

Probabilistic Reasoning in the Semantic Web using Markov Logic

Abstract. Uncertainty is a characteristic of any relevant and complex knowledge source. In this thesis, we explored the use of Markov Logic, a novel representation formalism that combines first-order logic with probabilistic graphical models, to learn and reason about uncertainty in the Semantic Web. We explored several ways to acquire this uncertainty automatically from several sources, and applied them in relevant tasks to the Semantic Web domain.

Keywords: Probabilistic Reasoning, Semantic Web, Markov Logic.

1 Introduction

The Semantic Web [1] envisions a world where agents share and transfer structured knowledge in an open and semi-automatic way. In most of the cases, this knowledge is characterized by uncertainty. However, Semantic Web languages do not provide any means of dealing with this uncertainty; they are mainly based on crisp logics, unable of dealing with partial and incomplete knowledge. Reasoning in the Semantic Web resigns to a deterministic process of verifying if statements are true or false.

In the last years, some efforts [2] have been made in representing and reasoning with uncertainty in the Semantic Web. These works are mainly focused on how to extend the logics behind Semantic Web languages to the probabilistic/possibilistic/fuzzy logics, or on how to combine these languages with probabilistic formalisms like Bayesian Networks. In all of these approaches, this is achieved by annotating the ontologies with some kind of uncertainty information about its axioms, using this information to perform uncertainty reasoning. Nevertheless, several questions arise: how can reasoning be done efficiently with this uncertainty information? Where to get this uncertainty information?

In this thesis, we presented solutions for both questions. The solution for the first question is Markov Logic [3], a new promising approach to reasoning with uncertainty. In this type of logic, there is no right and wrong world; there are multiple worlds with different degrees of probability. This is done by combining logic and probability in the same representation, and then using efficient learning and inference algorithms. For the second question, several solutions were developed:

- If the ontologies are annotated with some kind of uncertainty information, like probabilities, Markov Logic can be used to reasoning about this information;
- If the ontology contains individuals, those individuals can be used to automatically learn the uncertainty of the ontology;

- If the ontology does not comprise uncertainty information or individuals, both resources can be automatically learned by analyzing textual resources and web search engines.

We developed a system, *Incerto*, which explores the capabilities of Markov Logic for the Semantic Web. This system was applied in several interesting tasks, like reasoning about automatically learned ontologies and social networks analysis.

The main contributions of this thesis are:

- The application of Markov Logic for learning and reasoning about uncertainty in the Semantic Web;
- The development of several techniques for learning automatically the uncertainty of ontologies;
- The development of a new technique to parameterize Markov Logic networks with probabilities;
- The development of a new technique to learn the probability of ontology axioms by using web search engines;
- The development of *Incerto*, and its application to several Semantic Web domains.

The next section describes the concepts of Semantic Web and Markov Logic, followed by a description of the most relevant related work. Section 3 introduces our proposed approach, with several experimental results. We finalize this paper by describing general conclusions and future directions of this work.

2 State of the Art

In this section, we present the two most important concepts for this work, Semantic Web and Markov Logic, and review the most important related work.

2.1 Semantic Web

The Semantic Web [1] tries to fill the knowledge gap between human and machines by adding background knowledge to the Web, allowing machines to infer the real meaning of objects. This background knowledge is usually expressed by ontologies, i.e., sets of knowledge terms for some particular topic, including the vocabulary, semantic interconnections, and rules of logic/inference of those terms.

The most prominent markup language proposed by the W3C to model ontologies in the Semantic Web is the *Web Ontology Language*¹ (OWL). OWL provides an expressive shared vocabulary to represent knowledge in the Semantic Web. This vocabulary allows expressing axioms about classes, properties, and individuals of the domain. In this paper, we will focus on OWL2² [4], the new version of OWL

¹ <http://www.w3.org/2004/OWL/>

² <http://www.w3.org/TR/owl2-quick-reference/>

proposed by the W3C, which subsumes the decidable subsets of the original OWL (OWL DL and OWL Lite).

