Simulator genetskog algoritma za ciljano pretraživanje na internetu

Dražen Drašković
Katedra za računarsku tehniku i informatiku
Univerzitet u Beogradu - Elektrotehnički fakultet
Beograd, Srbija
drazen.draskovic@etf.bg.ac.rs

Sadržaj - **Simulator ima zadatak da izvrši pretragu indeksa veb stranica. Kod standardnog načina pretraživanja rangiraju se sve stranice koje odgovaraju upitu. Zatim se stranice sa najvišim rangom prikazuju korisniku. Kod genetskog algoritma postupak pretrage je drugačiji: rangira se samo deo stranica koje odgovaraju upitu, oslanjajući se na činjenicu da su rezultati pretrage sa visokim rangom često međusobno povezani. Sistem koristi dva različita metoda za otkrivanje sličnih stranica: veze između stranica i kategorije kojima stranice pripadaju. Kao primer korišćenja sistema uzete su srpska i bosanska podbaza sajta Wikipedia. Ovakav način pretrage ne garantuje pronalaženje najboljeg rezultata, ali ukoliko se primeni na veliki domen pretrage, vrlo brzo se nakon rangiranja malog procenta rešenja iz domena, dolazi do rezultata visokog ranga.**

Ključne reči: genetski algoritmi, ciljano pretraživanje

I. UVOD

Internet pretraživanje je proces pronalaženja veb stranica koje odgovaraju korisnikovom upitu. Upit se sastoji od ključnih reči koje korisnik unosi. Veb pretraživač predstavlja ekspertski sistem koji na osnovu ključnih reči predviđa šta je to što korisnik traži. Tokom procesa internet pretraživanja koristi se indeks veb stranica. Ovaj indeks sadrži sve relevantne informacije o stranicama i omogućava aplikaciji da bez posećivanja veb stranica odgovori na korisnikov upit. U ovom simulatoru prikazan je algoritam na primeru pretraživanja stranica svetske enciklopedije Wikipedia. Stranice koje odgovaraju upitu često imaju vezu ka drugim stranicama koje takođe dobro odgovaraju upitu. Zahvaljujući vezama između stranica prilikom pretrage se kada se pronađe neka stranica koja veoma dobro odgovara upitu, dohvataju i stranice sa kojima je data stranica povezana i ispituje njen rang. Nešto slično se dešava i kada su u pitanju kategorije. Kada stranica sa visokim rangom pripada nekoj kategoriji, dohvataju se stranice iz date kategorije i ispituje se njihov rang.

Za kreiranje indeksa veb stranica koriste se posebni programi - veb roboti. Ovi roboti predstavljaju specijalizovane programe, čiji je zadatak da u indeks upišu što više stranica. Dakle, kada veb robot naiđe na stranicu koja još uvek nije uneta u indeks, on prvo tu stranicu dohvati sa interneta. Zatim, u drugom koraku stranica se analizira i iz nje se izdvajaju sve ključne reči. U trećem koraku se stranica sa ključnim rečima unosi u indeks. Prilikom indeksiranja pojedinačne strane, ne pamti se samo stranica u originalnom obliku, već se kreira i takozvani invertovani indeks. Invertovani indeks predstavlja preslikavanje neke ključne reči u stranice na kojima se ta ključna reč pojavljuje. Ovi indeksi omogućavaju aplikacijama za internet pretraživanje da vrlo brzo odgovore na zahtev korisnika. Umesto pretrage svih indeksiranih stranica sa ključnim rečima iz upita, kod invertovanih indeksa se na osnovu ključnih reči dobija spisak svih stranica koje sadrže neku od datih ključnih reči.

