

Pretraživači

Dušan Ječmenica

Napravili ste sajt, prebacili ga na server i ponosni na svoje delo konačno konstatujete da je veliki posao gotov. A da li je? Zависи od toga da li ste prilikom izrade sajta na umu imali opake robote iz bliske budućnosti, u narodu poznate kao - pretraživači.

Cilj svakog web dizajnera je da njegov sajt bude lako dostupan populaciji kojoj je namenjen, a to se postiže na više načina: usmeno, preko e-maila, TV, štampe, mailing liste, forumima, razmenom banner-a i linkova itd., ali jedan način je naprosto nezaobilazan: prijavljivanje sajta na pretraživače. Ustvari, u poslednje vreme tu i nema više šta da se prijavljuje, pošto pretraživači automatski sa vremena na vreme indeksiraju vaš sajt - pa je zato bitno kvalitetno napraviti sajt u smislu da ga pretraživač pronade i **rangira što više**. Ali baš to predstavlja najveći problem - zato što svaki pretraživač ima neku svoju "filozofiju" kojom se vodi kada indeksira sajtove.

Čini mi se da se web dizajneri malo bave problematikom pretraživača: obično se sve završava konstatacijom "samo stavim META tagove i problem je rešen". Međutim tu se pravi kardinalna greška, jer to ne samo da nije dovoljno, nego i kao što ćemo kasnije videti - Google upošte ne konstatuje `<META NAME="keywords" CONTENT="...">`.

U ovoj seriji članaka pokušaću da pokažem kako pretraživači rade, zašto je Google bitan, i kako da vaš sajt bude što bolje rangiran na pretraživačima

Vrste pretraživača



Kada pretražujete internet preko nekog pretraživača, vi ustvari vršite pretragu po njegovoj bazi podataka, pa prema tome što više sajtova pretraživač ima u svojoj bazi - to je bolji. Zato je glavna podela pretraživača baš prema tom kriterijumu: kako se pune njihove baze i vrši rangiranje u njima. Ta podela bi izgledala ovako

Crawlers - (*Spiders, Robots*) su pretraživači koji automatski "krstare" webom, indeksiraju i ažuriraju svoje baze podataka ("*listings*") pa korisnici kasnije pretražuju po njima. Znači, sve je bazirano na algoritmu indeksiranja - što je algoritam kvalitetniji, baza je optimizovanija. (Google, Alltheweb, AltaVista)

Human Directories - ažuriranje i pretragu vrše ljudi (zaposleni): postoji formular na svakom od tih "humanoidnih" pretraživača koji treba popuniti, i kada se zahtev razmotri, onda se odlučuje da li će sajt upošte biti indeksiran, i ako hoće - određuje mu se mesto u određenoj sekciji - direktorijumu. (Yahoo, Open Directory)

Hibridni pretraživači - kombinacija ova dva gore navedena.

Meta Pretraživači - to su pretraživači koji pretražuju druge pretraživače i rezultate filtriraju (ponovljene rezultate prikazuju samo jednom) i rangiraju ih. Neki od njih se i prave u vidu aplikativnog softvera koga možete downloadovati i startovati kao aplikaciju, a ne kroz web browser, kao npr *Copernic*. (www.copernic.com)

Kako rade?

"Smatra se da je sajt dobro rangiran ako se nalazi na prvoj strani rezultata pretrage..."

Čitava svrha ove priče je da saznamo kako *crawlers* pretraživači rade, i pogotovu kako se rangirati što više kod njih, jer "humanoidni" pretraživači i nisu baš u sferi našeg interesovanja upravo zato što kod njih ne postoji neki određen algoritam kojim se rangiranje vrši, nego sve zavisi od mišljenja osobe koja razmatra zahtev da vaš sajt bude uvršten u bazu dotičnog pretraživača. Zato ćemo se upravo najviše pozabaviti *crawlers*-ima, a specijano Google-om, trenutno najboljim svetskim pretraživačem.

Grubo posmatrano, *crawlers* pretraživači se sastoje iz tri dela: samog "**pauka**" (*spider, robot*), **baze podataka** (*index, catalog*), i **softvera za pretragu i rangiranje rezultata** (*page searcher and indexer*). *Spider* radi 24 časa dnevno - on je neumoran. Obilazi web stranice, kad naiđe na neku novu, onda poseti i linkove koji vode od nje (pokušavajući da uvrđi strukturu samog sajta). Ti podaci se prosleđuju bazi podataka, koja zajedno sa softverom za pretragu i rangiranje određuje prioritet stranice u rezultatima buuće pretrage.

