Title: Structured Data Extraction from Unstructured Text

Author: Bc. Peter Kóša

Department: Department of Software Engineering

Supervisor: Mgr. Martin Nečaský, Ph.D., Department of Software Engineering

Abstract:

In the last 20 years, there has been an ever-growing amount of information present on the Internet and in published texts. However, this information is often in a non-structured format and this causes various problems such as the inability to efficiently search in diverse collections of texts (medical reports, ads, etc.). To overcome these problems, we need efficient tools capable of automatic processing, extracting the important information and storing of these results in some form for later reuse. The purpose of this thesis is to compare existing solutions as well as to compare them with our solution, which was created in the scope of software project SemJob. The SemJob project is introduced and the reader can therefore obtain knowledge about its inner structure and workings.

Keywords:

structured data extraction, extraction rules, ontologies, (semi)automatic wrapper induction