statistic in academia. The problem of finding equivalent instances in this case is usually referred to as "coreference resolution" or "instance unification".

The most common way of representing the results of instance unification is to declare the objects to be logically equivalent. This has the consequence that all properties of one resource are also properties of the other resource. However, this implementation has several drawbacks. First, it often represents a **logical**

overcommitment. Only in very limited cases are we absolutely *certain* that two instances are equivalent, in most cases we only have some partial evidence that two descriptions refer to the same object (e.g. name similarity). Further, it is not possible to **distinguish between different levels of confidence** in similarity relations. Moreover, transitivity of equivalence often causes an **undesired propagation of equivalence** over similarity relations.

In this paper, we will introduce an alternative to complete instance unification, which allows for reasoning over gradually weakening notions of similarity. Our tentative solution is an extension to standard DL that can be used for defining approximations of concepts without increasing the complexity of the language.

We illustrate the use of this language in openacademia,[1] an open source web-based system for collecting, aggregating and querying publication metadata in a group or community setting. openacademia offers an interactive, AJAX-based search interface for querying publications by a combination of facets. Query results can be visualized in a number of ways, including the possibility to generate various dynamic HTML representations that can be easily inserted into personal homepages or institutional publication pages. Integrating descriptions of similar persons and publications is an important task of this system, which is in this context sometimes called *smushing* [5].

In the subsequent sections, we introduce the language for defining approximations, and apply it to model different levels of similarity of persons and publications in openacademia. The flexible instance unification using Rough DL illustrates how already rather simple mechanisms for dealing with uncertainty in a qualitative way can be used to elegantly solve practical problems on the Web.

Of course, we do not claim that our practical problem could not have been solved by other, for example more quantitative formalisms. However, we believe that the simplicity of our approach makes it an attractive alternative for dealing with uncertainty on the Semantic Web.

2 Rough DL

In [10] we presented a new paradigm to represent and reason about similarity of instances in a qualitative way called *rough Description Logics* ($\mathcal{RDL}$).[2] This language is an obvious candidate for modeling similarity and reasoning about classes of de-referenced objects. Here, we introduce an adapted version of $\mathcal{RDL}$, which is based on similarity rather than equivalence relations.

Definition 1. *A relation is called a* similarity (or tolerance) relation *if it is reflexive and symmetric. An* equivalence relation *is a transitive similarity relation.*

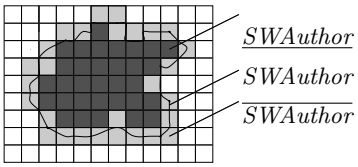
As equivalence relations extend similarity relations, we define Rough DL using the latter. We will use the notation $\overline{C}^{sim}$ and $\underline{C}_{sim}$ to describe the approximations of a class C with respect to a similarity relation sim. We will omit the

[1] `http://www.openacademia.org`

[2] As the relation between OWL and Description Logics is well established, we only introduce rough DL. The extension to rough OWL is conceptually trivial.

lower-script and upper-script sim whenever the choice of relation is irrelevant. The intuition regarding $\overline{C}^{sim}$ is that it denotes the set of all elements that are **possibly** in C, whereas $\underline{C}_{sim}$ is meant to describe all elements **definitively** in C. Often such an operator is useful when C cannot be specified in a crisp way. By way of the approximation(s) we can at least restrict C with an upper, and a lower, bound.

An illustrating example The following picture illustrates the general idea. In the spirit of Rough Set theory [7], two concepts approximate an under-specified, vague, concept as particular sub- and super-concepts. Suppose that we want to define *SWAuthor* as the class of all Semantic Web authors. This is a vague concept.



Each square denotes a set of domain elements, which cannot further be discerned by some available criterion at hand. The encircling line denotes the set of Semantic Web authors, i.e., the vague concept which we are incapable to formally define. If we capture this lack of criteria to discern between two objects as a indiscernibility relation `indis`, we can formalize the upper approximation as the authors that are indiscernible from at least one Semantic Web author.

$$\overline{\texttt{SWAuthor}} \equiv \{aut_1 \mid \exists\ aut_2\texttt{: indis(}aut_1\texttt{,}aut_2\texttt{) \& }aut_2 \in \texttt{SWAuthor}\}.$$

Similarly, we can define the lower approximation as the set of authors containing all, and only those authors, for which it is known that all indiscernible authors must be Semantic Web authors.

$$\underline{\texttt{SWAuthor}} \equiv \{aut_1 \mid \forall\ aut_2\texttt{: indis(}aut_1\texttt{,}aut_2\texttt{) } \rightarrow aut_2 \in \texttt{SWAuthor}\}$$

In our picture, the upper approximation is depicted as the union of the dark squares (the lower approximation), and the gray squares, the boundary. Note that in our example, following the literature on Rough Sets, the similarity of objects is determined by the indiscernibility of resources. This is an equivalence relation, which makes it appropriate to denote the sets of indiscernible instances as disjoint squares.

This intuition suggests two uses for Rough DL: first as a modeling language for representing vague knowledge and, secondly, as a language to query over similarity in a domain.

Modeling vague concepts Even if it is impossible to formally define a concept such as *SWAuthor*, we can often specify the approximations. The class of Semantic Web authors cannot be defined in a crisp way, but it is easy to think of an upper approximation (the possible Semantic Web authors, e.g. all authors having published in a Semantic Web conference or Journal). Rough DL semantics enforce restrictions on the class `SWAuthor` indirectly. We will discuss modeling with Rough DL concepts later in more detail.

Qualitative querying over similarities In the case of instance unification Rough DL can be used for querying classes of objects, and objects that were identified as being similar. Suppose we have a particular author `author` who is uniquely identifiable, say via his FOAF profile. Now, an algorithm Alg for object de-referencing creates a relation sim_{Alg} of pairs $(\texttt{author}_1, \texttt{author}_2)$. Based on this relation each algorithm for referencing induces a set $\texttt{Possibly}_{Alg}(\texttt{author})$ for each author `author`, i.e. a set of objects of the domain $\mathcal{U}$ which possibly correspond to this particular author `author`, with the formal definition:

$$\texttt{Possibly}_{Alg}(\texttt{author}) = \{i \in U \mid \exists j \in U : (i,j) \in sim_{Alg} \ \& \ j \in \texttt{oneOf}(\texttt{author})\},$$

which corresponds almost exactly to the formal semantics of an upper approximation. Most of the remainder of this paper will be about using Rough DL for querying ontologies with explicit similarities.

2.1 Semantics of Rough DL

Using this property we can define the semantics of the approximations formally:

Definition 2. *Let a rough interpretation be a triple $\mathcal{I} = (U, R^{\sim}, \cdot^{\mathcal{I}})$, where U is a universe, $\cdot^{\mathcal{I}}$ an interpretation function, and $R^{\sim}$ an equivalence relation over U. The function $\cdot^{\mathcal{I}}$ maps $\mathcal{RDL}$ concepts to subsets and role names to relations over the domain U. It extends to the new constructs as follows:*

- $(\overline{C})^{\mathcal{I}} = \{i \in U \mid \exists j \in U : (i,j) \in R^{\sim} \ \& \ j \in C^{\mathcal{I}}\}$
- $(\underline{C})^{\mathcal{I}} = \{i \in U \mid \forall j \in U : (i,j) \in R^{\sim} \rightarrow j \in C^{\mathcal{I}}\}$

The semantics of the lower approximation is defined as usual as the dual operator $\underline{\texttt{C}}_{sim} = \neg\overline{\neg C}^{sim}$ with its respective semantics. Depending on the specifics of the similarity relation, these semantics enforce powerful terminological consequences. In [10] we discuss a number of them, here we have to restrict ourselves to two relatively simple examples: Given an ontology $\mathcal{O} = \{\overline{\texttt{SWAuthor}}^{eq} \sqsubseteq \texttt{Author}\}$ where eq is an equivalence relation, it follows that $\mathcal{O} \models \overline{\texttt{SWAuthor}}^{eq} \sqsubseteq \underline{\texttt{Author}}_{eq}$. What does this mean? It means that if any possible Semantic Web author is an author, it must be a typical author. Another example is the non-existence of a definitively non-typical Semantic Web author. Let the non-typical Semantic Web authors be defined as the Semantic Web authors that are not typical Semantic Web authors, i.e. we add $\texttt{NTSWAuthor} \sqsubseteq \texttt{SWAuthor} \sqcap \neg\underline{\texttt{SWAuthor}}_{eq}$ to $\mathcal{O}$. Rough DL semantics implies that there can be no definitely non-typical Semantic Web authors, i.e. that $\mathcal{O} \models \underline{\texttt{NTSWAuthors}}_{eq} = \bot$.

Related to these semantic consequences is the question of reasoning support, i.e. the existence of tools that can calculate consequences such as the ones discussed above in reasonable time. This points to the nice property of Rough DL being a conservative extension of OWL, in the sense that any Rough DL ontology can be translated into a logically equivalent OWL ontology. This means that reasoning for our language comes for free, as we can use standard reasoners to calculate class hierarchies, consistency and all instances of a particular class. The latter is the reasoning most needed in openacademia.

We can now relate these semantics of Rough DL to our example of instance unification given before. If we identify the class `TBL` with the singleton set `t.b-l`, i.e. define `TBL` to be equivalent to `oneOf(t.b-l)`, the set $\texttt{Possibly}_{Alg}\texttt{(t.b-l)}$ semantically corresponds to the upper approximation of `TBL` according to the indiscernibility relation sim_{Alg}.

Adapting Rough DL for openacademia As already mentioned, for application of Rough DL in openacademia we need a slightly different formalism from the one introduced in [10] and described above. First, we use similarity relations in addition to equivalence relations, and secondly, we want to apply different similarity relations and approximations, as well as hierarchies on both.

Similarity versus equivalence In openacademia, approximations based on similarity relations are used in addition to equivalence relations. Some smushing algorithms indeed produce equivalences, e.g. when two instances are identified through equivalence of the value of an inverse functional property. But even in logically weaker cases, there will be methods which indicate most likely equivalence between objects.

On the other hand, there are weaker methods, which will give indications for similarity, and which are not transitive. A simple example is edit distance: a similarity between two instances which is defined by an edit distance smaller or equal to 1 is non-transitive.

Hierarchies on similarities & approximations In an application such as openacademia there is no unique best way to identify co-reference of instances. This means that there will usually be several algorithms, such as the ones described in the following section, which produce several possible similarity relations. Often inclusion properties of such relations are easily created, and are often even more meaningful than quantitative values. In an RDF(S) based framework we can make use of hierarchies on relations to specify confidence in smushing algorithms in a qualitative way.

A simple logical consequence of specifying similarities in hierarchies of properties is that it implies hierarchies of the approximations. More concretely, suppose that two algorithms A_1 and A_2 produce two similarities $\texttt{oa:similarTo}A_1$ and $\texttt{oa:similarTo}A_2$ where, by construction, $\texttt{individual}_1$ $\texttt{oa:similarTo}A_1$ $\texttt{individual}_2$ implies that $\texttt{individual}_1$ $\texttt{oa:similarTo}A_2$ $\texttt{individual}_2$. This is a typical case, as smushing algorithms often include results of other algorithms. In this case, it can be shown that an upper approximation based on $\texttt{oa:similarTo}A_1$ is more specific than an upper approximation based on $\texttt{oa:similarTo}A_2$.

We make use of this property to construct hierarchies of approximations based on the underlying similarity relations, which can be very useful for controlled query relaxation.

3 Using Rough DL in openacademia

The current interface of openacademia allows to query for publications using combinations of different criteria, such as “author”, “title”, “year”, “type” and

"group". With respect to the author criteria, users can provide a string that is matched to a part of the author name.

This is a suboptimal solution when one wants to have precise control over the search results, as there is no way to distinguish between publications of different authors with a similar name. Regardless of instance unification, the result will be a mix of publications of possibly different persons.

A first requirement for controlled instance unification is to search by the URI of the resource instead of its label. For example, the system could search by label and return a list of publications. Then, the user could select a specific instance of an author in the result list, whose URI is used for the subsequent searches.

openacademia currently uses several methods to determine similarity between authors and publications, which are sometimes called *smushing algorithms*.

1. The most certain way to determine the equivalence of two resources is by comparing the values for their `owl:InverseFunctionalProperty`'s. When two resources have the same value for such a property they can be considered as equivalent. The FOAF-specification defines a number of properties, including `foaf:mbox` and `foaf:homepage` as inverse functional.
2. Another method is based on the comparison of the labels of resources. In openacademia we use several heuristics with different certainty to determine possible equivalences. For example, we consider instances of `foaf:Person` as unification candidates if both their first and last names match exactly (i.e. the string is identical), or if their last name and their initials match, or if their last name and first name are within a certain edit-distance.
3. An alternative method exploits the similarity of related resources. If, e.g., two instances of `swrc:Publication` are determined to be equivalent, we assume that the resources in the author-list are also equivalent.

Different from related work that focuses on learning the rules of smushing (e.g. [1]), smushing is an iterative reasoning process in openacademia. The instance matches found in one iteration can be used to discover new matches in subsequent iterations. Iterative reasoning is even a requirement if the smushing rules are co-dependent, such as the case when one would like to infer similarities of persons based on similarities of publications and vice versa, similarities of publications based on similarities of their authors.

To reflect the fact that different algorithms have a different certainty, we add *similarity* statements of the form $\texttt{individual}_1$ `oa:similarTo`X $\texttt{individual}_2$ to the repository. Each individual must be identified by a URI and `oa:similarTo`X is a similarity relation of instances returned by one of the algorithms discussed.

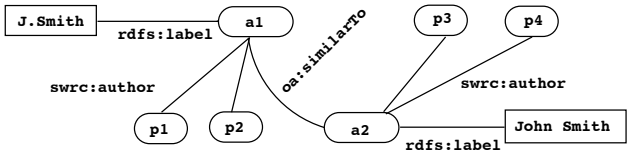
3.1 Benefits of Rough DL

In the context of openacademia, Rough DL has two functions: it offers an appealing conceptual framework for querying over similarities, and it provides the possibility to model vague concepts, such as typical Semantic Web authors. We will discuss both features in more detail.

Querying over similarities Using similarities for object-dereferencing is a relatively obvious proposal as the problems of ad-hoc solutions (such as using `owl:sameAs`) are known. However, most people decide to use the slight over-commitment of `owl:sameAs` in order to be able to continue to use the automatic reasoning support for OWL, i.e. in order to avoid having to deal with similarities explicitly.

Rough DL offers an attractive alternative, as it provides a conceptually elegant framework for querying an RDF(S) repository including similarity relations by two simple operators ($\bar{\cdot}$) and ($\underline{\cdot}$), the approximations. Let us start with a simple example of how Rough DL can help in order to formulate concise queries over graphs including similarities.

Suppose we have identified two resources, `a1` and `a2`, each of type `foaf:Person`. Each resource is connected via a `swrc:author` property to a number of resources of type `swrc:Publication` (`p1`...`p4`). Besides this, `a1` and `a2` each have an `rdfs:label`. One of the similarity heuristics discovered a similarity between the labels of `a1` and `a2`, which is represented by a property `oa:similarTo` between both resources.



The obvious query to take from this graph is to find all publications of the resource `a1`, uniquely identified by the URI `<http://www.uni1.edu/~personA/pubs.bib#john_smith>`, and every resource that is similar.

Formulating this as instance checking in Rough DL is simply to ask for all instances of class $\overline{\texttt{oneOf}\{\ldots\texttt{\#john\_smith}\}}$ (for the resource). Already for such a simple Rough DL query, the corresponding SeRQL query requires explicit knowledge of the structure of the graph, and of the similarity relation used.

```
SELECT distinct Pub FROM
  {Pub}    swrc:author {Person},
  {Person} oa:nameSimilarTo3
       {<http://www.uni1.edu/~personA/pubs.bib#john_smith>}
```

Another example exploits similarities between publications. Suppose that we add different kinds of similarity relations between publications. For example, two publications could be connected via a property `oa:hasJointAuthors` if they share at least two authors. Or, another possibility, they can be connected via `oa:hasRelatedKeywords` if their keywords are related according to some metric, e.g. because the keywords are semantically close to each other in some topic hierarchy. One could even add similarity relations based on the textual overlap of the abstract.

Now, the Rough DL framework allows queries for upper approximations — according to a specific type of similarity — of publications that fulfill specific criteria. For example, if we use the "author overlap" similarity, we could query for

the upper approximation of papers with "OWL" as a keyword by simply asking for the instances of the Rough DL class $\overline{\texttt{restriction(keyword hasValue("OWL"))}}$. When we have `oa:hasJointAuthors` in place as a similarity relation we get as the result all papers of authors that together have published papers with OWL as a keyword. Again, the corresponding SeRQL query is simple, but we could easily use a different similarity measure without requiring any change for the user.

```
SELECT DISTINCT Pub FROM {Pub}
   oa:hasJointAuthors {Other}, WHERE
{Other} IN (SELECT Pub FROM {Pub} swrc:keyword {"OWL"})
```

Modeling vague concepts in OA Up to now we discussed the use of Rough DL for formulating queries over an RDF(S) repository with similarity relations. But of course, the language can also be used to model vague concepts directly in the repository. Imagine that one wants to model Semantic Web conferences and authors, obviously terms that describe vague concepts for which it will be impossible to find commonly agreed upon definitions. What is possible, on the other hand, is to define approximations for both classes. Most people would agree that there are three prototypical Semantic Web conferences: the International and European Semantic Web conferences (ISWC and ESWC), and the World-Wide Web conference (WWW). Defining the lower approximation of a class `SWconference` can then be done simply as $\underline{\texttt{SWconference}} = \underline{\texttt{WWW}} \sqcup \underline{\texttt{ISWC}} \sqcup \underline{\texttt{ESWC}}$ where the conferences are described as classes (e.g. ISWC) containing at least one uniquely identifying resource (e.g., `ns:iswc`). Here, we chose the lower approximations of the resources for the conferences as we want to avoid ambiguity through spelling variants or other forms of synonymy (e.g. ESWC also refers to the Electronic Sports World Cup).

Semantic Web authors can now be defined as authors having published at *possible* Semantic Web conference, i.e.

$$\texttt{SWauthor} = \exists\ \texttt{publishedIn}.\overline{\texttt{SWconference}}$$

Even though Rough DL is a conservative extension of OWL DL this example shows that modeling the same information without approximation operators would be extremely cumbersome. Using Rough DL, and the reasoning machinery that comes for free, thanks to the translation back to OWL, allows queries such as for all possible Semantic Web authors $i : \overline{\texttt{SWauthor}}$?, which even for this simple example is non-trivial on a larger data set.

Furthermore, for the Rough DL fragment built on the OWL DL dialect, the usual reasoning services, such as query entailment, satisfiability checking or subsumption hierarchies can be easily calculated.

For openacademia querying with Rough DL is the more prominent application, and we have not yet pursued modeling of rough concepts in openacademia. Technically, and conceptually, it is easy to add Rough DL axioms to the Sesame repository. By adding rules, part of the OWL semantics can be captured, but completeness cannot be achieved. A more detailed study of this, e.g., considering the alternative semantics proposed in [12], is outside the scope of this paper.

Therefore, although we are also convinced that modeling approximate concepts will significantly improve the ease of use of openacademia, the focus in this paper will be on querying from now on.

3.2 Technical Issues

The application of Rough DL in openacademia amplifies a discrepancy of two Knowledge Representation frameworks that is present in many practical approaches to the Semantic Web. With query languages such as SeRQL or SPARQL for RDF(S) ontologies and the advent of robust and fast RDF repositories, efficient data access is now made possible even for very large data sets. On the other hand, the expressivity of OWL makes it possible to model ontological knowledge in very elegant ways, which is needed for many realistic applications. Unfortunately, theoretically both paradigms are less easily integrated than one would hope for, and than could be expected at first glance.

1. A first issue is the use of a query-language such as SeRQL for a repository containing OWL statements.
2. The second problem to be addressed is the question of Open- versus Closed-World Assumption.

For lack of space both issues can only be discussed briefly.

First, to make use of the best of both worlds, many people include OWL ontologies in RDF repositories, and query those with traditional RDF query languages. The problem with such an approach is that completeness cannot be guaranteed in general. For openacademia we do the same: particular knowledge is represented in OWL (e.g. defining a property as transitive or functional), but SeRQL is used for querying. In the case of openacademia, however, the problem of incompleteness can be circumvented. This is done by including parts of the OWL semantics in the Sesame inference engine (in the style of [12]), and by encoding parts in the queries.

Secondly, querying a triple store such as Sesame usually employs a Close-World assumption, i.e. a universally quantified statement is evaluated as true if all known instances in the relation have the required property. This is different than DL, where an Open-World Assumption is taken. For the lower approximation this means that the DL interpretation differs from the interpretation of a natural SeRQL encoding. As a DL query for lower approximations will only return relevant results when there are explicit universal statements (or approximations) in the ontology, our current research focuses on the use of the upper approximation for querying.

4 Case study

To illustrate the benefits of Rough DL descriptions in practice, we show the effect of applying different approximations in openacademia. For this, we use the approximation interface as shown in Figure 1. This interface translates a restricted set of Rough DL queries into SeRQL.

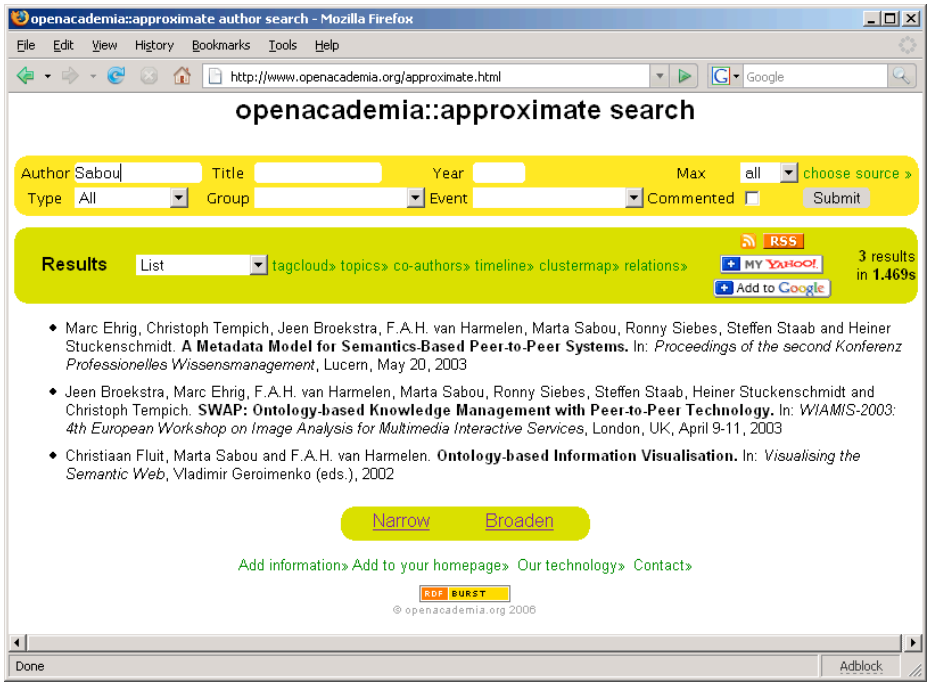


Fig. 1. The approximation interface, allowing for broadening or narrowing the author concept.

We apply Rough DL for author similarity, using a hierarchy of similarities. Suppose we want to query for publications of "Marta Sabou" with several entries in our database. We start with a URI `<http://www.uni1.edu/~personA/pubs.bib#marta_sabou>` in our own BibTeX files, which gives high confidence that the resource represents the right person.

No approximation Without approximation, a query for the publications of this resource results in 3 publications, which were all specified in `<http://www.uni1.edu/~personA/pubs.bib>`.

Exploiting inverse functional properties Smushing adds `oa:nameSimilarTo1` statements between all resources with the same value for an inverse functional property. As Marta's email is listed in her FOAF-profile, an `oa:nameSimilarTo1` statement is added between the original resource and `<http://www.uni1.edu/swhome/person/marta>`, an RDF representation of a personnel database. When querying for the upper approximation of the resource, the search results now include the publications on Marta's homepage.

Exact match of fullname A second level of approximation uses the label of the resources of type `foaf:Person`. When the first and lastname of a person exactly match, a `oa:nameSimilarTo2` is added. This results in similarity statements between the original resource and `<http://www.uni2.edu/~personB/`

biblio/bnaic2002.bib#marta_sabou> and <http://www.uni3.edu/~personC/pubs.bib#marta_sabou>. When using this property as similarity in the Rough DL framework, the search result contains 7 publications.

Exact match of lastname and initial The next level of approximation exploits the oa:nameSimilarTo3 statements that are added when both the lastnames match and the initial of one resource matches the first character of the firstname of another resource. This results in similarity statements to the resource <http://www.uni3.org/~personD/publications.bib#m_sabou> with the label "M. Sabou" and the resource <http://www.uni1.edu/swhome/person/marta> with the label "M.R. Sabou", yielding in one new publication.

Fuzzy match on fullname The final level of approximation uses n-gram distance between labels of two resources and adds oa:nameSimilarTo4 when the distance is above some threshold. In our data set there is such a statement between the original resource and <http://www.uni4.edu/publications/ins.bib#martha_sabou>, which has "Mart**h**a Sabou" as label. This again added one additional publication, resulting in 9 publications.

Note that we only discussed the *additional search results* in the description above. However, when exploiting the similarities between the authors in the search, we also get duplicate resources for publications for which we apply a similar strategy to combine publication resources.

5 Conclusions

Summary Rough DL is a conservative extension of DL, i.e. an extension of DL with new operators for modeling vague concepts, that does not increase the expressive power of the original language. We show that this language is suitable for reasoning over similarities or equivalences introduced into an ontology through co-reference resolution. Rough DL provides a qualitative way of representing vague concepts, and to reason and query over similarities. By applying Rough DL to openacademia we show that AI techniques can elegantly solve practical problems on the web.

To make the Rough DL version of openacademia robust and efficient for large collections, e.g., crawled on the WWW, the application has initially been restricted to querying. Large scale experiments with Rough DL modeling are planned as future work to evaluate scalability of this theoretically promising framework.

Related Work The related work covers modeling vagueness in ontologies, most prominently in combining fuzzy logic with Semantic Web research, as exemplified in [9]. Some of this work is based on Straccia's paper on fuzzy Description Logics, e.g., [11]. Vagueness of concepts is expressed as a degree of membership. Rough DL advocates a simpler, *qualitative*, approach to domains where there is no way of quantifying membership of the class but well-defined upper and lower approximations. The difference is intrinsically in the type of vagueness of particular concepts. On the querying side, there have also been efforts to integrate

querying over similarities into a standard RDF querying language, e.g., [2]. The language described there, iRDQL, has implicit functionality to query for objects with a certain similarity.

There, however, lies the biggest difference to our approach, which focuses on qualitative modeling of vagueness and querying over similarities. For some domains and particular applications, such as for access to distributed data sources, this approach can be more appropriate. This does not just hold for bibliographic data, but for any data integration where the identity of resources cannot always be established with absolute certainty, and where qualitative querying over similarities can provide a fine-grained access to collections.

Future Work The application of Rough DL to openacademia is a first step towards achieving the full potential of the language. Currently, SeRQL queries are automatically created for narrowing or broadening search results. A next step will be to extend querying to more expressive Rough DL queries, and to integrate Rough DL in the ontology. Together with such an extension of the functionality we will have to undertake a detailed investigation of the scalability of the system, and a qualitative and quantitative analysis of the effects on the querying results in openacademia.

References

1. Niraj Aswani, Kalina Bontcheva, and Hamish Cunningham. Mining Information for Instance Unification. In *ISWC 2006*, 2006.
2. Abraham Bernstein and Christoph Kiefer. Imprecise RDQL: towards generic retrieval in ontologies using similarity joins. In *SAC 2006*, 2006.
3. D.G. Feitelson. On identifying name equivalences in digital libraries. *Information Research*, 9(4):192, 2004.
4. R. Guha and A. Garg. Disambiguating People in Search. *TAP: Building the Semantic Web*, 2003.
5. T.P. Martin. Searching and smushing on the semantic web – challenges for soft computing. In *FLINT 2001 – New Directions in Enhancing the Power of the Internet*, pages 3–8, December 2001.
6. H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. Identity uncertainty and citation matching. *Advances in Neural Information Processing (NIPS)*, 2003.
7. Z. Pawlak. Rough sets. *Int. Journal of Computer and Information Sciences*, 11:341–356, 1982.
8. Allen Renear and David Dubin. Towards Identity Conditions for Digital Documents. Technical Report UIUCLIS-2003/2+ EPRG, 2003.
9. Elie Sanchez. *Fuzzy Logic and the Semantic Web*. Elsevier, April 5 2006.
10. Stefan Schlobach, Michel Klein, and Linda Peelen. Description logics with approximate definitions: Precise modeling of vague concepts. In *Proc. of the 20th Int. Joint Conf. on Art. Intel., IJCAI 07*, Hyderabad, India, January 2007.
11. Umberto Straccia. Reasoning with fuzzy description logics. *J. of AI Research*, 14:137–166, 2001.
12. Herman J. ter Horst. Completeness, decidability and complexity of entailment for rdf schema and a semantic extension involving the owl vocabulary. *J. of Web Semantics*, 3(2-3):79–115, 2005.

Optimizing the Crisp Representation of the Fuzzy Description Logic $\mathcal{SROIQ}$

Fernando Bobillo, Miguel Delgado, and Juan Gómez-Romero

Department of Computer Science and Artificial Intelligence, University of Granada
C. Periodista Daniel Saucedo Aranda, 18071 Granada, Spain
Phone: +34 958243194; Fax: +34 958243317
Email: fbobillo@decsai.ugr.es, mdelgado@ugr.es, jgomez@decsai.ugr.es

Abstract. Classical ontologies are not suitable to represent imprecise nor uncertain pieces of information. Fuzzy Description Logics were born to represent the former type of knowledge, but they require an appropriate fuzzy language to be agreed and an important number of available resources to be adapted. This paper faces these problems by presenting a reasoning preserving procedure to obtain a crisp representation for a fuzzy extension of the logic $\mathcal{SROIQ}$ which uses Gödel implication in the semantics of fuzzy concept and role subsumption. This reduction allows to reuse a crisp representation language as well as currently available reasoners. Our procedure is optimized with respect to the related work, reducing the size of the resulting knowledge base, and is implemented in DeLorean, the first reasoner supporting fuzzy OWL DL.

1 Introduction

Description Logics (DLs for short) [1] are a family of logics for representing structured knowledge which have proved to be very useful as ontology languages. For instance, $\mathcal{SROIQ}(\mathbf{D})$ [2] is the subjacent DL of OWL 1.1., a recent extension of the standard language OWL[1] which is its most likely immediate successor.

Nevertheless, it has been widely pointed out that classical ontologies are not appropriate to deal with imprecise and vague knowledge, which is inherent to several real-world domains. Since fuzzy logic is a suitable formalism to handle these types of knowledge, several fuzzy extensions of DLs can be found in the literature (see [3] for an overview).

Defining a fuzzy DL brings about that crisp standard languages are no longer appropriate, new fuzzy languages need to be used, and hence the large number of resources available need to be adapted to the new framework, requiring an important effort. An additional problem is that reasoning within (crisp) expressive DLs has a very high worst-case complexity (e.g. NExpTime in $\mathcal{SHOIN}$) and, consequently, there exists a significant gap between the design of a decision procedure and the achievement of a practical implementation [4].

