



BARN | Beogradska Akademija
Računarskih Nauka

Internet i informacione tehnologije

Seminarski rad:
Domaći search engine

Predmetni nastavnik:
Goran Radić

Student:
Siniša Ristović

19.12.2005.

SADRŽAJ:

UVOD.....	3
SEARCH ENGINE	4
OPŠTA PODELA	4
WEB DIREKTORIJUM	4
KROLERSKI ILI SPIDER PRETRAŽIVAČI	5
META PRETRAŽIVAČI	6
PROBLEMI PRETRAŽIVAČA.....	6
GOOGLE.....	8
NAPREDNO PRETRAŽIVANJE	8
DOMAĆI SEARCH ENGINE.....	9
PROBLEMI DOMAĆIH PRETRAŽIVAČA	9
WWW.YUSEARCH.COM.....	9
WWW.YU.....	9
WWW.SRBKO.COM	10
WWW.TRZIM.COM	10
WWW.ALADIN.CO.YU	10
WWW.KRSTARICA.COM.....	10
WWW.POGODAK.CO.YU.....	10
WWW.GOOGLE.CO.YU.....	11
TABELA PERFORMANSI.....	11
INDEKS TABELA:	12
INDEKS SLIKA:	12
REFERENCE:	13

UVOD

Svet je postao veliko tržište, a s obzirom da živimo u eri informatike i moderne komunikacione tehnologije, prava i tačna informacija znači sve: Zauzeti što bolje mesto i što duže opstati na tržištu. Pojavom globalne mreže veliki broj informacija je postao dostupan široj populaciji, čime se i konkurencija na tržištu zaoštrava jer biva mnogo lakše postati i ostati u "toku", pratiti trendove i prilagođavati svoj proizvodni i uslužni asortiman korisnicima. Naravno, tu su i "obični" korisnici koji Internet koriste za zabavu i kontakt sa rođacima i prijateljima iz udaljenih mesta, što doprinosi da Internet postaje mreža globalnih razmera a ujedno i naša svakodnevna potreba. Procenjuje se da danas na mreži postoji oko 2 000 000 web site-ova. Još mnogo pre ove cifre javlja se potreba za servisom koji bi mogao da umesto korisnika pretraži postojeće stranice i u mnogome olakša potragu i dolazak do željenih informacija. Iz cele te priče dolazi do pojave i razvoja *search engine-a* bez koga bi danas bilo bukvalno ne moguće koristiti moć Interneta.



Slika 1.

SEARCH ENGINE

Prvi veći pretraživač na web-u jeste Yahoo.com, koga su podigla dva studenta sa idejom da naprave top listu svojih omiljenih site-ova. U početku se na Yahoo-u našlo samo 10 site-ova, zatim 20, pa 50,100 i kako se broj povećavao rodila se ideja da naprave kategorije i da sve site-ove razvrstaju po njima. Takvi pretraživači nazivaju se "web direktorijumu" i o njima će biti reči kasnije.

Prvi ozbiljniji pretraživač jeste Altavista.com, koji je počeo kao Yahoo, ali uskoro razvija način indeksiranja i postaje "krolerski" pretraživač.

OPŠTA PODELA

Opšta podela pretraživača je izvršena na osnovu načina indeksiranja web strana. Ako je neka strana indeksirana od strane pretraživača to znači da je njen deo ili ceo sadržaj, na ovaj ili onaj način, ubačen u bazu tog pretraživača i da će iz nje biti pozivan, prilikom upita koji odgovara indeksiranom sadržaju. Iz toga nastaje podela na:

- web direktorijume
- krolerske ili spider pretraživače i
- meta pretraživače.

