Abstract

The thesis treats automatic extraction of semantic data from Web pages. Within this broad problem, it focuses on finding values of data figures within the page presenting certain entity (e.g. price of a laptop). The main idea we wanted to evaluate is that a figure can be found using its context in the page: the words that surround it and values of the attributes of the containing HTML tags, class attribute in particular.

Our research revealed there are two types of contemporary solutions of this problem: either the author of the Web page must inline semantic information inside the markup of the page or there are commercial tools that can be trained to parse a particular page format (targetting pages from a single Web domain).

We examined the possibilities of developing a general solution that would – for given entity – find its properties across the Web domains using text analysis and machine learning. The naïve algorithm had about 30% accuracy, the learning algorithms had the accuracy between 40 and 50% in finding the properties. Despite the accuracy is not acceptable for a final solution, we believe it confirms the potential of the idea.

Keywords: Web pages data extraction