An alternative is to represent fuzzy DLs using crisp DLs and to reduce reasoning within fuzzy DLs to reasoning within crisp ones. This has several advantages:

[1] http://www.w3.org/TR/owl-features

– There would be no need to agree a new standard fuzzy language, but every developer could use its own language expressing fuzzy $\mathcal{SROIQ}$, as long as he implements the reduction that we describe.
– We will continue using standard languages with a lot of resources available, so the need (and cost) of adapting them to the new fuzzy language is avoided.
– We will continue using the existing crisp reasoners. We do not claim that reasoning will be more efficient, but it supposes an easy alternative to support early reasoning in future fuzzy languages. In fact, nowadays there is no reasoner fully supporting a fuzzy extension of OWL DL.

Under this approach an immediate practical application of fuzzy ontologies is feasible, because of its tight relation with already existing languages and tools which have proved their validity.

Although there has been a relatively significant amount of works in extending DLs with fuzzy set theory ([3] is a good survey), the representation of them using crisp description logics has not received such attention. The first efforts in this direction are due to U. Straccia, who considered fuzzy $\mathcal{ALCH}$ [5] and fuzzy $\mathcal{ALC}$ with truth values taken from an uncertainty lattice [6]. F. Bobillo et al. extended Straccia's work to $\mathcal{SHOIN}$, including fuzzy nominals and fuzzy General Concept Inclusions (GCIs) with a semantics given by Kleene-Dienes (KD) implication [7]. Finally, G. Stoilos et al. extended this work to $\mathcal{SROIQ}$ [8]. This paper improves the latter work providing the following contributions:

– We provide a full representation, differently from [8] which do not show how to reduce qualified cardinality restrictions, local reflexivity concepts in expressions of the form $\rho(\exists S.Self, \lhd\gamma)$ nor negative role assertions.
– [5, 8] force GCIs and Role Inclusion Axioms (RIAs) to be either true or false, but we will allow them to be verified up to some degree by using Gödel implication in the semantics.
– We improve one of their starting points (the reduction presented in [5]) by reducing the number of new atomic elements and their corresponding axioms.
– We show how to optimize some important GCIs.
– We present DeLorean, our implementation of the reduction and the first implemented reasoner supporting fuzzy $\mathcal{SHOIN}$.

The remainder is organized as follows. Section 2 describes a fuzzy extension of $\mathcal{SROIQ}$ and discusses some logical properties. Section 3 depicts a reduction into crisp $\mathcal{SROIQ}$, whereas Section 4 presents our implementation of the procedure. Finally, in Section 5 we set out some conclusions and ideas for future work.

2 Fuzzy $\mathcal{SROIQ}$

In this section we define $f\mathcal{SROIQ}$, which extend $\mathcal{SROIQ}$ to the fuzzy case by letting *(i)* concepts denote fuzzy sets of individuals and *(ii)* roles denote fuzzy binary relations. Axioms are also extended to the fuzzy case and some of them hold to a degree. The following definition combines [7–9], but we will use Gödel implication in the semantics of GCIs and RIAs.

Syntax. $f\mathcal{SROIQ}$ assumes three alphabets of symbols, for concepts, roles and individuals. The concepts of the language (denoted C or D) can be built inductively from atomic concepts (A), atomic roles (R), top concept $\top$, bottom concept $\bot$, named individuals (o_i), universal role (U) and simple roles (S, which will be defined below) according to the following syntax rule, where n, m are natural numbers ($n \geq 0, m > 0$) and $\alpha_i \in [0,1]$: $C, D \rightarrow A \mid \top \mid \bot \mid C \sqcap D \mid C \sqcup D \mid \neg C \mid \forall R.C \mid \exists R.C \mid \{\alpha_1/o_1, \ldots, \alpha_m/o_m\} \mid (\geq n\ S.C) \mid (\leq n\ S.C) \mid \exists S.Self$. Notice that the only difference with the crisp case is the presence of fuzzy nominals [7]. Complex roles are built using the syntax rule $R \rightarrow R_A \mid R^- \mid U$.

A fuzzy Knowledge Base (fKB) comprises two parts: the extensional knowledge, i.e. particular knowledge about some specific situation (a fuzzy Assertional Box or ABox K_A with statements about individuals) and the intensional knowledge, i.e. general knowledge about the application domain (a fuzzy Terminological Box or TBox K_T and a fuzzy Role Box or RBox K_R).

In the rest of the paper we will assume $\bowtie \in \{\geq, <, \leq, >\}$, $\alpha \in (0,1]$, $\beta \in [0,1)$ and $\gamma \in [0,1]$. Moreover, for every operator $\bowtie$ we define *(i)* its symmetric operator $\bowtie^-$ defined as $\geq^-=\leq, >^-=<, \leq^-=\geq, <^-=>$, and *(ii)* its negation operator $\neg \bowtie$, defined as $\neg \geq=<, \neg >=\leq, \neg \leq=>, \neg <=\geq$.

A *fuzzy ABox* consists of a finite set of *fuzzy assertions*. A fuzzy assertion can be an inequality assertion $\langle a \neq b \rangle$, an equality assertion $\langle a = b \rangle$ or a constraint on the truth value of a concept or role assertion, i.e. an expression of the form $\langle \Psi \bowtie \alpha \rangle$, where Ψ is an assertion of the form $a\!:\!C$ or $(a,b)\!:\!R$.

A *fuzzy TBox* consists of *fuzzy GCIs*, which constrain the truth value of a GCI i.e. they are expressions of the form $\langle \Omega \geq \alpha \rangle$ or $\langle \Omega > \beta \rangle$, where $\Omega = C \sqsubseteq D$.

A *fuzzy RBox* consists of a finite set of role axioms, which can be *fuzzy RIAs* $\langle w \sqsubseteq R \geq \alpha \rangle$ or $\langle w \sqsubseteq R > \beta \rangle$ for a role chain $w = R_1 R_2 \ldots R_n$, or any other of the role axioms from the crisp case: *transitive* $trans(R)$, *disjoint* $dis(S_1, S_2)$, *reflexive* $ref(R)$, *irreflexive* $irr(S)$, *symmetric* $sym(R)$ or *asymmetric* $asy(S)$. As in the crisp case, role axioms cannot contain U and every RIA should be $\prec$-regular for a regular order $\prec$. A RIA $\langle w \sqsubseteq R \rhd \gamma \rangle$ is *$\prec$-regular* if $R = R_A$ and *(i)* $w = RR$, or *(ii)* $w = R^-$, or *(iii)* $w = S_1 \ldots S_n$ and $S_i \prec R$ for all $i = 1, \ldots, n$, or *(iv)* $w = RS_1 \ldots S_n$ and $S_i \prec R$ for all $i = 1, \ldots, n$, or *(v)* $w = S_1 \ldots S_n R$ and $S_i \prec R$ for all $i = 1, \ldots, n$.

Simple roles are inductively defined: *(i)* R_A is simple if does not occur on the right side of a RIA, *(ii)* R^- is simple if R is, *(iii)* if R occurs on the right side of a RIA, R is simple if, for each $\langle w \sqsubseteq R \rhd \gamma \rangle$, $w = S$ for a simple role S.

A fuzzy axiom τ is *positive* (denoted $\langle \tau \rhd \alpha \rangle$) if it is of the form $\langle \tau \geq \alpha \rangle$ or $\langle \tau > \beta \rangle$, and *negative* (denoted $\langle \tau \lhd \alpha \rangle$) if it is of the form $\langle \tau \leq \beta \rangle$ or $\langle \tau < \alpha \rangle$. $\langle \tau = \alpha \rangle$ is equivalent to the pair of axioms $\langle \tau \geq \alpha \rangle$ and $\langle \tau \leq \alpha \rangle$.

Notice that negative GCIs or RIAs are not allowed, because they correspond to negated GCIs and RIAs respectively, which are not part of crisp $\mathcal{SROIQ}$.

Semantics. A fuzzy interpretation $\mathcal{I}$ is a pair $(\Delta^\mathcal{I}, \cdot^\mathcal{I})$ consisting of a non empty set $\Delta^\mathcal{I}$ (the interpretation domain) and a fuzzy interpretation function $\cdot^\mathcal{I}$ mapping *(i)* every individual onto an element of $\Delta^\mathcal{I}$, *(ii)* every concept C onto a function $C^\mathcal{I} : \Delta^\mathcal{I} \rightarrow [0,1]$, *(iii)* every role R onto a function $R^\mathcal{I} : \Delta^\mathcal{I} \times \Delta^\mathcal{I} \rightarrow [0,1]$.

$C^{\mathcal{I}}$ (resp. $R^{\mathcal{I}}$) denotes the membership function of the fuzzy concept C (resp. fuzzy role R) w.r.t. $\mathcal{I}$. $C^{\mathcal{I}}(a)$ (resp. $R^{\mathcal{I}}(a,b)$) gives us the degree of being the individual a an element of the fuzzy concept C (resp. the degree of being (a,b) an element of the fuzzy role R) under the fuzzy interpretation $\mathcal{I}$. We do not impose unique name assumption, i.e. two nominals might refer to the same individual. For a t-norm $\otimes$, a t-conorm $\oplus$, a negation function $\ominus$ and an implication function $\rightarrow$, the fuzzy interpretation function is extended to complex concepts and roles as follows:

$$
\begin{aligned}
\top^{\mathcal{I}}(a) &= 1 \\
\bot^{\mathcal{I}}(a) &= 0 \\
(C \sqcap D)^{\mathcal{I}}(a) &= C^{\mathcal{I}}(a) \otimes D^{\mathcal{I}}(a) \\
(C \sqcup D)^{\mathcal{I}}(a) &= C^{\mathcal{I}}(a) \oplus D^{\mathcal{I}}(a) \\
(\neg C)^{\mathcal{I}}(a) &= \ominus C^{\mathcal{I}}(a) \\
(\forall R.C)^{\mathcal{I}}(a) &= \inf_{b \in \Delta^{\mathcal{I}}} \{R^{\mathcal{I}}(a,b) \rightarrow C^{\mathcal{I}}(b)\} \\
(\exists R.C)^{\mathcal{I}}(a) &= \sup_{b \in \Delta^{\mathcal{I}}} \{R^{\mathcal{I}}(a,b) \otimes C^{\mathcal{I}}(b)\} \\
\{\alpha_1/o_1, \ldots, \alpha_m/o_m\}^{\mathcal{I}}(a) &= \sup_{i \mid a \in \{o_i^{\mathcal{I}}\}} \alpha_i \\
(\geq 0\ S.C)^{\mathcal{I}}(a) &= \top^{\mathcal{I}}(a) = 1 \\
(\geq m\ S.C)^{\mathcal{I}}(a) &= \sup_{b_1,\ldots,b_m \in \Delta^{\mathcal{I}}} [(\otimes_{i=1}^{n} \{S^{\mathcal{I}}(a,b_i) \otimes C^{\mathcal{I}}(b_i)\}) \bigotimes (\otimes_{j<k} \{b_j \neq b_k\})] \\
(\leq n\ S.C)^{\mathcal{I}}(a) &= \inf_{b_1,\ldots,b_{n+1} \in \Delta^{\mathcal{I}}} [(\otimes_{i=1}^{n+1} \{S^{\mathcal{I}}(a,b_i) \otimes C^{\mathcal{I}}(b_i)\}) \rightarrow (\oplus_{j<k} \{b_j = b_k\})] \\
(\exists S.Self)^{\mathcal{I}}(a) &= S^{\mathcal{I}}(a,a) \\
(R^-)^{\mathcal{I}}(a,b) &= R^{\mathcal{I}}(b,a) \\
U^{\mathcal{I}}(a,b) &= 1
\end{aligned}
$$

A fuzzy interpretation $\mathcal{I}$ satisfies (is a model of):

- *(i)* $\langle a{:}C \geq \alpha \rangle$ iff $C^{\mathcal{I}}(a^{\mathcal{I}}) \geq \alpha$,
- *(ii)* $\langle (a,b){:}R \geq \alpha \rangle$ iff $R^{\mathcal{I}}(a^{\mathcal{I}}, b^{\mathcal{I}}) \geq \alpha$,
- *(iii)* $\langle a \neq b \rangle$ iff $a^{\mathcal{I}} \neq b^{\mathcal{I}}$,
- *(iv)* $\langle a = b \rangle$ iff $a^{\mathcal{I}} = b^{\mathcal{I}}$,
- *(v)* $\langle C \sqsubseteq D \geq \alpha \rangle$ iff $\inf_{a \in \Delta^{\mathcal{I}}} \{C^{\mathcal{I}}(a) \rightarrow D^{\mathcal{I}}(a)\} \geq \alpha$,
- *(vi)* $\langle R_1 \ldots R_n \sqsubseteq R \geq \alpha \rangle$ iff $\sup_{b_1 \ldots b_{n+1} \in \Delta^{\mathcal{I}}} \bigotimes [R_1^{\mathcal{I}}(b_1,b_2), \ldots, R_n^{\mathcal{I}}(b_n,b_{n+1})] \rightarrow R^{\mathcal{I}}(b_1,b_{n+1}) \geq \alpha$,
- *(vii)* $trans(R)$ iff $\forall a,b \in \Delta^{\mathcal{I}}, R^{\mathcal{I}}(a,b) \geq \sup_{c \in \Delta^{\mathcal{I}}} R^{\mathcal{I}}(a,c) \otimes R^{\mathcal{I}}(c,b)$,
- *(viii)* $dis(S_1,S_2)$ iff $\forall a,b \in \Delta^{\mathcal{I}}, S_1^{\mathcal{I}}(a,b) \otimes S_2^{\mathcal{I}}(a,b) = 0$,
- *(ix)* $ref(R)$ iff $\forall a \in \Delta^{\mathcal{I}}, R^{\mathcal{I}}(a,a) = 1$,
- *(x)* $irr(S)$ iff $\forall a \in \Delta^{\mathcal{I}}, S^{\mathcal{I}}(a,a) = 0$,
- *(xi)* $sym(R)$ iff $\forall a,b \in \Delta^{\mathcal{I}}, R^{\mathcal{I}}(a,b) = R^{\mathcal{I}}(b,a)$,
- *(xii)* $asy(S)$ iff $\forall a,b \in \Delta^{\mathcal{I}}$, if $S^{\mathcal{I}}(a,b) > 0$ then $S^{\mathcal{I}}(b,a) = 0$,
- *(xii)* a fKB iff it satisfies each element in fK_A, fK_T and fK_R.

In cases $(i), (ii)$ similar definitions can be given for $> \beta$, $\leq \beta$ and $< \alpha$, whereas in cases $(v), (vi)$ a similar definition can be given for $> \beta$.

Notice that individual assertions are considered to be crisp

In the rest of the paper we will only consider fKB satisfiability, since (as in the crisp case) most inference problems can be reduced to it [10].

Some logical properties. It can be easily shown that $f\mathcal{SROIQ}$ is a sound extension of crisp $\mathcal{SROIQ}$, because fuzzy interpretations coincide with crisp interpretations if we restrict the membership degrees to $\{0, 1\}$.

In the fuzzy DLs literature, the notation $f_i\mathcal{DL}$ has been proposed [11], where i is the fuzzy implication function considered. Here in after we will concentrate on $f_{KD}\mathcal{SROIQ}$, restricting ourselves to the Zadeh family: minimum t-norm, maximum t-conorm, Łukasiewicz negation and KD implication, with the exception of GCIs and RIAs, where we will consider Gödel implication. This choice comes from the fact that KD implication specifies a t-norm, a t-conorm and a negation which make possible the reduction to a crisp KB, as we will see in Section 3 (other fuzzy operators are not suitable for a similar reduction).

However, the use of KD implication in the semantics of GCIs and RIAs brings about two counter-intuitive effects: *(i)* in general concepts (and roles) do not fully subsume themselves and *(ii)* crisp subsumption (holding to degree 1) forces some fuzzy concepts and roles to be interpreted as crisp [7].

Another common semantics which could be considered is the one based on Zadeh's set inclusion ($C \sqsubseteq D = \forall x \in \Delta^{\mathcal{I}}, C^{\mathcal{I}}(x) \leq D^{\mathcal{I}}(x)$) as in [10, 12], but it forces the axioms to be either true or false. For example, under this semantics it is not possible that concept $Hotel$ subsumes concept Inn with degree 0.5.

Gödel implication solves the afore-mentioned problems and is suitable for a classical representation as we will see in Section 3. Moreover, for GCIs of the form $\langle C \sqsubseteq D \geq 1\rangle$, the semantics is equivalent to that of Zadeh's set inclusion.

It is possible to transform concept expressions into a semantically equivalent *Negation Normal Form* (NNF), which is obtained by pushing in the usual manner negation in front of atomic concepts, fuzzy nominals and local reflexivity concepts. In the case of $\neg(\geq 0\ S)$, it could be replaced by $\bot$ since it is an inconsistent concept. In the following, we will assume that concepts are in NNF.

Irreflexive, transitive and symmetric role axioms are syntactic sugar for every R-implication (and consequently it can be assumed that they do not appear in fKBs) due to the following equivalences:

- $irr(S) \equiv \langle \top \sqsubseteq \neg\exists S.Self \geq 1\rangle$,
- $trans(R) \equiv \langle RR \sqsubseteq R \geq 1\rangle$,
- $sym(R) \equiv \langle R \sqsubseteq R^{-} \geq 1\rangle$.

3 An Optimized Crisp Representation for Fuzzy $\mathcal{SROIQ}$

In this section we show how to reduce a $f_{KD}\mathcal{SROIQ}$ fKB into a crisp KB, similarly as in [5, 7, 8]. The procedure preserves reasoning, so existing $\mathcal{SROIQ}$ reasoners could be applied to the resulting KB. First we will describe the reduction and then we will provide an illustrating example. The basic idea is to create some new crisp concepts and roles, representing the α-cuts of the fuzzy concepts and relations, and to rely on them. Next, some new axioms are added to preserve their semantics and finally every axiom in the ABox, the TBox and the RBox is represented, independently from other axioms, using these new crisp elements.

Adding (an optimized number of) new elements. Let A^{fK} and R^{fK} be the set of atomic concepts and atomic roles occurring in a fKB $fK = \langle fK_A, fK_T, fK_R \rangle$. In [5] it is shown that the set of the degrees which must be considered for any reasoning task is defined as $N^{fK} = X^{fK} \cup \{1 - \alpha \mid \alpha \in X^{fK}\}$, where $X^{fK} = \{0, 0.5, 1\} \cup \{\gamma \mid \langle \tau \bowtie \gamma \rangle \in fK\}$. This also holds in $f_{KD}\mathcal{SROIQ}$, but note that it is not necessarily true when other fuzzy operators are considered. Without loss of generality, it can be assumed that $N^{fK} = \{\gamma_1, \ldots, \gamma_{|N^{fK}|}\}$ and $\gamma_i < \gamma_{i+1}, 1 \leq i \leq |N^{fK}| - 1$. It is easy to see that $\gamma_1 = 0$ and $\gamma_{|N^{fK}|} = 1$.

Now, for each $\alpha, \beta \in N^{fK}$ with $\alpha \in (0, 1]$ and $\beta \in [0, 1)$, for each $A \in A^{fK}$ and for each $R_A \in R^{fK}$, two new atomic concepts $A_{\geq\alpha}, A_{>\beta}$ and two new atomic roles $R_{\geq\alpha}, R_{>\beta}$ are introduced. $A_{\geq\alpha}$ represents the crisp set of individuals which are instance of A with degree higher or equal than α i.e the α-cut of A. The other new elements are defined in a similar way. The atomic elements $A_{>1}, R_{>1}, A_{\geq 0}$ and $R_{\geq 0}$ are not considered because they are not necessary, due to the restrictions on the allowed degree of the axioms in the fKB (e.g. we do not allow GCIs of the form $C \sqsubseteq D \geq 0$). Note that [5, 7] consider $A_{\geq 0}$ and $R_{\geq 0}$.

The semantics of these newly introduced atomic concepts and roles is preserved by some terminological and role axioms. For each $1 \leq i \leq |N^{fK}| - 1, 2 \leq j \leq |N^{fK}| - 1$ and for each $A \in A^{fK}$, $T(N^{fK})$ is the smallest terminology containing these two axioms: $A_{\geq\gamma_{i+1}} \sqsubseteq A_{>\gamma_i}$, $A_{>\gamma_j} \sqsubseteq A_{\geq\gamma_j}$. Similarly, for each $R_A \in R^{fK}$, $R(N^{fK})$ is the smallest terminology containing $R_{\geq\gamma_{i+1}} \sqsubseteq R_{>\gamma_i}$ and $R_{>\gamma_i} \sqsubseteq R_{\geq\gamma_i}$.

In contrast to previous works, which use two more atomic concepts $A_{\leq\beta}, A_{<\alpha}$ and some additional axioms $(2 \leq k \leq |N^{fK}|)$ [5, 7]:

$$A_{<\gamma_k} \sqsubseteq A_{\leq\gamma_k}, \qquad A_{\leq\gamma_i} \sqsubseteq A_{<\gamma_{i+1}}$$
$$A_{\geq\gamma_k} \sqcap A_{<\gamma_k} \sqsubseteq \bot, \qquad A_{>\gamma_i} \sqcap A_{\leq\gamma_i} \sqsubseteq \bot$$
$$\top \sqsubseteq A_{\geq\gamma_k} \sqcup A_{<\gamma_k}, \qquad \top \sqsubseteq A_{>\gamma_i} \sqcup A_{\leq\gamma_i}$$

we use $\neg A_{>\gamma_k}$ rather than $A_{\leq\gamma_k}$ and $\neg A_{\geq\gamma_k}$ instead of $A_{<\gamma_k}$, since the six axioms above follow immediately from the semantics of the crisp concepts as Proposition 1 shows:

Proposition 1. *If $A_{\geq\gamma_{i+1}} \sqsubseteq A_{>\gamma_i}$ and $A_{>\gamma_k} \sqsubseteq A_{\geq\gamma_k}$ hold, then the followings axioms are verified:*

(1) $\neg A_{\geq\gamma_k} \sqsubseteq \neg A_{>\gamma_k}$ *(2)* $\neg A_{>\gamma_i} \sqsubseteq \neg A_{\geq\gamma_{i+1}}$
(3) $A_{\geq\gamma_k} \sqcap \neg A_{\geq\gamma_k} \sqsubseteq \bot$ *(4)* $A_{>\gamma_i} \sqcap \neg A_{>\gamma_i} \sqsubseteq \bot$
(5) $\top \sqsubseteq A_{\geq\gamma_k} \sqcup \neg A_{\geq\gamma_k}$ *(6)* $\top \sqsubseteq A_{>\gamma_i} \sqcup \neg A_{>\gamma_i}$

(1) and (2) derive from the fact that in crisp DLs $A \sqsubseteq B \equiv \neg B \sqsubseteq \neg A$. (3) and (4) come from the law of contradiction $A \sqcap \neg A \sqsubseteq \bot$, while (5) and (6) derive from the law of excluded middle $\top \sqsubseteq A \sqcup \neg A$. Moreover, we do not introduce the axiom $A_{>0} \sqsubseteq A_{\geq 0}$; since $A_{\geq 0}$ is equivalent to $\top$ the axiom trivially holds.

Mapping fuzzy concepts, roles and axioms. Concept and role expressions are reduced using mapping ρ, as shown in Table 1. Axioms are reduced as in Table 2,

where σ maps fuzzy axioms into crisp assertions and κ maps fuzzy TBox (resp. RBox) axioms into crisp TBox (resp. RBox) axioms.

Notice that $\rho(R, \lhd\gamma)$ can only appear in a (crisp) negated role assertion. Notice also that expressions of the form $\rho(A, \geq 0), \rho(A, > 1), \rho(A, \leq 1), \rho(A, < 0)$ cannot appear, because there exist some restrictions on the degree of the axioms in the fKB. The same also holds for $\top$, $\bot$ and R_A. Besides, expressions of the form $\rho(U, \lhd\gamma)$ cannot appear either. Observe that the reduction preserves simplicity of the roles and regularity of the RIAs.

Our reduction of a fuzzy GCI $\langle C \sqsubseteq D \geq 1\rangle$ is equivalent to the reduction of a GCI under a semantics based on Zadeh's set inclusion proposed in [5], although it introduces some unnecessary axioms: $C_{\geq 0} \sqsubseteq D_{\geq 0}$ and $C_{>1} \sqsubseteq D_{>1}$.

Summing up, a fKB $fK = \langle fK_A, fK_T, fK_R\rangle$ is reduced into a KB $\mathcal{K}(fK) = \langle \sigma(fK_A), T(N^{fK}) \cup \kappa(fK, fK_T), R(N^{fK}) \cup \kappa(fK, fK_R)\rangle$.

Example 1. Let us consider the following fKB: $\{\langle sym(isCloseTo)\rangle, \langle (h_1, h_2) : isCloseTo \leq 0.75\rangle\}$. Firstly, $\langle sym(isCloseTo)\rangle$ is represented as the fuzzy RIA $\langle isCloseTo \sqsubseteq isCloseTo^- \geq 1\rangle$. Now, we have to compute the number of truth values which have to be considered: $X^{fK} = \{0, 0.5, 1, 0.75\}$, so $N^{fK} = \{0, 0.25, 0.5, 0.75, 1\}$.

Next, we create some new atomic concepts and roles, as well as some axioms preserving their semantics. $T(N^{fK}) = \emptyset$ and $R(N^{fK})$ will contain the following axioms: $isCloseTo_{\geq 1} \sqsubseteq isCloseTo_{>0.75}$, $isCloseTo_{>0.75} \sqsubseteq isCloseTo_{\geq 0.75}$, $isCloseTo_{\geq 0.75} \sqsubseteq isCloseTo_{>0.5}, isCloseTo_{>0.5} \sqsubseteq isCloseTo_{\geq 0.5}, isCloseTo_{\geq 0.5} \sqsubseteq isCloseTo_{>0.25}$, $isCloseTo_{>0.25} \sqsubseteq isCloseTo_{\geq 0.25}$ and $isCloseTo_{\geq 0.25} \sqsubseteq isClose\ To_{>0}$.

Finally, we map axioms in the ABox, TBox and RBox. Firstly, $\sigma(\langle (h_1, h_2) : isCloseTo \leq 0.75\rangle) = (h_1, h_2) : \neg isCloseTo_{>0.75}$. Then, $\kappa(\langle isCloseTo \sqsubseteq isCloseTo^- \geq 1\rangle) = \{isCloseTo_{>0} \sqsubseteq isCloseTo^-_{>0}, isCloseTo_{\geq 0.25} \sqsubseteq isCloseTo^-_{\geq 0.25}, isCloseTo_{>0.25} \sqsubseteq isCloseTo^-_{>0.25}$, $isCloseTo_{\geq 0.5} \sqsubseteq isCloseTo^-_{\geq 0.5}, isCloseTo_{>0.5} \sqsubseteq isCloseTo^-_{>0.5}$, $isCloseTo_{\geq 0.75} \sqsubseteq isCloseTo^-_{\geq 0.75}, isCloseTo_{>0.75} \sqsubseteq isCloseTo^-_{>0.75}$, $isCloseTo_{\geq 1} \sqsubseteq isCloseTo^-_{\geq 1}\}$. □

Optimizing GCI reductions. GCI reductions can be optimized in several cases:

- $\langle C \sqsubseteq \top \bowtie \gamma\rangle$ and $\langle \bot \sqsubseteq D \bowtie \gamma\rangle$ are tautologies, so their reductions are unnecessary in the resulting KB.
- $\kappa(\top \sqsubseteq D \bowtie \gamma) = \top \sqsubseteq \rho(D, \bowtie \gamma)$. Note that this kind of axiom appears in role range axioms i.e. C is the range of R iff $\top \sqsubseteq \forall R.C$ holds with degree 1.
- $\kappa(C \sqsubseteq \bot \bowtie \gamma) = \rho(C, > 0) \sqsubseteq \bot$. This appears when two concepts are disjoint i.e. C and D are disjoint iff $C \sqcap D \sqsubseteq \bot$ holds with degree 1.

Another optimization involving GCIs follows from the following observation. If the resulting TBox contains $A \sqsubseteq B$, $A \sqsubseteq C$ and $B \sqsubseteq C$, then $A \sqsubseteq C$ is unnecessary. This is very useful in concept definitions involving the nominal constructor. For example, the reduction of the definition $\kappa(C \sqsubseteq \{1/o_1, 0.5/o_2\}) = \{C_{>0} \sqsubseteq \{o_1, o_2\}, C_{\geq 0.5} \sqsubseteq \{o_1, o_2\}, C_{>0.5} \sqsubseteq \{o_1\}, C_{\geq 1} \sqsubseteq \{o_1\}\}$ can be optimized to: $\{C_{>0} \sqsubseteq \{o_1, o_2\}, C_{\geq 0.5} \sqsubseteq \{o_1\}\}$.

Table 1. Mapping of concept and role expressions.

x	y	$\rho(x, y)$
$\top$	$\rhd\gamma$	$\top$
$\top$	$\lhd\gamma$	$\bot$
$\bot$	$\rhd\gamma$	$\bot$
$\bot$	$\lhd\gamma$	$\top$
A	$\geq \alpha$	$A_{\geq\alpha}$
A	$> \beta$	$A_{>\beta}$
A	$\leq \beta$	$\neg A_{>\beta}$
A	$< \alpha$	$\neg A_{\geq\alpha}$
$\neg A$	$\bowtie \gamma$	$\rho(A, \bowtie^- 1-\gamma)$
$C \sqcap D$	$\rhd\gamma$	$\rho(C, \rhd\gamma) \sqcap \rho(D, \rhd\gamma)$
$C \sqcap D$	$\lhd\gamma$	$\rho(C, \lhd\gamma) \sqcup \rho(D, \lhd\gamma)$
$C \sqcup D$	$\rhd\gamma$	$\rho(C, \rhd\gamma) \sqcup \rho(D, \rhd\gamma)$
$C \sqcup D$	$\lhd\gamma$	$\rho(C, \lhd\gamma) \sqcap \rho(D, \lhd\gamma)$
$\exists R.C$	$\rhd\gamma$	$\exists\rho(R, \rhd\gamma).\rho(C, \rhd\gamma)$
$\exists R.C$	$\lhd\gamma$	$\forall\rho(R, \neg\lhd\gamma).\rho(C, \lhd\gamma)$
$\forall R.C$	$\geq \alpha$	$\forall\rho(R, > 1-\alpha).\rho(C, \geq \alpha)$
$\forall R.C$	$> \beta$	$\forall\rho(R, \geq 1-\beta).\rho(C, > \beta)$
$\forall R.C$	$\leq \beta$	$\exists\rho(R, \geq 1-\beta).\rho(C, \leq \beta)$
$\forall R.C$	$< \alpha$	$\exists\rho(R, > 1-\alpha).\rho(C, < \alpha)$
$\{\alpha_1/o_1, \ldots, \alpha_m/o_m\}$	$\bowtie \gamma$	$\{o_i \mid \alpha_i \bowtie \gamma, 1 \leq i \leq n\}$
$\neg\{\alpha_1/o_1, \ldots, \alpha_m/o_m\}$	$\bowtie \gamma$	$\rho(\{\alpha_1/o_1, \ldots, \alpha_m/o_m\}, \bowtie^- 1-\gamma)$
$\geq 0\ S.C$	$\bowtie \gamma$	$\rho(\top, \bowtie \gamma)$
$\geq m\ S.C$	$\rhd\gamma$	$\geq m\ \rho(S, \rhd\gamma).\rho(C, \rhd\gamma)$
$\geq m\ S.C$	$\lhd\gamma$	$\leq m-1\ \rho(S, \neg\lhd\ \gamma).\rho(C, \neg\lhd\ \gamma)$
$\leq n\ S.C$	$\geq \alpha$	$\leq n\ \rho(S, > 1-\alpha).\rho(C, > 1-\alpha)$
$\leq n\ S.C$	$> \beta$	$\leq n\ \rho(S, \geq 1-\beta).\rho(C, \geq 1-\beta)$
$\leq n\ S.C$	$\leq \beta$	$\geq n+1\ \rho(S, \geq 1-\beta).\rho(C, \geq 1-\beta)$
$\leq n\ S.C$	$< \alpha$	$\geq n+1\ \rho(S, > 1-\alpha).\rho(C, > 1-\alpha)$
$\exists S.Self$	$\rhd\gamma$	$\exists\rho(S, \rhd\gamma).Self$
$\exists S.Self$	$\lhd\gamma$	$\neg\exists\rho(S, \neg\lhd\gamma).Self$
R_A	$\geq \alpha$	$R_{\geq\alpha}$
R_A	$> \beta$	$R_{>\beta}$
R_A	$\leq \beta$	$\neg R_{>\beta}$
R_A	$< \alpha$	$\neg R_{\geq\alpha}$
R^-	$\rhd\gamma$	$\rho(R, \rhd\gamma)^-$
U	$\geq \alpha$	U
U	$> \beta$	U

Table 2. Reduction of the axioms.