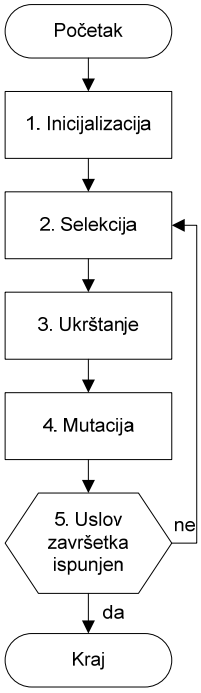
Za pretraživanje veb indeksa korisnik koristi veb pretraživač. Veb pretraživač je aplikacija koja omogućava pristup do podataka iz indeksa. Veb pretraživači od korisnika dobijaju upit, koji se sastoji od ključnih reči. Zatim se za ovaj upit, korišćenjem invertovanog indeksa dobija lista svih veb stranica koje sadrže bar jednu od ključnih reči koje je korisnik zadao. Sve stranice sa ove liste se rangiraju, na osnovu toga koliko dobro odgovaraju upitu. Na kraju se stranice koje najbolje odgovaraju upitu prve prikazuju korisniku.

II. OPIS GENETSKOG ALGORITMA

Genetski algoritam je tehnika pretraživanja, koja pripada klasi evolucijskih algoritama. Za potrebe genetskog algoritma, definiše se funkcija dobrote. Funkcija dobrote predstavlja preslikavanje jedne tačke iz domena pretrage u broj, koji predstavlja dobrotu te tačke, odnosno pokazuje koliko je data tačka dobro rešenje problema koji se posmatra. Sa ovako definisanom funkcijom dobrote problem pretraživanja može biti iskazan kao nalaženje onih tačaka u domenu pretraživanja u kojima funkcija dobrote ima maksimum. Upravo to je zadatak genetskih algoritama.[1][2]

Ideja rada ovih algoritama potiče od biološkog procesa evolucije. U procesu evolucije, jedinka koja je bolje

prilagođena uslovima ima veću verovatnoću preživljavanja i ukrštanja, a time i veću verovatnoću da svoje gene prenese u sledeću generaciju. [3] Selekcijom se odabiraju jedinke čiji se geni prenose u sledeću generaciju, a manipulacijom genetskog materijala stvaraju se nove jedinke. [4] Ovaj proces se ponavlja iz generacije u generaciju. Dijagram genetskog algoritma prikazan je na Sl. 1.



Slika 1: Dijagram genetskog algoritma

U prvom koraku genetskog algoritma populacija se popunjava proizvoljnim stranicama sa liste stranica koje odgovaraju upitu. Zatim se u drugom koraku ove stranice rangiraju. Uzeto je da populacija sadrži 1.25% od broja svih stranica koje odgovaraju upitu. Ukoliko po ovoj formuli dobijemo broj manji od deset, za veličinu populacije uzima se deset stranica. U trećem koraku proverava se da li je uslov za završetak algoritma ispunjen. Ukoliko je uslov ispunjen, algoritam prestaje sa radom, a korisniku se, od do tada rangiranih stranica prikazuju one sa najboljim rangom. Ukoliko uslov za završetak nije ispunjen, prelazi se na kreiranje sledeće generacije. U koracima četiri, pet i šest se iz skupa stranica koje odgovaraju upitu, a koje još uvek nisu rangirane, biraju stranice koje će konkurisati za ulazak u populaciju.

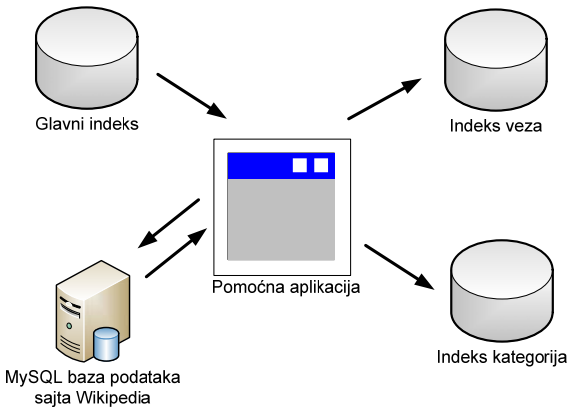
U četvrtom koraku se prvo kreira skup svih stranica koje su povezane makar sa jednom od stranica u populaciji. Ovim stranicama se dodeljuje važnost, koja zavisi od ranga stranica koje na njih ukazuju. Važnost stranice je veća ukoliko što više stranica iz populacije sa što boljim rangom ukazuje na datu stranicu. Iz ovog skupa se metodom jednostavne (rulet) selekcije biraju stranice koje će konkurisati za ulazak u populaciju. Važnost stranice dobija se kao suma rangova onih stranica koje se nalaze u populaciji, a koje sadrže vezu ka/od date stranice. Šansa da stranica bude izabrana u selekciji, proporcionalna je njenoj važnosti, ali sve stranice konkurišu u izboru.

