BELGRADE COLLEGE OF COMPUTER SCIENCES

INTERNET I INFORMACIONE TEHNOLOGIJE

Seminarski rad:

Pregled i uloge Internet pretrazivaca

Predmentni nastavnik:

Goran Radic

Student:

Marko Radovic 02/04

Datum predaje

31.12.2004

UVOD

Jaz izmedju Interneta i ostalih medija je sve veci, kao i sam Internet. Rastom Interneta, raste i njegov broj stranica, koji danas iznosi vise milijardi. Verovatno bi svako to protumacio kao vrlo koristan proces, iz razloga sto veceg broja informacija na Internetu, samim tim i sto vecoj korisnosti Interneta. Ali posle uobicajenog zakljucivanja dolazi na red dodatno, procesom rasta, Internet je sve vise komplikovaniji i tezi za snalazenje, kako za prosecnog, tako i za profesionalnog korisnika. Stoga, cinjenica je, da puni kapacitet Interneta mozemo iskoristiti, samo ukoliko imamo odgovarajucu pomoc. Tu, veoma potrebnu, pomoc cine Internet pretrazivaci (Internet search engines).

Internet pretrazivaci su specijalno kreirani sajtovi na Internetu, dizajnirani da pomognu nalazenje informacija lociranih na drugim sajtovima. Razlike izmedju Internet pretrazivaca su velike, ali postoje tri zadatka koje svaki od njih obavlja:

- pretrazuju Internet ili biraju delove Interneta prema vaznim recima;
- cuvaju index reci koje nadju i njihovu lokaciju;
- dozvoljavaju korisnicima da pretrazuju reci ili kombinacije reci nadjene u tom spisku

Prvi Internet pretrazivaci, su sadrzali index koji je iznosio nekoliko miliona stranica i dokumenata, i primili nekoliko hiljada pretraga dnevno. Danas, najbolji Internet pretrazivaci, sadrze index sastavljen od nekoliko milijardi stranica i dokumenata i sa desetinama miliona pretraga dnevno.

PRETRAZIVANJE INTERNETA

Pretrazivaci su jedan od nepredvidjenih rezultata distributerske komjuterske mreze, sada zvane World Wide Web (WWW ili web). WWW je samo jedna komponenta Interneta kao celine, ali uglavnom kada ljudi kazu Internet oni zapravo misle World Wide Web. U ranijim fazama WWW-a, jedni nacin sa se prosledi informacija od jednog kompjutera ka drugom je bio File Transfer Protocol ili FTP , ovaj nacin je zahtevao da jedan kompjuter zna adresu drugog. FTP klijent je onda mogao da kontaktira FTP server (Deamon) koristeci tacnu adresu, nakon cega bi pretrazivao i uzimao odabran deo dostupnog sadrzaja na distributerskoj masini. Tadasnje pretrazivanje, nije nalik danasnjem, sadrzaj prikazan kao obicna lista, bez grafickog pretrazivanja i linkova. Korisnik bi tada morao da prekine vezu i ponovo uspostavi vezu FTP klijenta ka drugoj masini da bi pretrazivao drugi sadrzaj (pritom je mogao da se poveze samo sa serverima cija adresa je poznata). Bilo je to tezak, komplikovan i radnicki zadatak gledano sa aspekta danasnjih standarda. Jedini nacin da neko nadje fajl, pogotovo nov fajl, bio je slanjem e-mail-a.

I pre nego sto je WWW postao najvidljiviji deo Interneta, postojali su Internet pretrazivaci sa ciljem lakseg snalazenja korisnika. To su bili, sada vec legendarni programi, gopher i Archie , koji su cuvali informacije spiskove locirane

na serverima, koji su bili konstantno poveazani sa Internetom. Iskljucivo njihovom zaslugom, ekstremno je smanjeno vreme nalazenja programa i dokumenata. U kasnim 80'-im, iskoristiti maksimum Interneta, znacilo je poznavanje programa: gopher, Archie, Veronika itd...