$\sigma(\langle a\!:\!C \bowtie \gamma\rangle)$	$a\!:\!\rho(C, \bowtie \gamma)$
$\sigma(\langle (a,b)\!:\!R \bowtie \gamma\rangle)$	$(a,b)\!:\!\rho(R, \bowtie \gamma)$
$\sigma(\langle a \neq b\rangle)$	$a \neq b$
$\sigma(\langle a = b\rangle)$	$a = b$
$\kappa(C \sqsubseteq D \geq \alpha)$	$\bigcup_{\gamma \in N^{fK} - \{0\} \mid \gamma \leq \alpha} \{\rho(C, \geq \gamma) \sqsubseteq \rho(D, \geq \gamma)\} \bigcup_{\gamma \in N^{fK} \mid \gamma < \alpha} \{\rho(C, > \gamma) \sqsubseteq \rho(D, > \gamma)\}$
$\kappa(C \sqsubseteq D > \beta)$	$\kappa(C \sqsubseteq D \geq \beta) \cup \{\rho(C, > \beta) \sqsubseteq \rho(D, > \beta)\}$
$\kappa(\langle R_1 \ldots R_n \sqsubseteq R \geq \alpha\rangle)$	$\bigcup_{\gamma \in N^{fK} - \{0\} \mid \gamma \leq \alpha} \{\rho(R_1, \geq \gamma) \ldots \rho(R_n, \geq \gamma) \sqsubseteq \rho(R, \geq \gamma)\} \bigcup_{\gamma \in N^{fK} \mid \gamma < \alpha}$ $\{\rho(R_1, > \gamma) \ldots \rho(R_n, > \gamma) \sqsubseteq \rho(R, > \gamma)\}$
$\kappa(\langle R_1 \ldots R_n \sqsubseteq R > \beta\rangle)$	$\kappa(\langle R_1 \ldots R_n \sqsubseteq R \geq \beta\rangle) \cup \{\rho(R_1, > \beta) \ldots \rho(R_n, > \beta) \sqsubseteq \rho(R, > \beta)\}$
$\kappa(dis(S_1, S_2))$	$dis(\rho(S_1, > 0), \rho(S_2, > 0))$
$\kappa(ref(R))$	$ref(\rho(R, \geq 1))$
$\kappa(asy(S))$	$asy(\rho(S, > 0)$

Theorem 1. *A $f_{KD}\mathcal{SROIQ}$ fKB fK is satisfiable iff $\mathcal{K}(fK)$ is satisfiable.*

Complexity. $|\mathcal{K}(fK)|$ is $O(|fK|^2)$ i.e. the resulting knowledge base is quadratic. The ABox is actually linear while the TBox and the RBox are both quadratic:

- $|N^{fK}|$ is linearly bounded by $|fK_A| + |fK_T| + |fK_R|$.
- $|\sigma(fK_A)| = |fK_A|$.
- $|T(N^{fK})| = 2 \cdot (|N^{fK}| - 1) \cdot |A^{fK}|$.
- $|\kappa(fK, \mathcal{T})| \leq 2 \cdot (|N^{fK}| - 1) \cdot |\mathcal{T}|$.
- $|R(N^{fK})| = 2 \cdot (|N^{fK}| - 1) \cdot |R^{fK}|$.
- $|\kappa(fK, \mathcal{R})| \leq 2 \cdot (|N^{fK}| - 1) \cdot |\mathcal{R}|$.

The resulting KB is quadratic because it depends on the number of relevant degrees $|N^{fK}|$. An immediate solution to obtain a KB which is linear in complexity is to fix the number of degrees which can appear in the knowledge base. From a practical point of view, in most of the applications it is sufficient to consider a small number of degrees, e.g. $\{0, 0.25, 0.5, 0.75, 1\}$.

It is easy to see that the complexity of the crisp representation is caused by fuzzy concepts and roles. Fortunately, in real applications not all concepts and roles will be fuzzy. Another optimization would be allowing to specify that a concept (resp. a role) is crisp. For instance, suppose that A is a fuzzy concept. Then, we need $N^{fK} - 1$ concepts of the form $A_{\geq \alpha}$ and another $N^{fK} - 1$ concepts of the form $A_{> \beta}$ to represent it, as well as $2 \cdot |N^{fK}| - 3$ axioms to preserve their semantics. On the other hand, if A is declared to be crisp, we just need one concept to represent it and no new axioms. The case for crisp roles is similar.

An interesting property of the procedure is that the reduction of an ontology can be reused when adding new axioms. In fact, for every new axiom τ, the reduction procedure generates only one new axiom or a (linear in size) set of axioms if τ does not introduce new atomic concepts nor new atomic roles and, in case τ is a fuzzy axiom, if it does not introduce a new degree of truth.

4 Implementation: DeLorean

Our prototype implementation of the reduction process is called DeLorean (DEscription LOgic REasoner with vAgueNess). It has been developed in Java with Jena API[2], the parser generator JavaCC[3], and using DIG 1.1 interface[4] to communicate with crisp DL reasoners. Currently the logic supported is $f_{KD}\mathcal{SHOIN}$ (OWL DL), since DIG interface does not yet support full $\mathcal{SROIQ}$.

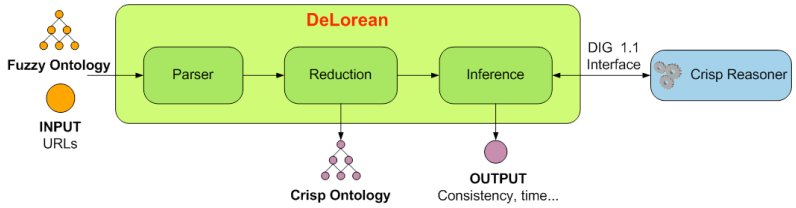


Fig. 1. Architecture of DeLorean reasoner.

Figure 1 illustrates the architecture of the system. The *Parser* reads an input file with a fuzzy ontology and translates it into an internal representation. As we have remarked in the Introduction, we could use any language to encode the fuzzy ontology, as long as the Parser can understand the representation and the reduction is properly implemented; consequently we will not get into details of our particular choice. In the next step, the *Reduction* module implements the procedure described in Section 3, building a Jena model from which an OWL file with an equivalent crisp ontology is created. Finally, the *Inference* module tests this ontology for consistency, using any crisp reasoner through the DIG interface. The *User interface* allows the user to introduce the inputs and shows the result of the reasoning and the elapsed time.

We have carried out some experiments in order to evaluate our approach in terms of reasoning, that is, in order to check that the results of the reasoning tasks over the crisp ontology were the expected. The aim of this section is not to perform a full benchmark, which could be the topic of a forthcoming work. Nevertheless, we will show some performance examples to show that our approach is feasible and the increment of time for small ontologies when using a limited number of degrees of truth is acceptable. In any case, optimizations are crucial.

We considered the Koala ontology[5], a sample $\mathcal{ALCON}(\mathcal{D})$ ontology with 20 named classes, 15 anonymous classes, 4 object properties, 1 datatype property (which we have omitted) and 6 individuals. We extended its axioms with random (lower bound) degrees and we used Pellet reasoner through the DIG interface.

[2] `http://jena.sourceforge.net/`
[3] `https://javacc.dev.java.net`
[4] `http://dl.kr.org/dig/`
[5] `http://http://protege.cim3.net/file/pub/ontologies/koala/koala.owl`

Table 3 shows the influence of the number of degrees on the time of the reduction process as well as on the time (in seconds) of a classification test over the resulting crisp ontology.

Table 3. Influence of the number of degrees in the performance of DeLorean.

Number of degrees	crisp	3	5	7	9	11
Reduction time	-	1.18	6.28	23.5	64.94	148.25
Reasoning time	0.56	0.98	1.343	2.88	4.28	6.47

It can be observed that the reduction time is quite large with respect to the reasoning time. We recall that DeLorean is currently just a prototype, so the implementation of the reduction process should be optimized. Moreover, as already discussed in the previous section, the reduction can be reused and hence needs to be computed just once. Regarding the reasoning time, the increment of complexity when the fuzzy ontology contains 3 or 5 degrees can be assumed.

5 Conclusions and Future Work

In this paper we have shown how to reduce a fuzzy extension of $\mathcal{SROIQ}$ with fuzzy GCIs and RIAs (under a novel semantics using Gödel implication) into $\mathcal{SROIQ}$. We have enhanced previous works by reducing the number of new elements and axioms. We have also presented DeLorean, our implementation of this reduction procedure which is, to the very best of our knowledge, the first reasoner supporting fuzzy $\mathcal{SHOIN}$ (and hence and eventually fuzzy OWL DL). The very preliminary experimentation shows that our approach is feasible in practice when the number of truth degrees is small, even for our non-optimized prototype. This work means an important step towards the possibility of dealing with imprecise and vague knowledge in DLs, since it relies on existing languages and tools.

In general, Gödel implication provides better logical properties than KD, but KD for example allows to reason with *modus tolens* [7]. A representation language could allow the use of two types of GCIs and RIAs $\sqsubseteq_{KD}$ y $\sqsubseteq_{G}$ (with semantics based on KD and Gödel implications respectively) similarly as [13] which allows three types of subsumption. This way, the ontology developer would be free to choose the better option for his own needs. [7] shows how to reduce GCIs under KD semantics, and RIAs can be reduced similarly.

Future work could include to compare DeLorean with other fuzzy DL reasoners, although they support different languages and features and, as far as we know, there does not exist any significant fuzzy knowledge base. We will also allow the definition of crisp concepts and roles in the fuzzy language. Finally, the reasoner will be extended to $f_{KD}\mathcal{SROIQ}$ (and hence OWL 1.1) as soon as DIG 2.0 interface is available, so it is independent of any concrete crisp reasoner.

Acknowledgements

This research has been partially supported by the project TIN2006-15041-C04-01 (Ministerio de Educación y Ciencia). Fernando Bobillo holds a FPU scholarship from Ministerio de Educación y Ciencia. Juan Gómez-Romero holds a scholarship from Consejería de Innovación, Ciencia y Empresa (Junta de Andalucía).

References

1. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.F.: The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press (2003)
2. Horrocks, I., Kutz, O., Sattler, U.: The even more irresistible $\mathcal{SROIQ}$. In: Proceedings of the 10th International Conference of Knowledge Representation and Reasoning (KR-2006). (2006) 452–457
3. Lukasiewicz, T., Straccia, U.: An overview of uncertainty and vagueness in description logics for the semantic web. Technical Report INFSYS RR-1843-06-07, Institut für Informationssysteme, Technische Universität Wien (2006)
4. Sirin, E., Cuenca-Grau, B., Parsia, B.: From wine to water: Optimizing description logic reasoning for nominals. In: Proceedings of the 10th International Conference of Knowledge Representation and Reasoning (KR-2006). (2006) 90–99
5. Straccia, U.: Transforming fuzzy description logics into classical description logics. In: Proceedings of the 9th European Conference on Logics in Artificial Intelligence (JELIA-04). Volume 3229 of Lecture Notes in Computer Science., Springer-Verlag (2004) 385–399
6. Straccia, U.: Description logics over lattices. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **14**(1) (2006) 1–16
7. Bobillo, F., Delgado, M., Gómez-Romero, J.: A crisp representation for fuzzy $\mathcal{SHOIN}$ with fuzzy nominals and general concept inclusions. In da Costa, P.C.G., Laskey, K.B., Laskey, K.J., Fung, F., Pool, M., eds.: Proceedings of the 2nd ISWC Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2006). Volume 218., CEUR Workshop Proceedings (2006)
8. Stoilos, G., Stamou, G.: Extending fuzzy description logics for the semantic web. In: Proceedings of the 3rd International Workshop on OWL: Experiences and Directions (OWLED 2007). (2007)
9. Straccia, U.: A Fuzzy Description Logic for the Semantic Web. In: Fuzzy Logic and the Semantic Web. Volume 1 of Capturing Intelligence. Elsevier Science (2006) 73–90
10. Straccia, U.: Reasoning within fuzzy description logics. Journal of Artificial Intelligence Research **14** (2001) 137–166
11. Stoilos, G., Stamou, G., Tzouvaras, V., Pan, J.Z., Horrocks, I.: Fuzzy OWL: Uncertainty and the semantic web. In: Proceedings of the International Workshop on OWL: Experience and Directions (OWLED 2005). (2005)
12. Stoilos, G., Straccia, U., Stamou, G., Pan, J.Z.: General concept inclusions in fuzzy description logics. In: Proceedings of the 17th European Conference on Artificial Intelligence (ECAI 06). (2006) 457–461
13. Ma, Y., Hitzler, P., Lin, Z.: Algorithms for paraconsistent reasoning with OWL. In Franconi, E., Kifer, M., May, W., eds.: Proceedings of the 4th European Semantic Web Conference (ESWC 2007). (2007) 399–413

Approximate Measures of Semantic Dissimilarity under Uncertainty

Nicola Fanizzi, Claudia d'Amato, Floriana Esposito

Dipartimento di Informatica
Università degli studi di Bari
Campus Universitario
Via Orabona 4, 70125 Bari, Italy
{fanizzi, claudia.damato, esposito}@di.uniba.it

Abstract. We propose semantic distance measures based on the criterion of approximate discernibility and on evidence combination. In the presence of incomplete knowledge, the distance measures the degree of belief in the discernibility of two individuals by combining estimates of basic probability masses related to a set of discriminating features. We also suggest ways to extend this distance for comparing individuals to concepts and concepts to other concepts.

1 Introduction

In the context of reasoning in the Semantic Web, a growing interest is being committed to alternative inductive procedures extending the scope of the methods that can be applied to concept representations. Among them, many are based on a notion of similarity such as *case-based reasoning* [4], *retrieval* [3], *conceptual clustering* [7] or *ontology matching* [6]. However this notion is not easily captured in a definition, especially in the presence of uncertainty.

As pointed out in the seminal paper [2] concerning similarity in Description Logics (DL), most of the existing measures focus on the similarity of atomic concepts within simple hierarchies. Besides, alternative approaches are based on related notions of *feature* similarity or *information content*. All these approaches have been specifically aimed at assessing concept similarity.

In the perspective of crafting inductive methods for the aforementioned tasks, the need for a definition of a semantic similarity measure for *individuals* arises, that is a problem that so far received less attention in the literature. Some dissimilarity measures for individuals in specific DL representations have recently been proposed which turned out to be practically effective for the targeted inductive tasks [3], however they are still partly based on structural criteria which determine also their main weakness: they can hardly scale to complex languages.

We devised a new family of dissimilarity measures for semantically annotated resources, which can overcome the aforementioned limitations [8]. Our measures are mainly based on Minkowski's measures for Euclidean spaces [18] induced by means of a method developed in the context of relational *machine learning* [14].

We extend the idea a notion of *discernibility* borrowed from *rough sets* theory [13] which aims at the formal definition of vague sets (concepts) by means of their approximations. In this paper, we propose (semi-)distance measures based on semantic discernibility and on evidence combination [16, 5, 15].

Namely, the measures are based on the degree of discernibility of the input individuals with respect to a committee of features, which are represented by concept descriptions expressed in the concept language of choice. One of the advantages of these measures is that they do not rely on a particular language for semantic annotations. However, these new measures are not to be regarded as absolute, since they depend both on the choice (and cardinality) of the features committee and on the knowledge base they are applied to. These measures can easily be computed based on statistics on individuals that are likely to be maintained by knowledge base management systems designed for storing instances (e.g. [10]), which can determine a potential speed-up in the measure computation during knowledge-intensive tasks.

Furthermore, we also propose a way to extend the presented measures to the case of assessing concept similarity by means of the notion of *medoid* [11], i.e., in a categorical context, the most centrally located individual in a concept extension w.r.t. a given metric.

The remainder of the paper is organized as follows. In the next section, we recall the basics of approximate semantic distance measures for individuals in a DL knowledge base. Hence, we extend the measures with a more principled treatment of uncertainty based on evidence combination. Conclusions discuss the applicability of these measures in further works,

2 Semantic Distance Measures

Since our method is not intended for a particular representation, in the following we assume that resources, concepts and their relationships may be defined in terms of a generic representation that may be mapped to some DL language with the standard model-theoretic semantics (see the handbook [1] for a thorough reference).

In this context, a *knowledge base* $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ contains a *TBox* $\mathcal{T}$ and an *ABox* $\mathcal{A}$. $\mathcal{T}$ is a set of concept definitions. $\mathcal{A}$ contains assertions concerning the world state. The set of the individuals (resources) occurring in $\mathcal{A}$ will be denoted with $\mathsf{Ind}(\mathcal{A})$. Each individual can be assumed to be identified by its own URI. Sometimes, it could be useful to make the *unique names assumption* on such individuals.

As regards the inference services, our procedure requires performing *instance-checking* and the related service of *retrieval*, which will be used for the approximations.

2.1 A Simple Semantic Metric for Individuals

Aiming at distance-based tasks, such as clustering or similarity search, we have developed a new measure with a definition that totally depends on semantic

aspects of the individuals in the knowledge base [8], following ideas borrowed from metric learning in clausal spaces [14].

Indeed, for our purposes, we needed functions to measure the (dis)similarity of individuals. However individuals do not have a syntactic (or algebraic) structure that can be compared. Then the underlying idea is that. on a semantic level, similar individuals should behave similarly with respect to the same concepts. A way for assessing the similarity of individuals in a knowledge base can be based on the comparison of their semantics along a number of dimensions represented by a set of concept descriptions (henceforth referred to as the *committee*). Particularly, the rationale of the measure is to compare them on the grounds of their behavior w.r.t. a given set of concept descriptions, say $\mathsf{F} = \{F_1, F_2, \ldots, F_k\}$, which stands as a group of discriminating *features* expressed in the language taken into account.

We begin with defining the behavior of an individual w.r.t. a certain concept in terms of projecting it in this dimension:

Definition 2.1 (projection function). *Given a concept $F_i \in \mathsf{F}$, the related* projection function

$$\pi_i : \mathsf{Ind}(\mathcal{A}) \mapsto \{0, 1/2, 1\}$$

is defined:
$\forall a \in \mathsf{Ind}(\mathcal{A})$

$$\pi_i(a) := \begin{cases} 1 & \mathcal{K} \models F_i(a) \\ 0 & \mathcal{K} \models \neg F_i(a) \\ 1/2 & \textit{otherwise} \end{cases}$$

The case of $\pi_i(a) = 1/2$ corresponds to the case when a reasoner cannot give the truth value for a certain membership query. This is due to the *Open World Assumption* normally made in this context. Hence, as in the classic probabilistic models uncertainty is coped with by considering a uniform distribution over the possible cases.

Now the discernibility functions related to the committee concepts which compare the two input individuals w.r.t. these concepts through their projections:

Definition 2.2 (discernibility function). *Given a feature concept $F_i \in \mathsf{F}$, the related* discernibility function

$$\delta_i : \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A}) \mapsto [0, 1]$$

is defined as follows:
$\forall (a, b) \in \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A})$

$$\delta_i(a, b) = |\pi_i(a) - \pi_i(b)|$$

Finally, a whole family of distance functions for individuals inspired to Minkowski's distances L_p [18] can be defined as follows [8]:

Definition 2.3 (dissimilarity measures). *Let $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ be a knowledge base. Given a set of concept descriptions $\mathsf{F} = \{F_1, F_2, \ldots, F_k\}$, a family of dissimilarity measures $\{d_p^{\mathsf{F}}\}_{p \in \mathbb{N}}$, contains functions*

$$d_p^{\mathsf{F}} : \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A}) \mapsto [0, 1]$$

defined
$\forall (a, b) \in \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A})$:

$$d_p^{\mathsf{F}}(a, b) := \frac{L_p(\pi_i(a), \pi_i(b))}{|\mathsf{F}|} = \frac{1}{k} \sqrt[p]{\sum_{i=1}^{k} \delta_i(a, b)^p}$$

Note that k depends on F and the effect of the factor $1/k$ is just to normalize the norms w.r.t. the number of features that are involved. It is worthwhile to recall that these measures are not absolute, then they should be also be considered w.r.t. the committee of choice, hence comparisons across different committees may not be meaningful. Larger committees are likely to decrease the measures because of the normalizing factor yet these values is affected also by the degree of redundancy of the features employed.

2.2 Example

Let us consider a knowledge base in a DL language made up of a TBox:
$T = \{$ $\mathsf{Female} \equiv \neg \mathsf{Male}$,
$\mathsf{Parent} \equiv \forall \mathsf{child}.\mathsf{Being} \sqcap \exists \mathsf{child}.\mathsf{Being}$,
$\mathsf{Father} \equiv \mathsf{Male} \sqcap \mathsf{Parent}$,
$\mathsf{FatherWithoutSons} \equiv \mathsf{Father} \sqcap \forall \mathsf{child}.\mathsf{Female}$
$\}$
and of an ABox:
$A = \{$ $\mathsf{Being}(\mathsf{ZEUS}), \mathsf{Being}(\mathsf{APOLLO}), \mathsf{Being}(\mathsf{HERCULES}), \mathsf{Being}(\mathsf{HERA})$,
$\mathsf{Male}(\mathsf{ZEUS}), \mathsf{Male}(\mathsf{APOLLO}), \mathsf{Male}(\mathsf{HERCULES})$,
$\mathsf{Parent}(\mathsf{ZEUS}), \mathsf{Parent}(\mathsf{APOLLO}), \neg\mathsf{Father}(\mathsf{HERA})$,
$\mathsf{God}(\mathsf{ZEUS}), \mathsf{God}(\mathsf{APOLLO}), \mathsf{God}(\mathsf{HERA}), \neg\mathsf{God}(\mathsf{HERCULES})$,
$\mathsf{hasChild}(\mathsf{ZEUS}, \mathsf{APOLLO}), \mathsf{hasChild}(\mathsf{HERA}, \mathsf{APOLLO})$,
$\mathsf{hasChild}(\mathsf{ZEUS}, \mathsf{HERCULES})$,
$\}$
Suppose $\mathsf{F} = \{F_1, F_2, F_3, F_4\} = \{\mathsf{Male}, \mathsf{God}, \mathsf{Parent}, \mathsf{FatherWithoutSons}\}$. Let us compute the distances (with $p = 1$):
$d_1^{\mathsf{F}}(\mathsf{ZEUS}, \mathsf{HERA}) = (|1-0| + |1-1| + |1-1| + |0-0|)\,/4 = 1/4$
$d_1^{\mathsf{F}}(\mathsf{HERA}, \mathsf{APOLLO}) = (|0-1| + |1-1| + |1-1| + |0-1/2|)\,/4 = 3/8$
$d_1^{\mathsf{F}}(\mathsf{APOLLO}, \mathsf{HERCULES}) = (|1-1| + |1-0| + |1-1/2| + |1/2-1/2|)\,/4 = 3/8$
$d_1^{\mathsf{F}}(\mathsf{HERCULES}, \mathsf{ZEUS}) = (|1-1| + |0-1| + |1/2-1| + |1/2-0|)\,/4 = 1/2$
$d_1^{\mathsf{F}}(\mathsf{HERA}, \mathsf{HERCULES}) = (|0-1| + |1-0| + |1-1/2| + |0-1/2|)\,/4 = 3/4$
$d_1^{\mathsf{F}}(\mathsf{APOLLO}, \mathsf{ZEUS}) = (|1-1| + |1-1| + |1-1| + |1/2-0|)\,/4 = 1/8$

2.3 Discussion

It is easy to prove that these dissimilarity functions have the standard properties for semi-distances [8]:

Proposition 2.1 (semi-distance). *For a fixed feature set* F *and* $p > 0$*, given any three instances* $a, b, c \in \mathsf{Ind}(\mathcal{A})$*. it holds that:*

1. $d_p^{\mathsf{F}}(a, b) \geq 0$ *and* $d_p^{\mathsf{F}}(a, b) = 0$ *if* $a = b$
2. $d_p^{\mathsf{F}}(a, b) = d_p^{\mathsf{F}}(b, a)$
3. $d_p^{\mathsf{F}}(a, c) \leq d_p^{\mathsf{F}}(a, b) + d_p^{\mathsf{F}}(b, c)$

This measure is not a distance since it does not hold that $a = b$ if $d_p^{\mathsf{F}}(a, b) = 0$. This is the case of *indiscernible* individuals with respect to the given committee of features F. However, if the *unique names assumption* were made then one may define a supplementary dimension for the committee (a sort of meta-feature F_0) based on equality, such that:
$\forall (a, b) \in \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A})$

$$\delta_0(a, b) := \begin{cases} 0 & a = b \\ 1 & a \neq b \end{cases}$$

and then

$$d_p^{\mathsf{F}}(a, b) := \frac{1}{k+1} \sqrt[p]{\sum_{i=0}^{k} \delta_i(a, b)^p}$$

The resulting measures are distance measures.

Compared to other proposed dissimilarity measures [2, 3], the presented functions do not depend on the constructors of a specific language, rather they require only (retrieval or) instance-checking for computing the projections through class-membership queries to the knowledge base.

The complexity of measuring he dissimilarity of two individuals depends on the complexity of such inferences (see [1], Ch. 3). Note also that the projections that determine the measure can be computed (or derived from statistics maintained on the knowledge base) before the actual distance application, thus determining a speed-up in the computation of the measure. This is very important for algorithms that massively use this distance, such as all instance-based methods.

So far we made the assumption that F may represent a sufficient number of (possibly redundant) features that are able to discriminate really different individuals. The choice of the concepts to be included – *feature selection* – may be crucial. Therefore, we have devised specific optimization algorithms founded in *randomized search* which are able to find optimal choices of discriminating concept committees [8, 7].

The fitness function to be optimized is based on the *discernibility factor* [13] of the committee. Given the whole set of individuals $\mathsf{Ind}(\mathcal{A})$ (or just a hold-out sample to be used to induce an optimal measure) $HS \subseteq \mathsf{Ind}(\mathcal{A})$ the fitness

function to be maximized is:

$$\text{DISCERNIBILITY}(\mathsf{F}, HS) := \sum_{(a,b)\in HS^2} \sum_{i=1}^{k} \delta_i(a,b)$$

However, the results obtained so far with knowledge bases drawn from ontology libraries [7, 9] show that (a selection) of the (primitive and defined) concepts is often sufficient to induce satisfactory dissimilarity measures.

3 Dissimilarity Measures Based on Uncertainty

The measure defined in the previous section deals with uncertainty in a uniform way: in particular, the degree of discernibility of two individuals is null when they have the same behavior w.r.t. the same feature, even in the presence of total uncertainty of class-membership for both. When uncertainty regards only one projection, then they are considered partially (possibly) similar.

We would like to make this uncertainty more explicit[1]. One way to deal with uncertainty would be considering intervals rather than numbers in [0,1] as a measure of dissimilarity. This is similar to the case of imprecise probabilities [17].

In order to extend the measure, we propose an epistemic definition based on rules for combining evidence [5, 15]. The ultimate aim is to assess the distance between two individuals as a combination of the evidence that they differ based on some selected features (as in the previous section).

The distance measure that is to be defined is again based on the degree of belief of discernibility of individuals w.r.t. such features. To this purpose the probability masses of the basic events (class-membership) have to be assessed. However, in this case we will not treat uncertainty in the classic probabilistic way (uniform probability). Rather, we intend to take into account uncertainty in the computation.

The new dissimilarity measure will be derived as a combination of the degree of belief in the discernibility of the individuals w.r.t. each single feature. Before introducing the combination rule (that will have the measure as a specialization), the basic probability assignments have to be considered, especially for the cases when instance-checking is not able to provide a certain answer.

As in previous works [3], we may estimate the concept extensions recurring to their retrieval [1], i.e. the individuals of the ABox that can be proved to belong to a concept. Thus, in case of uncertainty, the basic probabilities masses for each feature concept, can be approximated[2] in the following way:

[1] We are referring to a notion of *epistemic* (rather than *aleatory*) probability [15], which seems more suitable for our purposes. See Shafer's introductory chapter in [16] on this distinction.

[2] In case of a certain answer received from the reasoner, the probability mass amounts to 0 or 1.

$\forall i \in \{1, \ldots, k\}$

$$m_i(\mathcal{K} \models F_i(a)) \approx |\mathsf{retrieval}(F_i, \mathcal{K})|/|\mathsf{Ind}(\mathcal{A})|$$
$$m_i(\mathcal{K} \models \neg F_i(a)) \approx |\mathsf{retrieval}(\neg F_i, \mathcal{K})|/|\mathsf{Ind}(\mathcal{A})|$$
$$m_i(\mathcal{K} \models F_i(a) \vee \mathcal{K} \models \neg F_i(a)) \approx 1 - m_i(\mathcal{K} \models F_i(a)) - m_i(\mathcal{K} \models \neg F_i(a))$$

where the $\mathsf{retrieval}(\cdot, \cdot)$ operator returns the individuals which can be proven to be members of the argument concept in the context of the current knowledge base [1]. The rationale is that the larger the (estimated) extension the more likely is for individuals to belong to the concept. These approximated probability masses become more precise as more information is acquired. Alternatively, these masses could come with the ontology as a supplementary for of prior knowledge.