WEB DIREKTORIJUM

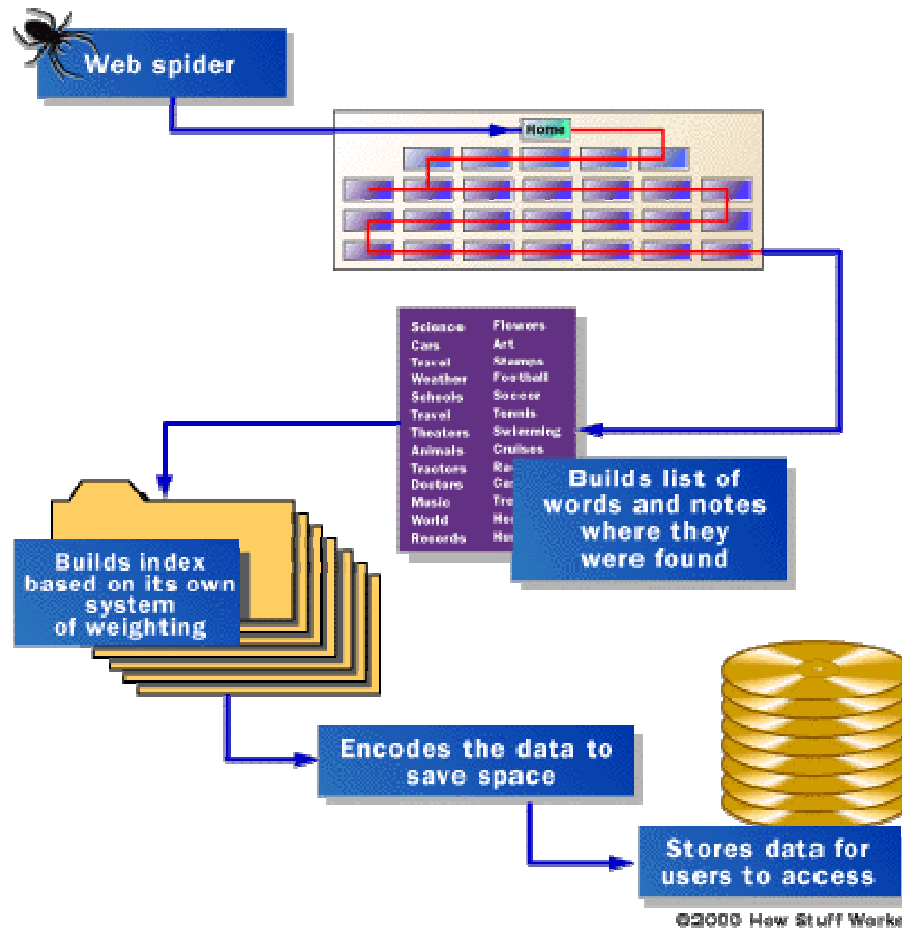
Web direktorijumi predstavljaju, uslovno rečeno, najprimitivniji servis za podršku korisnicima prilikom pretrage Interneta. Naime, to je prva vrsta search engine-a, i kao takva jasno je da je prevaziđena u mnogo tome.

Indeksiranje na web direktorijumima vrši sam čovek ili u većem slučaju više ljudi, odnosno fizički se u kategorije odnosno direktorijume ubacuju adrese odedenih site-ova. Naravno, to ima svoje prednosti ali i mane.

Mana ovog načina indeksiranja je očigledna. Neophodni su ogromni ljudski resursi da bi svojom bazom mogli da pariraju ostlim vstama pretraživača. To nije isplativo tako da su baze ovih pretraživača uglavnom male i najčešće se koriste kao specijalizovani pretraživači, odnosno kao na primr pretraživači polovnih automobila, pretraživači za sport itd. Nikada se ne indeksira ceo sadržaj site-a već samo ključne reči, naslov i reči u domenu.

Prednosti ovih pretraživača je u tome što čovekov faktor utiče na sadržaj direktorijuma tako da je gotovo nemoguće da za određeni upit kao rezultat pretrage dobijete nešto što nema veze sa traženom reči, što je ipak moguće kod krolerskih pretraživača. Ovo je i jedini razloog zbog čega se ovi pretraživači još uvek koriste.

Pošto savršenstvo ne postoji, i kod ovog pretraživača imamo mane koje su u zavisnosti od same organizacije i mogućnosti ljudi koji rade na njemu, veće ili manje. Jedna od mana je što uvek postoji mogućnost da kao rezultat pretrage dobijete nešto što nema veze sa upitom, ali to su problemi o kojima će kasnije biti reči.



Slika 4.

META PRETRAŽIVAČI

Pitaje je da li ovakav pretraživač možemo uopšte nazvati pretraživačem, jer u najvećem broju slučajeva ne poseduje svoju bazu već koristi tuđe, najčešće baze krolerskih pretraživača. Odnosno, omogućava nam da pretražimo pomoću naprednog pretraživača a da zaobiđemo njegov interfejs, što i nije neka prednost, ali jeste to što meta pretraživači najčešće koriste baze od više pretraživača, tako da je veća verovatnoća da za određeni upit dobijemo traženi rezultat. Međutim, tu se javlja problem dupliranja pogodaka, zbog čega je potrebno programirati određene filtere koji će taj problem rešavati.