Pre nego sto pretrazivac moze dati rezultat (lokaciju fajla ili dokumenta), on mora da ga nadje. Da bi nasao informaciju na milijardama stranica Web-a, pretrazivac uposljava specijalno softverske robote, takozvane paukove (spiders), koji prave listu reci nadjenih na Internet sajtovima. Proces paukovog pravljenja liste se naziva gmizanje Web-om (Web crawling). U cilju pravljenja sto korisnije liste reci, paukovi pretrazuju mnogo stranica.

Pocetak paukovog pretrazivanja stranica, su liste sastavljene od mnogo pristupanim serverima i veoma popularnim stranicima. Pauk pocinje od popularnog sajta, sastavljajuci index sa njegovih stranica i prateci svaki link, nadjen na pocetnom sajtu. Na ovaj nacin, sistem zasnovan na paukovima brzo pocinje putovanje kroz Web, sireci se preko njegovih siroko primenjenih delova.

Google.com je nastao kao akademski pretrazivac. Njegov inicijalni sistem koristi vise paukova, obicno tri istovremeno. Svaki pauk moze da odrzi 300 veza sa Web stranicama otvorenim i isto vreme. Pri svojim maksimalnim performansama, koristeci cetiri pauka, sistem moze da odgmize preko 100 stranica po sekundi, stvarajuci oko 600 kilobajta podataka svake sekunde.

Odrzavajuci sve brzim znacilo je napraviti sistem koji bi hranio pauke neophodnim informacijama. Prvobitni sistem Google-a sadrzao je server posvecen dostavljanje URL-a paucima. Google je imao svoj sopstveni DNS (Domain Name Server prevodi ime servera u adresu), iz razloga sto bi se u suprotnom oslanjao na Internet servis provajdera za DNS. Rezultat je svodjenje odlaganja na minimum, tj. veca brzina.

Kada Google-ov pauk pretrazuje HTML stranica, on belezi dve stvari:

- Reci na stranici
- Lokaciju reci

Reci koje se pojavlju u naslovu, podnaslovu, meta tagovima i drugim vaznim pozicijama su zabelezene za specijalno razmatranje u toku sledece korisnicke pretrage. Google-ov pauk pravi index svake znacajne reci na stranici, ostavljajuci clanove "a", "an" i "the". Drugi pauci koriste drugacije pristupe.

Drugacije pristupi su pokusaj da pauk radi brze, dozvoli korisniku da pretrazuje efikasnije ili oboje. Na primer neki paukovi ce cuvati trag reci u naslovu i podnaslovu i linkovima, zajedno sa sto najucestalije koriscenim recima na stranici i svaku rec u prvih dvadeset redova teksta. Tvrdi se da Lycos koristi ovakav pristup prilikom pretrage Web-a paucima.

Drugi sitemi, kao AltaVista, idu u drugom pravcu, uzimajuci svaku rec na stranici, uklucujuci "a", "an", "the" i druge "nebitne" reci. Korak do savrsenstva u ovom pristupu je

Meta tagovi dozvoljavaju vlasniku stranice da naznace kljucne reci i koncept po kojem ce stranica bi zabelezena. Ovo moze biti korisno, pogotovo u slucajevoma u kojim reci na stranicimmogu imati dvostruko ili trostruko znacenje; meta tagovi mogu da vode pretrazivac u biranju pravog znacenja za rec. Ipak, tu je i opasnost u prekomernom oslanjanju na meta tagove, jer nemarni i beskrupulozni vlasnici stranica mogu dodati meta tagove koji odgovaraju veoma popularnim temama, pritom nemajuci nikakve veze sa sadrzajem stranice. Da bi se zastitili, pauci ce uporediti meta tagove sa sadrzajem stranice, odbijajuci meta tagove koji koji se ne poklapaju sa recima na stranici.