As in the previous section, we define a discernibility function related to a fixed concept which measures the amount of evidence that two input individuals may be separated by that concept:

Definition 3.1 (discernibility function). *Given a feature concept $F_i \in \mathsf{F}$, the related* discernibility function

$$\delta_i : \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A}) \mapsto [0, 1]$$

is defined as follows:
$\forall (a, b) \in \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A})$

$$\delta_i(a, b) := \begin{cases} m_i(\mathcal{K} \models \neg F_i(b)) & \text{if } \mathcal{K} \models F_i(a) \\ m_i(\mathcal{K} \models F_i(b)) & \text{else if } \mathcal{K} \models \neg F_i(a) \\ \delta_i(b, a) & \text{else if } \mathcal{K} \models F_i(b) \vee \mathcal{K} \models \neg F_i(b) \\ 2 \cdot m_i(\mathcal{K} \models F_i(a)) \cdot m_i(\mathcal{K} \models \neg F_i(b)) & \text{otherwise} \end{cases}$$

The extreme values $\{0, 1\}$ are returned when the answers from the instance-checking service are certain for both individuals. If the first individual is an instance of the i-th feature (resp., its complement) then the discernibility depends on the belief of class-membership to the complement concept of the other individual. Otherwise, if there is uncertainty for the former individual but not for the latter, the function changes its perspective, swapping the roles of the two individuals. Finally, in case there were uncertainty for both individuals, the discernibility is computed as the chance that they may belong one to the feature concept and one to its complement,

The combined degree of belief in the case of discernible individuals, assessed using the *mixing combination rule* [12, 15], can give a measure of the semantic distance between them.

Definition 3.2 (weighted average measure). *Given an ABox $\mathcal{A}$, a dissimilarity measure for the individuals in $\mathcal{A}$*

$$d^{\mathsf{F}}_{avg} : \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A}) \mapsto [0, 1]$$

is defined as follows:
$\forall (a,b) \in \mathsf{Ind}(\mathcal{A}) \times \mathsf{Ind}(\mathcal{A})$

$$d^{\mathsf{F}}_{avg}(a,b) := \sum_{i=1}^{k} w_i \delta_i(a,b)$$

The choices for the weights are various. The most straightforward one is, of course, considering uniform weights: $w_i = 1/k$. Another one is

$$w_i = \frac{u_i}{u}$$

where

$$u_i = \frac{1}{|\mathsf{Ind}(A) \setminus \mathsf{retrieval}(F_k, \mathcal{K})|} \text{ and } u = \sum_{i=1}^{k} u_i$$

It is easy to see that this can be considered as a generalization of the measure defined in the previous section (for $p = 1$).

3.1 Discussion

It can be proved that function has the standard properties for semi-distances:

Proposition 3.1 (semi-distance). *For a fixed choice of weights $\{w_i\}_{i=1}^{k}$, function d^{F}_{avg} is a semi-distance.*

The underlying idea for the measure is to combine the belief of the dissimilarity of the two input individuals provided by several sources, that are related to the feature concepts. In the original framework for evidence composition the various sources are supposed to be independent, which is generally unlikely to hold. Yet, from a practical viewpoint, overlooking this property for the sake of simplicity may still lead to effective methods, as the Naïve Bayes approach in Machine Learning demonstrates.

It could also be criticized that the subsumption hierarchy has not been explicitly involved. However, this may be actually yielded as a side-effect of the possible partial redundancy of the various concepts, which has an impact on their extensions and thus on the related projection function. A tradeoff is to be made between the number of features employed and the computational effort required for computing the related projection functions.

The discriminating power of each feature concept can be weighted in terms of information and entropy measures. Namely, the degree of information yielded by each of these features can be estimated as follows:

$$H_i(a,b) = -\sum_{A \subseteq \Theta} m_i(A) \log(m_i(A))$$

where 2^{Θ}, w.r.t. the *frame of discernment*[3] [16, 15] $\Theta = \{D, \overline{D}\}$. then, the sum

$$\sum_{(a,b) \in HS} H_i(a, b)$$

provides a measure of the utility of the discernibility function related to each feature which can be used in randomized optimization algorithms.

3.2 Extensions

Following the rationale of the average link criterion used in agglomerative clustering [11], the measures can be extended to the case of concepts, by recurring to the notion of medoids.

The *medoid* of a group of individuals is the individual that has the highest similarity w.r.t. the others. Formally. given a group $G = \{a_1, a_2, \ldots, a_n\}$, the medoid is defined:

$$\text{medoid}(G) = \operatorname*{argmin}_{a \in G} \sum_{j=1}^{n} d(a, a_j)$$

Now, given two concepts C_1, C_2, we can consider the two corresponding groups of individuals obtained by retrieval $R_i = \{a \in \mathsf{Ind}(\mathcal{A}) \mid \mathcal{K} \models C_i(a)\}$, and their resp. medoids $m_i = \text{medoid}(R_i)$ for $i = 1, 2$ w.r.t. a given measure d_p^{F} (for some $p > 0$ and committee F). Then the function for concepts can be defined as follows:

$$d_p^{\mathsf{F}}(C_1, C_2) := d_p^{\mathsf{F}}(m_1, m_2)$$

Similarly, if the distance of an individual a to a concept C has to be assessed, one could consider the nearest (resp. farthest) individual in the concept extension or its medoid. Let $m = \text{medoid}(\mathsf{retrieval}(C))$ w.r.t. a given measure d_p^{F}. Then the measure for this case can be defined as follows:

$$d_p^{\mathsf{F}}(a, C) := d_p^{\mathsf{F}}(a, m)$$

Of course these approximate measures become more and more precise as the knowledge base is populated with an increasing number of individuals.

4 Concluding Remarks

We have proposed the definition of dissimilarity measures over spaces of individuals in a knowledge base. The measures are not language-dependent, differently from other previous proposals [3], yet they are parameterized on a committee of concepts. Optimal committees can be found via randomized search methods [8].

[3] Here D stands for the case of discernible individuals w.r.t. F_i, $\overline{D}$ for the opposite case, and some probability mass may be assigned also to the uncertain case represented by $\{D, \overline{D}\}$.

Besides, we have extended the measures to cope with cases of uncertainty by means of a simple evidence combination method.

One of the advantages of the measures is that their computation can be very efficient in cases when statistics (on class-membership) are maintained by the KBMS [10]. As previously mentioned, the subsumption relationships among concepts in the committee is not explicitly exploited in the measure for making the relative distances more accurate. The extension to the case of concept distance may also be improved. Hence, scalability should be guaranteed as far as a good committee has been found and does not change also because of the locality properties observed for instances in several domains (e.g. social or biological networks).

A refinement of the committee may become necessary only when a degradation of the discernibility factor is detected due to the availability of somewhat new individuals. This may involve further tasks such as *novelty* or *concept drift* detection.

4.1 Applications

The measures have been integrated in an instance-based learning system implementing a nearest-neighbor learning algorithm: an experimentation on performing semantic-based retrieval proved the effectiveness of the new measures, compared to the outcomes obtained adopting other measures [3]. It is worthwhile to mention that results where not particularly affected by feature selection: often using the very concepts defined in the knowledge base provides good committees which are able to discern among the different individuals [9].

We are also exploiting the implementation of these measures for performing conceptual clustering [11], where (a hierarchy of) clusters is created by grouping instances on the grounds of their similarity, possibly triggering the induction of new emerging concepts [7].

4.2 Extensions

The measure may have a wide range of application of distance-based methods to knowledge bases. For example, logic approaches to ontology matching [6] may be backed up by the usage of our measures, especially when concepts to be matched across different terminologies are known to share a common set of individuals. Ontology matching could be a phase in a larger process aimed at data integration. Moreover metrics could also support a process of *(semi-)automated classification* of new data also as a first step towards ontology evolution.

Another problem that could be tackled by means of dissimilarity measures could be the *ranking* of the answers provided by a *matchmaking* algorithm based on the similarity between the concept representing the query and the retrieved individuals.

Acknowledgments

The authors would like to thank the anonymous reviewers for their suggestions.

References

[1] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic Handbook*. Cambridge University Press, 2003.

[2] A. Borgida, T.J. Walsh, and H. Hirsh. Towards measuring similarity in description logics. In I. Horrocks, U. Sattler, and F. Wolter, editors, *Working Notes of the International Description Logics Workshop*, volume 147 of *CEUR Workshop Proceedings*, Edinburgh, UK, 2005.

[3] C. d'Amato, N. Fanizzi, and F. Esposito. Reasoning by analogy in description logics through instance-based learning. In G. Tummarello, P. Bouquet, and O. Signore, editors, *Proceedings of Semantic Web Applications and Perspectives, 3rd Italian Semantic Web Workshop, SWAP2006*, volume 201 of *CEUR Workshop Proceedings*, Pisa, Italy, 2006.

[4] M. d'Aquin, J. Lieber, and A. Napoli. Decentralized case-based reasoning for the Semantic Web. In Y. Gil, E. Motta, V. Benjamins, and M. A. Musen, editors, *Proceedings of the 4th International Semantic Web Conference, ISWC2005*, number 3279 in LNCS, pages 142–155. Springer, 2005.

[5] D. Dubois and H. Prade. On the combination of evidence in various mathematical frameworks. In J. Flamm and T. Luisi, editors, *Reliability Data Collection and Analysis*, pages 213–241. 1992.

[6] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer, 2007.

[7] N. Fanizzi, C. d'Amato, and F. Esposito. Evolutionary conceptual clustering of semantically annotated resources. In *Proceedings of the 1st International Conference on Semantic Computing, IEEE-ICSC2007*, pages 783–790, Irvine, CA, 2007. IEEE Computer Society Press.

[8] N. Fanizzi, C. d'Amato, and F. Esposito. Induction of optimal semi-distances for individuals based on feature sets. In *Working Notes of the International Description Logics Workshop, DL2007*, volume 250 of *CEUR Workshop Proceedings*, Bressanone, Italy, 2007.

[9] N. Fanizzi, C. d'Amato, and F. Esposito. Instance-based query answering with semantic knowledge bases. In R. Basili and M.T. Pazienza, editors, *Proceedings of the 10th Congress of the Italian Association for Artificial Intelligence, AI*IA2007*, volume 4733 of *LNAI*, pages 254–265, Rome, Italy, 2007. Springer.

[10] I. R. Horrocks, L. Li, D. Turi, and S. K. Bechhofer. The Instance Store: DL reasoning with large numbers of individuals. In V. Haarslev and R. Möller, editors, *Proceedings of the 2004 Description Logic Workshop, DL 2004*, volume 104 of *CEUR Workshop Proceedings*, pages 31–40. CEUR, 2004.

[11] A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[12] C. K. Murphy. Combining belief functions when evidence conflicts. *Decision Support Systems*, 29(1-9), 2000.

[13] Z. Pawlak. *Rough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publishers, 1991.

[14] M. Sebag. Distance induction in first order logic. In S. Džeroski and N. Lavrač, editors, *Proceedings of the 7th International Workshop on Inductive Logic Programming, ILP97*, volume 1297 of *LNAI*, pages 264–272. Springer, 1997.

[15] K. Sentz and S. Ferson. Combination of evidence in Dempster-Shafer theory. Technical Report SAND2002-0835, SANDIATech, April 2002.
[16] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.
[17] P. Walley. *Statistical reasoning with imprecise probabilities*. Chapman and Hall, London, 1991.
[18] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search – The Metric Space Approach*. Advances in Database Systems. Springer, 2007.

Using the Dempster-Shafer theory of evidence to resolve ABox inconsistencies

Andriy Nikolov, Victoria Uren, Enrico Motta and Anne de Roeck

Knowledge Media Institute, The Open University, Milton Keynes, UK
{a.nikolov, v.s.uren, e.motta, a.deroeck}@open.ac.uk

Abstract. Automated ontology population using information extraction algorithms can produce inconsistent knowledge bases. Confidence values assigned by the extraction algorithms may serve as evidence helping to repair produced inconsistencies. The Dempster-Shafer theory of evidence is a formalism, which allows appropriate interpretation of extractors' confidence values. The paper presents an algorithm for translating the subontologies containing conflicts into belief propagation networks and repairing conflicts based on the Dempster-Shafer plausibility.

1 Introduction

One of the approaches for ontology population considers using automatic information extraction algorithms to annotate natural language data already available on the Web [1, 2]. Unsupervised information extraction algorithms do not produce 100% correct output, which may lead to inconsistency of the whole knowledge base produced in this way. Also information extracted from different sources can be genuinely contradictory. So when performing fusion of knowledge extracted from different sources it is important to resolve such inconsistencies automatically or provide the user with a ranking of conflicting options estimating how likely each statement is to be wrong. Extraction algorithms can often estimate the reliability of their output by attaching confidence values to produced statements [3]. Uncertain reasoning using these confidence values can help to evaluate the plausibility of statements and rank the conflicting options. Most of the ongoing research in the field of applying uncertain reasoning to the Semantic Web focuses on fuzzy logic and probabilistic approaches. Fuzzy logic was designed to deal with representation of vagueness and imprecision. This interpretation is not relevant for the problem occurring during population of crisp OWL knowledge bases, where we need to assess the likelihood for a statement to be true or false. The probabilistic approach is more appropriate for dealing with such problems. However, as stated in [4], axioms of probability theory are implied by seven properties of belief measures. One of them is *completeness*, which states that "a degree of belief can be assigned to any well-defined proposition". However, this property cannot be ensured when dealing with confidence degrees assigned by extractors, because they do not always carry information about the probability of a statement being *false*. The Dempster-Shafer theory of

evidence [5] presents a formalism that helps to overcome this problem. It allows belief measurements to be assigned to sets of propositions, thus specifying explicitly degrees of ignorance. In this paper, we describe an algorithm for resolving conflicts using the Dempster-Shafer belief propagation approach.

2 Related Work

There are several studies dealing with inconsistency handling in OWL ontologies, among others [6] and [7]. The general algorithm for the task of repairing inconsistent ontologies consists of two steps:

- Ontology diagnosis: finding sets of axioms, which contribute to inconsistency;
- Repairing inconsistencies: changing/removing the axioms most likely to be erroneous.

Choosing the axioms for change and removal is a non-trivial task. Existing algorithms working with crisp ontologies (e.g., [7]) utilize such criteria as syntactic relevance (how often each entity is referenced in the ontology), impact (the influence of removal of the axiom on the ontology should be minimized) and provenance (reliability of the source of the axiom). The last criterion is especially interesting for the automatic ontology population scenario since extraction algorithms do not extract information with 100% accuracy. A study described in [8] specifies an algorithm which utilizes the confidence value assigned by the extraction algorithm. The strategy employed there was to order the axioms according to their confidence and add them incrementally, starting from the most certain one. If adding the axiom led to inconsistency of the ontology then a minimal subconsistent ontology was determined and the axiom with the lowest confidence was removed from it. A disadvantage of such a technique is that it does not take into account the impact of an axiom: e.g., when an axiom violates several restrictions, it does not increase its chances to be removed. Also it does not consider the influence of redundancy: if the same statement was extracted from several sources, this should increase its reliability. Using uncertain reasoning would provide a more sound approach to rank potentially erroneous statements and resolve inconsistencies.
In the Semantic Web domain the studies on uncertain reasoning are mostly focused on two formalisms: probability theory and fuzzy logic. Existing implementations of fuzzy description logic [9, 10] are based on the notion of fuzzy set representing a vague concept. The uncertainty value in this context denotes a membership function $\mu_F(x)$ which specifies the degree to which an object x belongs to a fuzzy class F. Probabilistic adaptations of OWL-DL include Bayes OWL[11] and PR-OWL [12]. However, both of these formalisms do not fully reflect the properties of the problems we are dealing with in the fusion scenario. In [4] a framework for choosing an appropriate uncertainty handling formalism was presented. The framework is based on the following seven properties of belief measurements:

1. *Clarity*: Propositions should be well-defined.
2. *Scalar continuity*: A single real number is both necessary and sufficient for representing a degree of belief.
3. *Completeness*: A degree of belief can be assigned to any well-defined proposition.
4. *Context dependency*: The belief assigned to a proposition can depend on the belief in other propositions.
5. *Hypothetical conditioning*: There exists some function that allows the belief in a conjunction of propositions to be calculated from the belief in one proposition and the belief in the other proposition given that the first proposition is true.
6. *Complementarity*: The belief in the negation of a proposition is a monotonically decreasing function of the belief in the proposition itself.
7. *Consistency*: There will be equal belief in propositions that have the same truth value.

It was proven that accepting all seven properties logically necessitates the axioms of probability theory. Alternative formalisms allow weakening of some properties. Fuzzy logic deals with the case when the *clarity* property does not hold, i.e., when concepts and relations are vague. Such an interpretation differs from the one we are dealing with in the fusion scenario, where the ontology TBox contains crisp concepts and properties. Confidence value attached to a type assertion $ClassA(Individual1)$ denotes a degree of belief that the statement is true in the real world rather than the degree of inclusion of the entity $Individual1$ into a fuzzy concept $ClassA$. This makes fuzzy interpretation inappropriate for our case.

Probabilistic interpretation of the extraction algorithm's confidence may lead to a potential problem. If we interpret the confidence value c attached to a statement returned by an extraction algorithm as a Bayesian probability value p, we, at the same time, introduce a belief that the statement is false with a probability $1 - p$. However, the confidence of an extraction algorithm reflects only the belief that the document supports the statement and does not itself reflect the probability of a statement being false in the real world. Also while statistical extraction algorithms ([13]) are able to assign a degree of probability to each extracted statement, rule-based algorithms ([14, 15]) can only assign the same confidence value to all statements extracted by the same rule based on the rule's performance on some evaluation set. Any extraction produced by a rule with a low confidence value in this case will serve as a negative evidence rather than simply lack of evidence. This issue becomes more important if the reliability of sources is included into analysis: it is hard to assign the conditional probability of a statement being false given that the document supports it. It means that the *completeness* property does not always hold.

The Dempster-Shafer theory of evidence [5] allows weakening of the completeness property. Belief can be assigned to sets of alternative options rather than only to atomic elements. In the case of binary logic, it means that the degree of ignorance can be explicitly represented by assigning a non-zero belief to the

set {true;false}. On the other hand, it still allows the Bayesian interpretation of confidence to be used, when it is appropriate (in this case the belief assigned to the set {true;false} is set to 0). This paper presents an algorithm for resolving inconsistencies by translating the inconsistency-preserving subset of ontology into the Dempster-Shafer belief network and choosing the axioms to remove based on their plausibility. We are not aware of other studies adapting the Dempster-Shafer approach to the Semantic Web domain.
Alternative approaches to uncertainty representation, which were not applied so far to ontological modelling, include probability intervals [16] and higher-order probability [17]. However, the first of these approaches uses min and max operators for aggregation, which makes it hard to represent cumulative evidence, and the second focuses on resolving different kinds of problems (namely expressing probability estimations of other probability estimations). There are also other approaches to belief fusion in the Semantic Web (e.g., [18] and [19]). These studies deal with social issues of representing trust and provenance in a distributed knowledge base and focus on the problem of establishing the certainty of statements asserted by other people. These approaches, however, do not focus on resolving the inconsistencies and just deal with direct conflicts (i.e., statement A is true vs statement A is false). They do not take into account ontological inference and mutual influence of statements in the knowledge base. In this way, they can be considered complementary to ours.

3 The Dempster-Shafer Belief Theory

Dempster-Shafer theory of evidence differs from the Bayesian probability theory as it allows assigning beliefs not only to atomic elements but to sets of elements as well. The base of the Dempster's belief distribution is the frame of discernment (Ω) - an exhaustive set of mutually exclusive alternatives. A belief distribution function (also called mass function or belief potential) $m(A)$ represents the influence of a piece of evidence on subsets of Ω and has the following constraints:

- $m(\oslash) = 0$ and
- $\sum_{A \subseteq \Omega} m(A) = 1$

$m(A)$ defines the amount of belief assigned to the subset A. When $m(A) > 0$, A is referred to as a focal element. If each focal element A contains only a single element, the mass function is reduced to be a probability distribution. Mass also can be assigned to the whole set of Ω. This represents the uncertainty of the piece of evidence about which of the elements in Ω is true. In our case each mass function is defined on a set of variables $D = \{x_1, ..., x_n\}$ called the domain of m. Each variable is boolean and represents an assertion in the knowledge base. For a single variable we can get degree of support $Sup(x) = m(\{true\})$ and degree of plausibility $Pl(x) = m(\{true\}) + m(\{true; false\})$. Plausibility specifies how likely it is that the statement is false. Based on plausibility it is possible to select from a set of statements the one to be removed.

4 Building Belief Networks

Our algorithm consists of four steps:

1. Inconsistency detection.
   At this stage we select the subontology containing all axioms contributing to an inconsistency.
2. Constructing a belief network.
   At this stage the subontology found at the previous step is translated into a belief network.
3. Assigning mass distributions.
   At this stage we assign mass distribution functions to nodes.
4. Belief propagation.
   At this stage we propagate uncertainties through the network and update the confidence degrees of ABox statements.

4.1 Illustrating Example

In order to illustrate our algorithm, we use an example from the banking domain. Supposedly, we have an ontology describing credit card applications, which defines two disjoint classes of applicants: reliable and risky. In order to be reliable, an applicant has to have UK citizenship and evidence that (s)he was never bankrupt in the past. For example, the TBox contains the following axioms:
T1: $RiskyApplicant \sqsubseteq CreditCardApplicant$
T2: $ReliableApplicant \sqsubseteq CreditCardApplicant$
T3: $RiskyApplicant \sqsubseteq \neg ReliableApplicant$
T4: $ReliableApplicant \equiv \exists wasBankrupt.False \sqcap \exists hasCitizenship.UK$
T5: $\top \sqsubseteq \leq 1 wasBankrupt$ ($wasBankrupt$ is functional)
The ABox contains the following axioms (with attached confidence values):
A1: $RiskyApplicant(Ind1) : 0.7$
A2: $wasBankrupt(Ind1, False) : 0.6$
A3: $hasCitizenship(Ind1, UK) : 0.4$
A4: $wasBankrupt(Ind1, True) : 0.5$
As given, the ontology is inconsistent: the individual $Ind1$ is forced to belong to both classes $RiskyApplicant$ and $ReliableApplicant$, which are disjoint, and the functional property $wasBankrupt$ has two different values. If we choose to remove the axioms with the lowest confidence values, it will require removing A3 and A4. However, inconsistency can also be repaired by removing a single statement A2. The fact that A2 leads to the violation of two ontological constraints should increase the likelihood it is wrong.

4.2 Inconsistency Detection

The task of the inconsistency detection step is to retrieve all minimal inconsistent subontologies (MISO) of the ontology and combine them. As defined in [6], an ontology O' is a minimal inconsistent subontology of an ontology O, if $O' \subseteq O$

and O' is inconsistent and for all O'' such that $O'' \subset O' \subseteq O$, O'' is consistent. OWL reasoner Pellet [7] is able to return the MISO for the first encountered inconsistency in the ontology. To calculate all MISO ${O'}_1, ..., {O'}_n$ in the ontology we employ Reiter's hitting set tree algorithm [20]. After all conflict sets were identified, the next step involves constructing belief networks from each set. If for two subontologies ${O'}_i \cap {O'}_j \neq \oslash$ then these two subontologies are replaced with $O' = {O'}_i \cup {O'}_j$.
For our illustrating example, the conflict detection algorithm is able to identify two conflict sets in this ontology: the first, consisting of {T3, T4, A1, A2, A3} (individual $Ind1$ belongs to two disjoint classes), and the second {T5, A2, A4} (individual $Ind1$ has two instantiations of a functional property). The statement A2 belongs to both sets and therefore the sets are merged.

4.3 Constructing Belief Networks

The networks for propagation of Dempster-Shafer belief functions (also called valuation networks) were described in [21]. By definition the valuation network is an undirected graph represented as a 5-tuple: $\{\Psi, \{\Omega_X\}_{X \in \Psi}, \{\tau_1, ..., \tau_n\}, \downarrow, \otimes\}$, where Ψ is a set of variables, $\{\Omega_X\}_{X \in \Psi}$ is a collection of state spaces, $\{\tau_1, ..., \tau_n\}$ is a collection of valuations (belief potentials of nodes), $\downarrow$ is a marginalization operator and $\otimes$ is a combination operator. In our case Ψ consists of ABox assertions, every $\{\Omega_X\}_{X \in \Psi} = \{0; 1\}$ and $\{\tau_1, ..., \tau_n\}$ are created using rules described below. The operators are used for propagation of beliefs and are described in the following subsections. The network contains two kinds of nodes: variable nodes corresponding to explicit or inferred ABox assertions and valuation nodes representing TBox axioms. Variable nodes contain only one variable, while valuation nodes contain several variables.
Translation of an inconsistent subontology into a belief propagation network is performed using a set of rules (Table 1). Each rule translates a specific OWL-DL construct into a set of network nodes and links between them. Rules 1 and 2 directly translate each ABox statement into a variable node. Other rules process TBox axioms and create two kinds of nodes: one valuation node to represent the TBox axiom and one or more variable nodes to represent inferred statements. Such rules only fire if the network already contains variable nodes for ABox axioms, which are necessary to make the inference. For example, a rule processing the class equivalence axiom (Rule 4) is interpreted as the following: "If there is a node N_1 representing the type assertion $I \in X$ and an $owl : equivalentClass$ axiom $X \equiv Y$, then:

- Create a node N_2 representing the assertion $I \in Y$;
- Create a node N_3 representing the axiom $X \sqsubseteq Y$;
- Create links between N_1 and N_3 and between N_3 and N_2."

If a rule requires creating a node, which already exists in the network, then the existing node is used.
Applying the rules described above to our illustrating example (rules 1, 2, 4, 5, 6, 9, 20) will result in the following network (Fig. 1).

Table 1. Belief network construction rules

N	Pre-conditions	Nodes to create	Links to create
1	$I \in X$	$N_1 : I \in X$	
2	$R(I_1, I_2)$	$N_2 : R(I_1, I_2)$	
3	$N_1 : I \in X$, $X \sqsubseteq Y$	$N_2 : I \in Y$, $N_3 : X \sqsubseteq Y$	(N_1,N_3),(N_3,N_2)
4	$N_1 : I \in X$, $X \equiv Y$	$N_2 : I \in Y$, $N_3 : X \sqsubseteq Y$	(N_1,N_3),(N_3,N_2)
5	$N_1 : I \in X$, $X \sqsubseteq \neg Y$	$N_2 : I \in Y$, $N_3 : X \sqsubseteq \neg Y$	(N_1,N_3),(N_3,N_2)
6	$N_1 : I \in X$, $X \sqcap Y$	$N_2 : I \in X \sqcap Y$, $N_3 : X \sqcap Y$, $N_4 : I \in Y$	(N_1,N_3),(N_4,N_3), (N_3,N_2)
7	$N_1 : I \in X$, $X \sqcup Y$	$N_2 : I \in X \sqcup Y$, $N_3 : X \sqcup Y$, $N_4 : I \in Y$	(N_1,N_3),(N_4,N_3), (N_3,N_2)
8	$N_1 : I \in X$, $\neg X$	$N_2 : I \in \neg X$, $N_3 : \neg X$	(N_1,N_3),(N_3,N_2)
9	$\top \sqsubseteq \leq 1R$, $N_1 : R(I, o_1)$, $N_2 : R(I, o_2)$	$N_3 : \top \sqsubseteq \leq 1R$	(N_1,N_3),(N_2,N_3)
10	$\top \sqsubseteq \leq 1R^-$, $N_1 : R(I_2, I_1)$, $N_2 : R(I_3, I_1)$	$N_3 : \top \sqsubseteq \leq 1R^-$	(N_1,N_3),(N_2,N_3)
11	$R \equiv R^-$, $N_1 : R(I_1, I_2)$	$N_2 : R \equiv R^-$, $N_3 : R(I_2, I_1)$	(N_1,N_2),(N_2,N_3)
12	$R \equiv Q$, $N1 : R(I_1, I_2)$	$N_2 : R \equiv Q$,$N_3 : Q(I_1, I_2)$	(N_1,N_2),(N_2,N_3)
13	$R \sqsubseteq Q$,$N_1 : R(I_1, I_2)$	$N_2 : R \sqsubseteq Q$, $N_3 : Q(I_1, I_2)$	(N_1,N_2),(N_2,N_3)
14	$R \equiv Q^-$,$N1 : R(I_1, I_2)$	$N_2 : R \equiv Q^-$, $N_3 : Q(I_2, I_1)$	(N_1,N_2),(N_2,N_3)
15	$Trans(R)$, $N_1 : R(I_1, I_2)$, $N_2 : R(I_2, I_3)$	$N_3 : Trans(R)$, $N_4 : R(I_1, I_3)$	(N_1,N_3),(N_2,N_3), (N_3,N_4)
16	$\leq 1.R$, $N_1 : R(I_1, o_1)$, $N_2 : R(I_1, o_2)$	$N_3 :\leq 1.R$, $N_4 : I \in \leq 1.R$	(N_1,N_3),(N_2,N_3), (N_3,N_4)
17	$\geq 1.R$, $N_1 : R(I_1, o_1)$, $N_2 : R(I_1, o_2)$	$N_3 :\geq 1.R$, $N_4 : I \in \geq 1.R$	(N_1,N_3),(N_2,N_3)
18	$= 1.R$, $N_1 : R(I_1, o_1)$, $N_2 : R(I_1, o_2)$	$N_3 : I \in = 1.R$	(N_1,N_3),(N_2,N_3)
19	$\forall R.X$, $N_1 : R(I_1, I_2)$, $N_2 : I_2 \in X$	$N_3 : \forall R.X$, $N_4 : I_1 \in \forall R.X$	(N_1,N_3),(N_2,N_3), (N_3,N_4)
20	$\exists R.X$, $N_1 : R(I_1, I_2)$, $N_2 : I_2 \in X$	$N_3 : \exists R.X$, $N_4 : I_1 \in \exists R.X$	(N_1,N_3),(N_2,N_3), (N_3,N_4)
21	$\exists R.\top \sqsubseteq X$, $N_1 : R(I_1, I_2)$, $N_2 : I_1 \in X$	$N_3 : \exists R.\top \sqsubseteq X$	(N_1,N_3),(N_2,N_3)
22	$\top \sqsubseteq \forall R.X$, $N_1 : R(I_1, I_2)$, $N_2 : I_2 \in X$	$N_3 : \top \sqsubseteq \forall R.X$	(N_1,N_3),(N_2,N_3)

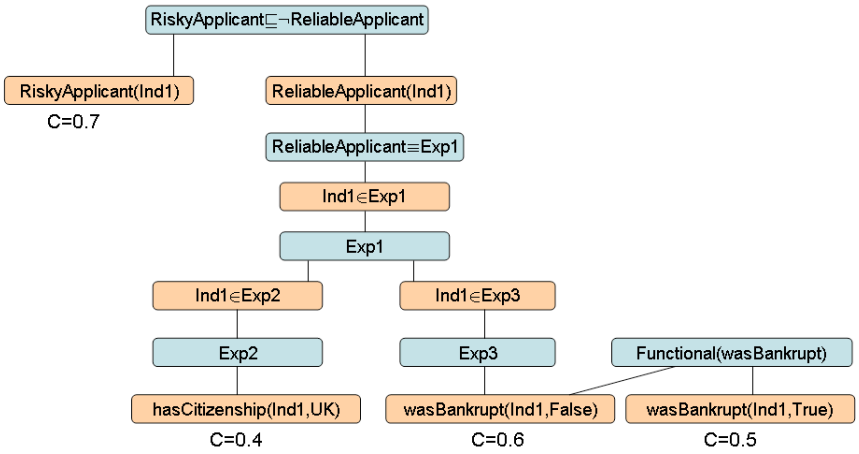


Fig. 1. Belief network example ($Exp1 = \exists wasBankrupt.False \sqcap \exists hasCitizenship.UK$, $Exp2 = \exists hasCitizenship.UK$, $Exp3 = \exists wasBankrupt.False$)

4.4 Assigning Mass Distributions

After the nodes were combined into the network, the next step is to assign the mass distribution functions to the nodes. There are two kinds of variable nodes: (i) nodes representing statements supported by the evidence and (ii) nodes representing inferred statements. Initial mass distribution for the nodes of the first type is assigned based on their extracted confidence. If a statement was extracted with a confidence degree c, it is assigned the following mass distribution: $m(True) = c, m(True; False) = 1 - c$. It is possible that the same statement is extracted from several sources. In this case, multiple pieces of evidence have to be combined using Dempster's rule of combination.
Nodes created artificially during network construction are only used for propagation of beliefs from their neighbours and do not contain their own mass assignment. Valuation nodes specify the TBox axioms and are used to propagate beliefs through the network. For the crisp OWL ontologies only mass assignments of 0 and 1 are possible. The principle for assigning masses is to assign the mass of 1 to the set of all combinations of variable sets allowed by the corresponding axiom. Table 2 shows the mass assignment functions for OWL-DL T-Box axioms [1].

