Chapter 16
FOL Learning for Knowledge Discovery in Documents

Stefano Ferilli
Università degli Studi di Bari, Italy

Floriana Esposito
Università degli Studi di Bari, Italy

Marenglen Biba
Università degli Studi di Bari, Italy

Teresa M.A. Basile
Università degli Studi di Bari, Italy

Nicola Di Mauro
Università degli Studi di Bari, Italy

ABSTRACT

This chapter proposes the application of machine learning techniques, based on first-order logic as a representation language, to the real-world application domain of document processing. First, the tasks and problems involved in document processing are presented, along with the prototypical system DOMINUS and its architecture, whose components are aimed at facing these issues. Then, a closer look is provided for the learning component of the system, and the two sub-systems that are in charge of performing supervised and unsupervised learning as a support to the system performance. Finally, some experiments are reported that assess the quality of the learning performance. This is intended to prove to researchers and practitioners of the field that first-order logic learning can be a viable solution to tackle the domain complexity, and to solve problems such as incremental evolution of the document repository.

INTRODUCTION

After some years in which it was seen as an area of interest only for research, Machine Learning (ML for short) techniques and systems have started to gain progressive credit in the everyday Computer Sci-

DOI: 10.4018/978-1-60566-766-9.ch016

ence landscape, and to be used for facing several real-world problems where classical systems show their limits. When ML systems work in real-world domains, they must typically deal with features such as complexity, need for efficiency and continuous adaptation to change. Optionally, humans may need to understand the inferred models in order to check and validate them. While the numeric and statistical setting, traditionally studied in the literature, can ensure efficiency, the other requirements call for more powerful representation formalisms. First-Order Logic (FOL for short) representation and processing techniques are more suitable than attribute-value and propositional ones for dealing with complexity and providing human understandability of the inferred models; moreover, they can deal with change and adaptation as well.

This work presents a suite of symbolic FOL learning algorithms and systems, that can serve as larger system components for satisfying these requirements in accomplishing real-world tasks. Their FOL representation language allows to effectively and efficiently deal with complex domains, yielding uniform human-understandable descriptions for observations and models. The FOL learning system INTHELEX tackles all of these requirements in a supervised setting. Being based on an incremental algorithm, it can update and refine the inferred models, instead of learning new ones from scratch, in order to account for new available evidence. Interestingly, its incremental abilities are not limited to examples processing, but allow to add to the theory and handle even completely new classes as soon as their instances come up. Conversely, when new observations become available for which a target classification is not given, an unsupervised setting is needed. In such a case, another module of the suite, that implements a recently developed similarity framework for FOL (Horn clause) representations, allows to face the problem.

This chapter focuses on Document Processing and Management, a real-world application domain that is gaining increasing interest in recent years, due to the progressive digitization of information sources. It involves almost all of the above requirements: need for quick response to the users' requests, complexity due to the high variability of documents, need for the librarians of checking and validating the inferred models, continuous flow of new documents. DOMINUS is a general-purpose document management framework that is able to process documents in standard electronic formats in order to recognize the type they belong to and their significant components based on their layout structure, and to selectively extract relevant information to be used for semantic indexing and later retrieval. Possible specific applications of the framework include support for the Semantic Web on Internet documents and content-based document management in organizations and libraries. Its architecture includes a module that provides Machine Learning services that support the different tasks involved in document processing and management. The FOL techniques presented in this chapter were embedded in such a module to provide the core functionality for making the framework powerful and flexible enough to deal with real-world cases.

In the following, after presenting the tasks and problems involved in Digital Libraries management, and how they have been tackled in DOMINUS, a more technical presentation of the incremental FOL framework exploited for document classification and understanding will be provided. The basics of the first-order logic setting will be recalled, and the similarity assessment on which the framework is based will be presented. The supervised and unsupervised learning modules and their cooperation will be then discussed, followed by experiments on a real-world dataset that show how the proposed framework can successfully and effectively be exploited as a viable solution to the incremental extension of documents and document classes in a digital library.