U petom koraku se, slično kao i u četvrtom, prvo kreira skup stranica. Tada skup stranica sadrži one stranice koje dele kategoriju sa nekom od, do tog trenutka rangiranih stranica. Ovde je umesto populacije uzet skup svih do tada rangiranih stranica, kako bi i za slučajeve kada je veličina populacije mala bilo moguće naći povezanost između stranica preko njihovih kategorija. Stranicama iz skupa se dodeljuje važnost, koja zavisi od ranga stranica sa kojima dele kategoriju. Što više do tada rangiranih stranica, sa što boljim rangom deli kategoriju sa datom stranicom, to je važnost ove stranice veća. I iz ovog skupa se metodom jednostave (rulet) selekcija biraju stranice koje će konkurisati za ulazak u populaciju. Važnost stranice izračunava se kao suma rangova svih do tada rangiranih stranica, koje dele jednu ili više kategorija sa datom stranicom. Šansa da stranica bude izabrana direktno je proporcionalna njenoj važnosti.

U šestom koraku vrši se mutacija. Skupu stranica koje će konkurisati za ulazak u populaciju dodaju se proizvoljno izabrane, do tada nerangirane stranice. Ova operacija omogućava da algoritam ne kovergira prerano ka lokalnom rešenju (skupu međusobno povezanih stranica ili skupu stranica iz jedne kategorije). Zatim se stranice dobijene tokom koraka četiri, pet i šest rangiraju. One stranice koje imaju veći rang od stranica iz populacije, dodaju se u populaciju, a one sa manjim rangom budu označene kao rangirane. Nakon toga se ponovo proverava uslov završetka algoritma. Uslov je zadovoljen ukoliko je polovina stranica koje odgovaraju upitu rangirana, ili ukoliko nije bilo promene u populaciji u poslednjih dvadeset generacija algoritma.

III. PRIMER PRETRAŽIVANJA

Kao primer korišćenja sistema uzet je indeks sajta Wikipedia i to njegove dve baze: srpska i bosanska. Ove podbaze uzete su zbog svoje veličine. Dovoljno su velike da verodostojno predstave proces internet pretraživanja, ali su i dovoljno male da dozvole pokretanje sistema sa personalnog računara. Za kreiranje glavnih indeksa ovih podbaza korišćen je alat Lucene-search [5], dok su pomoćni indeksi (indeksi veza i indeksi kategorija) kreirani posebnom pomoćnom aplikacijom koja je razvijena za potrebe sistema. Lucene biblioteka omogućava upisivanje dokumenata (u slučaju ovog sistema dokument predstavlja jednu veb stranu) sa proizvoljnim atributima u indeks, zatim čitanje i pretragu indeksa, parsiranje korisničkog upita (dobijanje ključnih reči na osnovu upita) i rangiranje dokumenata na osnovu ključnih reči. Sve ove mogućnosti sistem koristi kako bi efikasno izvršavao pretragu indeksa. Način rada pomoćne aplikacije dat je na Sl. 2.



Slika 2: Način pretraživanja indeksa

Pomoćna aplikacija simulatora čita podatke o veb stranicama iz glavnog indeksa. Zatim za svaku pročitanu stranicu, upitom u bazu podataka sajta Wikipedia dobijaju se stranice koje su sa datom stranicom povezane. Takođe, upitom u bazu podataka dohvata se i lista kategorija kojima data stranica pripada. Ove informacije zapisuju se u pomoćne indekse. Ovako kreirani pomoćni indeksi omogućavaju glavnoj aplikaciji da vrlo efikasno dobije informacije o povezanosti stranica, a time i da efikasno obavlja pretraživanje.