OWL2 is based on the Description logic $SROIQ(D)$ [4]. Description logics [5] are a family of logical languages specially designed to model terminological domains. Formulas in Description Logics are composed by two symbols: *concepts* (i.e., sets of individuals) and *roles* (i.e., relationships between individuals). A relevant feature of Description Logics is their separation of knowledge bases in two distinct parts: the intensional knowledge in the form of a terminology, called *Terminological Box* (TBox), and the extensional knowledge, called *Assertional Box* (ABox). The TBox provides the vocabulary, in terms of concepts and rules, of the knowledge base. This is usually done by defining concepts using the logical equivalence constructor (e.g., $Woman \equiv Person \sqcap Female$). The ABox uses the TBox vocabulary to make assertions about individuals (e.g. $Woman(ANNA)$).

2.2 Markov Logic

Markov Logic [3] combines first-order logic and probabilistic graphical models (Markov networks) in the same representation. The main idea behind Markov Logic is that, unlike first-order logic, a world that violates a formula is not invalid, but only less probable. This is done by attaching weights to first-order logic formulas: the higher the weight, the bigger is the difference between a world that satisfies the formula and one that does not, other things being equal. These sets of weighted formulas are called Markov Logic Networks (MLNs). Given a set of constants (i.e., individuals) of the domain and an interpretation, the groundings of the formulas in an MLN can generate a Markov network by adding a variable for each ground atom, an edge if two ground atoms appear in the same formula, and a feature for each grounded formula. The probability distribution of the network is defined as

$$P(X = x) = \frac{1}{Z} \exp \left(\sum_{i=1}^F w_i n_i(x) \right), \quad (1)$$

where F is the number of formulas in the MLN, $n_i(x)$ is the (binary) number of true groundings of F_i in the world x , w_i is the weight of F_i , and Z is a normalizing constant.

Formulas' weights can be learned generatively from example data by maximizing the pseudo-log-likelihood [3] of that data, while efficient inference can be done using approximate inference algorithms, such as the MC-SAT [3].

2.3 Related Work

In this section, we review the most important related work to this thesis. Refer to [2] for a complete overview of works related to the problem of uncertainty and vagueness in the Semantic Web.

PR-OWL [6] is a probabilistic generalization of OWL based on *multi-entity Bayesian networks* (MEBNs) [6]. MEBN logic combines Bayesian probability theory

with first order logic, constructing Bayesian networks from parameterized fragments representing the probabilistic knowledge about a collection of related hypotheses.

Ding et al. [7] proposed *BayesOWL*, a framework to represent and reason about OWL uncertain knowledge using Bayesian networks. They provide a set of rules and procedures to translate OWL DL concept taxonomies into Bayesian networks. The resulting Bayesian network preserves the semantics of the original ontology, and supports ontology reasoning both within and across ontologies.

Henrik and Norbert [8] describe work on probabilistic reasoning in two subsets of OWL Lite. They translate restricted OWL Lite ontologies into *Datalog*, a subset of first-order logic, and use a probabilistic extension of Datalog, *pDatalog*, to do probabilistic inference over the ontology.

Predoiu and Stuckenschmidt [9] use *Bayesian Description Logic Programs*, a formalism that joins *Description Logic Programs* (DLP), a subset of Datalog, with a fragment of *Bayesian Logic Programs*. In this representation, statements are translated to DLP rules with an attached probability. This way, a Bayesian network can be built from those annotated rules, providing a complete specification of the desired probability distribution.

3 Learning and Reasoning about Uncertainty in the Semantic Web

The first step in use Markov Logic capabilities to reason about uncertainty in the Semantic Web is to transform Semantic Web's representation languages, in our case OWL2, into Markov Logic Networks (MLNs). As seen in the previous section, a MLN is composed by a set of weighted first-order logic formulas. So, we must define where these formulas and weights come from.

Formulas. OWL2 is based on the Description Logic *SROIQ(D)* [4]. One characteristic of Description Logics languages is that they follow a *model-theoretic* semantics [5], and therefore can (in most of the cases) be interpreted as formulas in first-order logic. The main idea behind this interpretation is that concepts correspond to unary predicates, roles to binary predicates, and individuals correspond to constants. In our case, *SROIQ(D)* can be easily interpreted as first order formulas:

Table 1. OWL2 examples of interpretation as first-order logic formulas.

