Spletni pajek

Enes Fejzoski, Marko Kofol, Elian Mugerli

April 5, 2024

Abstract

To poročilo dokumentira izgradnjo spletnega pajka v sklopu študijskega predmeta Iskanje in Ekstrakcija Podatkov s Spleta. Glavni namen projekta je bil ustvariti program, ki omogoča avtomatizirano iskanje, pregledovanje in pridobivanje informacij s spletnih strani.

1 Uvod

V okviru študijskega predmeta Iskanje in Ekstrakcija Podatkov s Spleta smo se soočili z izzivom implementacije spletnega pajka. Glavni cilj naloge je bil razviti program, ki omogoča avtomatizirano iskanje, pregledovanje in zbiranje informacij s spletnih strani. Ta program, znan tudi kot spletni pajek ali web crawler, je osrednji element pri pridobivanju podatkov s spleta za različne namene, kot so analiza trga, sledenje trendom, odkrivanje spletnih ranljivosti in izdelava spletnih indeksov.

Spletni pajek deluje na način, da sistematično pregleduje spletne strani in sledi povezavam med njimi. Vsako obiskano spletno stran analizira, pridobi njeno vsebino ter izlušči in shrani želene podatke. S tem omogoča avtomatizirano in učinkovito pridobivanje informacij s širokega spektra spletnih virov. Ta proces omogoča hitro in zanesljivo pridobivanje podatkov za nadaljnjo analizo in uporabo v različnih kontekstih.

V nadaljevanju poročila bomo podrobneje predstavili implementacijo spletnega pajka, opisali ključne funkcionalnosti, ki smo jih razvili, ter analizirali izzive in rezultate našega projekta.

2 Delovanje pajka

Spletni pajek deluje na principu raziskovanja in pridobivanja podatkov s spletnih strani. Glavni cilj pajka je obiskati ciljne spletne strani, pridobiti njihovo vsebino ter izluščiti koristne informacije za nadaljnjo obdelavo. Pajek deluje po načelu več niti, kar pomeni, da lahko hkrati obdeluje več spletnih strani, kar pripomore k hitrejšemu in učinkovitejšemu pregledovanju.

Najprej se inicializira skrbnik pajka (CrawlerManager), ki določa število delovnih niti (workers) in začetne spletne naslove (initial seed), ki jih je potrebno obdelati. Skrbnik vključuje tudi območje (frontier), ki predstavlja množico spletnih naslovov, ki jih je treba pregledati.

Nato sledi delovanje pajka (Crawler), ki ima nalogo obdelati posamezne spletne strani. Vsaka nit pajka prevzame eno od spletnih strani iz območja (frontier) in jo pregleda. Med obdelavo pajek preveri pogoje iz robots.txt datoteke za določeno domeno, preveri dovoljenost obiska spletne strani ter preveri morebitne prepovedi. Po pregledu, pajek pridobi vsebino spletne strani in iz nje izlušči povezave do drugih spletnih strani ter slike. Pridobljene povezave se dodajo nazaj v območje (frontier) za nadaljnje obdelovanje, medtem ko se slike shranijo v podatkovno bazo za nadaljnjo analizo.

3 Težave

Med izvajanjem projekta smo se soočili z več izzivi, ki so vplivali na potek implementacije in delovanje programa. Na začetku smo imeli težave s postavitvijo in konfiguracijo podatkovne baze, kar je zahtevalo dodatno časovno in tehnično vlaganje, saj nismo imeli zadostnih izkušenj pri upravljanju s podatkovnimi bazami. Poleg tega smo med razvojem programa naleteli na različne robne primere, ki niso bili ustrezno obravnavani v našem programu, kar je privedlo do nepredvidenih težav in napak v delovanju spletnega pajka. Še en izziv, s katerim smo se srečali, je bil povezan s tabelo link, kjer

smo opazili težave pri pravilnem povezovanju strani. Med zadnjim pogonom programa pa se je tudi prekinila povezava do podatkovne baze, kar je povzročilo nepričakovano ustavitev delovanja pajka. Kljub več poskusom ponovne vzpostavitve povezave nam to ni uspelo, kar je privedlo do prekinjenega delovanja programa. Posledično nam ni uspelo zbrati zaželjeno število končnih strani.

4 Zaključek

Kljub trudu in naporom smo se v projektnem procesu soočili z izzivi, ki so vplivali na uspešnost razvoja spletnega pajka. Med izvajanjem smo se srečali s številnimi ovirami, vendar smo kljub temu uspeli ustvariti delujoč spletni pajek, ki je sposoben iskanja, pregledovanja in zbiranja podatkov s spletnih strani.

Na žalost se je med zadnjim pogonom programa pojavila težava s prekinjeno komunikacijo z bazo podatkov, kar je povzročilo, da smo namesto pričakovanih 50.000 strani pridobili le okoli 200. Kljub temu, da nismo dosegli načrtovane količine podatkov, smo se iz te izkušnje veliko naučili.