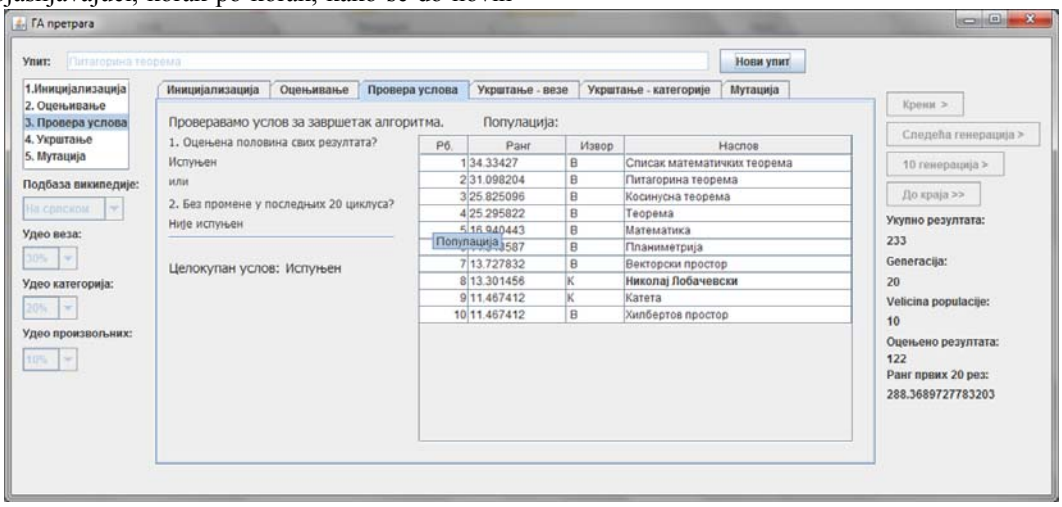
Simulator ima dva interfejsa: za pretragu i za demonstraciju rada genetskog algoritma. Pošto su ova dva interfejsa dizajnirana tako da omoguće različite nivoe uvida u interne procene pretraživanja, mogu da zadovolje potrebe različitih korisnika. Interfejs za pretragu predstavlja minimalistički pristup dizajnu korisničkih interfejsa. Ovaj interfejs, po ugledu na poznate internet pretraživače, sadrži polje za unos upita, kontrolnu tablu za pokretanje pretraživanja i listu rezultata pretrage. Korisniku omogućava da posmatrajući samo najbolje rezultate do kojih se do tog trenutka u pretraživanju došlo, brzo i jednostavno izvrši pretragu. Interfejs za demonstraciju rada genetskog algoritma omogućava korisniku da prati proces pretraživanja sa mnogo više detalja. Objašnjavajući, korak po korak, kako se do novih rešenja dolazi, ovaj interfejs daje korisniku informacije o tome kako algoritam radi. Interfejs ima zadatak da korisniku objasni kako genetski algoritmi funkcionišu i kako je pomoću njih moguće vršiti internet pretraživanje. Izgled korisničkog interfejsa po završetku izvršavanja algoritma, prikazan je na Sl. 3.

Simulator čine tri celine: biblioteka klasa, koja sadrži implementaciju genetskog algoritma za internet pretraživanje, pomoćnu aplikaciju, koja služi za kreiranje pomoćnih indeksa, i glavnu aplikaciju, koja omogućava korisniku da pretražuje sajt Wikipedia. Klase iz biblioteke omogućavaju programeru da primeni ovaj algoritam na proizvoljan problem pretrage. Biblioteka sadrži četiri paketa, koji grupišu klase u logičke celine:

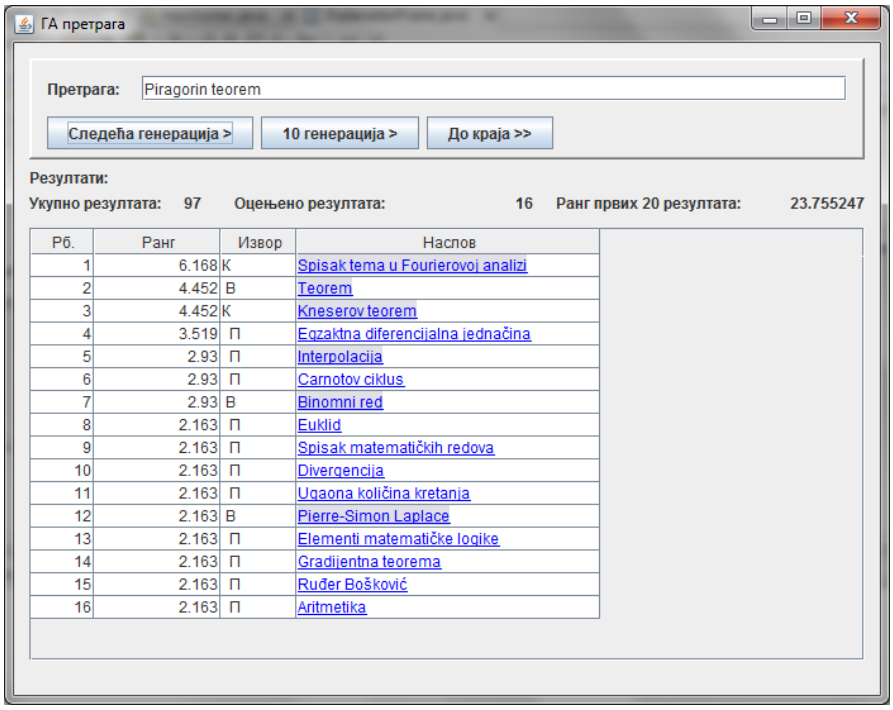
- klase koje predstavljaju interfejs ka indeksima, i omogućavaju ostatku sistema da čita podatke iz glavnog i pomoćnih indeksa;
- klase koje su zadužene za implementaciju koraka selekcije genetskog algoritma;
- klase koje implementiraju metod rulet selekcije da bi se odredile stranice koje konkurišu za ulazak u populaciju:
- klase koje implementiraju sam algoritam i koriste ostatak biblioteke za realziaciju pojedinačnih koraka.

Pomoćna aplikacija omogućava kreiranje pomoćnih indeksa za pretraživanja sajta Wikipedia. Aplikacija se povezuje na server baze podataka sajta Wikipedia i izvršavajući upite nad bazom, dobija informacije o međusobnoj povezanosti veb stranica. Zatim ove informacije upisuje u indeks veza i indeks kategorija. Ove indekse koristi glavna aplikacija kako bi efikasnije izvršavala pretragu.

Glavni deo simulatora je aplikacija koja koristi biblioteku klasa za pretraživanje podbaza sajta Wikipedia. Sama aplikacija ima dva korisnička interfejsa. Prvi interfejs omogućava brzu i jednostavnu pretragu sajta Wikipedia. Ovaj interfejs je prikazan na Sl. 4.



Slika 3: Izgled korisničkog interfejsa na kraju izvršavanja simulacije



Slika 4: Korisnički interfejs nakon pokretanja aplikacije i pretrage

Na vrhu se nalazi polje u koje korisnik unosi upit za pretragu. Klikom na jednu od ponuđenih opcija za pretraživanje korisnik može da prati rezultate pretrage. Korisnik može da prati rad algoritma generaciju po generaciju, u koracima od po deset generacija, ili da direktno ode na kraj algoritma. U tabeli ispod kontrolnog prostora korisniku se prikazuju trenutno najbolje rangirane stranice, pri čemu se prikazuje i način na koji je algoritam do stranice došao: P - proizvoljnim izborom stranica, V - preko veze sa nekom drugom stranicom, K - preko zajedničke kategorije sa nekom drugom stranicom. Klikom na naziv stranice korisnik može otvoriti datu veb stranicu u podrazumevanom veb pregledaču.