In our example, we assign distributions based on the extractor's confidence values to the variable nodes representing extracted statements: A1:(m(1)=0.7, m({0;1})=0.3), A2: (m(1)=0.6, m({0;1})=0.4), A3: (m(1)=0.4, m({0;1})=0.6),

[1] For nodes allowing multiple operands (e.g., intersection or cardinality) only the case of two operands is given. If the node allows more than two children, then number of variables and the distribution function is adjusted to represent the restriction correctly

Table 2. Belief distribution functions for valuation nodes

N	Node type	Variables	Mass distribution
1	$X \sqsubseteq Y$	$I \in X$, $I \in Y$	m({0;0}, {0;1}, {1;1})=1
2	$X \equiv Y$	$I \in X$, $I \in Y$	m({0;0},{1;1})=1
3	$X \sqsubseteq \neg Y$	$I \in X$, $I \in Y$	m({0;0},{0;1},{1;0})=1
4	$X \sqcap Y$	$I \in X$, $I \in Y$, $I \in X \sqcap Y$	m({0;0;0},{0;1;0},{1;0;0},{1;1;1})=1
5	$X \sqcup Y$	$I \in X$, $I \in Y$, $I \in X \sqcup Y$	m({0;0;0},{0;1;1},{1;0;1},{1;1;1})=1
6	$\neg X$	$I \in X$, $I \in \neg X$	m({0;1},{1;0})=1
7	$\top \sqsubseteq \leq 1R$	$R(I, o_1)$, $R(I, o_2)$	m({0;0},{0;1},{1;0})=1
8	$\top \sqsubseteq \leq 1R^-$	$R(I_2, I_1)$, $R(I_3, I_1)$	m({0;0},{0;1},{1;0})=1
9	$R \equiv R^-$	$R(I_1, I_2)$, $R(I_2, I_1)$	m({0;0},{1;1})=1
10	$R \equiv Q$	$R(I_1, I_2)$, $Q(I_1, I_2)$	m({0;0},{1;1})=1
11	$R \sqsubseteq Q$	$R(I_1, I_2)$, $Q(I_1, I_2)$	m({0;0},{0;1},{1;1})=1
12	$R \equiv Q^-$	$R(I_1, I_2)$, $Q(I_2, I_1)$	m({0;0},{1;1})=1
13	$Trans(R)$	$R(I_1, I_2)$, $R(I_2, I_3)$, $R(I_1, I_3)$	m({0;0;0},{0;0;1},{0;1;0},{0;1;1}, {1;0;0},{1;0;1},{1;1;1})=1
14	$\leq 1.R$	$R(I_1, o_1)$, $R(I_1, o_2)$, $I_1 \in \leq 1.R$	m({0;0;1},{0;1;1},{1;0;1},{1;1;0})=1
15	$\geq 1.R$	$R(I_1, o_1)$, $R(I_1, o_2)$, $I_1 \in \geq 1.R$	m({0;0;0},{0;1;1},{1;0;1},{1;1;1})=1
16	$= 1.R$	$R(I_1, o_1)$, $R(I_1, o_2)$, $I_1 \in = 1.R$	m({0;0;0},{0;1;1},{1;0;1},{1;1;0})=1
17	$\forall R.X$	$R(I_1, I_2)$, $I_2 \in X$, $I_1 \in \forall R.X$	m({0;0;1},{0;1;1},{1;0;0},{1;1;1})=1
18	$\exists R.X$	$R(I_1, I_2)$, $I_2 \in X$, $I_1 \in \exists R.X$	m({0;0;1},{0;1;1},{1;0;0},{1;1;1})=1
19	$\exists R.\top \sqsubseteq X$	$R(I_1, I_2)$, $I_1 \in X$	m({0;0}, {0;1}, {1;1})=1
20	$\top \sqsubseteq \forall R.X$	$R(I_1, I_2)$, $I_2 \in X$	m({0;0}, {0;1}, {1;1})=1

A4: (m(1)=0.5, m({0;1})=0.5). The valuation nodes obtain their distributions according to the rules specified in the Table 2: T3 (rule 3), T4 (rules 2, 4, 18) and T5 (rule 7).

4.5 Belief Propagation

The axioms for belief propagation were formulated in [22]. The basic operators for belief potentials are marginalization $\downarrow$ and combination $\otimes$. Marginalization takes a mass distribution function m on domain D and produces a new mass distribution on domain $C \subseteq D$.

$$m^{\downarrow C}(X) = \sum_{Y^{\downarrow C}=X} m(Y)$$

For instance, if we have the function m defined on domain $\{x, y\}$ as $m(\{0;0\}) = 0.2$, $m(\{0;1\}) = 0.35$, $m(\{1;0\}) = 0.3$, $m(\{1;1\}) = 0.15$ and we want to find a marginalization on domain $\{y\}$, we will get $m(0) = 0.2 + 0.3 = 0.5$ and $m(1) = 0.35 + 0.15 = 0.5$. The combination operator is represented by the Dempster's rule of combination:

$$m_1 \otimes m_2(X) = \frac{\sum_{X_1 \cap X_2 = X} m_1(X_1) m_2(X_2)}{1 - \sum_{X_1 \cap X_2 = \oslash} m_1(X_1) m_2(X_2)}$$

Belief propagation is performed by passing messages between nodes according to the following rules:

1. Each node sends a message to its inward neighbour (towards the root of the tree). If $\mu^{A \to B}$ is a message from A to B, $N(A)$ is a set of neigbours of A and the potential of A is m_A, then the message is specified as a combination of messages from all neighbours except B and the potential of A:

$$\mu^{A \to B} = (\otimes\{\mu^{X \to A} | X \in (N(A) - \{B\}) \otimes m_A\})^{\downarrow A \cap B}$$

2. After a node A has received a message from all its neighbors, it combines all messages with its own potential and reports the result as its marginal.

As the message-passing algorithm assumes that the graph is a tree, it is necessary to eliminate loops. All valuation nodes constituting the loop are replaced by a single node with the mass distribution equal to the combination of mass distributions of its constituents. The marginals obtained after propagation for the nodes corresponding to initial ABox assertions will reflect updated mass distributions. After the propagation we can remove the statement with the lowest plausibility from each of the MISO found at the diagnosis stage.
Calculating the beliefs for our example gives the following Dempster-Shafer plausibility values for ABox statements: Pl(A1)=0.94, Pl(A2)=0.58, Pl(A3)=0.8, Pl(A4)=0.65. In order to make the ontology consistent it is sufficient to remove from both conflict sets an axiom with the lowest plausibility value (A2). In this example, we can see how the results using Dempster-Shafer belief propagation differ from the Bayesian interpretation. Bayesian probabilities, in this case, are calculated in the same way as Dempster-Shafer support values. If we use confidence values as probabilities and propagate them using the same valuation network we will obtain the results: P(A1)=0.66, P(A2)=0.35, P(A3)=0.32 and P(A4)=0.33. In this scenario, we would remove A3 and A4 because of the negative belief bias. Also we can see that all three statements A2, A3 and A4 will be considered wrong in such a scenario (resulting probability is less than 0.5). The Dempster-Shafer approach provides more flexibility by making it possible to reason about both support ("harsh" queries) and plausibility ("lenient" queries).

5 Conclusion and Future Work

In this paper, we described how the Dempster-Shafer theory of evidence can be used for dealing with ABox-level inconsistencies produced by inaccurate information extraction. It would be interesting to investigate if the capabilities of the Dempster-Shafer uncertainty representation (e.g., explicit representation of ignorance) can be utilized for knowledge modelling at the TBox level. In [23] it was shown that the Dempster-Shafer approach may lead to problems when it is used to represent uncertainty of inferencing rules (i.e., TBox-level) and not only of pieces of evidence (ABox assertions). These problems occur if the ontology contains contradictory pieces of knowledge, and are caused by the fact that

the Dempster-Shafer approach does not distinguish pieces of evidence regarding specific individuals from generic rules applicable to all individuals. It will be interesting to investigate if these problems can be avoided when modelling description logic axioms.
The algorithm described in the paper focuses on only one aspect of provenance information: confidence values assigned by extraction algorithms. However, such an approach has its limitations: for instance, we know that rule-based extractors tend to repeat their errors when applied to several documents. Such reoccurring errors lead to erroneous inconsistency resolution if interpreted as independent pieces of evidence. In order to improve the quality of the fusion procedure, it would be useful to take into account other kinds of provenance information, in particular, the reliability of the extraction algorithm itself, the reliability of sources from which statements were extracted, and the timestamp reflecting when each statement was produced. This we consider our primary direction for the future work.

6 Acknowledgements

This work was funded by the X-Media project (www.x-media-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-026978.

References

1. Welty, C., Murdock, W.: Towards knowledge acquisition from information extraction. In: 5th International Semantic Web Conference, Athens, GA, USA, November 5-9, 2006. (2006) 709–722
2. Popov, B.: KIM - a semantic platform for information extraction and retrieval. Natural language engineering **10** (2004) 375–392
3. Iria, J.: Relation extraction for mining the Semantic Web. In: Dagstuhl Seminar on Machine Learning for the Semantic Web, Dagstuhl, Germany (2005)
4. Horvitz, E.J., Heckerman, D.E., Langlotz, C.P.: A framework for comparing alternative formalisms for plausible reasoning. In: AAAI-86. (1986) 210–214
5. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)
6. Haase, P., van Harmelen, F., Huang, Z., Stuckenschmidt, H., Sure, Y.: A framework for handling inconsistency in changing ontologies. In: Proceedings of the International Semantic Web Conference (ISWC2005). (2005) 353–367
7. Kalyanpur, A., Parsia, B., Sirin, E., Grau, B.C.: Repairing unsatisfiable concepts in OWL ontologies. In: ESWC2006. (2006) 170–184
8. Haase, P., Volker, J.: Ontology learning and reasoning - dealing with uncertainty and inconsistency. In: International Workshop on Uncertainty Reasoning for the Semantic Web (URSW). (2005) 45–55
9. Stoilos, G., Stamou, G., Tzouvaras, V., Pan, J.Z., Horrocks, I.: Fuzzy OWL: Uncertainty and the semantic web. In: International Workshop of OWL: Experiences and Directions, ISWC 2005. (2005)

10. Straccia, U.: Towards a fuzzy description logic for the Semantic Web (preliminary report). In: 2nd European Semantic Web Conference (ESWC-05). Number 3532 in Lecture Notes in Computer Science, Crete, Springer Verlag (2005) 167–181
11. Ding, Z., Peng, Y.: A probabilistic extension to ontology language OWL. In: 37th Hawaii International Conference On System Sciences (HICSS-37). (2004)
12. da Costa, P.C.G., Laskey, K.B., Laskey, K.J.: PR-OWL: A Bayesian ontology language for the semantic web. In: Workshop on Uncertainty Reasoning for the Semantic Web, ISWC 2005. (2005)
13. Li, Y., Bontcheva, K., Cunningham, H.: SVM based learning system for information extraction. In: Deterministic and Statistical Methods in Machine Learning. (2005) 319–339
14. Ciravegna, F., Wilks, Y.: Designing Adaptive Information Extraction for the Semantic Web in Amilcare. In: Annotation for the Semantic Web. (2003)
15. Zhu, J., Uren, V., Motta, E.: ESpotter: Adaptive named entity recognition for web browsing. In: Professional Knowledge Management Conference (WM2005). (2005)
16. de Campos, L.M., Huete, J.F., Moral, S.: Uncertainty management using probability intervals. In: Advances in Intelligent Computing IPMU '94. (1994) 190–199
17. Gaifman, H.: A theory of higher order probabilities. In: TARK '86: Proceedings of the 1986 conference on Theoretical aspects of reasoning about knowledge, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1986) 275–292
18. Richardson, M., Agrawal, R., Domingos, P.: Trust management for the Semantic Web. In: 2nd International Semantic Web Conference, ISWC 2003, Sanibel Island, Florida (2003) 351–368
19. Nickles, M.: Social acquisition of ontologies from communication processes. Journal of Applied Ontology (2007)
20. Reiter, R.: A theory of diagnosis from first principles. Artificial Intelligence **32**(1) (1987) 57–95
21. Shenoy, P.P.: Valuation-based systems: a framework for managing uncertainty in expert systems. In: Fuzzy logic for the management of uncertainty. John Wiley & Sons, Inc., New York, NY, USA (1992) 83–104
22. Shenoy, P.P., Shafer, G.: Axioms for probability and belief-function propagation. In: Readings in uncertain reasoning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1990) 575–610
23. Pearl, J.: Bayesian and belief-function formalisms for evidential reasoning: a conceptual analysis. In: Readings in uncertain reasoning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1990) 575–610

An Ontology-based Bayesian Network Approach for Representing Uncertainty in Clinical Practice Guidelines

Hai-tao Zheng, Bo-Yeong Kang, and Hong-Gee Kim*

Biomedical Knowledge Engineering Laboratory
Dentistry College, Seoul National University
28 Yeongeon-dong, Jongro-gu, Seoul, Korea

Abstract. Clinical Practice Guidelines (CPGs) play an important role in improving the quality of care and patient outcomes. Although several machine-readable representations of practice guidelines implemented with semantic web technologies have been presented, there is no implementation to represent uncertainty with respect to activity graphs in clinical practice guidelines. In this paper, we are exploring a Bayesian Network(BN) approach for representing the uncertainty in CPGs based on ontologies. Based on the representation of uncertainty in CPGs, when an activity occurs, we can evaluate its effect on the whole clinical process, which, in turn, can help doctors judge the risk of uncertainty for other activities, and make a decision. A variable elimination algorithm is applied to implement the BN inference and a validation of an aspirin therapy scenario for diabetic patients is proposed.

1 Introduction

Clinical Practice Guidelines (CPGs) play an important role in improving the quality of care and patient outcomes; therefore, the task of clinical guideline-sharing across different medical institutions is a prerequisite to many EMR (Electronic Medical Record) applications including medical data retrieval [18], medical knowledge management [7], and clinical decision support systems (CDSSs) [13]. To facilitate clinical guideline-sharing, GLIF (GuideLine Interchange Format) and SAGE (Standards-based Sharable Active Guideline Environment) have been the focus of extensive research [12]. GLIF is a semantic web based standard for representing clinical guidelines [15] and SAGE is an interoperable guideline execution engine, which encodes the content of the clinical guideline to an ontology representation, and executes the ontology through the functions of a CIS (clinical information system) [17].

Most previous approaches using GLIF and SAGE are designed to proceed from one step to the next only if there is no uncertain data in the former step [13]. However, this expectation is unrealistic in practice. For example, a guideline, which requires risk factors for heart disease to be assessed, needs to proceed

* Corresponding author: hgkim@snu.ac.kr

even when the information about this item is uncertain. In the clinical process, uncertain data can be (1) data stemming from unreliable sources (e.g., a patient can not remember the results of his/her last glucose test); (2) data not obtainable (e.g., no historical data on familial diabetes); and (3) data not yet collected (e.g., levels of serum glucose today) [14]. If data represented in CPGs is uncertain, the activities that handle these uncertain data become uncertain as well. For instance, in CDSS systems, when using the diabetes clinical guideline, it is necessary to get the family history for evaluating the risk of insulin therapy. However, in the real hospital environment, clinicians cannot easily obtain all the needed data for his/her health care activity. Based on these issues, the goal of this paper is to construct an approach to represent the uncertainty in CPGs and help doctors judge the risk of these uncertainties in the clinical process. Uncertainty in CPGs means that activity graphs that CPGs are composed of contain uncertain activities.

As a model for uncertainty, Bayesian Networks (BNs) occupy a prominent position in many medical decision making processes and statistical inference [11, 3, 2]. However, there have been few reports on applying BNs to the representation of uncertainty in CPGs. Therefore, to address this issue, we propose an ontology-based representation of uncertainty in CPGs by using BNs.

In this paper, we first introduce BNs, then we describe the use of BNs for the medical domain, and review previous work on applying semantic web technology to model CPGs in section 2; Section 3 elaborates the mechanism of encoding uncertainty into a CPG ontology; Section 4 describes a scenario validation based on BN inference; Section 5 discusses the conclusions and future work.

2 Background and Related Work

2.1 Bayesian Network

There are several models that are used to represent uncertainty, such as fuzzy-logic, BNs, etc. Generally, a BN of n variables consists of a DAG (Direct Acyclic Graph) of n nodes and a number of arcs. Nodes X_i in a DAG correspond to random variables, and directed arcs between two nodes represent direct causal or influential relations from one variable to the other. The uncertainty of the causal relationship is represented locally by the CPT (Conditional Probability Table). $P(X_i|pa(X_i))$ associated with each node X_i, where $pa(X_i)$ is the parent set of X_i. Under the conditional independence assumption, the joint probability distribution of $\mathbf{X} = (X_1, X_2, ..., X_n)$ can be factored out as a product of the CPTs in the network, namely, the chain rule of BN: $P(\mathbf{X}) = \prod_i P(X_i|pa(X_i))$. With the joint probability distribution, BNs support, at least in theory, any probabilistic inference in the joint space. Besides the power of probabilistic reasoning provided by BNs themselves, we are attracted to BNs in this work for the structural similarity between the DAG of a BN and activity graphs of CPGs: both of them are directed graphs, and direct correspondence exists between many nodes and arcs in the two graphs. Moreover, BNs can be utilized to represent the uncertainty visually, provide inference effectively and facilitate human understanding

of CPGs. Considering the advantages of BNs, we apply BNs to represent the uncertainty in CPGs.

2.2 Bayesian Networks for the Medical Domain

Because BNs occupy a prominent position as a model for uncertainty in decision making and statistical inference, it has been applied to many medical decision support systems [11, 3, 2]. Atoui [3] adopted a decision making solution based on a BN that he trained to predict the risk of a cardiovascular event (infarction, stroke, or cardiovascular death) based on a set of demographic and clinical data. Aronsky [2] presented the development and the evaluation of a BN for the diagnosis of community-acquired pneumonia and he showed that BNs are an appropriate method to detect pneumonia patients with high accuracy. With respect to clinical guidelines, Mani [11] proposed BNs for the induction of decision tables and generated the guideline by these tables. However, although these method focus on predicting some feature or risk of disease by using BN inference, there has been no implementation to represent the uncertainty with respect to activity graphs in CPGs and to reason on the uncertainty to provide the probabilities of target activities, which is the focus of our approach.

2.3 Semantic Web for Clinical Practice Guideline

A representational form of clinical guideline knowledge, which promotes completeness and minimizes inconsistency and redundancy, is essential if we want to implement and share guidelines for computer-based applications. Semantic Web technology offers such sharable and manageable methodology for modeling CPGs. GLIF [15] and SAGE [17] are two good examples. For creation and maintenance of implementable clinical guideline specifications, an architecture is presented in [8]. This architecture includes components such as a rules engine, an OWL-based classification engine and a data repository storing patient data. Moreover, approaches for modeling clinical guidelines are discussed and they show that guideline maintenance is tractable when a CPG is represented in an ontology. Here, we apply an ontology to represent the uncertainty in CPGs because it is more extensible and maintainable than other methods such as relational databases.

3 Encoding Uncertainty into a CPG Ontology

Figure 1 depicts the overall procedure of the proposed method. Firstly, the original CPG is encoded into an ontology model that contains uncertainty features using BNs. For this, we propose a formal model of CPG Ontology to represent uncertainty and an algorithm to construct the CPTs (Conditional Probability Tables) of the BN. The CPG ontology can be shared and utilized in different clinical information systems. Then, when a user provides his/her observed evidence in the clinical process, the BN inference engine will load the CPG ontology

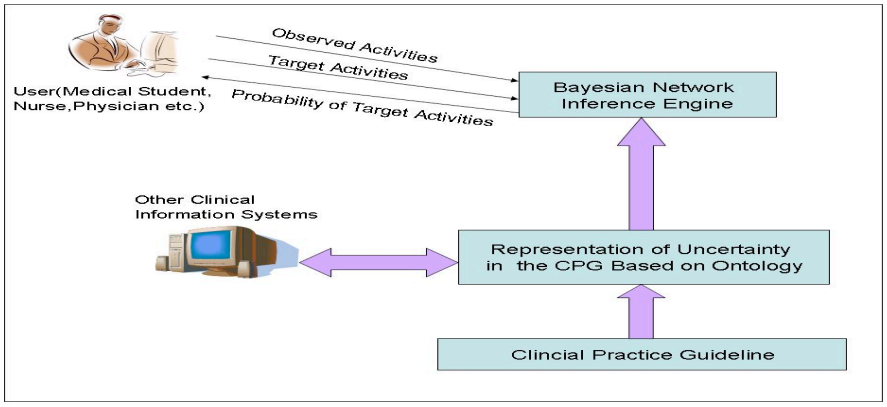


Fig. 1. The framework

as a BN and mark the nodes that are observed by the user in the BN. Based on the observed evidence, the BN inference engine can reason out the probabilities of target activities asked by the user. Given the reasoning results, the user can judge the risk of unobserved activities and make a further decision.

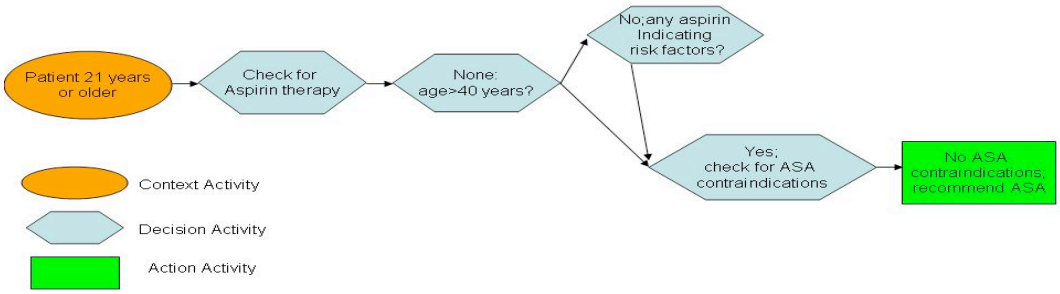


Fig. 2. Clinical practice guideline of aspirin therapy for diabetic patients(ASA means aspirin therapy)

3.1 Clinical Practice Guideline Ontology

CPGs typically include multiple recommendation sets represented as an activity graph that show the recommended activities during a clinical process and [4]. An activity graph describes the relationship between activities in the recommendation set as a process model. In this article, we use a single recommendation set in the SAGE diabetes CPG [1], which is an activity graph of aspirin therapy for diabetic patients, to illustrate how we represent the uncertainty in CPGs

based on ontology (Fig. 2). Typically, an activity graph contains three kinds of activities, i.e., context activity, decision activity, action activity. Each activity graph segment within a guideline begins with a context activity node that serves as a control point in guideline execution by specifying the clinical context for that segment. A decision activity node in the SAGE guideline model represents clinical decision logic by listing alternatives (typically subsequent action activity nodes), and specifying the criteria that need to be met to reach those nodes. An action activity node encapsulates a set of work items that must be performed by either a computer system or persons.

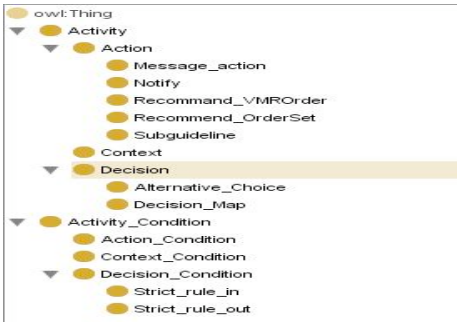


Fig. 3. Classes representation for clinical practice guideline

To represent activities in CPGs, we create the activity class that represents all the nodes in activity graph as shown in Figure 3. Because there are three kinds of activities, we construct an action class, a context class, and a decision class as sub classes of the activity class in the ontology respectively. In CPGs, activities may include internal conditions that restrict their execution. For example, for the decision activity "Yes;check for ASA(aspirin therapy) contraindications" (Fig. 2), there are many internal conditions, such as checking presence of family history, checking presence of hypertensive disorder etc., to make sure the ASA contraindications will be checked correctly. We encode these internal conditions of activity as an activity condition class in the ontology (Fig. 3).

A CPG Ontology with uncertainty features is defined as follows:

Definition 1. *(CPG Ontology) CPG Ontology* $O := \{C, I, Ps, cinst\}$*, with an activity class set* C*, an activity instance set* I*, a property set* Ps*, and an activity class instantiation function* $cinst : C \rightarrow 2^I$*.*

In CPG ontology, the activity instance set I represents the set of real activities that belong to activity classes accordingly. The property set Ps is proposed to represent the different attributes of activities in order to encode the features of the BN into ontology. The property set Ps is defined as follows:

Definition 2. *(Properties for uncertainty representation) Property Set* $Ps :=$ $\{cause, hasCondition, hasState, isObserved, hasPriorProValue, hasCondiProValue\}$*,*

has a property function cause : $I \rightarrow I$*, a property function hasCondition* : $I \rightarrow I$*, a property function hasState* : $I \rightarrow Boolean$*, a property function isObserved:* $I \rightarrow Boolean$*, a property function hasPriorProValue:* $I \rightarrow Float$*, and a property function hasCondiProValue:* $I \rightarrow Float$*.*

In CPGs, if the criteria associated with an activity node are satisfied, it will be successfully executed, which will cause the execution of subsequent nodes in the activity graph. Therefore, the relationship between activities is called the *cause* relationship. For example, in Figure 2, the context activity "Patient 21 years or older" causes the decision activity "Check for Aspirin therapy". To represent this relationship in the ontology, we construct the object property *cause* whose domain and range are activity class and activity condition class. The *hasCondition* property is proposed as inverse properties of the *cause* property, which describes the "parent" activities of an activity that *cause* its execution. For instance, the decision activity "Check for aspirin therapy" has the property *hasCondition* with value "Patient 21 years or older" activity that causes its execution. With the *hasCondition* property, users can easily figure out all the conditions that cause the execution of any activity. The *cause* property plays the role of "directed arc" and all the activity instances play the role of "node" in the DAG of BN. Another property, the *hasState* property, which has a boolean value range, is denoted as the state of the activity instance; the *isObserved* property shows if the activity instances have been observed or not.

Prior probability and conditional probability are two features that represent the uncertainty level of nodes in BNs. To encode prior probability and conditional probability of activity instances into the ontology respectively, *hasPriorProValue* property and *hasCondiProValue* property are employed. Let A, B be the instances of the activity class representing two concrete activities. We interpret $P(A = a)$ as the prior probability that a value a is a state of instance A and $P(B = b|A = a)$ as the conditional probability that when A has state a, B has state b. For example, when A is activity "Patient 21 years or older", B is activity "Check for Aspirin therapy", $P(A = true) = 0.5$ can be expressed as follows:

```
<Context rdf:ID="Patient_21_yo_or_older">
    <hasPriorProValue
    rdf:datatype="http://www.w3.org/2001/XMLSchema#float"
    >0.5</hasPriorProValue>
    <hasState
    rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
    >true</hasState>
    <cause rdf:resource="#Check_for_Aspirin_therapy"/>
</Context>
```

The conditional probability $P(B = true|A = true) = 1.0$ can be expressed as follows:

```
 <Decision rdf:ID="Check_for_Aspirin_therapy">
    <hasCondition>
        <Context rdf:ID="Patient_21_yo_or_older">
        <hasState rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
```

```
        >true</hasState>
        </Context>
    </hasCondition>
    <hasState rdf:datatype="http://www.w3.org/2001/XMLSchema#boolean"
    >true</hasState>
    <hasCondiProValue rdf:datatype="http://www.w3.org/2001/XMLSchema#float"
    >1.0</hasCondiProValue>
    <cause rdf:resource="#Check_fo_age_older_than_40"/>
</Decision>
```

3.2 Construction of Conditional Probability Tables

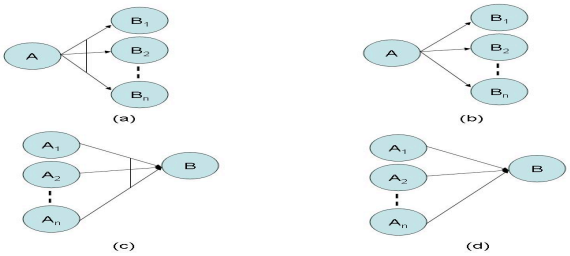


Fig. 4. Algorithm for constructing the conditional probability tables

In this section, we introduce an algorithm utilized to construct the CPTs of activity instances. After creating the properties to represent the uncertainty in the ontology, another important work is the construction of the CPTs for the BN because BN inference is based on the CPTs of each node in BN. For building the CPTs of activity instance in CPGs, we propose the encoding algorithm according to the features of CPGs. Each activity X_i in a CPG η has a corresponding activity instance Xo_i in CPG Ontology O, i.e., we mark the corresponding activity instance by adding the letter "o" to the activity variable.

This algorithm provides principles to initialize the CPTs of BNs. The instances in the definition mean the activity instances in the CPGs. Firstly, we assign prior probabilities to activity instances, i.e, only when the activity instances have no "parents" in BN, they have prior probabilities. When an activity A causes the set of activities $\{B_1, B_2, ..., B_n\}$ simultaneously, the conditional probability $P(B_i|A) = 1.0, (i = 1, ..., n)$ (Fig. 4(a)); when an activity A causes one of the activities $\{B_1, B_2, ..., B_n\}$, the conditional probability $P(B_i|A) = 1.0/n, (i = 1, ..., n)$ (Fig. 4(b)); when a set of activities $\{A_1, A_2, ..., A_n\}$ cause activity B together, then $P(B|A_1, A_2, ..., A_n) = 1.0$ (Fig. 4(c)); when one of the activities $\{A_1, A_2, ..., A_n\}$ can cause activity B, then $P(B|\overline{A_1, A_2, ..., A_n}) = 0.0$ (Fig. 4(d)).