OWL2 Expression	First-order logic formula
<i>SubClassOf</i> (CE_1, CE_2)	$\forall x : CE_1(x) \Rightarrow CE_2(x)$
<i>TransitiveProperty</i> (OPE)	$\forall x, y, z : OPE(x, y) \wedge OPE(y, z) \Rightarrow OPE(x, z)$
<i>ClassAssertion</i> (CE, a)	$CE(a)$

Weights. In the next sections, we explore several sources for acquiring weights. First, we explore the cases when the ontologies are already annotated with some kind of uncertainty values that can be interpreted as weights. Second, we explore the cases where ontologies do not have any type of uncertainty annotation available. If the

ontology contains individuals, we can use those individuals to learn the weights using the weight learning capabilities of Markov logic. In the cases where the ontologies do not have individuals, resources like textual corpus and web search engines can be used to learn individuals or derive automatically the probability of axioms.

3.1 Probabilistic Reasoning in Uncertainty-annotated Ontologies

Ontology axioms can be annotated with a value representing its uncertainty (usually a weight or a probability) (e.g., a certain class has $X\%$ probability of being subclass of another). This allows ontology engineers to build uncertain ontologies with their own knowledge about the domain. In fact, some tasks like learning and mapping ontologies [10] already automatically produce uncertain ontologies. In most of the cases, this uncertainty is represented as a probability. However, Markov Logic receives weights and not probabilities as its parameterization. For this purpose, we developed a new method to parameterize MLNs with probabilities, by changing the *discriminative weight learning* algorithm [3] to use the desired probabilities in the counting of true groundings of a formula. This way, the algorithm learns a weight that describes the desired probability.

The Body Gesture Experiment. Abbasi et al. [11] recorded the unintentional movements of students during a lecture, and manually labeled the movements in 7 gestures (*Head Scratch, Nose Itch, Lip Touch, Eye Rub, Chin Rest, Lip Zip, and Ear Scratch*), corresponding to 6 distinct affective states (*Recalling, Satisfied, Thinking, Tired, Bored, and Concentrating*). Based on their results, we developed a simple probabilistic ontology and used the proposed approach to predict the affective state of a person based on its gestures.

The Ontology Learning Experiment. Using the *lexico-syntactic* patterns defined by [12], we developed a simple system that receives the root of the taxonomy and uses a web search engine to infer its descendants until a pre-defined depth. Using the metrics of [13], each subsumption relation receives a probability describing the confidence on the asserted relation. Taxonomies about several domains (e.g., substances, cereals, hydrocarbons) were created, and, using the proposed approach, Markov Logic could be used to answer conditional queries about them.

3.2 Probabilistic Reasoning using Ontology Individuals

In the last section, we adopted the principle that ontologies were somehow annotated with some kind of uncertainty information. However, these situations only occur in restricted domains, mainly those where these ontologies were built automatically by machines. In all the other situations, other methods must be developed to acquire the uncertainty of the ontology. The most desirable method is to learn that uncertainty automatically, without user intervention. In this section, we study how this can be achieved by using ontology individuals. With this information, MC-SAT can be used to perform probabilistic inference over the ontology.

As previously noted, in Markov logic, formulas' weights can be learned generatively through example data. This example data is usually composed by individuals of the domain and their relations. In OWL2, individuals correspond to the ABox of the ontology, and therefore they can be used to learn the formulas' weights by interpreting them as ground atoms.

The Financial Experiment. In this experiment, we use a financial ontology, GoldDLP³, to assess the risk of certain financial operations. In this ontology, there is information about a bank that offers services like loans and credit cards to private persons. The ontology contains 116 class and property axioms and 297 individuals, mainly distributed between accounts, clients, credit cards, and loans. The main task in this experiment is to determine each loan's probability of being a problematic one. Using generative learning and MC-SAT, we found nine loans with a probability >90% of being a problematic one, while all the other loans have a probability between 45-48%. This result demonstrates, roughly speaking, that any loan has an associated probability of being a problematic loan.