To je ono što ste možda često čuli kao "page rank", odnosno na kom mestu će se naći vaš sajt ako se unese ključna reč koja ga opisuje. Smatra se da je sajt dobro rangiran ako se nalazi na **prvoj** strani rezultata pretrage (najčešće među prvih 10). Pošto i sami znate koliko postoji sajtova koji se bave istim temama, potrebno je dosta umeća i sreće da uspete da se baš vaš sajt pojavi među prvih deset u istoj oblasti.

Da li treba napomenuti kako je hardverski zahtevan svaki od tri dela *crawler-a*, pa ga pokreću čuda tehnike: multiprocesorski sistemi sa vodenim hlađenjem i super brzim hard-diskovima iza kojih stoje trostruki back-up sistemi. Jednostavno - tu nema mesta za grešku.



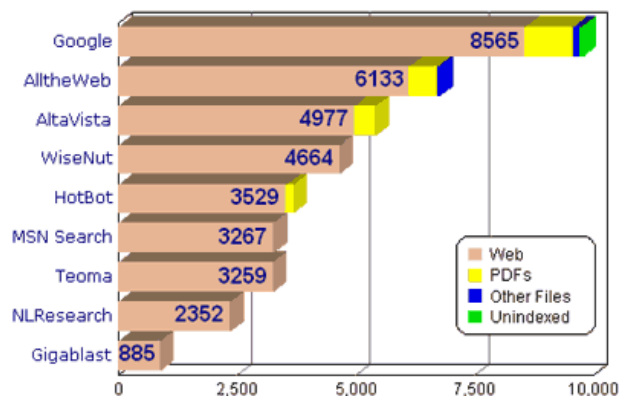
Inače, kada vaša web strana bude jednom posećena paukom ("spidered") on se vraća svaki mesec ili dva - da proveri da li je došlo do nekih promena, kako u tematici sajta, tako i u strukturi. Nemojte misliti da spider "razume" tematiku sajta: od reči koje pronađe, i od zavisnosti od njihove učestanosti pojavljivanja, "predlaže" softveru za pretragu i rangiranje ključne reči pomoću kojih će mnogi posetioci doći do vašeg sajta. Ponekad prođe neko vreme od posete *spider-a* do indeksiranja sajta, i ono se naziva "*index delay time*" i može da iznosi i do 10 dana. Znači, ključne reči pomoću kojih će se pronaći vaš sajt će se menjati iz meseca u mesec - ako se sadržaj prvih 100Kb vašeg sajta menja mesečno.

Kompleksnost

"Kada unesete neki termin, pretraživač generiše (za svoje potrebe) malo drugačiji termin radi što boljeg pronalazjenja i rangiranja rezultata..."

Ovde nećemo do detalja iznositi kako se indeksiraju i rangiraju stranice, to je vrlo kompleksna tehnologija - ukratko: **document processor** (normalizuje dokument, "razbija" ga na jednostavne delove, izoluje relevantne META tagove, identifikuje elemente koji su pogodni za indeksiranje, briše tzv. "stop" reči - veznike, predloge, itd), **query processor** (pretrvara tražene termine u znakove (*tokenizing*) prepoznaje tražene termine od eventualno unetih operatora, ponovo briše "stop" reči, kombinuje reči praveći izraz, upoređuje novodobijene termine sa bazom, pa ponovo proširuje traženi termin), **search and matching function** (set algoritama koji upoređuju dobijeni izraz iz prethodnog koraka sa bazom, ali po više kriterijuma - broju ponavljanja izraza u bazi, da li

se izraz nalazi u obliku naslova ili linkova, onda sledi ponovna opitimizacija dobijenih rezultata algoritmom koji je sličan document processor-u) i na kraju - *ranking capability* (set instrukcija kojima se dobijeni rezultati ređaju i to po specijalnom kriterijumu koji pretstavlja kombinaciju uslova: učestanost pojavljivanja generisane fraze u prethodnim koracima u dokumentima koji su ušli u "uži izbor", posećenost dotičnih dokumenata, datum njihovog formiranja, koliko linkova sa drugih domena vodi do njih, koliko ima izlaznih linkova, lokacija traženog termina, itd). I sve ovo za manje od sekunde!!! Znači, kada unesete neki termin, pretraživač generiše (za svoje potrebe) malo drugačiji termin radi što boljeg pronalazjenja i rangiranja rezultata. Naravno - ovo nekad ne funkcioniše baš najbolje, pa se onda uvek potkrade neki nerelevantni rezultat za pretragu.