PROBLEMI PRETRAŽIVAČA

Od pojave Interneta do danas web masteri se trudeda budu što bolje rangirani na listi pretraživača, jer neka ispitivanja govore da korisnik ide najčešće na prvi, drugi eventualno treći pogodak za traženi upit. Stoga su web masteri najveći neprijatelji pretraživača jer imaju različite interese. Ljudi koji prave pretraživač, prave ga tako da bude od koristi i da bude efkasan kako bi stekao poverenje i posećenost a samim tim i profit, dok je web masterima u

interesu da privuku korisnike Interneta na svoj site iako on nema veze sa potrebom korisnika. Zbog toga su i kroleri imali svoj razvitak od najjednostavnijih (ako je moguće reći za njih da su jednostavni) do ovih trenutno aktivnih koji uspeavaju da prevaziđu većinu problema koje im zadaju web masteri.

Najveći problem današnjih pretraživača jeste ogroman broj podataka i informacija i kako ih sortirati odnosno rangirati. Cilj je napisati pretraživač koji će svoje podatke u bazi sortirati na način na koji će biti od najveće pomoći korisnicima čak i u situaciji koje je vrlo česta a to je da ni sam korisnik ne zna šta traži. Odnosno, potrebno je napraviti algoritam koji će rangirati pogotke. U početku se rangiranje zasnivalo na tome koliko se puta tražena reč nalazi na strani. Što više puta, to bolje mesto na pretraživaču. Ali su web masteri došli na ideju da pušu ključne reči istom bojom kojom je i pozadina tako da se reči ne vide a kroleri ih vidi i broji. Uskoro je i taj problem bio rešen, bar kod ozbiljnijih pretraživača, i takvi site-ovi su stavljeni na crne liste.

Sledeća stvar koja se javlja, koja je za korisnike problem, a za pretraživače nije, jesu sponzori. Naime, pretraživači koji su se okrenuli malo više ka profitu su u rezultate pretrage ubacivli i sponzore i to vrlo dobro kotirane, iako nisu zadovoljavali upit korisnika. Takvi pretraživači su, logično gubili poverenje i posećenost a samim tim i sponzore.

Sledeći problem koji se javlja jeste dupliranje pogodaka odnosno jedan korisnik-više domena. Korisnik ne želi da na listi pogodaka vidi npr. www.link.com i www.link.net ukoliko su to potpuno iste prezentacije. Problem je u tome koji prikazati a koji ne. Svaki pretraživač to rešava na svoj način ali je Google u tome najbolji. O tome u daljem tekstu.

Ništa manji problem je više korisnika-jedna domen. To je karakteristično ja domaće provajdere koji svojim korisnicima omogućavaju postavku njihovih prezentacija na svom serveru pa imamo npr. www.eunet.co.yu/korisnici/2005/petarpetrovic...

GOOGLE

Nije moguće pričati o search engine-u a ne pomenuti trenutno najveće ime u toj oblasti. Google je jedan od najmlađih ozbiljnijih pretraživača. Poznato je da je duže vreme radio sa minusom ali je imao cilj i uspeo je u tome. Od samog početka do danas Google svoje sponzore odvaja od rezultata pretrage zbog čega je stekao poverenje ogromnog broja korisnika. Jedan podatak govori da 21% korisnika Interneta Google drži kao home page. Ovaj pretraživač je postao vodeći iz dva vrlo jednostavna razloga a tiču se najbitnije stvari kod pretraživanja a to je sortiranje. Naime ljudi koji su odgovorni za sortiranje su napravili algoritam takav da je veoma bitno mesto gde se nalazi tražena reč, odnosno da svako mesto ima svoju težinu. Tako, reči u domenu su najteže, zatim dolaze naslovi site-ova, pa naslovi stranica pa tek onda same reči kao sadržaj strane. Ovaj sistem je u mnogome unapredio rangiranje i okrenuo ga potpuno ka korisniku. Kao šlag na tortu, Google je uveo još jedan parametar za sortiranje a to je koliko linkova sa različitih site-ova vodi do tog site-a. Što više linkova sa što više site-ova, govori da je sam taj site vrlo ozbiljan i da ga treba visiko rangirati. Zbog ovih prednosti, vrlo jednostavnog interfejsa bez opterećujućih popapova i grfike i veoma brzog odziva Google je veoma brzo postao vodeći pretraživač stavljajući u drugi plan Altavistu i Yahoo. Treba pomenuti saradnju sa Alexa.com-om koji pravi rang listu najposećenijih site-ova što je još jedan merodavan kriterijum za uspešno sortiranje.

NAPREDNO PRETRAŽIVANJE

Pored standardnog, ozbiljniji pretraživači nam nude i napredno pretraživanje, kako bismo lakše došli do željene informacije. U tbeli su dati operatori koji nam u tome pomažu.