The Social Network Experiment. The objective of this experiment is to use Markov Logic to explore the relational structure of the *Advogato's*⁴ Friend of a Friend (FOAF) social network. We performed several link mining [14] tasks, such as:

- Link Prediction - Predict the existence of a link between two users based on the relations of the user with other users;
- Link-based Classification - Predict the experience group (*Apprentice*, *Journeyer*, or *Master*) of an user based on the relations of that user with other users;
- Link-based Cluster Analysis - Cluster users into groups that show similar relational characteristics.

The Non-Relational Experiment. In this experiment, we transformed some machine learning datasets into ontologies and performed classification tasks:

- Mushrooms Dataset⁵ - Predict the class of the mushrooms (*Edible* or *Poisonous*);
- Titanic Dataset⁶ - Predict if a certain passenger survived to the RMS Titanic accident.

3.3 Probabilistic Reasoning by Learning Individuals/Probabilities

In the last section, we have explored the use of ontology individuals to automatically learn the uncertainty of the ontology axioms and perform inference with that information. This feature proved to be useful in domains where there was no information about that uncertainty, or in complex domains where this uncertainty is hard to infer, specially for humans. However, there are domains that are uncertain but do not have any type of information that could help us infer its uncertainty. In these

³ <http://www.cs.put.poznan.pl/alawrynowicz/semintec.htm>

⁴ <http://advogato.org/>

⁵ <http://archive.ics.uci.edu/ml/datasets/Mushroom>

⁶ <http://stats.math.uni-augsburg.de/Mondrian/Data/Titanic.txt>

domains, we have to find other ways of gathering information to learn the uncertainty of the axioms. In this thesis, we explored two approaches to tackle this problem: learn individuals and learn probabilities.

3.3.1 Learning Individuals

Due to the enormous quantity of textual resources currently available, specially those present in the World Wide Web and available through web search engines, extracting ontology individuals from those sources has become a task of growing interest. This is the task studied in the field of ontology population [13]. Using unsupervised methods such as the ones proposed by [12], we developed a system that uses web search engine APIs to populate existing ontologies. Using this system and the approach of the previous section, we can populate ontologies and learn the uncertainty of the ontology's axioms using the learned individuals, and then perform probabilistic reasoning with that information.

The Disease Ontology Experiment. In this experiment, we automatically learned an ontology about diseases and their symptoms, using a web search engine, and performed clustering on it. At the end, we had an ontology composed by 140 diseases, 459 symptoms, and 671 symptoms assertions. Using this ontology, we performed link-based cluster analysis [14], creating clusters of similar diseases. Some of the clusters could be analyzed by their main symptoms. For example, all the diseases in one of the clusters had a common symptom, depression, while in other cluster all the diseases were related to the respiratory system.

3.3.2 Learning Probabilities

Other way to acquire the uncertainty of an ontology's axioms is to learn the uncertainty of the axioms directly, without the need to populate the ontology and use Markov Logic weight learning algorithms. In this section, we use semantic similarity techniques to perform this task.

Semantic similarity [15] is the process of finding the similarity between two words or entities. This is usually done by studying the *co-occurrence* between those words or entities in a textual corpus. The most used techniques use results of search engines to measure that similarity. This is usually done by counting the number of results returned by those search engines in specific queries related to the words or entities whose similarity we want to assert. In this thesis, we explored several metrics used for this purpose (e.g., [15]). However, we modified the format of the search engine queries so the metrics return the confidence value of the relation asserted in each axiom, instead of the general co-occurrence between the subject and object of the axiom.

The Animals Taxonomy Experiment. We used a web search engine to learn the probability of each axiom in one automatically learned taxonomy about animals. With that information, we used Markov Logic to perform several conditional queries about the uncertainty of the taxonomy's subsumption relations.

3.4 Incerto – A Probabilistic Reasoner for the Semantic Web

Using the ideas of the previous sections, we developed *Incerto*, a probabilistic reasoner for the Semantic Web based on Markov logic. The system was developed in Java, and is freely available through a LGPL license at <http://code.google.com/p/incerto>. The system interacts with several external components, such as Markov Logic reasoning engines and ontology processors, and can be accessed programmatically, visually, and with a command-line.