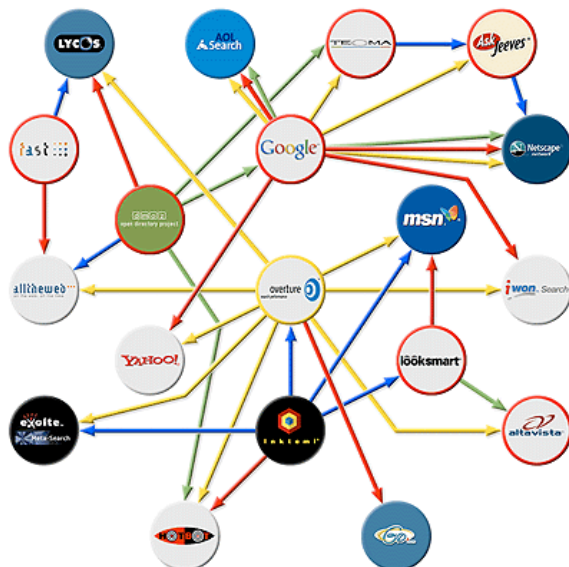
Slika 1: Google daje najviše relevantnih rezultata

Veličina je bitna

Koliko veliku bazu imaju određeni pretraživači možemo da vidimo na **slici 1**: za 25 različitih unetih kratkih reči u polje za pretragu - Google je ukupni pobednik, ostavljajući daleko ispred sebe sve ostale. Znači on ima najbolji *spidering* sistem i najbolju optimizaciju same baze.



Ono što je posebno interesantno je to da je svaki od poznatijih pretraživača pomalo - metapretraživač, jer se oslanja na drugog. To se najbolje vidi na **slici 2**, gde vidimo kako su pretraživači međusobno povezani. **Crvena** strelica prema nekom od njih znači da prima primarne rezultate pretrage od ovog drugog, **plava** da prima sekundarne rezultate pretrage, **zelena** da prima rezultate direktorijumske pretrage, a **žuta** da prima plaćene (sponzorisane) rezultate pretrage. Primetite da Google jedino prima direktorijumski listing od Open Directory, kao i da se većina oslanja na Google, bar kad su primarni rezultati pretrage u pitanju, MSN i Yahoo nisu resurs nikome, a Fast se ne oslanja ni na koga. Znači Google je čvrsto rešen da doprema informacije i u direktorijumskom obliku, a Fast (AllTheWeb), naprotiv čvrsto ostaje pri odluci da bude samo *crawler*.



Slika 2: Ko podržava koga?

Kako se rangiraju rezultati?

Ono što razlikuje *crawler*-e jedne od drugih je upravo *ranking capability*. Sada ćemo da vidimo kako on generalno funkcioniše, a nešto kasnije ćemo se detaljno pozabaviti Google-ovom tehnologijom.

Kada tražite film u video-klubu, a niste baš sigurni kako se zove, osoba koja radi na izdavanju filmova će vam uputiti par dodatnih pitanja: da li znate ko glumi, da li je film novijeg datuma, da li je akcija, triler ili drama itd... Međutim, nijedan pretraživač nema mogućnost takve interakcije (bar ne za sada), pa stoga mora da se pronade drugi način da se pogodi na šta je posetilac mislio. Zato se pomoću određenog algoritma utvrđuje koje stranice su najposećenije u toj oblasti, koliko spoljnih linkova sa drugih domena vodi ka njima, a potom se one prikazuju rangirane po prethodnim uslovima. To je kao da vam je pretraživač rekao: "Nisam baš siguran koju stranicu želite, ali ovo vam je spisak trenutno najposećenijih kojima odgovara tražena ključna reč u toj oblasti - pa je stranica koju tražite najverovatnije među njima, jer se pokazalo da se u 80% slučajeva pretraga završava baš na njima. Ako nije ovde - idite na sledeću stranu, itd...". Zašto je ovaj način prikazivanja rezultata kontroverzan? Zato što se svodi na ono poznato pitanje: šta je starije - kokoška ili jaje? Tj, kako sajt da postane posećen kad nikad nije među prvih deset u rezultatima pretrage, a kako da bude među prvih deset - kad nije posećen - i tako sve u krug. Kako je Google rešio taj problem (ili se bar misli da je rešio), kako on upošte funkcioniše, i zašto - ako vas nema na listingu Google-a - Yahoo vas upošte i ne indeksira, videćemo u sledećem broju.

Do tada, evo adresa poznatih pretraživača:

www.google.com
www.altavista.com
www.alltheweb.com
www.yahoo.com
www.msn.com
www.hotbot.com