Implementacija spletnega pajka

Skupina **DOMACI-NJOKI**: Niki Bizjak, Bojan Vrangeloski, Uroš Škrjanc

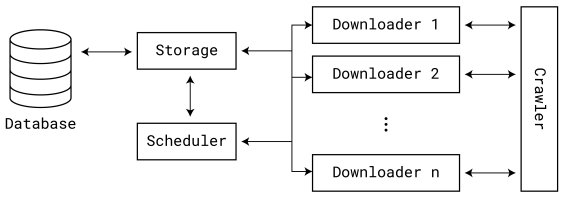
***Povzetek*—V poročilu o seminarski smo opisali strukturo našega spletnega pajka, njegovo delovanje in rezultate, ki smo jih dobili pri preizkušanju pajka.**

I. Uvod

Za prvo seminarsko nalogo pri predmetu Iskanje in ekstrakcija podatkov s spleta smo implementirali spletnega pajka, ki je obiskal in v lokalno podatkovno bazo prenesel vsebino spletnih strani na štirih spletnih mestih (`gov.si`, `evem.gov.si`, `e-uprava.gov.si`, `e-prostor.gov.si`). V nadaljevanju poročila smo opisali strukturo in delovanje našega pajka, predstavili rezultate naše naloge in težave, na katere smo naleteli tekom dela.

II. Struktura pajka

Pajka smo izdelali v programskem jeziku Python. Strukturno je pajek razdeljen na tri med seboj povezane komponente, ki so v programski kodi definirani kot objekti: razvrščevalnik, pajek in shramba.



Slika 1. Arhitektura implementiranega spletnega pajka

A. Razvrščevalnik

Razvrščevalnik (ang. scheduler) URL naslovov je implementiran kot podatkovna vrsta, do katere lahko obenem dostopa več različnih niti. Skrbi za seznam naslovov, ki jih mora naš pajek še obiskati in za etično delovanje pajka - zagotoviti je potrebno, da pajek do istega strežnika ne dostopa bolj pogosto, kot to strežnik dovoljuje.

Ob dodajanju novega URL naslova v vrsto, razvrščevalnik preveri, ali je bil ta URL naslov že prenesen. Prav tako zna iz URL naslova prebrati domeno, pridobiti IP naslov strežnika s to domeno in prenesti datoteko `robots.txt`.

Razvrščevalnik hrani seznam IP naslovov in čas zadnjega dostopa do določenega strežnika. Za vsak IP naslov vzdržuje vrsto URL naslovov, ki jih je potrebno še prenesti. Ko nit pajka od razvrščevalnika zahteva URL naslov, ta preveri kateri izmed strežnikov je bil obiskan najprej in iz vrste za ta strežnik vzame prvi URL naslov tega strežnika. Ko nit pridob URL naslov za prenos, se v razvrščevalniku IP naslov, ki pripada tej domeni, zaklene in se odklene šele po tem, ko nit razvrščevalniku sporoči, da je bil prenos uspešen.

Pri dodajanju URL naslovov v razvrščevalnik, najprej preverimo, ali je bila stran že obiskana. Če še ni bila, iz domene pridobimo IP naslov strežnika in URL shranimo na konec vrste z ustreznim IP naslovom. Implementiran pajek tako izvaja preiskovanje v širino (angl. breadth-first search) za posamezen IP naslov strežnika. Na tak način zagotovimo, da bo pajek v primernem času zagotovo obiskal vse strani, v danem trenutku pa pajek dostopa do strežnikov z različnimi IP naslovi, kar je v skladu z dogovorjenimi etičnimi načeli.

B. Pajek

Ob zagonu, pajek (ang. crawler) zažene več niti, vsaka od njih pa ima nalogo, da iz razvrščevalnika URL naslovov pridobi naslov, prenese stran in shrani podatke v podatkovno bazo. Pridobivanje vsebine strani izvajamo v dveh korakih:

- Najprej na strežnik izvedemo `HTTP HEAD` zahtevek. Na tak način pridobimo statusno kodo strežnika in morebitne preusmeritve med stranmi. Pajek v tem koraku tudi preveri kakšen je `Content-Type` strani. Če je vsebina strani `HTML`, potem nadaljujemo s prenosom, v nasprotnem primeru pa v tabelo `page_data` vstavimo ustrezen vnos s tipom datoteke.
- S pomočjo Selenium brskalnika obiščemo stran. Če pri prenašanju strani pride do prekoračitve vnaprej določenega časa, z obiskom prekinemo in URL naslov ponovno dodamo na konec vrste.