In this section, we study the scalability of Markov logic procedures in the Semantic Web domain, as implemented in *Incerto*. Our purpose is to measure the scalability of three distinct procedures:

- *Pre-processing*, composed by the load and transformation of ontologies in MLNs;
- *Weight Learning*, using the generative algorithm;
- *Inference*, using the MC-SAT algorithm.

For this purpose, those procedures were performed on seven distinct ontologies, each one with a varied number of individuals. This variation was made by randomly populating each ontology with a set of individuals (we tested with 1, 10, 100, 1.000, and 10.000 individuals), using these individuals to make an average of three assertions for each ontology class or property. Results can be seen in Fig. 1.

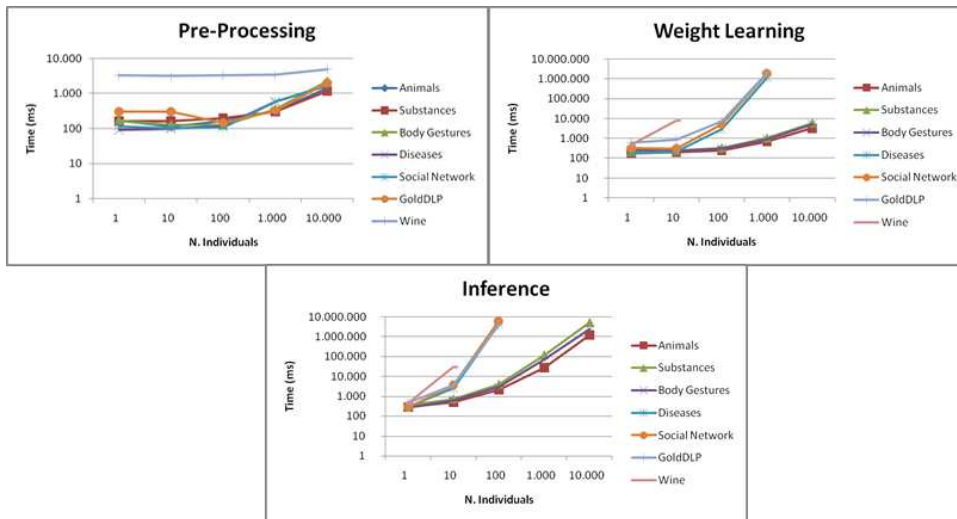


Fig. 1. Incerto scalability tests.

4 Conclusions

To realize the Semantic Web vision of a world where knowledge is the most important element, mechanisms must be developed to represent and reason about the uncertainty that this knowledge could arise. In this thesis, we explored the use of

Markov logic, a unifying representation of logic and probability, to learn and reason about uncertainty in the Semantic Web. We explored several ways to acquire this uncertainty automatically, testing them in relevant domains for the Semantic Web, such as ontology learning and social network analysis.

We think that our approach can be seen as an introductory step in providing a general Markov logic framework for the Semantic Web.

4.1 Future Work

In this thesis, we explored several ways to automatically learn the uncertainty of ontology axioms. However, more ideas could be explored:

- *Other ways of learning individuals.* We explored the use of textual resources to learn ontology individuals. Other way of populating ontologies is through the analysis of structured data, like relational databases or other ontologies. In this case, *mappings* [10] must be made between the structured data objects and the entities of the ontology.
- *Learn the uncertainties directly from textual corpus.* This is done by analyzing textual resources for patterns like “70% of A is B” or “Most of the A’s are B’s”. This can be done by using previously trained classifiers or general lexico-syntactic rules.
- *Use the structure of the ontology.* The structure of the ontology can provide interesting information about the uncertainty of its axioms. Some other works (e.g., [16]) already explored similar approaches in ontologies, however with distinct objectives than ours. The field of *network analysis* [17] can provide us with some interesting concepts that can be potentially transferred to our specific case.
- *Collective learning of weights.* The idea is to learn the weights collectively from multiple ontologies about the same domain. This task can be achieved by exploring techniques from collective learning fields, like *relational reinforcement learning* [18].
- *Trust propagation.* The idea is to use the propagation of *trust metrics* in groups to automatically learn the uncertainty of certain axioms. This idea was already applied in the Markov logic context [19].

