ROZŠÍRENÉ ZADANIE DIPLOMOVEJ PRÁCE

Názov práce: Koordinované získavanie a extrakcia dát z webových portálov cez spolupracujúce rozšírenia webových prehliadačov

Riešiteľ: Bc. Matej Perejda

Vedúci práce: RNDr. Peter Gurský, PhD.

Školiace pracovisko: ÚINF - Ústav informatiky

Ciele:

1. Porovnanie súčasných spôsobov extrakcie dát z webových portálov najmä z hľadiska schopnosti extrahovať dáta z dynamicky vytváraných webových stránok cez AJAX volania a schopnosti distribúcie procesu prehľadávania a extrakcie.
2. Obohatenie existujúceho rozšírenia webového prehliadača na anotáciu webových stránok o schopnosť prehľadávania a extrakcie dát z webu aj pre dynamické webové stránky simuláciou správania používateľa.
3. Návrh a vytvorenie škálovateľného servera koordinujúceho spoluprácu viacerých inštancií vytvoreného rozšírenia webového prehliadača z cieľa 2.
4. Otestovanie korektnosti a škálovateľnosti vytvoreného riešenia extrakciou reálnych webových portálov.

Popis:

Cieľom projektu Kapsa je vytváranie katalógu produktov internetových obchodov za účelom ich vzájomného porovnávania na základe rôznych vlastností, ponuky predajcov či spokojnosti zákazníkov. Využitím rozšírenia webových prehliadačov s názvom Exago dokážeme interaktívne anotovať webové produktové katalógy, následne vytvoriť sadu pravidiel na extrahovanie atribútov (tzv. wrapper) a tieto pravidlá zasielať na server.

Jedným z cieľov našej práce je obohatiť rozšírenie Exago o koordinované prechádzanie webovým portálom, hľadanie stránok na extrakciu a odosielanie extrahovaných dát (napr. atribútov produktov, obrázkov, komentárov a nájdených odkazov na ďalšie produkty) na server. Z Exaga vytvoríme extraktor dát simulujúci užívateľskú aktivitu v internetových obchodoch, samostatne prechádzajúci tieto stránky a získavajúci z nich produktové informácie. Nainštalované rozšírenia budú riadené serverom, ktorý im jednotlivo pridelí úlohy na vyhľadávanie produktov. Ďalšou fázou

práce bude teda návrh, implementácia a nasadenie škálovateľného servera na koordináciu úloh extrakcie dát. Ukladaním informácií o úlohách na spracovanie do distribuovanej databázy umožníme, v prípade veľkého množstva pripojených klientov, nasadenie totožnej inštancie servera na iný stroj. Tým sa koordinácia prideľovania URL adries klientom rozdelí a zrýchli.

Riešenie tejto práce ponúka vylepšenie pôvodného spôsobu získavania a extrahovania dát z webových stránok v projekte Kapsa. Jednotlivé úlohy sa rozdistribuujú medzi viaceré stroje s rôznymi IP adresami, čím sa odľahčí a zrýchli práca na serveri. Mnohonásobný prístup k stránkam internetových obchodov v krátkom časovom intervale z jednej IP adresy môže viesť k blokovaniu prístupu zo strany správcu daného internetového obchodu. Distribúciou úloh medzi viaceré stroje sa dokážeme vyhnúť potencionálnemu zablokovaniu prístupu.

V práci sa tiež budeme zaoberať porovnaním vlastností existujúcich webových scraperov s rozšírením Exago, prípadne výhodami a nevýhodami ich použitia. Posledná fáza práce bude venovaná testovaniu korektnosti a škálovateľnosti nášho rozšírenia extrakciou atribútov produktov z reálnych internetových obchodov.

Literatúra:

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

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

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

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

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

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

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