

Ludovic Denoyer
Université Paris VI

Corpus based Structure Mapping Application to XML documents corpora

We address the problem of learning to map automatically flat and semi-structured documents onto a mediated target XML schema. This problem is motivated by the recent development of IR applications for searching and analyzing semi-structured document sources and corpora. Academic research has mainly dealt with homogeneous collections. In practical applications, data come from multiple heterogeneous sources and querying such collections requires to define a mapping or correspondence between the different document formats. Automating the design of such mappings has rapidly become a key issue for these applications. We propose a machine learning approach to this problem where the mapping is learned from pairs of input and corresponding target documents provided by a user. The mapping process is formalized as a Markovian Decision Process, and training is performed through a classical machine learning framework known as reinforcement learning. The resulting model is able to cope with complex mappings while keeping a linear complexity. We describe a set of experiments on several corpora representative of different mapping tasks and show that the method is able to learn mappings with a high accuracy on different corpora.