Drugi interfejs glavne aplikacije omogućava korisniku da detaljno isprati svaki korak pretrage. Korisnik može da pretražuje i srpsku i bosansku podbazu sajta Wikipedia, pa može i da uporedi rad algoritma primenjenog na domene pretrage različitih veličina. Pre pokretanja pretrage korisnik unosi svoj upit i odabira podbazu koju želi da pretražuje. Korisnik može podesiti i vrednosti za izbor udela u populaciji (udeo veza, udeo kategorija, udeo proizvoljnih) koje utiču na selekcioni pritisak, a time i na brzinu konvergencije algoritma. Ukupan zbir sve tri vrednosti daje procenat populacije koji će u jednoj generaciji biti zamenjen novim stranicama. Što je ukupni procenat veći, to algoritam brže konvergira. Sa povećanjem brzine konvergencije povećava se i efikasnost algoritma, ali i šansa da se algoritam zaustavi u lokalnom minimumu. Nakon što postavi željena podešavanja i unese upit, korisnik može da pokrene simulaciju. Korisnik na raspolaganju ima četiri moda rada algoritma: pokretanje algoritma korak po korak, generaciju po generaciju, po deset generacija odjednom ili da pokrene izvršavanje sve dok uslov završetka nije ispunjen. U svakom trenutku, korisnik je obavešten o napredovanju algoritma, prema ranije opisanim koracima.

IV. Zaključak

Opisani simulator pretraživanja omogućava korisnicima sa različitim potrebama efikasnu pretragu i detaljan uvid u implementaciju genetskog algoritma. Na primeru sajta Wikipedia, pokazana je mogućnost praktične primene genetskog algoritma za internet pretragu. Indeksiranjem dve različite podbaze sajta Wikipedia korisniku je omogućeno da uporedi rezultate svojih upita dobijene za domene pretrage različitih veličina, ali i da analizira kako se performanse algoritma menjaju sa promenom veličine domena pretrage. Kako sistem predstavlja samu bazu algoritma pretraživanja, moguća su poboljšanja i nadogranje.

Ovo istraživanje pokazalo je sve prednosti i nedostatke genetskih algoritmima primenjenih na indeksirane podatke na internetu. U nekim slučajevima standardni načini pretrage davali su bolje rešenje, ali sve zavisi od veličine domena pretrage. Treba uzeti u obzir da rad genetskih algoritama nikada ne garantuje nalaženje stvarnog maksimuma, pa ovaj vid pretrage može biti korišćen samo u slučajevima kada je takva neizvesnost dopuštena. Dalje istraživanje biće orijentisano ka podacima koji nisu indeksirani i genetskim algoritmima primenjenih na tzv. veb krolere (eng. *web crawlers*) u internet pretraživanju.

Zahvalnica

Ovaj rad podržalo je Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije kroz projekat tehnološkog razvoja 32047.

LITERATURA

[1] Draskovic, D., Nikolic, B., Milutinovic, V., "A classification of mutational approaches for genetic search", IEEE International Conference on Industrial Technology, Athens, March 2012, pages 260-264, ISBN: 978-1-4673-0340-8

[2] Draskovic, D., Milutinovic, V., "Hybrid approaches to mutation in genetic search algorithms", 6th IEEE International Conference Intelligent Systems, Sofia, September 2012, pages 336-340, ISBN: 978-1-4673-2276-8

[3] Sivaraj, R., Ravichandran, T., "A Review Of Selection Methods In Genetic Algorithm", International Journal of Engineering Science and Technology (IJEST), 2011.

[4] Hackett, P., "A Comparison of Selection Methods Based on the Performance of a Genetic Program Applied to the Cart-pole Problem", Griffith University, 1995.

[5] Gospodnetic, O., "Lucene in action", Manning Publications, 2004.

ABSTRACT

The aim of the simulator is to perform search index webpages. For a standard search methods, we look at the ranking of all pages that match the query. Then the sites with the highest ranking display to the user. In the genetic algorithm search procedure is different: it ranks only part of the page that match the query, relying on the fact that the search results with a high rank often interrelated. The system uses two different methods to detect similar pages: links between pages and category of page. As an example of using the system are taken Serbian and Bosnian Wikipedia subdatabase. This kind of search is stochastic and does not guarantee to provide the best results, but if applied in a domain search, soon after ranking small percentage of solutions in the field, there is a high level of results.

SIMULATOR OF GENETIC ALGORITHM FOR TARGET SEARCH ON THE INTERNET

Dražen Drašković, BSc EE, MSc EE