**ROZŠÍRENÉ ZADANIE DIPLOMOVEJ PRÁCE**

**Názov práce:** Webový scraper v rozšírení prehliadača s poloautomatickou anotáciou

**Riešiteľ:** Bc. Dávid Varga

**Vedúci práce:** RNDr. Peter Gurský, PhD.

**Školiace pracovisko:** ÚINF - Ústav informatiky

**Ciele:**

1. Vytvorenie a implementovanie modelu pre anotovanie ľubovoľných objektov z webu.
2. Vytvorenie a implementovanie spôsobu navštevovania stránok obsahujúcich anotovaný objekt do webového crawlera.
3. Zakomponovanie filtrovania objektov spĺňajúcich zadané kritériá do webového crawlera.

**Popis:**

Exago je nástroj pre anotáciu internetových obchodov, ktorý je implementovaný ako rozšírenie pre prehliadač Google Chrome. Jeho cieľom je umožniť anotovanie pomocou klikania myšou, pričom využíva generovanie XPath-ov a regulárnych výrazov.

Hlavným cieľom tejto práce je prerobiť nástroj Exago na všeobecný webový scraper, ktorý bude schopný extrahovať ľubovoľné objekty z webu. V tejto fáze je potrebné navrhnúť spôsob anotovania:

1. viacerých objektov zobrazených na jednej stránke,
2. jedného objektu, ktorého informácie sa nachádzajú na jednej stránke,
3. jedného objektu, ktorého informácie sa nachádzajú na viacerých stránkach.

Taktiež je potrebné vyriešiť spôsob anotovania komplexných atribútov objektu, ako napríklad rôzne zoznamy, časti HTML kódu, obrázky, tabuľky, súbory na stiahnutie a iné neatomické atribúty. Tento spôsob anotovania má byť užívateľsky prívetivý a má si zachovať vlastnosť anotovania iba pomocou klikania myšou. Z tohto dôvodu je potrebné vykonať detailnú analýzu vyše 50 súčasných webových scraperov, pričom sa sústredíme na postup anotovania, ktorý využívajú, a detailne skúmať všetky funkcionality týchto scraperov.

Ďalším cieľom je nájsť čo najefektívnejší spôsob prechádzania stránok anotovaného portálu. Potrebujeme teda nájsť všetky stránky na danom portáli, kde sa nachádzajú objekty, ktoré chceme extrahovať, a prechádzať pritom čo najmenší počet stránok. Jeden z návrhov je umožniť pridávanie nápovedí počas anotovania, aby sa obmedzilo zbytočné navštevovanie irelevantných stránok.

Väčšina súčasných webových scraperov funguje tak, že používateľ zadá spôsob extrakcie dát, a následne podľa tohto spôsobu je vykonaná extrakcia. Táto práca sa chce priblížiť k takému spôsobu anotovania, kde používateľ zadá typ dát, aký chce získať, a následne si webový scraper automaticky vyberie spôsob ako extrahovať tieto dáta.

**Literatúra:**

Liu, Bing: Web Data Mining: Exploring Hyperlinks, Contents,and Usage Data. Second Edition, ISBN 978-3-642-19459-7, Springer, 2011

Kushmerick, N.: Wrapper induction: efficiency and expressiveness. Artificial Intelligence, 118:15-68, 2000.

Muslea, I., Minton, S. and Knoblock, C.: A hierarchical approach to wrapper induction. Agents-99, 1999.

Cohen, W., Hurst, M., and Jensen, L.: A flexible learning system for wrapping tables and lists in HTML documents. WWW-2002, 2002.

Hsu, C.N., Dung, M.T.: Generating finite-state transducers for semi-structured data extraction from the Web. Information Systems. 23(8): 521-538, 1998.

Chabaľ, V: Poloautomatická extrakcia komentárov z produktových katalógov. Diplomová práca. Košice 2014

Crescenzi, V., Mecca, G., Merialdo,P.: Roadrunner: Towards automatic data extraction from large web sites. In Proceedings of VLDB 2001, pp. 109-118.

**Články:**

Peter Gurský, Matej Perejda, and Dávid Varga: Semiautomatic annotation of e-shops , ITAT 2018

Rudolf Pavel, Peter Gurský: Focused Web Crawling of Relevant Pages on e-shops, ITAT 2017