Název práce: Interaktivní procházení webu a extrakce dat

Autor: Bc. Petr Fejfar

E–mailová adresa autora: pfejfar@gmail.com

Katedra: Katedra distribuovaných a spolehlivých systémů

Vedoucí práce: Mgr. Pavel Ježek, Ph.D., Katedra distribuovaných a spolehlivých systémů

Abstrakt: Tato práce se zaměřuje na problematiku automatického procházení stránek a extrakce dat v kontextu moderních webových aplikací, obsahujících vysoké množství aplikační logiky implementované v prohlížeči pomocí JavaScriptu.

V práci je provedena analýza moderních webových stránek, spolu s technikami, které jsou bežně používany k extrakci dat. Na základě této analýzy jsme navrhli nástroj, který moderní webové stránky prochází na základě instrukcí zadaných uživatelem pomocí grafického prostředí. Narozdíl od ostatních nástrojů na procházení a extrakci dat z moderních webových stránek, náš nástroj umožnuje práci uživatelům, kteří nemají zkušenosti s programováním.

Navhrhovaný nástroj je implementován jako webová aplikace a využívá protokolu WebDriver pro automatizaci více prohlžečů pro procházení a extrakci dat z webových stránek pomocí uživatelem definovaných posloupností instrukcí. Náš nástroj umožňuje uživateli prozkoumat aktuální stav prohlížeče extrahujícího data zobrazením aktuálně prochazené stránky. Toto umožní uživatelům vyhledávat a ladit chyby jejich posloupností instrukcí, tak aby extrahovaly data, které mají extrahovat.

Výstupem této práce je návrh a následná implementace nástroje pro extrakci dat z moderních webových stránek pro uživatele bez schopnosti programovat. Tento nástroj umožní sběr dat, který dříve nebyl možný. Tyto data mohou být využity pro další analýzu nebo jako vstupní data do dalších systému.

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