With the initialization of CPTs, we have finished constructing the BN from an ontology that represents the uncertainty in CPGs, namely, the activity graphs

containing uncertain activities are represented using a BN. When a BN inference engine loads this ontology, the ontology will be converted to a BN for BN inference. All the instances of the activity class and the activity condition class are translated into the node of the BN whose properties are also converted from properties of these instances in ontology accordingly. In the BN, an arc is drawn between nodes if the corresponding two activity instances are related by a *cause* property, with the direction from the activity instance that has *cause* property to the value of this property. CPTs in the BN are also easily obtained from property *hasCondiProValue* and property *hasPriorProValue* of corresponding activity instances.

Algorithm 1 Construct CPTs(CPG η, CPG Ontology O)

for each activity X_i in η **do**
  if X_i can be successfully executed **then**
    Set property *hasState* of activity instance Xo_i in O with value true
  end if
  if there is no activity that causes the execution of X_i **then**
    Set property *hasCondition* of activity instance Xo_i in O with value null
    Set property $hasPriorProValue$ of Xo_i with value 0.5
  end if
end for
if activity A in η cause the execution of activities $\bigcap_{i=1}^{n} B_i$ **then**
  Set property $hasCondiProValue$ of activity instance Bo_i in O with value 1.0
  Set activity instance Ao as the value of property *hasCondition* of activity instance Bo_i in O
  Set property *hasState* of activity instance Ao and Bo_i in O with value true
end if
if activity A in η cause the execution of activities $\bigcup_{i=1}^{n} B_i$ **then**
  Set property $hasCondiProValue$ of activity instance Bo_i in O with value $1.0/n$
  Set activity instance Ao as the value of property *hasCondition* of activity instance Bo_i in O
  Set property *hasState* of activity instance Ao and Bo_i in O with value true
end if
if activities $\bigcap_{i=1}^{n} A_i$ in η cause the execution of activities B **then**
  Set property $hasCondiProValue$ of activity instance Bo in O with value 1.0
  Set activity instances $Ao_1, ..., Ao_n$ as the value of property *hasCondition* of activity instance Bo in O
  Set property *hasState* of activity instance Ao_i and Bo in O with value true
end if
if activities $\bigcup_{i=1}^{n} A_i$ in η cause the execution of activities B **then**
  Set property $hasCondiProValue$ of activity instance Bo in O with value 0.0
  Set activity instances $Ao_1, ..., Ao_n$ as the value of property *hasCondition* of activity instance Bo in O
  Set property *hasState* of activity instance Ao_i with value false
  Set property *hasState* of activity instance Bo with value true
end if

4 A Scenario Validation Based on Bayesian Network Inference

We apply the variable elimination algorithm [9, 5] to perform BN inference. To verify the feasibility of our approach, a scenario of aspirin therapy for a diabetic patient is proposed. Based on this scenario, we apply our ontology-based BN approach to represent the uncertainty in CPGs and carried out the BN inference based on this BN.

4.1 Bayesian Network Inference

There are a lot of algorithms that manipulate BNs to produce posterior values [16, 10]. The variable elimination algorithm [9, 5] and the bucket elimination algorithm [6] are focused on algebraic operations. Since algebraic schemes like variable and bucket elimination compute marginal probability values for a given set of variables that is suitable for inference on observed evidence, we apply the variable elimination algorithm to implement the BN inference on the uncertainty of CPGs.

We assume all random variables have a finite number of possible values. Set of variables are denoted in bold; for instance, $\mathbf{X}$. The set of all variables that belong to $\mathbf{X}$ but do not belong to $\mathbf{Y}$ is indicated by $\mathbf{X}\backslash\mathbf{Y}$. The expression $\sum_{\mathbf{X}} f(\mathbf{X}, \mathbf{Y})$ indicates that all variables in $\mathbf{X}$ are summed out from the function $f(\mathbf{X}, \mathbf{Y})$. Denoted by $P(X)$ is the *probability density of* X: $P(x)$ is the probability measure of the event $\{X = x\}$. Denoted by $P(X|Y)$ is the probability density of X conditional on values of Y.

Given a BN, the event E denotes the *observed evidence* in the network. Denoted by $\mathbf{X}_E$ is the set of observed variables. Inferences with BNs usually involve the calculation of the posterior marginal for a set of query variables $\mathbf{X}_q$. The posterior of $\mathbf{X}_q$ given E is:

$$P(\mathbf{X}_q|E) = \frac{P(\mathbf{X}_q, E)}{P(E)} = \frac{\sum_{\mathbf{X}\backslash\{\mathbf{X}_q,\mathbf{X}_E\}} P(\mathbf{X})}{\sum_{\mathbf{X}\backslash\mathbf{X}_E} P(\mathbf{X})} \tag{1}$$

The detail of variable elimination algorithm can be found in [9, 5].

4.2 A Validation of an Aspirin Therapy Scenario for Diabetic Patients

We demonstrate the validity of our approach by applying an experiment to the CPG of aspirin therapy for diabetic patients (Fig. 2). Let us consider a scenario:

Scenario 1 *A user(medical student, nurse or physician etc.) is trying to apply aspirin therapy for a diabetic patient using the diabetes CPG. When he/she tries to check the aspirin risk factors, he/she can get a few observed evidence, such as observations of hypertensive disorder, tobacco user finding, hyperlipidemia, and myocardial infarction. In this case, the user wants to evaluate target activities*

that he is concerned about in this CPG. In this way, he hopes the results can help him understand the effect of the observed evidence on the target activities during the whole clinical process.

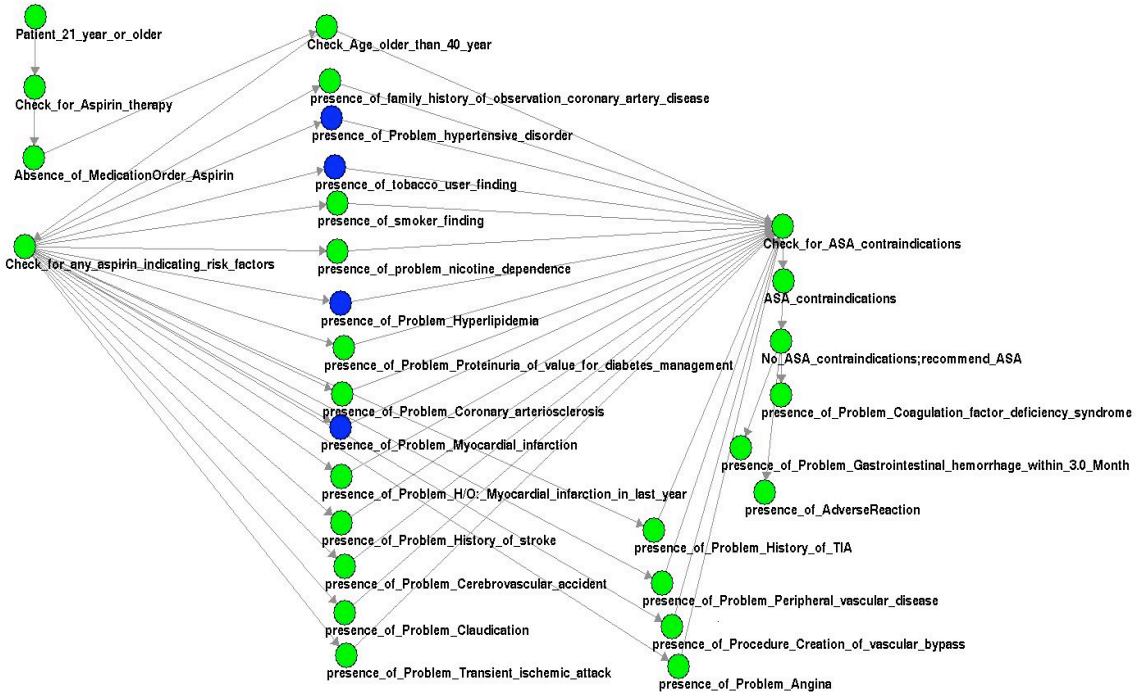


Fig. 5. An ontology based Bayesian network of aspirin therapy for diabetic patients derived from figure 2 (blue nodes are the observed ones)

In the scenario, the CPG of aspirin therapy for diabetic patients is used. Since there are some uncertain activities with respect to the activity graph in this CPG, we can apply our ontology-based BN approach to represent this uncertainty. Details are described in Section 3. Figure 5 shows the ontology-based BN representing the uncertainty in the CPG of aspirin therapy for diabetic patients.

After loading the ontology-based BN, the BN inference engine can process BN inference when the user provides his/her observed evidence, such as observations of hypertensive disorder, tobacco user finding, hyperlipidemia, and myocardial infarction in this scenario (Fig. 5). If the user selects the target activities, the BN inference engine can calculate the probability of them by using the variable elimination.

For example, after the user got the observed evidence of some aspirin risk factors, he wants to know the probability of activity "No ASA (aspirin therapy) contraindications; recommend ASA" to help him to judge whether or not his ob-

servations of aspirin risk factors are adequate. In the BN inference engine, since the activity instance "presence of problem hypertensive disorder" is observed, its property *isObserved* is set *true* and the property *hasState* is set $false$. Similarly, the activities instances "presence of problem myocardial infarction", "presence of tobacco user finding", and "presence of problem hyperlipidemia" are also set in the same manner. With CPTs in this BN, equation 1 (Section 4.1) is applied to calculate the probability of activity instance "No ASA contraindications; recommend ASA" :

$$P(\mathbf{X}_q|E) = \frac{P(\mathbf{X}_q, E)}{P(E)} = \frac{\sum_{\mathbf{X}\setminus\{\mathbf{X}_q,\mathbf{X}_E\}} P(\mathbf{X})}{\sum_{\mathbf{X}\setminus\mathbf{X}_E} P(\mathbf{X})} = 0.775$$

where $\mathbf{X}_q$ ={"No ASA contraindications; recommend ASA"}, and E={"presence of problem hypertensive disorder" = $false$,"presence of problem myocardial infarction" = $false$, "presence of tobacco user finding" = $false$, "presence of problem hyperlipidemia"= $false$ }.

In another case, when the user wants to get the uncertain degree of activity instance "presence of problem coagulation factor deficiency syndrome", he can choose this target activity instance based on the observed evidence E. Through BN inference, we can obtain:

$$P(\mathbf{X}_q|E) = \frac{P(\mathbf{X}_q, E)}{P(E)} = 0.6425$$

where $\mathbf{X}_q$={"presence of problem coagulation factor deficiency syndrome"} and E is the same as the above case.

The results in the two cases show high probabilities for the target activities, which suggest the user can make a decision to go ahead based on the observed evidence. When we consult several medical experts with this scenario, their opinions are coincident with these results, which shows the feasibility of our approach.

5 Conclusion and future work

In this paper, we contribute an ontology-based BN approach to represent the uncertainty in CPGs. With this uncertain representation in ontology, computers can: (1) calculate the uncertainty of target activities in CPGs; (2) remind users of the missing important data or event items, which should be observed in the clinical process; (3) simulate the clinical process under uncertain situations, which can be applied to e-learning systems in medical schools.

In the future, we are planning to combine our approach with a real CIS environment and apply uncertain clinical data to our application. A more comprehensive evaluation based on real clinical data should also be carried out.

Acknowledgements

This study was supported by a grant of National Project for Information Technology Advancement, Ministry of Information and Communication, and the

Interoperable EHR Research and Development Center(A050909), Ministry of Health & Welfare, Republic of Korea.

References

1. Sage diabetes guideline. *http://sage.wherever.org/cpgs/diabetes/diabetes_html/phtml.html*.
2. D. Aronsky and P. J. Haug. Diagnosing community-acquired pneumonia with a bayesian network. In *Proc AMIA Symp*, pages 632–636, 1998.
3. H. Atoui, J. Fayn, F. Gueyffier, and P. Rubel. Cardiovascular risk stratification in decision support systems:a probabilistic approach. application to phealth. *Computers in Cardiology*, 33:281–284, 2006.
4. J. Campbell, S. Tu, J. Mansfield, J. Boyer, J. McClay, C. Parker, P. Ram, S. Scheitel, and K. McDonald. The sage guideline model:a knowledge representation framework for encoding interoperable cpgs. *Stanford Medical Informatics Report SMI-2003-0962*, 2003.
5. F. G. Cozman. Generalizing variable elimination in bayesian networks. In *Workshop on Probabilistic Reasoning in Artificial Intelligence*, pages 27–32, 2000.
6. R. DECHTER. Bucket elimination: A unifying framework for probabilistic inference. *In M. I. Jordan,editor, Learning in Graphical Models,MITPress*, pages 75–104, 1999.
7. G. Hripcsak. Writing arden syntax medical logic modules. *Comput Biol Med*, 24:331–363, 1994.
8. V. Kashyap, A. Morales, and T. Hongsermeier. Creation and maintenance of implementable clinical guideline specifications. In *ISWC 2005*, 2005.
9. Z. N. L and P. D. Exploiting causal independence in bayesian network inference. *Artificial Intelligence Research*, pages 301–328, 1996.
10. S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical SocietyB*, 50(2):157–224, 1988.
11. S. Mani and M. J. Pazzani. Guideline generation from data by induction of decision tables using a bayesian network framework. *JAMIA supplement*, pages 518–522, 1998.
12. A. H. Morris. Developing and implementing computerized protocols for standardization of clinical decisions. *Annal of Internal Medicine*, 132(5):373–383, 2000.
13. M. Musen, S. Tu, A. Das, and Y. Shahar. Eon:a component-based approach to automation of protocol-directedtherapy. *J Am Med Inform Assoc*, 2:367–388, 1996.
14. L. Ohno-Machado. Representing uncertainty in clinical practice guidelines. In *An Invitational Workshop: Towards Representations for Sharable Guidelines*, March 2000.
15. L. Ohno-Machado, S. N. Murphy, D. E. Oliver, R. A. Greenes, and G. O. Barnett. The guideline interchange format: A model for representing guidelines. *Journal of the Americal Medical Informatics Association*, 5(4):357–372, Jul 1998.
16. J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, SanMateo,California, 1988.
17. P. Ram, D. Berg, S. Tu, G. Mansfield, Q. Ye, R. Abarbanel, and N. Beard. Executing clinical practice guidelines using the sage execution engine. *Medinfo*, pages 251–255, 2004.
18. P. Stoufflet, L. Ohno-Machado, S. Deibel, D. Lee, and R. Greenes. Geode-cm:a state-transition framework for clinical management. In *20th Annu Symp Comput Appl Med Care*, page 924, 1996.

Uncertainty Issues in Automating Process Connecting Web and User

Alan Eckhardt[1] , Tomáš Horváth[2], Dušan Maruščák[1], Róbert Novotný[2], Peter Vojtáš[1]

[1] Charles University, Prague,
[2] P. J. Šafárik University Košice
{alan.eckhardt, dusan.maruscak, peter.vojtas}@mff.cuni.cz,
{tomas.horvath, robert.novotny}@upjs.sk

Abstract. We are interested in replacing human processing of web resources by automated processing. Based on an experimental system we identify uncertainty issues which make this process difficult for automated processing. We show these uncertainty issues are connected with Web content mining and user preference mining. We conclude with a discussion of possible future development heading to an extension of web modeling standards with uncertainty features.

Keywords: Uncertainty modeling, Uncertain reasoning, World Wide Web, Web content mining, User profile mining,

1 Introduction

The amount of data accessible on Web is a great challenge for web search systems. Using these data (and information and knowledge hidden in them) can be a competitive advantage both for companies and individuals. Hence Web search systems form a part of different systems ranging from marketing systems, competitors and/or price tracking systems to private decision support systems.

The main vision of Semantic Web [3] is to automate some web search activities that a human is able to do personally, but they are time-consuming or tedious. Using this automation of human search will speed up the process of searching, find a wider range of resources and when necessary soften and optimize our search criteria.

We quote the Uncertainty Reasoning for the World Wide Web (URW3) Incubator Group charter [22]: "*…as work with semantics and services (on the Web) grows more ambitious, there is increasing appreciation of the need for principled approaches to representing and reasoning under uncertainty. In this Charter, the term «uncertainty» is intended to encompass a variety of forms of incomplete knowledge, including incompleteness, inconclusiveness, vagueness, ambiguity, and others. The term «uncertainty reasoning» is meant to denote the full range of methods designed for representing and reasoning with knowledge when Boolean truth values are unknown, unknowable, or inapplicable. Commonly applied approaches to uncertainty reasoning*

include probability theory, Dempster-Shafer theory, fuzzy logic, and numerous other methodologies." In this paper we are using term "uncertainty" in this wider (generic) understanding and we would like to contribute to these efforts (for related discussion see [23]).

In this paper we concentrate especially to issues connected with replacing human abilities on the web by software. From this point of view, some sorts of uncertainty are not "human_to_machine_web" specific, like faulty sensors, input errors, data recorded statistically, medical diagnosis, weather prediction, gambling etc. These are difficult for human alone and also outside the web.

According to Turtle and Croft [18], uncertainty in information retrieval can be found especially in three areas: "*Firstly, there is the problem of the representation and annotation of a resource (service). Difficulties arise also in case when attempting to represent the degree to which a resource is relevant to the task. The second problem is the representation of the kind of information, action, that a user needs to retrieve, perform (this need is especially difficult since it typically changes during the session). Thirdly, it is necessary to match user needs to resource concepts.*"

In our opinion, these areas of uncertainty apply also to our case, when replacing human activities on the web by software. Specific tasks connected to these three problems are depicted in Figure 1 and we will discuss them in this paper.

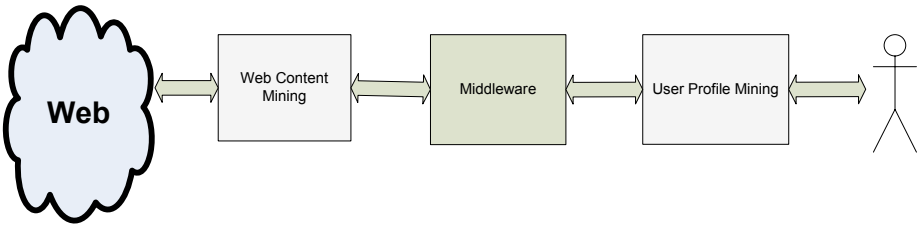


Fig. 1. Schema of an automated process connecting Web and User

Our goal is to discuss uncertainty issues based on a system integrating the whole chain of tools from the Web to the user. The uncertainty problem here appears as a problem of two inductive procedures. Two types of data mining that appear in these systems will be discussed here. One is Web content mining and second is user profile (preference) mining. Middleware will do the matching part and query evaluation optimization.

1. 1 Motivating example

As a motivating example, assume that we have users looking for a hotel in a certain region. The amount of data is huge and they are distributed over several sites. Moreover users have different preferences which are soft and difficult to express in a standard query language.

From the middleware point of view, there is no chance to evaluate user's query over all data. For middleware we have decided to use Fagin threshold algorithm [10],

which can find best (top-k) answers without looking to all objects. Fagin algorithm works under following assumptions. First, we have the access to objects (in our case hotels) in different lists ordered by user particular attribute ordering, equipped by a numerical score ranging from 0 to 1, e.g. $f_1(x)$ = cheap(x), f_2(x) = close(x),… Second, we have a combination function computing total fuzzy preference value of an object based on preference values of attributes, e.g. @(x) = ((3*cheap(x) + close(x))/4).

In the practical application we have to consider different users with possible different attribute orderings f_1^u, f_2^u and combination functions @^{u}. These represent the overall user preference @^{u}(f_1^u, f_2^u) and the user profile for this task. The task for the user profile mining part is to find these particular attribute orderings and the combination function (using user's ranking of a sample of hotels).

On the web side of our system, the information of vendors, companies or advertisement is very often presented using Web pages in a structured layout containing data records. These serve for company presentation and are assumed to be mainly visited by a potential customer personally.

Structured data objects belong to very important type of information on the Web for systems dealing with competitor tracking, market intelligence or tracking of pricing information from sources like vendors.

We need to bring this data to our middleware. Due to the size of Web, the bottleneck is the degree of automation of data extraction. We have to balance the tradeoff between the degree of automation of Web data extraction and the amount of user (administrator) effort which is needed to train data extractor for a special type of pages (increasing precision).

First restriction we make is that we consider Web pages containing several structured data records. This is usually the case of Web pages of companies and vendors containing information about products and services and, in our case, hotels. Main problem is to extract data and especially attribute values to middleware.

Although we use a system which has the modules in experimental implementation, we do not present this system here. Our main contributions are

- Identification of some uncertainty issues in web content mining system and extracting attribute values from structured pages with several records
- Identification of some uncertainty issues in user profile model and using profile mining methods
- Discussion of coupling of these systems via a middleware based on Fagin threshold algorithm complemented by various storage and querying methods

We point to uncertainty issues by inserting **(UNC)** in the appropriate place in the text.

2 Uncertainty in Web Content Mining

In this section we describe our experience with a system for information extraction from certain types of web pages and try to point out places where uncertainty occurred.

Using our motivation as a running example, imagine a user looking for a hotel in a certain location. A relevant page for a user searching for hotels can look as on Figure 2. Comparing more similar pages would increase the chance of finding the best hotel. An automated tool would enhance this search.

Fig.2. A typical Web page containing several records

For structured Web data extraction it is possible to use semiautomatic systems like Lixto [1], Stalker [16] or WIEN [15]. These require user preannotated pages, which are used in the training process. Moreover, they are most suitable for pages, which have dynamic content, but relatively fixed structure.

Our solution is based on different approach. Instead of training techniques we use the automatic discovery of data regions which encompass multiple similar data records on the page. This is supported by an extraction ontology [9], which is used to extract the values from data records. There are many ways how to search for similar records in source tree. The system IEPAD [4] uses the features of Patricia tree (radix tree) to find the repeating sequences. This system is outperformed by the MDR system [5] which operates directly on the DOM tree of input in which it searches for repeating node sequences with same parent. However, both methods search for objects of interest in the whole web document. This can be time consuming and, as

we have experienced, it surprisingly decreases precision. Furthermore, these systems do not extract attribute values from data records.

In this paper we consider a system as a sequence of both data record extraction and attribute value selection, with possibility of ontology starting almost from scratch (e.g. user search key words).

The system will be described in several phases, which are described in the following sections.

2.1 Data Regions and Data Records Discovery

The first step in the extraction process is the retrieval of relevant web pages. For automatic localization of such resources we use the system Egothor (see [11], [12] and [24]), which is an open-source, high-performance, full-featured text search engine. This system is used for downloading the HTML source codes of relevant pages.

In the next step we build a DOM model [21] of the web page under considerations. This model is used for both data region and data records extraction. Figure 2 shows an example of relevant web page. This page contains summary information about three hotels, i. e. three data records. All of them form a single data region. Our goal is to automatically discover this data region and records within. (It should be noted that the discovery process is not limited to the single-region pages).

To reduce the search space and to increase precision, we prune the input DOM tree, omitting elements which do not contain any textual information in their subtrees. An example of such tree is shown on the Figure 3 – the numbers in black circles represent the relevance of the particular node. Zero-weighted nodes are omitted from the data record search. **(UNC1)** To identify nodes with relevant information in the sub-tree is the first uncertainty problem we point out in our system.

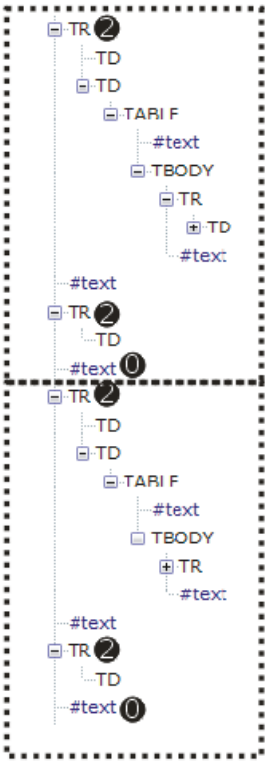


Fig.3. DOM subtree

Next, we use breadth first partial tree alignment to detect data regions and records by taking element tuples, triples etc. and comparing their corresponding subtrees by various metrics (e. g. the tree Levenshtein distance) **(UNC2)** To tune the similarity measures for discovery of similar tags is another uncertainty problem in our system.

Most often every repeated sequence of tags discovered in section 2.1 makes up a real data record (a single hotel). All attributes of this record can be found in one sub-tree and we can proceed to the attribute extraction using the ontology. However, the non-contiguous data records can pose a problem in the region discovery phase.

Typically a data record constitutes a single visual region, nevertheless in the HTML code can two or more records occur in a single table, which means that attributes of these records have a common subtree. It is therefore necessary to identify non-contiguous data records and separate attributes of these records **(UNC3)**.

2.2 Attribute Values Extraction

As we have mentioned before, we use an ontology to extract the actual attribute values of product in the page. This ontology is dynamic – it starts from the scratch, containing user search keywords, and subsequently it evolves with new key words and typical values (using standard vocabularies). It is represented in OWL syntax with additional annotation properties and allows the specification of values extraction parameters: e. g. a regular expression which can be used to match the attribute values, an explicit enumeration of possible attribute values, or the tuning parameters (such as maximum or minimum attribute value length). It is evident, that the richer ontology leads to better results in the extraction process. An example of ontology specification can be seen on Figure 4:

```
<owl:DatatypeProperty rdf:ID="hasPrice">
   <rdfs:domain rdf:resource="#Hotel"/>
   <p1:maxLength
       rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
       10
   </p1:maxLength>
  <p1:pattern rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
     (\$)? ?[\d]{1,10} ?(.){1,3}
   </p1:pattern>
    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
      PRICE
    </rdfs:label>
</owl:DatatypeProperty>
```

Fig.4. An example of ontology

The extraction process can be improved in various ways. We have experimented with data extraction from the detail page (which is usually linked to the summary page), including OCR usage and a special technique based on text difference algorithm and style sheet analysis for better attribute value extraction. Additionally it is possible to employ the approximate regular expression matching, which allows to detect and repair mistyped or mismatched attribute values.

3 Middleware

3.1 Semantic Web infrastructure

User preference mining is done locally and assumes the extracted data are stored in middleware. Extracted data have to be modeled on an OWA (open world assumption) model, and hence traditional database models are not appropriate. We are compatible with a semantic web infrastructure described in [20]. The storage is based on the ideas of Data Pile described in [2]. A typical schema of record resembles a RDF statement with some statements about this statement (nevertheless we do not need reification here).

resource	*attribute*	*value*	*Extracted_from*	*Extracted_by*	*Using_Ontology*
Hotel1	Price	V1	URL1.html	Tool1	O1
Hotel1	Distance	D1	URL1.html	Tool1	O1

If a value of an attribute is missing, for our middleware system it means that a record is missing (thus implementing OWA). Note that we have records without any uncertainty degree attached. Any application can evaluate it according to the remaining values (e. g. it can be known that Tool1 is highly reliable on extracting price, but less on distance).

To know what we are looking for and which attribute values to extract we need to know user interest. For middleware we moreover need to know the ordering of particular attributes and the combination function.

3.2 Usage of user profiles as the user preference model

One possibility to model user preferences is to use user profiles. We work with the assumption that we have a set of user profiles $P_1,\ldots,P_k$ and we know the ideal hotel for each profile. These profiles may be created as the clusters of users or manually by an expert in the field (a hotel-keeper in our example). Manual creation is more suitable because we will know more details about user, but it is often impossible. Independently of the way profiles are created, we have ratings of hotels associated with each profile, thus knowing the best and worst hotels for that profile.

We propose computing the distance d_i of user User1 profile U_1 from each profile P_i in following way

$$\mathrm{d}_i = \frac{\sum_{j=1,\ldots,n} \left| \mathrm{R}ating(User1, o_j) - \mathrm{R}ating(P_i, o_j) \right|}{n} \tag{1}$$

Equation (1) represents the average difference between the user's rating of an object o_j and profile's P_i's rating.

The ideal hotel for the user can be computed as an average of ideal hotels for each profile $\mathrm{P_i}$, weighted by the inverse of distance d_i (see (2)). The average is computed on attributes of hotels. Formally,

$$\text{IdealHotel}(\text{User1}) = \frac{\sum_{i=1,\ldots,k} \text{IdealHotel}(P_i)/d_i}{\sum_{i=1,\ldots,k} 1/d_i} \quad (2)$$

Then, IdealHotel(User1) is the weighted centroid of profiles' best hotels. An example of data, user profiles' best hotel and user's best hotel is on Figure 5. User's best hotel is clearly closest to Profile 3.

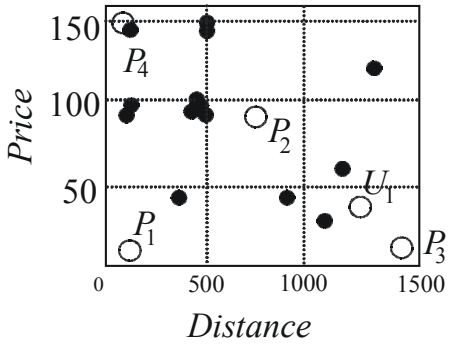


Fig. 5. Positions of best hotels for the user profiles and for the user

After the computation of the ideal hotel for the user, we will use it for computing ratings of remaining hotels. Disadvantage of this user model is that it cannot be used in the Fagin threshold algorithm.

4 Uncertainty in User Preference Mining

In our meaning, user preferences are expressed in the form of classification rules, where the values of attributes are assigned their *grades* corresponding to the orderings of the domains of these attributes. The higher the grade is, the more appropriate (preferable) the value of an attribute is for the given user. This form of grading corresponds to *truth values* well-known in fuzzy community and thus the orderings correspond to *fuzzy functions*.

The combination function can be represented by a fuzzy aggregation function (see [10]). Fuzzy aggregation functions are monotone functions of n variables, with the range of the unit interval [0, 1] of real numbers (in practical applications we use only a finite part of it).

Main assumption of our learning of the user preferences is that we have a (relatively small) sample of objects (hotels) evaluated by the user. We would like to learn his/her preferences from this sample evaluation. The point is to use this learned user preference to retrieve top-*k* objects from a much larger amount of data. Moreover, using the user sample evaluation, we do not have to deal with the problem of matching the query language and document language. These ratings are a form of

QBE – querying by example.

There are many approaches to user modeling, one of the most used is collaborative filtering method [28]. Our method is content based filtering – it uses information about attributes of objects.

4.1 Learning Local Preferences

In [7] and [8] we have described several techniques of learning user's preferences of particular attributes **(UNC5)** represented by fuzzy functions f_1, f_2, … on attribute domains. These techniques use regression methods. A problem occurs here. There can be potentially a big number of hotels of one sort (e.g. cheap ones) but the detection of user preference (cheap, medium or expensive) should not be influenced by the number of such hotels. Regression typically counts number of objects. We have introduced a special technique of discretization to get the user's true local preference (for details see [7] and [8]).

Another approach not using regression is the following. The view of the whole domain of attribute *Price* is in Figure 6. We can see that with increasing price, the rating is decreasing. This can be formalized (details are out of the scope of this paper)

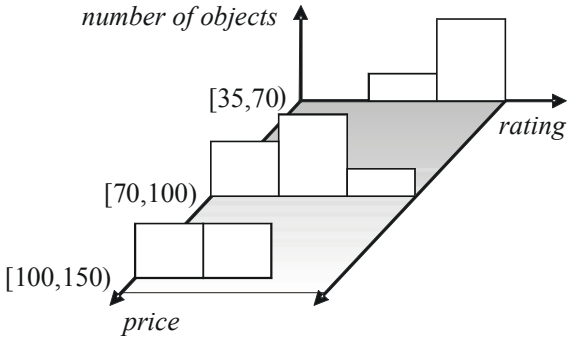


Fig. 6. Ratings for whole attribute domain

and we have experimented also with this possibility. These methods also give local preference in the form of a fuzzy function (here *small*, *cheap*,…) and hence are usable for Fagin Threshold algorithm.

