Crawlovanie a extrakcia relevantných častí webových portálov

RNDr. Peter Gurský, PhD; Rudolf Pavel

Prírodovedecká fakulta UPJŠ v Košiciach

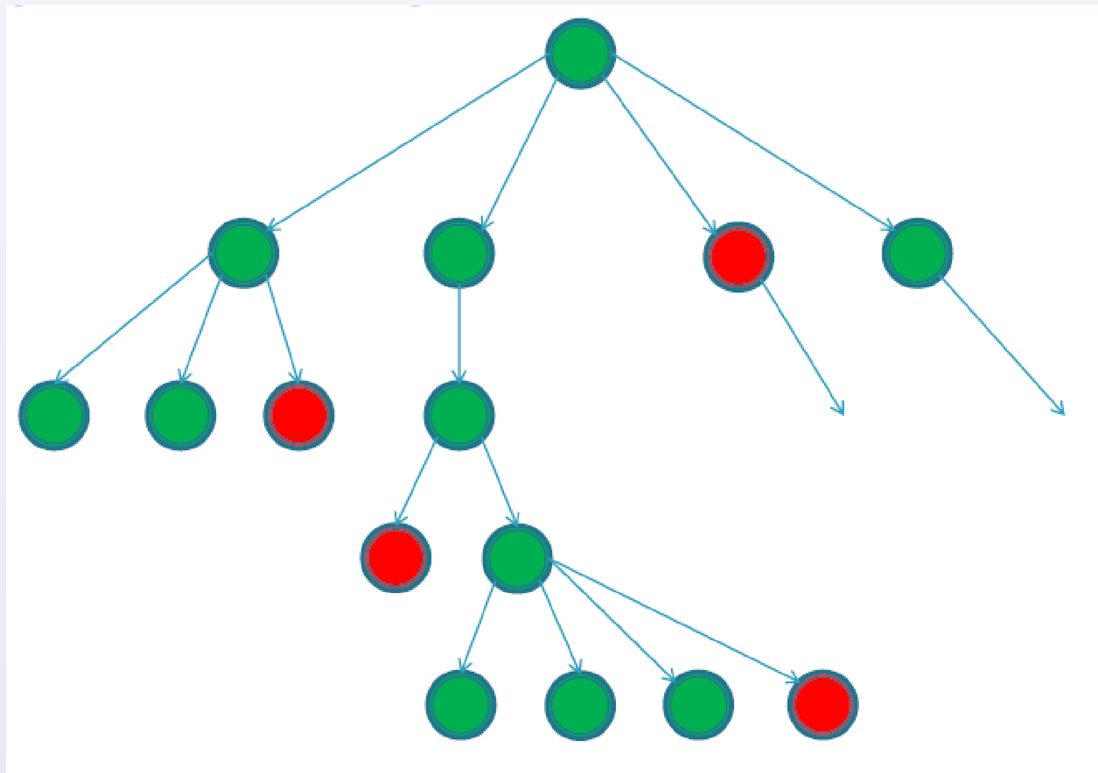
Abstrakt

Práca je súčasťou školského projektu kapsa. Zaoberá sa crawlovaním webových portálov internetových obchodov a následnou extrakciou dát z týchto portálov. Kľúčovou časťou práce je zabezpečiť extrakciu relevantných častí webových portálov za pomoci automatického orezávania prehľadávania webového portálu prostredníctvom analýzy úspešnosti predchádzajúcich vetiev.

Návrh riešenia

- Využívame crawler Crawler4j, ktorý je už v Kapse implementovaný.
- Pre získavanie elementov a ich XPathov z HTML kódu používame knižnice Xsoup a Jsoup.
- Na simuláciu kliknutí na webovej stánke používame Selenium browser automation.
- Pri crawlovaní zisťujeme element a XPathy každého prekliku.
- Dáta počas crawlovania ukladáme do tabuľky v databáze.
- Po skončení crawlovania sa metódou bottom – up v tabuľke označia relevantné a slepé vetvy.
- Potom sa metódou top – down pre všetky relevantné vetvy vypočítajú generalizované XPathy.
- Pri ďalších iteráciách crawlovania sa budú crawlovať len stránky, ktorých elementy nám vrátia generalizované XPathy nad HTML kódom.

Model časti stromu prehľadávania



- Zelenou farbou: relevantné vetvy
- Červenou farbou: slepé vetvy

Databáza

id	Parent Id	url	XPath_Id	isDetailPage	generalizedXPathId
1	0	http://penazenkyshop.sk/	0	3	1709448165
2	1	http://penazenkyshop.sk/cart/	1	0	0
3	1	http://penazenkyshop.sk/eshop/login	2	0	0
4	1	http://penazenkyshop.sk/kozene-opasky/c1022	3	2	7365840
5	4	http://penazenkyshop.sk/kozeny-opasok-sh-hnedy-10h/p665947c1023	4	1	2541916
6	4	http://penazenkyshop.sk/kozeny-opasok-sh-2016-chrom/p666233c1024	5	1	-1485511120
7	4	https://penazenkyshop.sk/registration/	6	0	0

id	idURL	XPath
1	1	//body/div[2]/div[1]
2	2	//body/div[2]/div[3]/div[2]
3	3	//body/div[2]/div[4]/div[3]
4	3	//body/div[2]/div[4]/div[4]

id	My Hash	idDownload	Generalized_Xpath
1	7365840	358	//body/div[2]/div[2]/div/a
2	7365840	358	//body/div[2]/div[2]/div[3]/div/a
3	2541916	358	//body/div[3]/div/section/div/div[3]/a

- Generalizácia Xpathov v tabuľke.
- Množinu generalizovaných Xpathov identifikujeme na základe ich hashu.
- Z XPathov: //body/div[2]/div[4]/div[3] a //body/div[2]/div[4]/div[4] získame generalizáciou //body/div[2]/div[2]/div/a

Poďakovanie

Ďakujem RNDr. Petrovi Gurskému, PhD., za cenné rady pri tvorbe práce.