Pretraga	Operator	Primer
Uključiti u pretragu	+	Da bi našli Ministarstvo pravde u Beogradu <i>Search: Ministarstvo pravde + Beograd</i>
Uključiti jedna ili drugi u pretragu	OR	Da bi našli let za Podgoricu ili Tivat <i>Search: Let za Podgoricu OR Tivat ('OR' velikim slovima)</i>
Pretraga opsega brojeva	..	Da bi našli telefon između 100 i 200e <i>Search: Telefon 100e..200e</i>
Pretraga sinonima	~	Da bi smo našli kulinarske informacije <i>Search: ~kuvar</i>
Pretraga domena	:	Da bi smo našli određenu temu na site-u <i>Search: ict:www.bccs.edu.yu</i>

Tabela 1.

DOMAĆI SEARCH ENGINE

Domaći Internet je voma mali. Procenjuje se na oko 50 000 site-ove. Donekle ta brojka se može opravdati lošom finansijskom situacijom građana, ali podatak koji zapanjuje je da samo 3% stanovništva izvan Beograda koristi Internet. Globalna mreža u našoj zemlji dostupna je korisnicima od 1998., dok je u 1997. postojala samo Akademska mreža.

Domaći pretraživač predstavlja posebnu vrstu specijalizovanog pretraživača, koji je ograničen na govorno područje. Sve opšte vrste pretraživača, o kojima je bilo reči, mogu se naći i kod nas, što znači da i svi problemi koje imaju oni imaju i domaći, uz još mnogo jezičkih problema.

PROBLEMI DOMAĆIH PRETRAŽIVAČA

Možda najveći problem domaćih pretraživača jeste odrediti koji je to domaći site. Da li su to svi site-ovi koji su na .yu domenu? Postoji veliki broj site-ova na engleskom jeziku koji se nalaze na .yu domenu. Oni nikako ne spadaju u domaće site-ove. Skoro polovna naših siteova se nalazi na .com domenu. Njih treba indeksirati. Tu su i ostali domeni: .net, .org, ... Dolazimo da zaključka da su domaći site-ovi oni koji su pisani za naše govorno područje. Ali kako napisati algoritam za kroler koji pravi razliku između srpskog i bosanskog koji su veoma slični jezici? U takvim problemima, odnosno rešenjima tih problema se i vidi vrednost pretraživača.

Ali ni to nije sve. Neophodno je pomenuti da smo jedna od vrlo retkih zemalja koja ima dva pisma. Ćirilicu koja je službeno pismo i latinicu koja je mnogo više zastupljena na web-u zbog sličnosti sa alfabetom. To znači da bi naš pretraživač morao da nam za upit na latinici izlista i rezultate na ćirilici i obrnuto, što je zaista komplikovana procedura. Međutim, ni to nije sve što se pisama tiče. Domaći pretraživač bi morao da nam za upit na latinici da i rezultate pretrage za taj isti upit na alfabetu, odnosno bez slova: č, ć, š, ž, đ. Nije potredno pominjati da se očekuje i obrnuto.

Sledeća jezička karakteristika nešeg govornog područja je u tome što za razliku od npr. Engleza koji nemaju ni jedan, mi imamo čak 7 padeža. Tako da je poželjno da pretraživač pored svih pomenutih mogućnosti, ima i sposobnost da prepozna koren reči. Izlišno je pričati to komplikovanosti ovog zahteva. Ukoliko pretraživač ne podržava ovu opciju, a korisnik otkuca "novine", ostaće uskraćen za stranice na kojima piše npr. "novinama", što nije dobro za popularnost.

WWW.YUSEARCH.COM

Yusearch je jedan od prvih domaćih pretraživača. Od samog nastanka do danas on je tipa web direktorijum. Bio je veoma popularan, ali je mu je posećenost znatno opala pojavom Krstarice. Nije rađeno na njegovom usavršavanju tako da ni danas nema podršku za latinicu i ćirilicu, a o drugim mogućnostima nema ni govora. Ne daje informacije o vremenu pretraživanja baze.

WWW.YU

Takođe jedan od prvih. Doživeo je sličnu sudbinu kao i yusearch što se tiče nadogradnje. Ima veoma malu bazu, u odosu na vreme koje je proveo kao pretraživač. Nije bio

posebno popularan. Ne podržava latinicu i ćirilicu, samim tim ni naprednije mogućnosti. Ni on ne daje informacije o vremenu pretraživanja baze.

WWW.SRBKO.COM

Iako vlasnici tvrde da ima krolersku bazu, na testu se nije tako pokazao. Radi kao najobičniji web direktorijum, čak ne ni kao meta. Podržava alfabet i latinicu. Ne daje informacije o vremenu trajanje pretrage.