4.2 Learning Combination Function

Second assumption of the Fagin's model [10] is to have a combination function @, which combines the particular attribute preference degrees f_1, f_2, … (local preferences) to an overall score – $@(f_1, f_2, \ldots)$ - according to which the top-k answers will be computed.

There are several ways to learn **(UNC6)** the combination functions and several models. It is an instance of classification trees with monotonicity constraints (see [17], more references to ordinal classification are presented).

We learn the aggregation function by the method of Inductive Generalized Annotated Programming (IGAP) described in [13, 14]. The result of IGAP is a set of Generalized Annotated Program rules in which the combination function has a form of a function annotating the head of the rule – here the quality of hotel:

```
User1_hotel(H) good in degree at least @( f1(x), f1(y), …)
IF User1_hotel_price(x) good in degree at least f1(x) AND
   User1_hotel_distance(y) good in degree at least f2(y)
```

Note that these are rules of generalized annotated programs [25].

5 The Implementation and Experiments

Our Web content mining system has a modular implementation which allows additional modules to be incorporated (e. g. querying with preference-based querying). Communication between modules is based on the traditional Observer/Listener design pattern. All modules, which require communication with other ones, have to implement a Listener interface. All listeners are bound to the central Bus, which manages the communication between them. Each listener can specify a range of broadcasted and received events, which will be supported by it.

We proposed and implemented the middleware system for performing top-k queries over RDF data. As a Java library, our system can be used either on the server side, for example in a Web service, or on the client side. In both cases, it gathers information from local or Web data sources and combines them into one ordered list. To avoid reordering each time a user comes with different ordering, we have designed a general method using B^+ trees to simulate arbitrary fuzzy ordering of a domain ([6]). There are several implemented classes for standard user scoring functions, and Fagin TA and NRA algorithms.

Detailed description of experiments is out of the scope of this paper. We can conclude that experiments have shown this solution is viable.

6 Conclusions and Future Work

Using an experimental implementation, in this paper we have identified several uncertainty challenges, when

(UNC1) identifying HTML nodes with relevant information in the sub-tree,
(**UNC2)** tuning similarity measures for discovery of similar tag subtrees,
(UNC3) identifying single data records in non-contiguous html source,
(UNC4) extracting attribute values
(UNC5) learning user's preferences of particular attributes
(UNC6) learn the user preference combination function.

We have experimented with some candidate solutions.

Models and methods in these experiments can be based on models of fuzzy description logic (FDL).

One possibility is to use a FDL with both concepts and roles fuzzified (see e. g. [26]). One problem of embedding FDL with fuzzy roles into OWL is that they consist of subject, predicate, object and the fuzzy value. This cannot be directly modeled by RDF data.

Second possibility is to use a FDL where only concepts are fuzzified and roles remain crisp (and hence both roles and fuzzy concepts can be modeled by RDF data). One such example is *fEL@* introduced in [27].

Acknowledgement. This work was supported in part by Czech projects 1ET 100300517 and 1ET 100300419 and Slovak projects VEGA 1/3129/06 and NAZOU.

References

1. Baumgartner, R., Flesca, S., Gottlob, G.: Visual Web Information Extraction. VLDB Conference, 2001
2. Bednárek, D., Obdržálek, D., Yaghob, J., Zavoral, F.: Data Integration Using DataPile Structure, In: Proceedings of the 9th East-European Conference on Advances in Databases and Information Systems, ADBIS 2005, Tallinn, ISBN 9985-59-545-9, 2005, 178-188
3. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. In: Scientific American Magazine, May 2001
4. Chang, C-H., Lui, S-L.: IEPAD: Information extraction based on pattern discovery. WWW-10, 2001.
5. Liu, B., Grossman, R., Zhai, Y.: Mining Data Records in Web Pages. In: Proc S IGKDD.03, August 24-27, 2003, Washington, DC, USA.
6. Eckhardt, A., Pokorny, J., Vojtas, P.. A system recommending top-*k* objects for multiple users preferences. In: 2007 IEEE Conference on Fuzzy Systems, IEEE 2007, 1101 – 1106
7. Eckhardt, A., Horváth, T., Vojtáš, P.: PHASES: A User Profile Learning Approach for Web Search. Accepted as short paper for WI'07 Web Intelligence Conference, November 2007, Fremont CA
8. Eckhardt, A., Horváth, T., Vojtáš, P.: Learning different user profile annotated rules for fuzzy preference top-k querying. Accepted for SUM'07 Scalable Uncertainty Management Conference, October 2007, Washington DC Area
9. Embley, D. W., Campbell, D. M., Smith, R. D., Liddle, S. W.: Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Documents. CIKM 1998, 52-59
10. Fagin R., Lotem A., Naor M.: Optimal Aggregation Algorithms for Middleware. In Proc. 20th ACM Symposium on Principles of Database Systems, 102-113, 2001
11. Galamboš L.: Dynamization in IR Systems. In: Proc. IIPWM '04 - Intelligent Information Processing And Web Mining, ed. M. A. Klopotek, Springer 2004, 297-310
12. Galamboš, L.: Semi-automatic stemmer evaluation, ibid. 209-218.
13. Gurský, P., Horváth, T. , Novotný, R., Vaneková, V., Vojtáš, P.: UPRE: User preference based search system, 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06), IEEE 2006, pp.841-844

14. Horváth, T., Vojtáš, P.: Ordinal Classification with Monotonicity Constraints. In. Proceedings of the 6th Industrial Conference on Data Mining (ICDM '06), Leipzig, Germany, 2006: LNAI 4065, Springer, 2006, ISBN 3-540-36036-0, p: 217-225
15. Kushmerick, N.: Wrapper induction: efficiency and expressiveness. Artificial Intelligence, 118:15-68, 2000
16. Muslea, I., Minton, S., Knoblock C..: A hierarchical approach to wrapper induction. Conf. on Autonomous Agents, 1999
17. Potharst, R., Feelders, A. J.: Classification trees for problems with monotonicity constraints. In: ACM SIGKDD Explorations Newsletter archive Volume 4 , Issue 1 (June 2002): ACM Press, 2002, p: 1-10
18. Turtle, H. R., Croft, W. B.: Uncertainty in Information Retrieval Systems. In: Proc. Second Workshop Uncertainty Management and Information Systems: From Needs to Solutions, Catalina, Calif., 1993 as quoted in S. Parsons. Current Approaches to Handling Imperfect Information in Data and Knowledge Bases. IEEE TKDE 8,3 (1996) 353-372
19. Vojtáš, P.: EL description logic with aggregation of user preference concepts. In: Duží, M. et al. Eds. Information modeling and Knowledge Bases XVIII, IOS Press, Amsterdam, 2007, 154-165
20. Yaghob, J., Zavoral, F.: Semantic Web Infrastructure using DataPile, In: Proc. 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Imtelligent Agent Technology, Butz, C. J. et al. (eds.), IEEE 2006, 630-633
21. W3C Document Object Model. http://www.w3.org/DOM/
22. Charter of W3C Uncertainty Reasoning for the World Wide Web Incubator Group, http://www.w3.org/2005/Incubator/urw3/charter
23. Wiki of W3C Uncertainty Reasoning for the World Wide Web XG Search: http://www.w3.org/2005/Incubator/urw3/wiki/FrontPage
24. http://www.egothor.org/
25. Kifer, M., Subrahmanian, V. S.: Theory of generalized annotated logic programming and its applications, J. Logic Programing, 12 (1992) pp 335–367
26. Stoilos, G.. Stamou, G., Tzouvaras, V., Pan, J. Z., Horrocks, I.: The fuzzy description logic f-shin. Proceedings of URSW, 2005.
27. Vojtáš, P.: EL description logic with aggregation of user preference concepts. M. Duží et al. Eds. Information modelling and Knowledge Bases XVIII, IOS Press, Amsterdam, 2007, 154-165
28. Aggarwal, C.: Collaborative Crawling: Mining User Experiences for Topical Resource Discovery, IBM Research Report, 2002.

Position Papers

Axiom-oriented Reasoning to Deal with Inconsistency Between Ontology and Knowledge Base

Tuan A. Luu[1], Tho. T Quan[1], Tru H. Cao[1] and Jin-Song Dong[2]

[1]Faculty of Computer Science and Engineering
Ho Chi Minh City University of Technology
Vietnam
[2] School of Computing
National University of Singapore
Singapore
qttho@cse.hcmut.edu.vn

Abstract. When deployed in practical applications, Ontologies and KBs often suffer various kinds of inconsistency, which limit the applications performances significantly. In this paper, we propose a framework to reason inconsistency between Ontology and KB and refine the inconsistency accordingly. To make our framework efficient, we only focus on reasoning a part responsible for the inconsistency, rather than the whole structures of Ontology and KB. Moreover, to improve the execution speed of algorithms employed in the framework, we also discuss an axiom-oriented strategy to reason on a reduced space of formula to be inferred in Ontology and KB.

1 Introduction

The Semantic Web [1] is developed as a concept of how computers, people, and the Web can work together more effectively than it is possible now. Ontology and Knowledge Base (KB) are two significant elements of the Semantic Web. However, when used in practical applications, Ontologies and KBs always suffer inconsistencies due to various reasons. In recent literature, there are two emerging approaches following this direction: to diagnose and repair inconsistency in Ontology by finding minimal inconsistent subset [2] ; and to reason in inconsistent Ontology and KB based on maximum consistent subset constructed [3] .

In this paper, we propose a framework to handle inconsistency between Ontology and KB. It is done by reasoning to find the part responsible for the inconsistency and then refining the detected inconsistencies accordingly. In addition, to reduce the complexity cost of algorithms employed in the framework, we also develop an axiom-oriented strategy to isolate and detect the axioms responsible for the inconsistency. The rest of the paper is organized as follows. Section 2 presents formal definitions of Ontology and Knowledge Base. Section 3 discusses inconsistency between Ontology and KB. In Section 4, the general framework for inconsistency detecting and repairing is given. Section 5 gives discussion of the axiom-oriented strategy to deal with inconsistency. Finally, Section 6 concludes the paper.

2 Ontology and Knowledge Base

Definition 1 (Ontology). An ontology is a structure $O = (C; T; R; A; \leq_C; \leq_T; \delta_R; \delta_A; \tau_T; S_A)$. It consists of disjoint sets of concepts (or classes) C, types T, relations R, attributes A, and values V. The partial orders $\leq_C$ (on C) and $\leq_T$ (on T) define a concept hierarchy and a type hierarchy, respectively. The function δ_R: $R \rightarrow C^2$ provides relation signatures (i.e., for each

relation, the function specifies which concepts may be linked by this relation); while the function δ_A: $A \rightarrow C \times T$ provides attribute signatures (for each attribute, the function specifies to which concept the attribute belongs and what is its data type); and τ_T: $T2^V$ is the assignment of values to types. S_A is a set of axioms, restrictions between concepts and attributes.

Example 1. We define *Football Ontology* $O = (C; T; R; A; \leq_C; \leq_T; \delta_R; \delta_A; \tau_T; S_A)$ where

C = {football-player, person, club, city }
$\leq_C$ = {football-player $\subseteq$ person}
T = {integer}
R = {live-in, locate-in, play-for, has-wife}
A = {age, height, weight}
δ_R = {live-in → football-player x city,live-in → person x city,locate-in → club x city, play-for → football-player x club, has-wife → football-player x football-player}
δ_A = {age→football-playerx integer,height → football-player x integer, weight → football-player x integer}
S_A = {(O_1) football-player(x) ∧ club(y) ∧ city (z) ∧ play-for(x, y) ∧ locate-in(y, z) → live-in(x, z) *// football player plays for club will live in the cty that the club locates*.
(O_2) football-player(x) ∧ city(y) ∧ city (z) ∧ live-in(x, y) ∧ live-in(x, z) → y = z *// football player is not living in more than one city*.
(O_3) football-player(x) ∧ has-wife(x, y)∧city(z) ∧ live-in(y, z) → live-in(x,z) *// football player who has wife will lives in the city will live in the same city as her wife's*.
(O_4) club(x) ∧ locate-in(x, z) ∧ club(y) ∧ locate-in(y, z) → x = y *// each city has not more than one club*.}

Definition 2 (Knowledge Base). A Knowledge Base (KB) is a structure $K = (C; R; A; I; V; \tau_C; \tau_R; \tau_A)$. It consists of disjoint sets of concepts (or classes) C, relations R, attributes A, individuals I and values V. The function τ_C: $C2^I$ is the assignment of instances to concepts), the function τ_R: $R \rightarrow 2^{I \times I}$ defines relations between instances, and τ_A: $A \rightarrow 2^{I \times V}$ defines attributes of instances.

Example 2. We define *Football KB as* K $= (C; R; A; I; V; \tau_C; \tau_R; \tau_A)$ where:

I = {Beckham, MU, Manchester, Liverpool, Chelsea, Maria)
τ_C = {(K_5) football-player (Beckham), (K_6) club (MU), (K_7) city (Manchester), (K_8) city (Liverpool), (K_9) club (Chelsea)}
τ_R = {(K_{10}) live-in (Beckham, Liverpool), (K_{11}) play-for (Beckham, MU), (K_{12}) locate-in (MU, Manchester), (K_{13}) has-wife (Beckham, Maria), (K_{14}) live-in (Maria, Manchester), (K_{15}) locate-in (Chelsea, Manchester)}
τ_A = { (K_{16}) age (Beckham, 30), (K_{17}) height (Beckham, 180),(K_{18}) weight (Beckham,80)}

3 Inconsistency between Ontology and KB

Although KB (containing concrete data) is always encoded with respect to an ontology (containing a general conceptual model of some domain knowledge), people may find it difficult to understand the logical meaning of the underlying ontology. Hence, people may fail to formulate precisely axioms, which are logically correct, or may specify contradictory statements.

Example 3. Between in *Football Ontology* and *Football KB* defined respectively in Example 1 and Example 2, from (K_5), (K_{10}), (K_{13}), and (K_{14}), we can infer that Beckham lives in Liverpool but has wife living in Manchester. However, from (O_3) we can see that Beckham must live in the same city with his wife. Thus, *Football Ontology* and *Football KB* are inconsistent.

4 Framework for Diagnosing and Repairing Inconsistency Between Ontology and KB

In this section, we present a framework to reason inconsistency between Ontology and KB. The framework is conducted by incorporating the algorithm for debugging

inconsistency proposed in [2] and the basic theory of finding the inconsistency introduced in [3] . As shown in Figure 1, the proposed framework consists of three steps as follows:

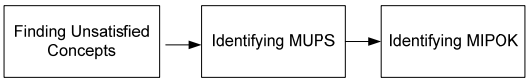


Figure 1. Framework for diagnosing and repairing inconsistency between Ontology and KB

- *Step 1*: It finds all unsatisfied concepts. An unsatisfied concept is a concept that does not have any individual for all models of Ontology and KB.
- *Step 2*: For every unsatisfied concept, we identify a minimal subset axioms and facts that are responsible for an inconsistency, called Minimal Unsatisfied Preserving Sub Ontology and KB (MUPS).
- *Step 3*: From the set of MUPS, we diagnose the smallest subsets of axioms and facts responsible for all inconsistencies, or Minimal Inconsistent Preserving Sub Ontology and KB (MIPOK). Finally, relying on this MIPOK, we will repair this Ontology and KB.

Example 4. We apply the proposed framework to deal with inconsistency between *Football Ontology* and *Football KB* given in Example 1 and Example 2. As a result, *Refined Football Ontology* is redefined as $O_R = (C; T; R; A; \leq_C; \leq_T; \delta_R; \delta_A; \tau_T; S_A)$, where:

C = {football-player, person, club, city }
$\leq_C$ = {football-player $\subseteq$ person}
T = {integer}
R = {live-in, locate-in, play-for, has-wife}
A = {age, height, weight}
δ_R = {live-in → football-player x city, live-in → person x city, locate-in → club x city, play-for → football-player x club, has-wife → football-player x football-player}
δ_A = {age → football-player x integer, height → football-player x integer, weight → football-player x integer}
S_A = {(O_1) football-player(x) ∧ club(y) ∧ city (z) ∧ play-for(x, y) ∧ locate-in(y, z) → live-in(x, z) // *football player plays for club will live in the city that the club locates*.
(O_2) football-player(x) ∧ city(y) ∧ city (z) ∧ live-in(x, y) ∧ live-in(x, z) → y = z // *football player is not living in more than one city.*
(O_3) football-player(x) ∧ has-wife(x, y) ∧ city (z) ∧ live-in(y, z) → live-in(x, z) // *football player who has wife will lives in the city will live in the same city as her wife's*.}

Refined Football KB is redefined *as* $K_R = (C; R; A; I; V; \tau_C; \tau_R; \tau_A)$ where:

I = {Beckham, MU, Manchester, Liverpool, Chelsea, Maria)
V = {30, 80, 180}
τ_C = {(K_5) football-player (Beckham), (K_6) club (MU), (K_7) city (Manchester), (K_8) city (Liverpool), (K_9) club (Chelsea)}
τ_R = {(K_{11}) play-for (Beckham, MU), (K_{12}) locate-in (MU, Manchester), (K_{13}) has-wife (Beckham, Maria), (K_{14}) live-in (Maria, Manchester),(K_{15}) locate-in (Chelsea, Manchester)}
τ_A = {(K_{16}) age (Beckham, 30), (K_{17}) height (Beckham, 180), (K_{18}) weight (Beckham, 80)}

5 Axiom-oriented Construction of MUPS

In [2] and [3] , the authors have proposed an algorithm to find MUPS, as presented in Figure 2. However, because we only focus on solving the inconsistency between Ontology and KB, i.e. inconsistency occurs in the relations between facts and axioms, so we can apply an axiom-oriented strategy in the selection function. It is carried out using the following selection rules.

Rule 1 (Axiom-Related Selection). Only add to the *final_set* mentioned in Algorithm 1

formulae that are not only directly relevant to this set but also directly relevant to at least an axiom in Ontology.

Rule 2 (Onto-KB Selection). Only consider the *subset* S_1 and *subset* T_1 mentioned in Algorithm 1 if the formulae in them occur in both Ontology and KB.

Algorithm 1. Finding MUPS of an unsatisfied concept *c*.

Input: Unsatisfied concept *c* with set of formulae $\sum$.
Output: set MUPS corresponding to *c*.
Process:

1: **set** $S = \{c\}$, *final_set* = Ø.
2: **from** S find set of formulae S' that is **directly relevant** to S.
3: **if** S' is consistent **then**
4: **set** $S = S'$.
5: **repeat**
6: Find new set of formulae S' that is **direct relevant** to S
7: **if** S' is consistent **then** $S = S'$
8: **until** c is inconsistent in S'
9: **end if**
10: **set** $T = S' - S$
11: **for** all subset T_1 of T and all subset S_1 of S
12: **if** c is inconsistent in $\{T_1 \cup S_1\}$ **then** *final_set* = *final_set* $\cup \{T_1 \cup S_1\}$
13: **end for**
14: *MUPS*($\sum$, *c*) := Minimality-Checking(*final_set*)
15: **return** *MUPS*($\sum$, c)

Figure 2. Algorithm for finding MUPS of an unsatisfied concept *c*.

Example 5. Consider *Football Ontology* and *Football KB* given in Example 1 and Example 2. The effectiveness of using axiom-oriented approach is demonstrated, as the numbers of subsets generated when calculated MUPS($\sum$, *football-player*) are 2^{32} and $2^{26} - 2^{21}$ in non axiom-oriented and axiom-oriented methods, respectively.

6 Conclusion

In this paper, we first introduced inconsistency occurring between Ontology and KB. Then, we proposed some refinements and improvements for an effective framework to solve the inconsistency between Ontology and KB in the reasonable complexity and time. Generally, our proposed framework only focuses on axioms, rather than the whole structure of ontology. Hence, our approach is highly potential in terms of reducing computational cost, as compared to similar existing work.

References

[1] Tim Berners-Lee, James Hendler and Ora Lassila. *The Semantic Web.* Scientific American, 284(5) , pages 34-43, May 2001.

[2] Stefan Schlobach and Zhisheng Huang, *Inconsistent Ontology Diagnosis: Framework and Prototype*. SEKT/2005/D3.6.1/v0.9, SEKT EU-IST-2003-506826, October 2005.

[3] Zhisheng Huang, Frank van Harmelen, and Annette ten Teije. *Reasoning with Inconsistent Ontologies*. Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI 2005), August 2005.

[4] Marc Ehrig, Peter Haase, Mark Hefke, Nenad Stojanovic. *Similarity for Ontologies – a Comprehensive Framework*. Proceedings of the 13th European Conference on Information Systems (ECIS 2005), May 2005.

[5] Steffen Staab and Heiner Stuckenschmidt. *Semantic Web and Peer-to-Peer*. Springer-Verlag Berlin Heidelberg, 2006.

Uncertain Reasoning for Creating Ontology Mapping on the Semantic Web

Miklos Nagy[1], Maria Vargas-Vera[2], and Enrico Motta[1]

[1] Knowledge Media Institute (KMi)
The Open University
Walton Hall, Milton Keynes
MK7 6AA, United Kingdom
mn2336@student.open.ac.uk,e.motta@open.ac.uk
[2] Department of Information Systems
Poznan University of Economics
al. Niepodleglosci 10, 60-967 Poznan, Poland
maria@kie.ae.poznan.pl

Abstract. Mapping ontologies with high precision on the Semantic Web is a challenging problem that needs to be addressed in various domains. One of the main problems with any mapping process, which needs to be applied on different domains is that it always has a certain degree of uncertainty associated with it. In this paper we introduce a method based on Dempster-Shafer theory that use uncertain reasoning over the possible mappings in order to select the best possible mapping without using any heuristic or domain specific rules.

1 Introduction

The problem of mapping two ontologies effectively and efficiently is a necessary precondition to integrate information on the Semantic Web. In recent years different research communities have proposed[1] a wide range of methods for creating such mappings. The proposed methods usually combine syntactic and semantic measures by introducing different techniques ranging from heuristics to machine learning. While these methods perform well in certain domains the quality of the produced mappings can differ from domain to domain depending on the specific parameters defined in the methods e.g. tuning similarity threshold.

We have developed a multi agent ontology mapping framework [2–4] in the context of Question-Answering over heterogeneous sources, where each agent can build mapping between a user's query and the ontology concepts. Our objective was to produce a ontology mapping method that does not depend on any fine tuned internal parameters for a specific domain or does not assume having large amount of data samples a-priory for machine learning or Bayesian probability assessment. Our hypothesis is that the correctness of different similarity mapping algorithms is always heavily dependent on the actual content and conceptual structure of these ontologies which are different even if two ontologies have been created on the same domain but with different purpose. Therefore

from the mapping point of view these ontologies will always contain inconsistencies, missing or overlapping elements and different conceptualisation of the same terms, which introduces a considerable amount of uncertainty into the mapping process. In this paper we introduce a novel method how these uncertainties can be harnessed in order to improve the correctness of the mappings.

2 Similarity

In order to assess similarity we need to compare all concepts and properties from $Ontology1$ to all concepts and properties in $Ontology2$. Our similarity assessments, both syntactic and semantic produce a sparse similarity matrix where the similarity between C_n from $Ontology1$ and C_m in $Ontology2$ is represented by a particular similariy measure between the i and j elements of the matrix as follows:

$$SIM := (s_{i,j})_{n \times m}$$
$$1 \leq i \leq n \; and \; 1 \leq j \leq m$$

where SIM represents a particular similarity assessment matrix, s is a degree of similarity that has been determined by a particular similarity e.g. Jaccard or semantic similarity measure. We consider each measure as an "expert" which assess mapping precision based on its knowledge. Therefore we assume that each similarity matrix is a subjective assessment of the mapping what needs to be combined into a coherent view. If combined appropriately this combined view provides a more reliable and precise mapping that each separate mapping alone. However one similarity measure or some technique can perform particularly well for one pair of concepts or properties and particularly badly for another pair of concepts or properties, which has to be considered in any mapping algorithm.

3 Belief over the mapping

In our ontology mapping method we assume that each expert carries only partial knowledge of the domain and can observe it from its own perspective where available prior knowledge is generally uncertain and subjective. In order to represent these subjective probabilities in our system we use the Dempster-Shafer theory of evidence [5], which provides a mechanism for modeling and reasoning uncertain information in a numerical way, particularly when it is not possible to assign belief to a single element of a set of variables. Missing data (ignorance) can also be modeled by Dempster-Shafer approach and additionally evidences from two or more sources can be combined using Dempster's rule of combination. The combined support, disbelief and uncertainty can each be separately evaluated. The main advantage of the Dempster-Shafer theory is that it provides a method for combining the effect of different learned evidences to establish a new belief by using Dempster's combination rule.

The following elements have been used in our system in order to model uncertainty:

Frame of Discernment(Θ) :finite set representing the space of hypothesizes. It contains all possible mutually exclusive context events of the same kind.

$$\Theta = \{H_1, ..., H_n, ...H_N\} \tag{1}$$

In our method Θ contains all possible mappings that have been assessed by the particular expert.

Evidence:available certain fact and is usually a result of observation. Used during the reasoning process to choose the best hypothesis in Θ. We observe evidence for the mapping if the expert detects that there is a similarity between C_n from O_1 and C_m in O_2.

Belief mass function (m): is a finite amount of support assigned to the subset of Θ. It represents the strength of some evidence and

$$\sum_{A \subseteq \Theta} m_i(A) = 1 \tag{2}$$

where m_i(A) is our exact belief in a proposition represented by A that belongs to expert i. The similarity algorithms itself produce these assignment based on different similarity measures. In practice we assess up to 8 inherited hypernyms similarities with different algorithms (considered as experts) which can be combined based on the combination rule in order to create a more reliable mapping. Once the combined belief mass functions have been assigned the following additional measures can be derived from the available information.

Belief: amount of justified support to A that is the lower probability function of Dempster, which accounts for all evidence E_k that supports the given proposition A.

$$belief_i(A) = \sum_{E_k \subseteq A} m_i(E_k) \tag{3}$$

An important aspect of the mapping is how one can make a decision over how different similarity measures can be combined and which nodes should be retained as best possible candidates for the match. To combine the qualitative similarity measures that have been converted into belief mass functions we use the Dempster's rule of combination and we retain the node where the belief function has the highest value.

Dempster's rule of combination:Suppose we have two mass functions $m_i(E_k)$ and $m_j(E_{k'})$ and we want to combine them into a global $m_{ij}(A)$. Following Dempster's combination rule

$$m_{ij}(A) = m_i \oplus m_j = \sum_{E_k E_{k'}} m_i(E_k) * m_j(E_{k'}) \tag{4}$$

where i and j represent two different experts.

The belief combination process is computationally very expensive and from an engineering point of view, this means that it not always convenient or possible

to build systems in which the belief combination process is performed globally by a single unit. Therefore, applying multi agent architecture is an alternative and distributed approach to the single one and in this case there is no more a single agent having the global view of the system, but each agent has partial view of it. This allows that the computational load can be divided among the agents of the group. Our algorithm takes all the concepts and its properties from the different external ontologies and assesses similarity with all the concepts and properties in the query graph.

4 Conclusions

Inconsistency and incompleteness are important problems that affect the Semantic Web therefore ontology mapping systems that operate in this environment should have the appropriate mechanisms to cope with these issues. The main contribution of our research is the use of Dempster-Shafer theory for assessing whether similar terms in different ontologies refer to the same or similar concepts. Our preliminary results have shown that using Dempster-Shafer theory is a promising approach and needs to be investigated further in ontology mapping context since in this form and context has not been done so far. We believe that this is because Dempste-Shafer combination rules can be unfeasible in domains with large number of variables. In our future research we will investigate how these optimization methods can be adapted and applied in our scenario with a dynamic multi agent environment where each agent has partial knowledge of the domain.

Acknowledgements: This research project has been supported by a Marie Curie Transfer of Knowledge Fellowship of the European Community's Sixth Framework Programme under the contract number MTKD-CT-2004-509766 (enIRaF).

References

1. Choi N.,Song I-Y., Han H.(2006). A survey on ontology mapping. SIGMOD Record 35(3): 34-41.
2. Vargas-Vera M., and Motta E. (2004). An Ontology-driven Similarity Algorithm. *KMI-TR-151*, Knowledge Media Institute, The Open University, UK.
3. Vargas-Vera M.,Motta E., and Domingue J. (2003). AQUA: An Ontology-Driven Question Answering System. *AAAI Spring Symposium, New Directions in Question Answering*, Stanford, USA.: AAAI Press.
4. Nagy M.,Vargas-Vera M., and Motta E. (2005) Multi-agent Ontology Mapping Framework in the AQUA Question Answering System. the Fourth International Mexican Conference on Artificial Intelligence (MICAI-2005), Lecture Notes in Artificial Intelligence LNAI 3789, Gelbukh, A de Albornoz and H. Terashima (Eds), pp. 70-79, Monterrey Mexico.
5. Shafer G. (1976). A Mathematical Theory of Evidence.: Princeton University Press.

A Fuzzy Ontology-Approach to improve Semantic Information Retrieval

Silvia Calegari[1] and Elie Sanchez[2]

[1] Dipartimento Di Informatica, Sistemistica e Comunicazione
Università di Milano – Bicocca
V.le Sarca 336/14, 20126 Milano (Italia)
calegari@disco.unimib.it
[2] LIF, Biomathematiques et Informatique Medicale
Faculte de Medecine (Universite Aix-Marseille II)
27 Bd Jean Moulin, 13385 Marseille Cedex5, (France)
elie.sanchez@medecine.univ-mrs.fr

Abstract. This paper shows how a Fuzzy Ontology based approach can improve semantic documents retrieval. After formally defining a Fuzzy Knowledge Base, it is discussed a special type of new non-taxonomic fuzzy relationships, called (semantic) correlations. These correlations, first assigned by experts, are updated after querying, or when a document has been inserted into a database. It is then introduced an Information Retrieval algorithm that allows to derive a unique path among the entities involved in the query in order to obtain maxima semantic associations in the knowledge domain.

1 Introduction: Fuzzy Ontology and Fuzzy Knowledge Base

Ontologies in the sense of a formal, explicit specification of a shared conceptualisation [1], constitute a key component of the Semantic Web, facilitating a machine processable representation of information. Two-valued-based logical methods are insufficient to handle ill-structured, uncertain or imprecise information encountered in real world knowledge. A tolerance for imprecision, by a positive use of Fuzzy Logic may be exploited to enhance the power of the Semantic Web [2, 3]. It has been shown that Fuzzy Logic allows to bridge the gap between human-understandable soft logic and machine-readable hard logic. Indeed there has been a natural integration of Fuzzy Logic in Ontology in order to define a new theoretical paradigm called Fuzzy Ontology [4, 5, 6].

Recently, an increasing number of approaches to Information Retrieval have proposed models based on concepts rather than on keywords. So that, in this work, ontologies have been combined to objects (stored in a database) in order to search new documents semantically correlated to user's query.