Other ideas, like using the structure of the ontology to guide the learning of ontology individuals [20] and the use of other Markov Logic reasoning algorithms (e.g., [21]), could also be used to improve the existing results.

References

- [1] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” *Scientific American*, vol. 284, 2001, pp. 28-37.
- [2] T. Lukasiewicz and U. Straccia, “Managing Uncertainty and Vagueness in Description Logics for the Semantic Web,” *Web Semantics Sci Serv Agents World Wide Web*, 2008.
- [3] P. Domingos, S. Kok, D. Lowd, H. Poon, M. Richardson, and P. Singla, “Markov Logic,” *Probabilistic Inductive Logic Programming*, 2008, pp. 92-117.

- [4] B.C. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider, and U. Sattler, "OWL 2: The next step for OWL," *Web Semantics: Science, Services and Agents on the World Wide Web*, 2008.
- [5] F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, and P.F. Patel-Schneider, *The Description Logic Handbook: Theory, Implementation, and Applications*, Cambridge University Press, 2007.
- [6] P.C.G. Costa and K.J. Laskey, "PR-OWL: A Bayesian Ontology Language for the Semantic Web," *Proceedings of the 1st Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2005)*, 2005, pp. 6-10.
- [7] Z. Ding, Y. Peng, and R. Pan, "BayesOWL: Uncertainty Modeling in Semantic Web Ontologies," *Soft Computing in Ontologies and Semantic Web*, 2006, pp. 3-29.
- [8] N. Henrik and F. Norbert, "Adding Probabilities and Rules to Owl Lite Subsets Based on Probabilistic Datalog," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 14, 2006, pp. 17-42.
- [9] L. Predoiu and H. Stuckenschmidt, "A probabilistic Framework for Information Integration and Retrieval on the Semantic Web," *Proc. of 3rd International Workshop on Database Interoperability (InterDB) in conjunction with the VLDB conference*, Vienna, Austria: 2007, pp. 23-28.
- [10] J. Euzenat and P. Shvaiko, *Ontology Matching*, Springer, 2007.
- [11] A. Rehman Abbasi, N. V. Afzulpurkar, and T. Uno, "Exploring Un-Intentional Body Gestures for Affective System Design," *Affective Computing*, InTech Education and Publishing, 2008.
- [12] M.A. Hearst, "Automatic acquisition of hyponyms from large text corpora," *Proceedings of the 14th conference on Computational linguistics-Volume 2*, Association for Computational Linguistics Morristown, NJ, USA, 1992, pp. 539-545.
- [13] L.K. McDowell and M. Cafarella, "Ontology-driven, unsupervised instance population," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, 2008, pp. 218-236.
- [14] L. Getoor and C.P. Diehl, "Link mining: a survey," *SIGKDD Explor. Newsl.*, vol. 7, 2005, pp. 3-12.
- [15] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring semantic similarity between words using web search engines," *Proceedings of the 16th international conference on World Wide Web*, Banff, Alberta, Canada: ACM, 2007, pp. 757-766.
- [16] C. Ramakrishnan, W.H. Milnor, M. Perry, and A.P. Sheth, "Discovering informative connection subgraphs in multi-relational graphs," *SIGKDD Explor. Newsl.*, vol. 7, 2005, pp. 56-63.
- [17] U. Brandes and T. Erlebach, *Network Analysis: Methodological Foundations*, Springer, 2005.
- [18] P. Tadepalli, R. Givan, and K. Driessens, "Relational Reinforcement Learning: An Overview," *Proceedings of the ICML'04 Workshop on Relational Reinforcement Learning*, 2004, pp. 1-9.
- [19] M. Richardson, "Learning and Inference in Collective Knowledge Bases," PhD Thesis, University of Washington, 2004.
- [20] F.M. Suchanek, M. Sozio, and G. Weikum, "SOFIE: A Self-Organizing Framework for Information Extraction," *Proceedings of the 18th International World Wide Web Conference*, Madrid, Spain: 2009.
- [21] P. Singla and P. Domingos, "Lifted first-order belief propagation," *Proceedings of the Twenty-Third National Conference on Artificial Intelligence*, Chicago, IL: AAAI Press, 2008.