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,
- editovanie existujúcej anotácie zo servera,
- výnimočnosť: označovanie viacerých atribútov naraz.

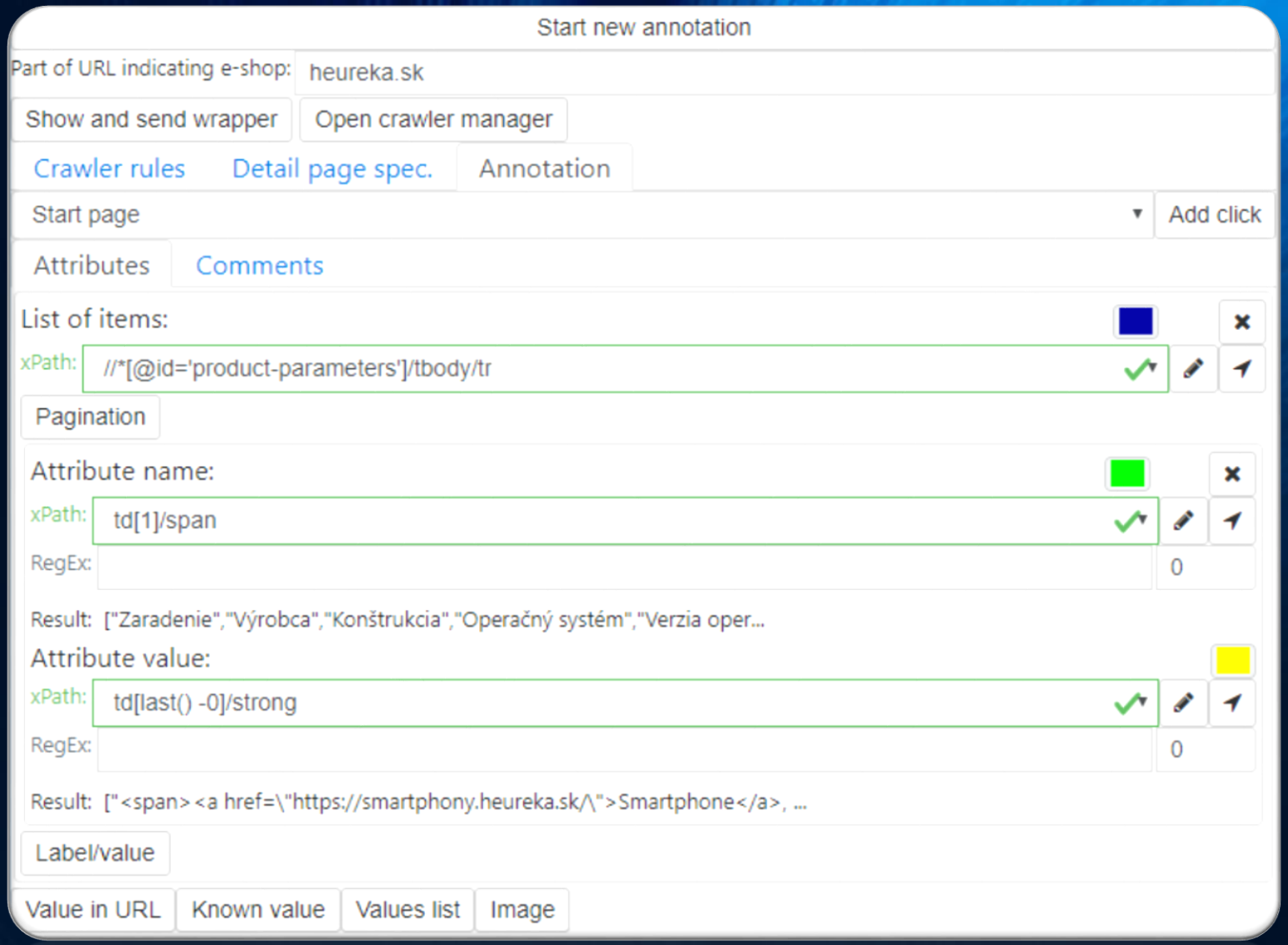
Parametry produktu

Súhrn

Zaradenie	Smartphone, Android telefon
Výrobca	Samsung
Konštrukcia	dotykové
Operačný systém	Android
Verzia operačného systému	Android 7.0 (Nougat)
Hmotnosť	155 g
Možnosť pamäťovej karty	áno
Pamäť RAM	4096 MB

Displej

Rozlíšenie displeja	2960 x 1440
Veľkosť displeja	5.8 "
Počet farieb	16 mil. farieb
Počet displejov	1


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:
//*[@id='product-parameters']/tbody/tr
Pagination
Attribute name:
xPath:
td[1]/span
RegEx:
0
Result: ["Zaradenie","Výrobca","Konštrukcia","Operačný systém","Verzia oper...
Attribute value:
xPath:
td[last() -0]/strong
RegEx:
0
Result: ["Smartphone, ...
Label/value
Value in URL
Known value
Values list
Image

Parametry produktu

Súhrn

Zaradenie	Smartphone, Android telefon
Výrobca	Samsung
Konštrukcia	dotykové
Operačný systém	Android
Verzia operačného systému	Android 7.0 (Nougat)
Hmotnosť	155 g
Možnosť pamäťovej karty	áno
Pamäť RAM	4096 MB

Displej

Rozlíšenie displeja	2960 x 1440
Veľkosť displeja	5.8 "
Počet farieb	16 mil. farieb
Počet displejov	1

Profesijná motivácia

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 práce

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.

Ciele diplomovej práce

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.

Ciele diplomovej práce

3. Návrh a vytvorenie škálovateľného servera koordinujúceho spoluprácu viacerých inštancií vytvoreného rozšírenia webového prehliadača z cieľa 2.

Ciele diplomovej práce

4. Otestovanie korektnosti a škálovateľnosti vytvoreného riešenia extrakciou reálnych webových portálov.

Postup práce

Postup práce

- vytvorenie prehľadu webových scraperov,
- pochopenie Exaga,
- rozšírenie Exaga o prechádzanie webovým portálom a hľadanie stránok na extrakciu,
- vytvorenie extraktora dát z webových stránok v Exagu,
- návrh škálovateľného servera na koordináciu úloh extrakcie,
- implementácia a nasadenie servera,
- koordinácia viacerých klientov prostredníctvom servera,
- testovanie.

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.

Ďakujem za pozornosť!

Otázky?