In this paper, the notion of Fuzzy Concept Network (FCN), introduced in [7], is extended incorporating Database Objects so that, concepts and documents can similarly be represented in the network. It is then introduced and described an Information Retrieval algorithm using an Object-Fuzzy Concept Network (O-FCN). This algorithm allows to derive a unique path among the entities involved in the query in order to obtain the maximum semantic associations in the knowledge domain.

It will now be introduced a formal Fuzzy Ontology (see also [4, 5]). This approach depends purely on an application choice. Indeed, we consider a formal Fuzzy Ontology as a quadruple $\mathbf{O}_{\mathcal{F}} = \{\mathbf{C}, \mathbf{R}, \mathbf{F}, \mathbf{A}\}$ where $\mathbf{C}$ is a set of fuzzy concepts, or entities indifferently. The set of entities of the fuzzy ontology will be indicated by $\mathbf{E}$. $\mathbf{R}$ is a set of fuzzy relations. Each $R \in \mathbf{R}$ is a n-ary fuzzy relation on the domain of entities $R : \mathbf{E}^n \mapsto [0, 1]$. In particular, $\mathbf{R} = \mathcal{T} \cup \mathcal{T}_{not}$ where $\mathcal{T}$ is the set of the taxonomic relations and $\mathcal{T}_{not}$ is the set of the non-taxonomic relations. $\mathbf{F}$ is a set of fuzzy relations on the set of entities $\mathbf{E}$ and a specific domain contained in $\mathcal{D} = \{integers, strings, ...\}$, and $\mathbf{A}$ is a set of axioms expressed in an proper logical language.

Note that even an OWL ontology "may" only include instances: we separated them in our approach, the advantage is that we can have one ontology and multiple instances that conform to it. Using this definition, it is possible to introduce the notion of Fuzzy Knowledge Base. Our definition is based on the vision of an ontology for the Semantic Web where knowledge is expressed in a Description Logic-based ontology as a triple $\langle \mathcal{T}, \mathcal{R}, \mathcal{A} \rangle$ where $\mathcal{T}$, $\mathcal{R}$ and $\mathcal{A}$ are respectively a TBox, RBox and ABox [8]. Thus, by using a fuzzy ontology the knowledge of a domain is defined in order to correspond to a Description Logic (DL) knowledge base.

Definition 1. *A Fuzzy Knowledge Base is a couple defined as:*

$$\mathbf{KB}_{\mathcal{F}} = (\mathbf{O}_{\mathcal{F}}, \mathcal{I})$$

where $\mathbf{O}_{\mathcal{F}}$ is a Fuzzy Ontology as previously defined and $\mathcal{I}$ is a set of instances associated with the fuzzy ontology. Furthermore, every concept $C \in \mathbf{C}$ is a fuzzy set on the domain of the instances defined as $C : \mathcal{I} \mapsto [0, 1]$.

In this context the set $\mathcal{I}$ is identified with the objects stored in the database, i.e. $\mathbf{O_{DB}} = \mathcal{I}$ and $C : \mathbf{O_{DB}} \mapsto [0, 1]$. In particular the set of objects can consist of documents, digital pictures, notes and so on, i.e. $\mathbf{O_{DB}} = \{\mathcal{D}, \mathcal{P}, \mathcal{N}, \dots\}$ where $\mathcal{D}$ is a set of documents, $\mathcal{P}$ is a set of digital pictures, and $\mathcal{N}$ is a set of notes, etc.

A new fuzzy relationship: Correlation. In the Semantic Web area of research, a crucial topic is to define a dynamic knowledge of a domain adapting itself to the context. *In order to achieve this aim, it is needed to handle the trade off between the correct definition of an object (given by the ontology structure) and the actual meaning assigned to the artifact by humans (i.e. the experience-based context assumed by every person according to his specific knowledge).*

In [7] it has been proposed a system that allows to achieve these objectives. It consists in the determination of a semantic correlation among the entities that are searched together, for example, in a query or when a document has been inserted into the database. In particular, a fuzzy weight on the correlations is also assigned during the definition of the ontological domain by an expert according to his/her experience. A *correlation* is a binary non-taxonomic fuzzy relation: $corr : \mathbf{E} \times \mathbf{E} \mapsto [0, 1]$, where $\mathbf{E} = \{e_1, e_2, \dots, e_n\}$ is the set of the entities contained in the ontology. This defines how the entities are linked semantically. The closer to 1 is the $corr$ value, the more the two considered entities are semantically associated.

In this way, the fuzzy ontology gives a solution to the trade off of the knowledge base and allows to dynamically adapt itself to the context in which it is introduced.

2 Information Retrieval Algorithm using O-FCN and its Evaluation

In [7] we introduced a Fuzzy Concept Network (FCN) to represent the dynamical behaviour of the fuzzy ontologies. In particular, the FCN representation lets us introduce a new semantic network based on the correlations defined in the fuzzy ontology. But an ontology allows to handle a complete knowledge base and so to make reasoning on the instances. In this work we extend this possibility by inserting directly in the FCN the objects of the domain stored into the database. In this way, we can reason directly with the elements of the specific application only visiting the FCN graph. In the following an extended FCN definition is given in order to insert the objects of the domain:

Definition 2. *An Object-Fuzzy Concept Network (O-FCN) is a weighted graph* $\mathcal{N}_{fo} = \{\mathbf{O_{DB}}, \mathcal{N}_\mathbf{f}\}$*, where* $\mathbf{O_{DB}}$ *is the set of the objects stored in the database and* $\mathcal{N}_f = \{\mathbf{E}, F, m\}$ *is a Fuzzy Concept Network (FCN). Each object is described by the entities of the FCN , i.e.* $\forall o_i \in \mathbf{O_{DB}}\ \mathbf{o_i} = \{\mathbf{e_1}, \ldots, \mathbf{e_n}\}$ *where* $\mathbf{e_i}, \ldots, \mathbf{e_n} \in \mathbf{E}$.

The set $\mathbf{O_{DB}}$ identifies all the information that is contained into the database, such as documents, digital pictures, videos, and so on.

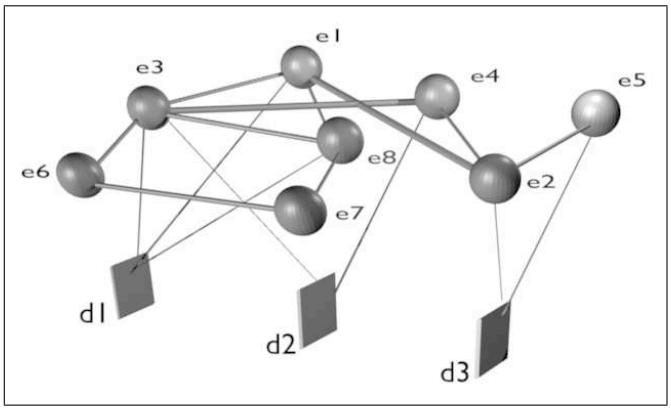


Fig. 1. A graphical representation of an Object-Fuzzy Concept Network.

In Fig. 1 it is given a 3D graphical representation of the prototype of a small O-FCN. The different thickness of the links identifies how strongly the entities are correlated. The thicker the link the more correlated are the two entities (i.e. the closer to 1 is the fuzzy value).

A recent application of Information Retrieval System (IRS) is the Semantic Web area of research. Indeed, the necessity of a better definition of IRS emerged in order to retrieve semantic information considered useful to a user query. Information Retrieval is a domain that involves the organization, storage, retrieval and display of information [9]. In order to extend the query vector it has been proposed a new algorithm based on

fuzzy ontology. When navigating the O-FCN it is possible to find semantic links among the concepts: for each term specified in the query, a unique path is defined at each step, corresponding to the maximum value correlation. A step-by-step brief description of this new algorithm is given below (see also Fig. 2):

'O-FCN'-IR Search (E_q : word vector)
1: 'O-FCN'-based E_q extension (pruning phase)
2: 'O-FCN'-based documents extraction
3: 'O-FCN'-based relevance calculation (cosine distance)
return ranking of the documents

Fig. 2. New Information Retrieval Algorithm using O-FCN

The O-FCN has been involved in all the steps of the algorithm in order to semantically enrich the results that were obtained. In this way, to retrieve documents it is easier to process than from the previous one that used only FCN [7]. The algorithm input is a vector E_q identifying the terms in the query. The first step (1) uses these terms to locate the unique path finding maximum correlation value among them. E_q is extended navigating the O-FCN recursively. Now, the "pruning phase" is directly inserted into the query extension algorithm. In this way, it is possible to find immediately the important entities, which are more semantically correlated w.r.t. the E_q set. In step (2) the O-FCN has been involved in order to directly extract the documents by the network. Whereas in the last step, O-FCN is used to calculate the relevance of the documents in order to sort them in decreasing order. The final score of a document is evaluated through a cosine distance among the weights of each entity. This is done for normalisation purposes. Such a value is finally sorted in order to obtain a ranking among the documents.

Evaluation A creative learning environment is the context chosen to test the new Information Retrieval algorithm based on O-FCN. In particular, the ATELIER (Architecture and Technologies for Inspirational Learning Environments) project has been involved. ATELIER is an EU-funded project that was part of the Disappearing Computer initiative. The aim of this project was to build a digitally enhanced environment, supporting a creative learning process in architecture and interaction design education. In this context, it emerges that the evolution of the O-FCN is mainly given by the words of the documents inserted in a hyper-media data base (HMDB) and from the entities written during the definition of a query by the students.
We have studied the dynamic evolution of the O-FCN examining 485 documents and 200 queries of the students. For each query a user had the opportunity to include up to 5 different concepts and the possibility to semantically enrich his/her requests using the following list of concept modifiers: *little, enough, moderately, quite, very, totally*.
The algorithm has been tested in two different situations: classical and fuzzy approaches. In the first case, the crisp situation has been reported assigning value 1.0 to the correlations values and without taking the concept modifiers into the queries of the students. Instead, in the last case, all the parameters described in this paper have been considered.

Fuzzy recall and fuzzy precision measures [10] are the parameters used in order to evaluate retrieval algorithms in these two different situations: crisp and fuzzy cases. In Table 1 it is reported the average values of fuzzy precision and fuzzy recall for the 200 queries performed in the two approaches. Retrieved documents are ranked up to a theta threshold (θ). In particular, we have chosen three values of θ ($0.35, 0.50$ and 0.75), to validate the algorithm in different situations.

Table 1. Average values of Fuzzy Precision and Fuzzy Recall in the fuzzy and crisp cases.

	Fuzzy Case		**Crisp Case**	
θ **value**	F. Precision	F. Recall	F. Precision	F. Recall
0.35	0.573	0.612	0.590	0.622
0.50	0.602	0.523	0.604	0.593
0.75	0.912	0.221	0.942	0.234

In Table 1, apparently, crisp approach is similar to the fuzzy one and the relevance of the obtained documents is more or less the same. Instead, in the fuzzy case it has been observed a better accuracy of relevant documents. Indeed, this result was derived from the analysis of *coefficient variance* based on fuzzy precision measure (here CV_P). In detail, $CV_P := (\frac{\sigma}{P_F}) \cdot 100$ where σ is the standard deviation calculated on the relevance of the documents and P_F is the fuzzy precision, and it is a useful statistic for comparing the degree of variation from one data series to another. In general, the larger this number, the greater the variability in the data. Figure 3 depicts the trend of CV_P between fuzzy and crisp approaches, for each query. In the fuzzy case we can observe higher CV_P values for the fuzzy case, for all the queries analysed. This means that the fuzzy case approach identifies more refinement and accuracy than the crisp case.

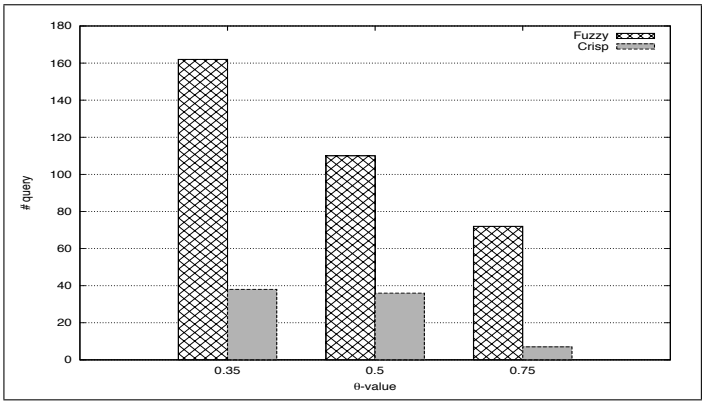


Fig. 3. Trend of CV_P value for each query.

3 Conclusion

It has been shown how the introduction of Fuzzy Ontologies, derived models and new structures, can improve an Information Retrieval System. More extensive developments will be shown in a forthcoming journal paper. The methodology allows to handle a trade off between the correct definition of an object, taken in the ontology structure, and the actual meaning assigned by individuals. So that it offers the opportunity to exploit an additional knowledge hidden in entities-documents relationships, or semantic correlations, after querying a database, but also to enrich the semantics of the system. After analysis, the obtained results for relevance presented a better accuracy in the fuzzy case than in the crisp one.

Acknowledgements

The work presented in this paper had been partially supported by the ATELIER project (IST-2001-33064). Particular thanks are due to Fabio Farina for his contribution in the Section of the numerical validations.

References

[1] Gruber, T.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition **5** (1993) 199–220
[2] Sanchez, E.: Fuzzy Logic and the Semantic Web. Capturing Intelligence. Elsevier (2006)
[3] Zadeh, L.: From Search Engines to Question-Answering Systems - The Problems of World Knowledge, Relevance, Deduction and Precisiation. In Sanchez, E., ed.: Fuzzy Logic and the Semantic Web. Capturing Intelligence. Elsevier (2006) 163–210
[4] Calegari, S., Ciucci, D.: Fuzzy Ontology, Fuzzy Description Logics and Fuzzy-OWL. In: Proceedings of WILF 2007. Volume 4578 of LNCS. (2007) In printing.
[5] Calegari, S., Ciucci, D.: Fuzzy Ontology and Fuzzy-OWL in the KAON Project. In: FUZZ-IEEE 2007. IEEE International Conference on Fuzzy Systems (2007) In printing.
[6] Sanchez, E., Yamanoi, T.: Fuzzy ontologies for the semantic web. In Larsen, H.L., Pasi, G., Arroyo, D.O., Andreasen, T., Christiansen, H., eds.: FQAS. LNCS 4027, Springer (2006) 691–699
[7] Calegari, S., Farina, F.: Fuzzy Ontologies and Scale-free Networks Analysis. International Journal of Computer Science and Applications **IV**(II) (2007) 125–144
[8] Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F., eds.: The Description Logic Handbook: Theory, Implementation, and Applications. In Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F., eds.: Description Logic Handbook, Cambridge University Press (2003)
[9] Salton, G., Mcgill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA (1986)
[10] Sanchez, E., Pierre, P.: Fuzzy Logic and Genetic Algorithms in Information Retrieval. In Yamakawa, T., ed.: Proceedings of the 3rd Int. Conf. on Fuzzy Logic, Neural Nets and Soft Computing, Jono Printing Co. (1994) 29–35

Trustworthiness-related Uncertainty of Semantic Web-style Metadata: A Possibilistic Approach

Paolo Ceravolo[1], Ermesto Damiani[1], and Cristiano Fugazza[1]

Department of Information Technology, University of Milan
via Bramante, 65 - 26013, Crema (CR), Italy
{damiani,ceravolo,fugazza}@dti.unimi.it

Abstract. We discuss the specific type of uncertainty deriving from the non-uniform *trustworthiness* of Semantic Web style metadata sources, arguing toward the feasibility of modal possibilistic reasoning based on trust assertions expressing such uncertainty.

1 Introduction

A cornerstone of the Semantic Web vision is the notion that resource descriptions can be modeled as Description Logics (DL) assertions. Indeed, many innovative applications enabled by the Semantic Web are based on the idea of reasoning on knowledge about network resources made available as Semantic Web-style metadata. However, intuition suggests that generalized manual annotation of Web resources is simply not feasible; and while automatic metadata generation is of paramount importance, manually validating (semi-)automatically generated assertions would require an effort comparable to manually writing metadata from scratch. In this scenario, performing approximate reasoning on Semantic Web metadata requires solving two major open problems related to their expressive power:

- *Non-uniform representation of uncertainty.* Current Semantic Web description languages cannot specify neither uncertainty degrees nor their semantics. Two main reasons motivate the introduction of explicit representation of uncertainty of Semantic Web metadata: (i) representing each assertion's degree of *fulfillment* on the part of the Web resource it describes (e.g., "the image at URL so-and-so is a high resolution one") [1], or (ii) each assertion's *importance* w.r.t. other assertions regarding the same resource. In principle, this type of uncertainty can be represented by stating the assertions in some kind of *fuzzy description logics* (fuzzy DL). Several fuzzy extensions to DLs are have been proposed [5], whose decidability property and deduction algorithms widely differ; choosing the right formalization for performing reasoning a given setting would require all the assertions involved to have a uniform semantics, quite a tall order for heterogeneous Web environments.
- *Lack of support for modalities.* Semantic Web description languages cannot express assertions belonging to different modalities, including *alethic* or *deontic* ones. Alethic rules are used to model necessities (e.g. implied by physical laws) which cannot be violated, even

in principle. For example, an alethic rule may state that an image file has a (single) date of creation. Deontic rules are used to model obligations (e.g., resulting from company policy) which ought to be obeyed, but may be violated in real world scenarios. For example, a deontic rule may state that all landscape images must carry the indication of the country where they were taken .

2 A Possibilistic Approach

While the two problems outlined above are hard to tackle in a generalized setting, they can be successfully approached in a restricted case, i.e. the specific type of uncertainty deriving from the non-uniform *trustworthiness* of Semantic Web metadata sources[2]. In a typical Semantic Web setting, assertions about network resources can be generated by different sources, including automatic extraction by autonomous software agents, as well as manual annotation by the data owner or other users. The degree associated to the assertions provided by a data source represents the trustworthiness of that source w.r.t. the specific assertion. We propose to express such a degree by stating a special purpose *trust assertion* expressing the level of trustworthiness of an ordinary Semantic Web assertion. Trust assertions follow the pattern "the (reified) assertion so-and-so has a level of trustworthiness of X", are built using their own reserved vocabulary (expressed as a suitable task ontology). More importantly, their associated degree has a uniform semantics which can be modeled as a possibility. Trust assertions can be used to rank assertions about a specific resource; also, the uniform semantics of the associated degree enables formalization using possibilistic fuzzy logics. More specifically, a modal possibilistic logics formulation can capture both the the missing modalities[4]; also, reasoning can be carried out using extensions of the tableaux methods available for ordinary modal logics[3].

References

1. P. Bosc, E. Damiani, and M.G. Fugini, "Fuzzy Service Selection in a Distributed Object-Oriented Environment, IEEE Trans. Fuzzy Systems, vol. 9, no. 5, Oct. 2001.
2. Paolo Ceravolo, Ernesto Damiani, Marco Viviani, "Bottom-Up Extraction and Trust-Based Refinement of Ontology Metadata," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 2, pp. 149-163, Feb. 2007.
3. D. Dubois, J. Lang, H. Prade, "Possibilistic Logics," In D. Gabbay, C. Hogger, J. Robinson (eds), Handbook of Logics for Artificial Intelligence and Logic Programming, Clarendon Press, 1994.
4. L. Godo, P. Hjek, and F. Esteva, "A fuzzy modal logic for belief functions," Fundam. Inf. 57, 2-4, pp. 127-146, Oct. 2003.
5. U. Straccia, "Towards a fuzzy description logic for the semantic web (preliminary report)," In proceedings of the second European Semantic Web Conference, Springer, 2005.

Extending Fuzzy Description Logics with a Possibilistic Layer

Fernando Bobillo, Miguel Delgado, and Juan Gómez-Romero

Department of Computer Science and Artificial Intelligence, University of Granada
Email: fbobillo@decsai.ugr.es, mdelgado@ugr.es, jgomez@decsai.ugr.es

Abstract. Classical ontologies are not suitable to represent imprecise nor uncertain pieces of information. As a solution we will combine fuzzy Description Logics with a possibilistic layer. Then, we will show how to perform reasoning by relying on classical existing reasoners.

Description Logics (DLs for short) are a family of logics for representing structured knowledge which have proved to be very useful as ontology languages. Nevertheless, it has been widely pointed out that classical ontologies are not appropriate to deal with imprecise, vague and uncertain knowledge, which is inherent to several real-world domains and Semantic Web tasks (e.g. the integration or merging of ontologies). Fuzzy and possibilistic logics have proved to be suitable formalisms to handle imprecise/vague and uncertain knowledge respectively. Fuzzy and possibilistic logics are orthogonal, the former handling degrees of truth and the latter handling degrees of certainty.

There exist several fuzzy and possibilistic extensions of DLs in the literature (see [1] for an overview). These extensions are appropriate to handle either vagueness or uncertainty, but handling both of them has not received such attention. An exception is [2], where every fuzzy set is represented using two crisp sets (its support and core) and then axioms are extended with necessity degrees. Although for some applications this representation may be enough (and the own authors suggest to consider more α-cuts), there is a loss of information which we will overcome here. Another related work combines fuzzy vagueness and probabilistic uncertainty with description logic programs [3].

We propose to build a layer to deal with uncertain knowledge on top of a fuzzy Knowledge Base (KB) defined as in [4], by annotating the axioms with possibility and necessity degrees, and to reduce it to a possibilistic layer over a crisp ontology. Interestingly, this makes possible to perform reasoning tasks relying on existing classical reasoners e.g. Pellet (`http://pellet.owldl.com`).

Syntax. A *possibilistic fuzzy knowledge base* pfK is a fuzzy KB where each fuzzy axiom τ (see [4] for details) is equipped with a possibility or necessity degree, $(\tau, \Pi\ \alpha)$ or $(\tau, N\ \alpha)$ respectively with $\alpha \in (0, 1]$. If no degree is specified, $N\ 1$ es assumed. Necessity degrees express to what extent a formula is necessary true, whereas possibility degrees express to what extent a formula is possible.

Semantics. Let $\Im$ be the set of all (fuzzy) interpretations. A *possibilistic interpretation* is a mapping $\pi : \Im \rightarrow [0, 1]$ such that $\pi(\mathcal{I}) = 1$ for some $\mathcal{I} \in \Im$. The intuition here is that $\pi(\mathcal{I})$ represents the degree to which the world $\mathcal{I}$ is possible.

$\mathcal{I}$ is impossible if $\pi(\mathcal{I}) = 0$ and fully possible if $\pi(\mathcal{I}) = 1$. The possibility of an axiom τ is defined as $Poss(\tau) = \sup\{\pi(\mathcal{I}) \mid \mathcal{I} \in \mathfrak{I}, \mathcal{I} \models \tau\}$ (where $\sup \emptyset = 0$), and the necessity is defined as $Nec(\tau) = 1 - Poss(\neg\tau)$. A possibilistic interpretation π satisfies a possibilistic axiom $(\tau, \Pi\gamma)$, denoted $\pi \models (\tau, \Pi\gamma)$, iff $Poss(\tau) \geq \gamma$ and a possibilistic axiom $(\tau, N\ \gamma)$, denoted $\pi \models (\tau, N\ \gamma)$, iff $Nec(\tau) \geq \gamma$.

Reasoning. B. Hollunder showed that reasoning within a possibilistic DL can be reduced to reasoning within a classical DL [5]. We will reduce here our possibilistic fuzzy DL to a possibilistic DL. A fuzzy KB fK can be reduced to a crisp KB $\mathcal{K}(fK)$ and every axiom $\tau \in fK$ is reduced to $\mathcal{K}(\tau)$, which can be an axiom or a set of axioms [4]. Adding degrees of certainty to fK formulae is equivalent to adding degrees of certainty to their reductions, as long as we also consider axioms preserving the semantics of the whole process (which are assumed to be necessarily true and do not have any degree of certainty associated). For every axiom $(\tau, \Pi\gamma) \in pfK$, $Poss(\tau) \geq \gamma$ iff $Poss(\mathcal{K}(\tau)) \geq \gamma$. Similarly, $(\tau, N\gamma) \in pfK$, $Nec(\tau) \geq \gamma$ iff $Nec(\mathcal{K}(\tau)) \geq \gamma$.

Example 1. The axiom $(\langle tom : High \geq 0.5\rangle, N\ 0.2)$ means that it is possible with degree 0.2 that *tom* can be considered a $High$ person with (at least) degree 0.5. It is reduced into $(\langle tom : High_{\geq 0.5}\rangle, N\ 0.2)$, meaning that it is possible with degree 0.2 that *tom* belongs to the crisp set $High_{\geq 0.5}$. The final crisp KB would also need some additional axioms (consequence of the reduction of the fuzzy KB): $High_{\geq 0.5} \sqsubseteq High_{>0}, High_{>0.5} \sqsubseteq High_{\geq 0.5}$ and $High_{\geq 1} \sqsubseteq High_{>0.5}$. □

Final remarks. [4] reduces a fuzzy KB to a crisp KB and reasoning is performed by computing a consistency test on the crisp KB. Our case is more difficult and needs to perform several entailment tests. Moreover, how to represent the possibilistic DL using a classical DL remains an open issue.

Acknowledgements. This research has been partially supported from Ministerio de Educación y Ciencia (under project TIN2006-15041-C04-01 and a FPU scholarship which holds F. Bobillo) and Consejería de Innovación, Ciencia y Empresa, Junta de Andalucía (under a scholarship which holds J. Gómez-Romero).

References

1. Lukasiewicz, T., Straccia, U.: An overview of uncertainty and vagueness in description logics for the semantic web. Technical Report INFSYS RR-1843-06-07, Institut für Informationssysteme, Technische Universität Wien (2006)
2. Dubois, D., Mengin, J., Prade, H.: Possibilistic Uncertainty and Fuzzy Features in Description Logic. A Preliminary Discussion. In: Fuzzy Logic and the Semantic Web. Volume 1 of Capturing Intelligence. Elsevier Science (2006) 101–113
3. Lukasiewicz T., Straccia U.: Description Logic Programs under Probabilistic Uncertainty and Fuzzy Vagueness. In Proceedings of ECSQARU 2007. To appear (2007)
4. Bobillo, F., Delgado, M., Gómez-Romero, J.: A crisp representation for fuzzy $\mathcal{SHOIN}$ with fuzzy nominals and general concept inclusions. In Proceedings of URSW 2006. Volume 218, CEUR Workshop Proceedings (2006)
5. Hollunder, B.: An alternative proof method for possibilistic logic and its application to terminological logics. In Proceedings of UAI'94 (1994) 327–335

A Pattern-based Framework for Representation of Uncertainty in Ontologies

Miroslav Vacura[1], Vojtěch Svátek[1], Pavel Smrž[2], and Nick Simou[3]

[1] University of Economics
vacuram|svatek@vse.cz
[2] Brno University of Technology
smrz@fit.vutbr.cz
[3] National Technical University of Athens
nsimou@image.ece.ntua.gr

Abstract. We present a novel approach to representing uncertain information in ontologies based on design patterns. We provide a brief description of our approach, present its use in case of fuzzy information and probabilistic information, and describe the possibility to model multiple types of uncertainty in a single ontology. We also shortly present an appropriate fuzzy reasoning tool and define a complex ontology architecture for well-founded handling of uncertain information.

Motivation for our research is the CARETAKER project[4] which comprises advanced approaches to recognition of multimedia data, which led us to problems of representing uncertain information.

Although fuzziness isn't, exactly said, type of uncertainty, we will in this example consider representing fuzzy information in the form of facts, i.e. A-Box from description logic (DL) point of view. The key principle of our approach to representing fuzzy information is the *separation* of crisp ontology from fuzzy information ontology. We allow the fuzzy ontology to be OWL Full and only suppose that the base ontology is OWL DL compliant. Regular OWL DL crisp reasoning tools can be applied to the base ontology, fuzzy reasoning tools (i.e. FiRE[5]) to fuzzy ontology.

Instantiation axioms in Fuzzy OWL [1] are assertions of form $\langle a : C \bowtie n \rangle$ – facts saying that individual a belongs to class C, n is level of certainty $(0, 1)$ and $\bowtie$ is one of $\{\leq, <, \geq, >\}$. We introduce a few constructs that enable us to model such axioms with uncertainty by ontology patterns. For each crisp axiom of base ontology we create a new individual belonging to class fuzzy-instantiation, which will have several properties attaching it to that crisp axiom in base ontology and implementing uncertainty. Properties fi-instance and fi-class characterize the membership of an individual person-1 to class problem-person. Property f-type defines the type of uncertainty relation ($\bowtie$) and datatype property f-value defines the level of uncertainty n (Fig. 1, individuals are grayed and classes are bright).

[4] http://www.ist-caretaker.org/
[5] http://www.image.ece.ntua.gr/∼nsimou

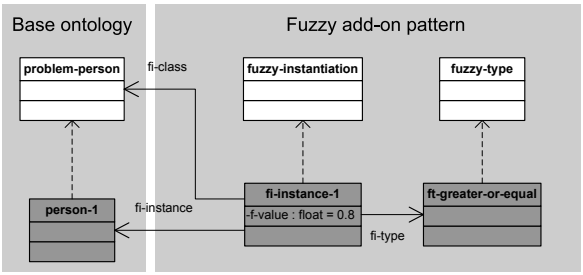


Fig. 1. Instantiation pattern

One of major advantages of our modeling approach is that it enables us to model various kinds of uncertainty in same ontology at the same time. Using approach described above we can define well-founded architecture of ontology that fully supports handling uncertainty – Uncertainty Modeling Framework (UMF): crisp ontology is aligned to foundational ontology (i.e. DOLCE) while fuzzy and i.e. probabilistic ontology are based on appropriate patterns of UMF. Such architecture is modularized, so these parts of ontology are separated to independent modules. On top of these ontologies there can be number of different specialized reasoners operating (Fig. 2).

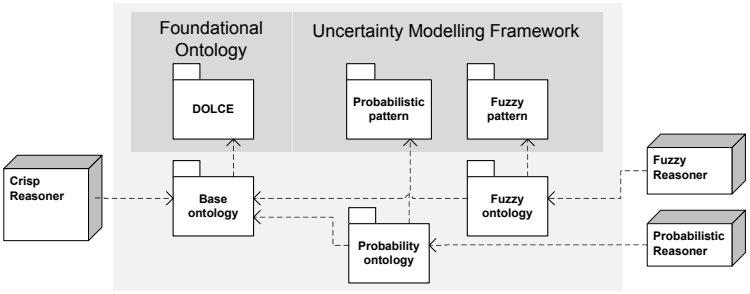


Fig. 2. Ontology architecture supporting reasoning with uncertainty.

More information can be found in full paper draft[6]. M. Vacura, V. Svátek and N. Simou are supported by the EC under FP6, project K-Space (no.: FP6-027026), and first two also by the Czech IGA VSE grant no.12/06. P. Smrž is supported by the EU FP6 project CARETAKER (no.: FP6-027231).

References

1. G. Stoilos, G. Stamou, V. Tzouvaras, J. Z. Pan, and I. Horrocks. Fuzzy OWL: Uncertainty and the Semantic Web. In *Proc. of the OWL-ED 2005).*

[6] `http://keg.vse.cz/_papers/2007/framew.pdf`

The 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference

November 11~15 2007
BEXCO, Busan KOREA