WWW.TRZIM.COM

Ovaj pretraživač teško da tako možemo nazvati. Čak nije ni tipična meta i ne krije to jer nas pita čiju bazu želimo da pretražimo (Yahoo, Google, Altavista...), što znači da ne može čak ni da uporedi dve baze pa da na osnovu toga sam sortira rezultate. Pun je linkova ka drugim specijalizovanim pretraživačima (automobili), zbog čega je trenutno popularan, što je i jedini razlog njegovog pominjanja.

WWW.ALADIN.CO.YU

U odnosu na pomenute, malo napredniji. Nema bazu već koristi Yahoo i Google. Prikazuje vreme trajanja pretrage i naravno znatno je veće od trajanja pretrage samog Google-a. Podržava alfabet i latinicu i odvojeno ćirilicu. Ukoliko je upit na latinici rezultat na latinici je boldovan a alfabet nije. Najveći broj pogodaka koji prikazuje je 99 000.

WWW.KRSTARICA.COM

Prvi ozbiljniji domaći pretraživač. Nastao pre Googla. U početku je radila kao meta, a sada ima i svoju bazu. Kao što je Altavista doživela ekspanziju uvođenjem krolera, tako je i krstarica postala najpopularniji domaći pretraživač koristeći baze Altaviste i Yahoo-a (koji je do tada i sam postao krolerski pretraživač). Međutim, ostao je istih mogućnosti do danas. Podržava oda pisma i alfabet ali odvojeno. Prikazuje vreme trajanja pretrage. U početku je bio bez popapova i reklamnih banera, sada je situacija drugačija. Ipak uspeva da zadrži treće mesto na listi domaćih pretraživača.

WWW.POGODAK.CO.YU

Najmađi domaći pretraživač. Delo Slovenaca koji su posle uspeha u svojoj zemlji i Hrvatskoj došli kod nas i napravili najozbiljniji pretraživač. Radi kao sinteza mete jer koristi bazu Googla, i kroler pretraživača jer ima i svoje krolere i svoju bazu koju administriraju ljudi, zbog čega podseća i na web direktorijum. U prevodu, izvukli su najbolje osobine svih pretraživača i napravili do sada najbolji search engine kod nas. Zbog svojih mogućnosti uveren sam da će uskoro postati najpopularniji domaći pretraživač. Treba pomenuti da je trenutno jedini koji za upit na jednom pismu daje rezultate i na drugom i to vrlo lepo sortirane. Da jedini prepoznaje koren reči, što znači da ima podršku za padeže.

WWW.GOOGLE.CO.YU

Trenutno najpopularniji domaći, kao i svetski, pretraživač jeste Google. Na svetskoj sceni ne vidim mu ozbiljnu konkurenciju, dok kod nas je samo pitanje vremena kada će Pogodak preuzeti tron. Srpska verzija Google-a ne podržava dvojezične rezultate pretrage, i veoma često kao pogodak "zaluta" hrvatski ili bosanski pa čak i ruski site, što kod Pogot-ka nije slučaj. Očigledno da se ljudi iz Google-a nisu dovoljno pozabavili tom tematikom, inče ne verujem da je to problem koji oni ne mogu da reše. U svakom slučaju, imaju znatno veću bazu od Pogodka i brži odziv.

TABELA PERFORMANSI

	Krstarica.com	Pogodak.co.yu	Google.co.yu	Aladin.com	Srbko.com
Latinica	podržava	podržava	podržava	podržava	ne podržava
Ćirilica	podržava	podržava	podržava	podržava	ne podržava
Latin+alfabet	ne podržava	podržava	podržava	podržava	ne podržava
Oba pisma	ne podržava	podržava	podržava	ne podržava	ne podržava
Padeži	ne podržava	podržava	ne podržava	ne podržava	ne podržava

Tabela 2.

Iz tabele vidimo da je trenutno najbolji domaći pretraživač Pogodak i da je pitanje trenutka kada će preuzeti vodeći poziciju od neprikosnovenog Google-a.

INDEKS TABELA:

Tabela 1. Napredno pretraživanje	8
Tabela 2. Tabela performansi	11

INDEKS SLIKA:

Slika 1. Web	3
Slika 2. Spider	5
Slika 3. Način rada search engine-a	5
Slika 4. Način rada spidera	6

REFERENCE:

- www.google.co.yu
- www.pogodak.co.yu
- www.krstarica.com
- www.aladin.com
- www.srbko.com
- www.trzim.com
- www.yu
- www.yusearch.com