## Prikupljanje podataka

World wide web je moguće predstaviti kao veoma veliki usmereni graf u kome web stranice predstavljaju čvorove, a linkovi između stranica predstavljaju grane grafa.

World wide web svakako predstavlja najznačajniji izvor podataka. Podaci na webu se najčešće pronalaze u formi web stranica, koje za cilj imaju da podatke predstave korisniku na način koji će za čoveka biti intuitivan. Ipak kada je u pitanju rad sa velikom količinom podataka koju prikuplja i obrađuje računar, ovaj način predstavljanja podataka nije naročito efikasan. Ipak, kako struktuirani grupisani podaci nisu uvek dostupni za prikupljanje, često dolazi do potrebe za prikupljanjem podataka sa web stranica. Jedna od najčešćih tehnika u prikupljanju podataka sa web stranica jeste korišćenje web crawler komponenti. Web crawler je softverska komponenta koja koristi grafovsku strukturu world wide weba u cilju omogućavanja pretrage međusobno povezanih stranica bez ljudske interakcije. Suština web crawler komponenti jeste pretraga html sadržaja web stranica iz kojeg je moguće prikupiti tražene podatke, kao i linkove ka daljim stanicama od interesa. Kako bi pretraga web stranica korišćenjem crawlera bila optimizovana, uvek je poželjno da struktura sajta koji se pretražuje bude smislena kao da je i html struktura samih stranica pogodna za web crawling. Ove osobine su veoma značajne i kada je u pitanju rangiranje web sajtova na raznim sistemima pretrage, kao što su Google ili Bing.

Web crawler komponenta ima za cilj pristupanje skupu međusobno povezanih stranica korišćenjem nekog od algoritama za obilazak grafova, i prikupljanje značajnih podataka tokom njihovog obilaska. Kada je u pitanju obilazak željenih stranica, prvo što je potrebno odrediti jeste skup početnih stranica obilaska. Prilikom ovog koraka je najvažnije pažljivo odabrati stranice iz kojih se putem linkova može stići do što većeg broja drugih stranica od značaja (Slika x.1). Ukoliko se ovaj korak ne izvrši na dobar način, čitav proces prikupljanja podataka može biti neuspešan.

//////////////////////////////////////////////////SLIKA X1///////////////////////////////////////////////////

Web crawler komponenta mora sadržati kolekciju linkova ka stranicama koje je potrebno posetiti, kao i kolekciju stranica koje su već posećene kako bi se izbeglo zaglavljivanje web crawlera u slučaju pojave konture u grafu (Slika x.2). Prilikom posete svakoj od stranica, link posećene stranice se najpre prebacuje iz skupa neposećenih u skup posećenih stranica, potom se sa stranice ekstraktuju svi potrebni podaci. Poslednji korak koji se obavlja na svakoj stranici je izvlačenje linkova ka ostalim stranicama, i njihovo smeštanje u kolekciju neposećenih ukoliko stranice na tim linkovima već nisu posećene. Kraj procesa prolaska kroz sajt nastupa u trenutku kada kolekcija neposećenih linkova više ne sarži ni jedan element.

//////////////////////////////////////////////////SLIKA X2///////////////////////////////////////////////////

Kako bi web crawler komponenta imala smisao, potrebno je implementirati segment za analizu html sadržaja web stranice, kao i parsiranje podataka sa iste. Prilikom posete svake od stranica komponenta zadužena za prikupljanje podataka izvršava tri koraka: pristupanje traženim html elementima, parsiranje vrednosti elemenata u podatke i struktuiranje i čuvanje podataka. Kada su u pitanju prva dva koraka, oni uveliko zavise od same html strukture stranice, što je veoma često najveći izazov u implementaciji komponenti za prikupljanje podataka. Kako bi pravilno prikupljanje podataka sa web stranice bilo moguće, sama stranica mora imati jasno definisanu strukturu. Takođe kako bi se omogućilo lakše preuzimanje podataka, web stranice često koriste identifikacione html atribute kao što su id ili data atributi, pomoću kojih web crawler komponente mogu direktno pristupiti željenim html elementima i njihovim vrednostima. Nakon prikupljanja traženih podataka sa stranice, poslednji korak jeste struktuiranje i čuvanje istih. Ovaj proces počinje dizajnom strukture podataka, koja zavisi od samog cilja istraživanja. Ove strukture podataka često se nazivaju i modeli. Modeli u svojoj strukturi sadrže atribute čije vrednosti bivaju određene na osnovu prikupljenih podataka iz prethodnih koraka. Prilikom kreiranja modela, potrebno je dodeliti vrednost svim atributima koji su dostupni i neophodni za dalje istraživanje. Nakon što su novi podaci prikupljeni i struktuirani u modele, najčešća praksa jeste njihovo skladištenje, kako bi se omogućio kasniji pristup bez potrebe za ponovnom analizom web sadržaja.

Kao što je moguće videti iz prethodnih nekoliko pasusa, upotreba web crawler komponenti je veoma efikasan način pravilnog prikupljanja podataka koji se nalaze u formatu web stranica, te njihovog struktuiranja i skladištenja za kasniju upotrebu.

## Indeksiranje podataka

## Pretraga indeksiranih fajlova

## ComSIS – IR – WebApi

## ComSIS – IR –Angular aplikacija