Applying Markov Logic to Document Annotation and Citation Deduplication

Jean Baptiste Faddoul*, Boris Chidlovskii*

*Xerox Research Centre Europe
6, chemin de Maupertuis 38240 Meylan-FRANCE

Abstract. Structured learning approaches are able to take into account the relational structure of data, thus promising an enhancement over non-relational approaches. In this paper we explore two document-related tasks in relational domains setting, the annotation of semi-structured documents and the citation deduplication. For both tasks, we report results of comparing relational learning approach namely *Markov logic*, to non-relational one namely *Support Vector Machines (SVM)*. We discover that increased complexity due to the relational setting is difficult to manage in large scale cases, where non-relational models might perform better. Moreover, our experiments show that in Markov logic, the contribution of its probabilistic component decreases in large scale domains, and it tends to act like First-order logic *(FOL)*.

1 Introduction

A large majority of existing machine learning models can be seen as non-relational. They represent each object as an isolated point in a space, and they learn prediction models using the features of each object. A new trend in statistical machine learning is represented by relational models that take into account the relational structure of data. The relations between objects are frequent in real world cases, and taking them into account may offer a potential performance improvement over non-relational models. On the other hand, real world data sets are large in the number of objects, the feature dimensions and even the relation numbers. So, from the scalability point of view, simpler non-relational models might scale better and might outperform complex ones. In this paper, we study relational and non-relational models on two different tasks, the annotation of semi-structured documents and the citation deduplication.

In our experiments, we use SVM as the best state of art non-relational model. As relational models, we applied *Markov logic* (introduced in Domingos and Richardson (2007)), which is a *Statistical Relational Learning (SRL)* approach that combines FOL and Markov networks. Markov logic has been chosen for its capacity to represent more complex relations than previously used models and a number of important results shown when tested on different domains (Culotta and McCallum (2006) and Kok and Yih (2009)).