Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Matevž Lapajne, Simon Klavžar in Tine Črnugelj

Spletni pajek

Domača naloga pri predmetu Iskanje in ekstrakcija podatkov s spleta
MENTOR: prof. dr. Marko Bajec, doc. dr. Slavko Žitnik

Pričujoče poročilo predstavlja implementacijo spletnega pajka pri predmetu Iskanje in ekstrakcija podatkov s spleta. Cilj naloge je bil zgraditi samostojni program za proces iskanja, pregledovanja in zbiranja podatkov spletnih strani, katerih naslov domene se konča z besedo »gov.si«.

1. Uvod

Spletni pajek deluje tako, da sledi povezavam med spletnimi stranmi in obišče vsako na seznamu, pridobiva njihovo vsebino in poizkuša izluščiti želene podatke s pomočjo razčlenjevalnika HTML. Podatke nato shranjuje v podatkovno bazo za nadaljnjo obdelavo ali uporabo. Uporablja se lahko za različne namene, kot so pridobivanje podatkov za analizo trga, odkrivanje spletnih ranljivosti, izdelavo spletnih indeksov in še več. Njihova uporaba pa je lahko tudi etično vprašljiva, saj lahko vplivajo na zasebnost in varnost uporabnikov spletnih strani. Zato je pomembno, da se spletni pajki uporabljajo v skladu z zakonodajo in etičnimi smernicami. Pri programski izvedbi smo sledili navodilom za programsko nalogo 1, v kateri nam je vsebovana shema za oblikovanje Crawldb bila zelo v pomoč pri razumevanju kompleksnih procesov.

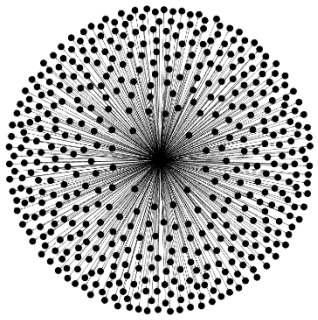
2. Struktura pajka

Pajek sprva začne pregledovati začetne spletne strani, tako da pogleda v domenske datoteke robots.txt iz katerih izlušči url povezave za katere upravitelj strani predpostavi, da se jih sme pregledovati. Iz te datoteke lahko dobi povezavo na datoteko sitemap, ki vsebuje seznam vseh strani na tej spletni domeni. V naslednjem koraku začne s pregledom vsebine teh, v katerih išče url naslove izhodnih strani in nadaljuje z izločevanjem ustreznih in nepodvojenih naslovov. Te se v naslednjem koraku, skupaj z drugimi podatki na strani kot so slike, vsebina strani, čas dostopa, itd. shranjujejo v lokalno bazo.

3. Statistika

	Site	Page	HTML	DUPLICATE	BINARY	FROTNIER	IMAGES
vse	181	369381	9557	6	0	359818	17807
gov.si	/	365274	9557	6	0	355711	17807
evem.gov.si	/	85	0	0	0	85	0
e-uprava.gov.si	/	249	0	0	0	249	0
e-prostor.gov.si	/	14	0	0	0	14	0

4. Vizualizacija



Slika 1 Prikazuje glavno povezavo strani z uporabo sita giant component v orodju gephi

5. Težave

Med delom smo naleteli na več težav. Največ težav smo imeli z implementacijo večnitnosti, kjer smo imeli težave s pravilnim upoštevanjem časovnih zakasnitev za dostop do domen oziroma strežnikov. Ker za naše težave nismo našli dobre rešitve, večnitnosti v končni verziji pajka nismo obdržali. Nekaj napak smo odkrili tudi med izvajanjem pajka in pregledovanjem baze. Pri shranjenih prvih tisoč strani smo opazili, da nismo pravilno preverjali niza »gov.si« v spletnih domenah, zato smo v tabelo z domenami (crawldb.site) dobili tudi domeno »facebook.com«. Napako pri pregledovanju domen smo nato odpravili. Prepoznali smo, da nekatere strani za njihov dostop zahtevajo certifikat in slednje izločili. Nekaterim slikam nismo znali določiti njihovega formata, za katere smo ugotovili, da so kodirane s pomočjo algoritma Base64 in ob dešifriranju predstavljajo slike za sledenje aktivnosti uporabnika na spletni strani.

6. Zaključek

Spletne strani državne uprave obsegajo veliko količino spletnih naslovov. K velikem številu strani prispevajo tudi prevodi strani v različne jezike, kar je razvidno tudi iz »sitemap« datotek. Pri nalogi nam je implementacija večnitnosti povzročila nepričakovano veliko težav, kar nam je precej upočasnilo tako testiranje kot tudi izvajanje pajka.