

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

Ústav informatiky PF UPJŠ

Autor: **Bc. Matej Perejda**

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

Kapsa

- katalóg produktov s anotáciou
- vznik na ÚINF UPJŠ KE
- vytváranie katalógu produktov ponúkaných e-shopmi
- porovnávanie produktov podľa rôznych vlastností, recenzií používateľov
a ponuky predajcov

Exago

- zásuvný modul, rozšírenie do webových prehliadačov,
- interaktívna anotácia webových produktových katalógov,
- automatické extrahovanie atribútov z popisov produktov,
- zaslanie výslednej anotácie na server (**wrapper**),
- editovanie existujúcej anotácie zo servera,
- generovanie XPath a regex (**nové**),
- výnimočnosť: označovanie viacerých atribútov naraz.

Parametry produktu

Súhrn	
Zaradenie	Smartphone, Android telefon
Výrobca	Samsung
Konstrukcia	dotykové
Operačný systém	Android
Verzia operačného systému	Android 7.0 (Nougat)
Hmotnosť	155 g
Možnosť pamätevej karty	áno
Pamäť RAM	4096 MB
Displej	
Rozlíšenie displeja	2960 x 1440
Veľkosť displeja	5.8"
Počet farieb	16 mil. farieb
Počet displejov	1

Exago 2

Start new annotation

Part of URL indicating e-shop: heureka.sk

Show and send wrapper Open crawler manager

Crawler rules Detail page spec. Annotation

Start page Add click

Attributes Comments

List of items:

xPath: `//*[contains(@class, 'js-param-table')]/table/tbody/tr`

Pagination

Attribute name:

xPath: `*[contains(@class, 'product-body__specification__para`

Regex: `(?<=)(\s)*(?=)`

Result: Zaradenie

Attribute value:

xPath: `td[2]`

Regex:

Result: `[\n <a href="https://smartph...`

Label/value

Value in URL Known value Values list Image

inšpektor tester + - akceptuj zmeny späť ďalej

paste do konzoly

Počet for-each úrovní: 0 1 2 3

Úrovne XPath-ov:

Modifikované úrovne XPath-ov:

Exago – interaktívna anotácia

Exago - GUI

Profesijná motivácia

- distribúcia úloh medzi viaceré stroje (odľahčenie servera),
- využitie JavaScript-u (Java nie je dobrá voľba),
- „viac strojov, viac IP adries“ (nepôsobiť ako zlodej dát).

Ciele diplomovej práce

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.

Postup práce

- vytvorenie prehľadu webových scraperov,
- pochopenie Exaga,
- článok - ITAT 2018 - Framework for Distributed Computing on the Web (Šiller, Kuchař),
- vytvorenie extraktora dát z webových stránok v Exagu,
- rozšírenie Exaga o prechádzanie webovým portálom a hľadanie stránok na extrakciu,
- návrh škálovateľného servera na koordináciu úloh extrakcie,
- implementácia a nasadenie servera,
- koordinácia viacerých klientov prostredníctvom servera,
- testovanie.

Zdroje a literatúra

Prehľad webových scraperov: [google.com](https://www.google.com), github.com, obchod Google Chrome, obchod Firefox Add-ons, oficiálne stránky nástrojov, scraping.pro, hongkiat.com/blog/web-scraping-tools/

- [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] Chabal', 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.

Ďakujem za pozornosť!

Otázky?