Learning of OWL Class Descriptions on Very Large Knowledge Bases

Sebastian Hellmann, Universität Leipzig, Germany
Jens Lehmann, Universität Leipzig, Germany
Sören Auer, Universität Leipzig, Germany

ABSTRACT

The vision of the Semantic Web is to make use of semantic representations on the largest possible scale - the Web. Large knowledge bases such as DBpedia, OpenCyc, GovTrack, and others are emerging and are freely available as Linked Data and SPARQL endpoints. Exploring and analysing such knowledge bases is a significant hurdle for Semantic Web research and practice. As one possible direction for tackling this problem, the authors present an approach for obtaining complex class descriptions from objects in knowledge bases by using Machine Learning techniques. They describe in detail how we leverage existing techniques to achieve scalability on large knowledge bases available as SPARQL endpoints or Linked Data. Their algorithms are made available in the open source DL-Learner project and we present several real-life scenarios in which they can be used by Semantic Web applications. [Article copies are available for purchase from InfoSci-on-Demand.com]

Keywords: Data Processing; Data Extraction; Information Filtering; Knowledge Base; Ontologies; Reasoning; RDF; Scalability; Semantic Web; SPARQL

INTRODUCTION

The vision of the Semantic Web is to make use of semantic representations on the largest possible scale - the Web. We currently experience that Semantic Web technologies are gaining momentum and large knowledge bases such as DBpedia (Auer et al., 2007), OpenCyc (Lenat, 1995), GovTrack (Tauberer, 2008) and others are freely available. These knowledge bases are based on semantic knowledge representation standards like RDF and OWL. They contain hundred thousands of properties as well as classes and an even larger number of facts and relationships. These knowledge bases and many more (ESWWiki, 2008) are available as Linked Data (Berners-Lee, 2006; Bizer, Cyganiak, & Heath, 2007) or SPARQL endpoints (Clark, Feigenbaum, & Torres, 2008).
Due to their sheer size, users of these knowledge bases, however, are facing the problem, that they can hardly know which identifiers are used and are available for the construction of queries. Furthermore, domain experts might not be able to express their queries in a structured form at all, but they often have a very precise imagination what kind of results they would like to retrieve. A historian, for example, searching in DBpedia for ancient Greek law philosophers influenced by Plato can easily name some examples and if presented a selection of prospective results he will be able to quickly identify false results. However, he might not be able to efficiently construct a formal query adhering to the large DBpedia knowledge base a priori.

The construction of queries asking for objects of a certain kind contained in an ontology, such as in the previous example, can be understood as a class construction problem: We are searching for a class description which subsumes exactly those objects adhering to our informal query (e.g. ancient Greek law philosophers influenced by Plato). Recently, several methods have been proposed for constructing ontology classes by means of Machine Learning techniques from positive and negative examples (Lehmann & Hitzler, 2007a, 2007b). These techniques are tailored for small and medium size knowledge bases, while they cannot be directly applied to large knowledge bases (such as the initially mentioned ones) due to their dependency on reasoning methods. In this paper, we present an approach for leveraging Machine Learning algorithms for learning of ontology class descriptions in large knowledge bases, in particular those available as SPARQL (Clark et al., 2008) endpoints or Linked Data. The scalability of the algorithms is ensured by reasoning only over “interesting parts” of a knowledge base for a given task. As a result users of large knowledge bases are empowered to construct queries by iteratively providing positive and negative examples to be contained in the prospective result set.

Overall, we make the following contributions:

- development of a flexible method for extracting relevant parts of very large and possibly interlinked knowledge bases for a given learning task,
- thorough implementation, integration, and evaluation of these methods in the DL-Learner framework (Lehmann, 2007)
- presentation of several application scenarios and examples employing some of the largest knowledge bases available on the Web.

In this article, we will first cover preliminaries, namely a quick introduction in Description Logics, OWL, and the learning problem we consider. We then briefly describe the underlying learning algorithm and present in detail how it can be applied on very large knowledge sources. We describe and evaluate our approach in several application scenarios in the following sections and, finally, conclude with some related work and an outlook on future work.

**OWL, DESCRIPTION LOGICS**

Description logics are a family of knowledge representation (KR) formalisms. They emerged from earlier KR formalisms like semantic networks and frames. Their origin lies in the work of Brachmann on structured inheritance networks (Brachman, 1978). Since then, description logics have enjoyed increasing popularity. They can essentially be understood as fragments of first-order predicate logic. They have less expressive power, but usually decidable inference problems and a user-friendly variable free syntax.

Description logics represent knowledge in terms of objects, concepts, and roles. Concepts formally describe notions in an application domain, e.g. we could define the concept of being a father as “a man having a child” (Father ≡ Man ⊓ ∃hasChild.⊤ in DL notation). Objects are members of concepts in the application domain and roles are binary relations between objects. Objects correspond
Related Content

GlobalMind: Automated Analysis of Cultural Contexts with Multicultural
[www.irma-international.org/article/globalmind-automated-analysis-cultural-contexts/2831/](www.irma-international.org/article/globalmind-automated-analysis-cultural-contexts/2831/)

IoT-Based Big Data: From Smart City towards Next Generation Super City Planning
[www.irma-international.org/article/iot-based-big-data/172421/](www.irma-international.org/article/iot-based-big-data/172421/)

Neural Networks for Content-Based Image Retrieval
[www.irma-international.org/chapter/neural-networks-content-based-image/28930/](www.irma-international.org/chapter/neural-networks-content-based-image/28930/)

GlobalMind: Automated Analysis of Cultural Contexts with Multicultural
[www.irma-international.org/article/globalmind-automated-analysis-cultural-contexts/2831/](www.irma-international.org/article/globalmind-automated-analysis-cultural-contexts/2831/)

Nested Optional Join for Efficient Evaluation of SPARQL Nested Optional Graph Patterns
[www.irma-international.org/chapter/nested-optional-join-efficient-evaluation/41658/](www.irma-international.org/chapter/nested-optional-join-efficient-evaluation/41658/)