The Semantic Web effort has steadily been gaining traction in the recent years. In particular, Web search companies are recently realizing that their products need to evolve towards having richer semantic search capabilities. Description logics (DLs) have been adopted as the formal underpinnings for Semantic Web languages used in describing ontologies. Reasoning under uncertainty has recently taken a leading role in this arena, given the nature of data found on the Web. In this paper, we present a probabilistic extension of the DL $\mathcal{EL}^{++}$ (which underlies the OWL2 EL profile) using Markov logic networks (MLNs) as probabilistic semantics. This extension is tightly coupled, meaning that probabilistic annotations in formulas can refer to objects in the ontology. We show that, even though the tightly coupled nature of our language means that many basic operations are data-intractable, we can leverage a sublanguage of MLNs that allows to rank the atomic consequences of an ontology relative to their probability values (called ranking queries) even when these values are not fully computed. We present an anytime algorithm to answer ranking queries, and provide an upper bound on the error that it incurs, as well as a criterion to decide when results are guaranteed to be correct.

1 Introduction

Recently, it has become apparent that Semantic Web formalisms must be able to cope with uncertainty in a principled manner. The Web contains many examples where uncertainty comes in [22]: as an inherent aspect of Web data (such as in reviews of products or services, comments in blog posts, weather forecasts, etc.), as the result of automatically processing Web data (for instance, analyzing a document’s HTML Document Object Model usually involves some degree of uncertainty), and as the result of integrating information from many different heterogeneous sources (such as in aggregator sites, which allow users to query multiple sites at once to save time). Finally, inconsistency and incompleteness are also ubiquitous as the result of over- and under-specification, respectively. To be applicable to Web-sized data sets, any machinery developed for dealing with uncertainty in these settings must be scalable. In this paper, we develop an extension of $\mathcal{EL}^{++}$ [1] by means of a probabilistic semantics based on Markov logic networks [20]. $\mathcal{EL}^{++}$ is a DL that combines tractability of several key reasoning problems with enough expressive power to model a variety of ontologies; for instance, it is expressive enough to model real-world ontologies such as the well-known SNOMED CT, large segments of the Galen medical knowledge base, as well as the Gene Ontology. Moreover, $\mathcal{EL}^{++}$ underlies the OWL2 EL profile, in which basic reasoning problems are solvable in polynomial time, and highly scalable implementations are available. One of the key aspects of the extension presented here is that it is tightly coupled, meaning that probabilistic annotations can refer to objects in the ontology, providing greater expressive power than similar efforts in the literature.

In the area of Web data extraction [16], uncertainty comes into play even for very basic tasks. Consider, as an example, the form-labeling problem, consisting of the association of blocks of text (i.e., the labels) to the corresponding fields in a Web form [6], illustrated in Figure 1, where three possible instances of the problem with decreasing likelihood of occurrence are shown. In the first case, the fields $f_1$ and $f_2$ have horizontally aligned blocks of text on their left and on their right; this case exemplifies the most typical situation where the blocks of text on the left of a field (denoted $l_1$ and $l_2$) are the actual labels, while those on the right (denoted $t_1$ and $t_2$) are either unrelated or carry additional information about the fields. The second case represents the symmetric situation, where the labels are on the right of the fields. This situation is typical for eastern Web sites where the content has right-to-left reading order (e.g., for Arabic). Another possibility is exemplified by the third case, where the labels occur in the north-west region of the field. In this