Pajek po prenosu strani s spleta izvede še naslednje korake procesiranja:

- Iz HTML vsebine strani izračuna razpršilno vrednost (angl. hash) in preveri, ali stran z enako vrednostjo v bazi že obstaja. V tem primeru, se v bazi URL naslov označi kot duplikat.
- Na strani poiščemo vse slike in jih vstavimo v tabelo `image` v podatkovni bazi.
- Na strani se poiščejo vse hiperpovezave. Razvrščevalnik za vsako hiperpovezavo najprej izračuna njeno kanonizirano obliko in skuša na podlagi poti v naslovu ugotoviti, kakšne vrste je datoteka. V primeru, da ima datoteka ustrezno končnico (npr. `.pdf`, `.docx`, ...), takega URL naslova ne dodamo v vrsto, temveč ustvarimo nov zapis v tabeli `page_data` v podatkovni bazi.
- Pridobljene hiperpovezave dodamo v razvrščevalnik.

C. Shramba

Shramba (ang. storage) je del pajka, ki skrbi za vpis zbranih podatkov in vsebin v relacijsko bazo podatkov. Baza je postavljena v okolju PostgreSQL, strukturi baze, ki je

bila predlagana v projektu, pa smo v tabeli `page` dodali še polje `page hash`, v katerem se shranjuje hash vrednost, izračunana z MD5 algoritmom. Hash vrednost smo shranjevali za preverjanje, ali spletna stran z določeno vsebino že obstaja.

Za vpisovanje, popravljanje in izpis podatkov iz podatkovne baze smo sprogramirali 25 funkcij, vsaka od teh funkcij zaklene dostop do podatkovne baze, da ne bi prišlo do pomanjkljivih ali večkratnih vpisov vsebin.

III. Delo na projektu

Kar se tiče samega dela na projektu, je bilo najtežje uskladiti večnitno delovanje pajka in čakanje pri pošiljanju zahtev za spletne strani, tako da pajek deluje stabilno in hkrati v skladu z etičnimi načeli. Vsaka takšna zahteva usklajeno delovanje tako med večimi nitmi samega pajka, kot tudi med vsemi komponentami, apliciranimi v pajku. Na težave smo naleteli tudi pri uporabi gonilnika Geckodriver za brskanje v ozadnju po spletnih straneh. Če je pajek prišel do strani, kjer je uproabnik lahko izbral certifikat, se je, ne glede na nastavljen 10 sekundni timeout, nit ustavila. To smo rešili tako....

IV. Rezultati

Naš spletni pajek je v 20 urah zbral podatke o 221 spletnih mestih, 47.631 spletnih straneh, 33.104 dokumentih in 392.447 slikovnih datotekah. V bazo som ujeli tudi 6 spletnih mest in 3.396 strani, ki so imele povezave s strani gov.si domene in ne gostutejo na tej domeni. Glede na to, da gre za kar veliko količino podatkov, predvsem s kvantitativnega vidika, smatramo, da je delovanje pajka zadovoljivo. Tudi same podatke smo uspeli s spletnih mest pridobiti tako, da so primerni za nadaljnjo uporabo, kar se nam zdi pomembno za to nalogo. Je pa, seveda, tudi še nekaj prostora za izboljšave, ki bi jih lahko, če bi imeli na razpolago nekaj več časa, tudi implementirali.

V spodnji tabeli navajamo rezultate.

	gov.si	evem.gov.si	e-uprava.gov.si	e-prostor.gov.si
Število spletnih mest	215	1	1	4
Število spletnih strani	47631	365	3020	1130
Število dvojnikov strani	10636	287	56	412
Število vseh dokumentov	33122	27	9	960
Število PDF dokumentov	23998	6	8	905
Število DOC dokumentov	2498	15	0	39
Število DOCX dokumentov	6197	6	1	12
Število PPT dokumentov	176	0	0	1
Število PPTX dokumentov	253	0	0	3
Število slik	425016	563	3199	5302
Povprečno število slik na stran	9	1,5	1	4,6
Povprečno število dokumentov na stran	0,8	0,08	0,01	0,9