Automatic Acquisition of Topic Hierarchies from Wikipedia

Ruben Dorado

June 2, 2016

1 Introduction

Wikipedia is a free-access encyclopedia, written collaborative by all people over the world. Wikipedia is targeted to a general public that can access the text content through a web interface. In this work, we propose the use of wikipedia content to obtain a topic hierarchy automatically given an article. For example, given the article "dog", the method should return the terms that are related to it such as: {"animal","canid","predator", "breed", ... }.

The project will cover the **phase** of concept modelling, since a model to acquire concepts will take into account. However, preprocessing and post processing phases are also required.

The **data** will consist of the wikipedia dump, which contains the full wikipedia in an xml file. This archive will be preprocessed to extract text and annotations. Both of them will be used to feed a data mining algorithm to obtain the topic hierarchy.

For this project, we will use the following **tools**:

- Wikitools: a tool developed R. Dorado to extract text and annotations from the Wikipedia
- Spark: a general engine for big data processing that is advertised as fast and scalable. This engine also contains built-in modules for machine learning.
- MLib: a machine learning library for Spark. It contains implementation of algorithms for tasks such as classification, clustering, dimensionality reduction, and feature extraction

- Epic: provides structured prediction of structural prediction for Scala. It contains modules for parsing, pos-tagging and named entity recognition.
- FACTORIE: a library for probabilistic graphical modelling that includes modules for inference, learning, optimization, and NLP.