Title: Interactive crawling and data extraction

Author: Bc. Petr Fejfar

Author's e–mail address: pfejfar@gmail.com

Department: Department of Distributed and Dependable Systems

Supervisor: Mgr. Pavel Ježek, Ph.D., Department of Distributed and Dependable Systems

Abstract: The subject of this thesis is Web crawling and data extraction from Rich Internet Applications (RIA).

The thesis starts with analysis of modern Web pages along with techniques used for crawling and data extraction. Based on this analysis, we designed a tool which crawls RIAs according to the instructions defined by the user via graphic interface. In contrast with other currently popular tools for RIAs, our solution is targeted at users with no programming experience, including business and analyst users.

The designed solution itself is implemented in form of RIA, using the WebDriver protocol to automate multiple browsers according to user–defined instructions. Our tool allows the user to inspect browser sessions by displaying pages that are being crawled simultaneously. This feature enables the user to troubleshoot the crawlers.

The outcome of this thesis is a fully design and implemented tool enabling business user to extract data from the RIAs. This opens new opportunities for this type of user to collect data from Web pages for use as primary data, as well data for further analysis.

Keywords: Web crawling, Web data extraction, Web scraping, AJAX, RIA, Rich Internet Application, browser automation