Sve ovo ukazuje da vlasnici stranica zele da budu ukljuceni u rezultate Internet pretrazivaca. Mnogo puta, vlasnik ne zeli da pauk pretrazi njegovu stranicu. Na primer, igra pravi nove, aktivne stranice svaki put kada se prikazu delovi stranice ili kada se novi link otvori. Ako Web pauk pristupi jednoj od ovakvih stranica i pocne da otvara sve linkove ka novim stranicama, igra bi mogla da pogresno da protumaci aktivnost kao ljudskog igraca visoke brzine i izmakne kontroli. Da bi izbegli ovakve situacije, razvijen je robot exclusion protocol, koji umetnut u pocetak stranice, kaze pauku da ostavi stranicu na miru, tj., da ne belezi reci na stranici niti da prati njene linkove.

PRAVLJENJE INDEKSA

Kada su paukovi zavrsili zadatak, nasli informacije na Web stranicama (treba naglasiti da je ovo zadatak koji se zapravo nikad ne zavrsi; konstanto menjanje prirode Web-a znaci da paukovi uvek gmizu), pretrazivac mora da sacuva informacije na nacin koji ih cini korisnim. Postoje dve kljucne komponente ukljucene u omogucavanje sakupljenih informacija dostupnim korisniku:

- Informacije sacuvane sa podacima
- Metod po kojem su informacije zabelezene

U najjednostvnijem slucaju, pretrazivac moze samo da sacuva rec i URL gde je nadjena. U realnosti, ovo bi bilo primenjeno u slucaju pretrazivaca ogranicene upotrebe, jer ne bi bilo nacina reci da li je rec iskoricena na bitan ili trivijalan nacin, da li je rec iskoriscena samo jednom ili vise puta, da li stranica sadrzi linkove ka drugim stranicama koje sadrze rec itd. Drugim recima ne postoji nacin uspostavljanja hijerarhije medju podacima prema korisnosti.

Da bi imali sto korisnije rezultate, pretrazivaci cuva vise nego rec i URL. Pretrazivac moze da sacuva broj pojavljivanja reci na stranici. Pretrazivac moze da dodeli tezinu (weight) svakom pristupu, sa povecavajucim vrednostima dodeljenim recima dok se prikazuju blizu vrha dokumenta, u podnaslovima, u linkovima, u meta tagovima ili u nazivu stranice. Svaki komercijalni pretrazivac ima drugaciju formulu za dodeljivanje tezine recima u svom spisku. Posledica su razliciti rezultati pretrazivanja prilikom koriscenja razlicitih pretrazivaca.

Bez obzira na preciznu kombinaciju dodatnih delova informacija sacuvanih od strane pretrazivaca, podaci ce biti kodirani sa ciljem ustede radnog prostora. Na primer, originalni Google papir opisuje koristeci 2 bajta od 8 bita svakog, da sacuva informacije tezina; bilo da rec pocinje velikim slovom, njen font, velicinu, poziciju i druge informacije da bi pomogli rangiranje pogodaka. Svaki faktor moze da uzme do 2 ili 3 bita u okviru 2-bitnog grupisanja (8 bita = 1 bajt). Kao rezultat, veliki deo informacija moze biti sacuvan u veoma zgusnutoj formi. Kada je informacija zgusnuta, spremna je za belezenje.

Index ima jednu ulogu: On omogucava da se informacije nadju sto brze. Postoje nekoliko nacina da se index napravi, ali jedan od najefikasnijih nacina je da kreiranje melanzna tabela (hash table). U melanziranju, formula je dodata dodatku numericke vrednosti svake reci. Formula je dizajnirana da ravnomerno raspodeli pristupe preko unapred odredjenog broja podela. Ovo numericko distibuiranje je drugacija forma distribucije reci preko alfabeta i to je kljuc efikasnosti melanzne tabela.

U Engleskom(najvise koriscenim jezikom na Web-u), postoje slova na koja pocinju mnoge reci i ona na koja pocinju manje reci. Naci cete, za preimer, da "M" deo recnika mnogo manji od "X" dela. Ova nejednakost znaci da nalazenje reci koja pocinje na veoma popularno slovo moze uzeti mnogo vise vremena nego nalazenje reci koja pocinje na manje popularno slovo. Melanziranje ujednacuje dotadasnju razliku i smanjuje prosecno vreme nalazenja pristupa. Takodje odvaja index od stvarnog pristupa. Melanzna tabela sadrzi melanzirane brojeve zajedno sa pointerima ka stvarnim podacima, koji se mogu sortirati na bilo koji nacin koji dozvoljava da budu sacuvani najefikasnije. Kombinacija efikasnog spiska i efikasno cuvanje cini mogucim da se rezultati dobiju brzo, cak i kad korisnik kreira komplikovanu pretragu.

ISTORIJA INTERNET PRETRAZIVACA

Prvi pretrazivac, otac (moze se slobodno reci i deda) svoh ostalih, stvoren je od strane Alana Intejdza (Alan Entage) u prostorijama McGill Univerziteta u Montrealu, ranih 90'-ih. Prva namera je bila da to bude lista slobodnih FTP fajlova i njihove adrese (arhiva). Alen je planirao da nazove ovu listu "Archives", ali naziv je zamenjen za Archie, zbog potreba korisnika Unix-a da koriste krace i ponekad sifrovane programe i nazive fajlova.

Archie bi pretrazivao anonimne FTP sajtove (anonimni FTP sajtovi su oni koji dozvoljavaju svetski pristup, oni koji ne koriste korisnicka imena i sifre za ogranicen pristup zatvorenim fajlovima) i pravio index u bazi podataka svih dostupnih fajlova. Korisnik koji trazi fajl je mogao da koristi jednostavan interfejs i neke regularne izraze poklapanja uzoraka da bi nasao lokaciju zeljenog fajla. Ovo je eliminisalo proces pretrazivanja mnogo sajtova u potrazi za zaljenim fajlom i u isto vreme bio prvi pravi skok u svet spiskova Internet izvora u svrshu pretrazivanja.

Kako se WWW razvijao, DNS (Domain Name Server) je video svetlo dana. Ovo je sistem pomocu koga tekst adresa ili URL (Universal Resource Locator) moze biti preobracena u numericki lokacioni indetifikator ili IP (Internet Protocol) adresu za odredjen kompjuter. Princip rada se zasniva na sledecem; svaki kompjuter ili komponenta na Internetu su locirani IP adresom, ovo je numericki identifikator koji je jedinstven za tu komponentu. Na primer: 192.34.55.44, ovo nije lako zapamtiti, tako da se stvorila potreba za jednostavnijim nacinom

Internet adresiranja. DNS postize to pretvaranjem ime domena (domain name) u IP adresu; kada se otkuca adresa web sajta u pretrazivacu, zahtev je posla lokalnom DNS serveru koji ga konvertuje u IP broj, nakon toga komjuter koristi taj IP da locira pravi web sajt.

Ovo je dovelo do povezivanja sa tekstom, klikom na link, DNS trazi odredjenu IP adresu i upucuje racunar direktno (web stranice kakvim ih danas znamo su tada stvorene). Od skromnog pocetka masivna mreza web-ovih dokumenata koju zovemo World Wide Web je stvorena.

World Wide Web Wanderer kreiran od strane Metja Greja (Matthew Gray) bio je prvi anonimni agent na web-u. Bio je dizajniran da prati rast Interneta, u pocetku je samo brojao web servere, posle je i uzimao i URL-ove. On je pravio prvu web-ovu bazu podataka i Matju ju je nazvao Wandex. Wandex je iskoriscavao povezanu prirodu web-a, prateci jedan link ka drugom. To je sasvim isti proces koji koriste Web Roboti danas (Robot je program koji automatski prenosi strukturu web-ovog hiperteksta u toku vracanja dokumenta i rekurzivno vracanje svih dokumenata koji su referencirani u tom dokumentu). Postojao je daleko

INFORMACIJE O VAZNIJIM PRETRAZIVACIMA

POJMOVI

- Size - približan broj URL adresa, koji sadrži indeks mašine za pretraživanje.

- Spider Class - Postoje dve vrste: Deep i Shallow. Deep Spider pretražiće sve stranice site-a na određenoj URL adresi. Shallow Spider pretraživanje radi na dva načina: pretražiće samo jednu stranu, definisanu URL adresom i stati, ili će pretražiti i ostale strane Web site-a, koje se nalaze na istom hijerarhijskom nivou.

- Meta Tag support - Ukazuje da li mašina za pretraživanje koristi meta tagove iz html koda. Mogući slučajevi su: da, ne, delimično.

- Frame support - Može li mašina za pretraživanje pretražiti site čiji se sadržaji nalaze u okviru frameset-a? Mogućnosti su da/ne.

- Image Map support - Da li mašina za pretraživanje čita URL adrese koje se nalaze u okviru Imagemap na strani klijenta? Mogućnosti su da/ne.

- Alt Text support - Da li mašina za pretraživanje indeksira alternativne tekstove korišćenih grafičkih elemenata? Mogućnosti su da/ne.

- HTML Comments - Da li mašina za pretraživanje čita i indeksira HTML komentare? Mogućnosti su da/ne.

- URL Searching - Da li mašina za pretraživanje može da pretraži naziv domena? Mogućnosti su da/ne.

- Embedded Directory - Da li mašina za pretraživanje ima i direktorijum na svom site-u? Mogućnosti su da/ne.

- Submission URL - URL našeg site-a, koji prijavljujemo mašini za pretraživanje.

DIREKTORIJUM

Direktorijum je sistem baze podataka (database system), zasnovan na ručnom unosu podataka. Zadatak Web administratora je da direktorijumu prijavi URL adresu. Pored adrese, direktorijumu se obezbeđuju i informacije o naslovu i kratkom sadržajem Web site-a. Direktorijumi najčešće ne posećuju prijavljeni site, mada nekoliko direktorijuma poseduje jednostavan spider, čiji je zadatak da proveri korektnost URL adrese.

Zajednički elementi mašina za pretraživanje i direktorijuma su mogućnost pretraživanja baze podataka i podrška za postavljanje upita korišćenjem sintakse matematičke logike.

Osnovna razlika između njih je način obezbeđenja podataka: mašine za pretraživanje to rade automatski, a direktorijumi ručno.

ENGINE / DIRECTORY	Pages	Meta Tag Support	Database Refresh	Average Submission Time	Frames Supported
EngineAltaVista	140 miliona	YES	Monthly or longer	1-3 days	NO
Engine Excite	55 miliona	NO	1-3 weeks	2-4 weeks	NO
Engine HotBot	110 miliona	YES	1 week	2-4 weeks	NO
Engine InfoSeek	30 miliona	YES	3 weeks	1-3 days	YES
Engine Lycos	30 miliona	Partial	1-2 weeks	1-3 weeks	YES
Engine Northern Light	80 miliona	Partial	2 weeks	2-3 weeks	NO
Engine WebCrawler	2 miliona	YES	Weekly	3-4 weeks	NO
Directory Yahoo	1 milion	NO	Manually added	6-8 weeks	YES

Tabela 1. Karakteristike direktorijuma Internet pretrazivaca

Zajednička karakteristika svih većih mašina za pretraživanje je da odbacuju stranice, kod kojih utvrde nekontrolisano ponavljanje istih pojmova (keywords stuffing).

ALTAVISTA

Alta je zapocela sa radom, prema vecini podataka, kasne 1995 u DEC-ovoj istrazivackoj laboratoriji u mestu zvanom Palo Alto. Ideja za ime AltaVista je dosla sa laboratorijske bele table ciji je sadrzaj bio polovicno izbrisan. Rec Alto (od punog naziva Palo Alto) stajala je pored reci Vista i neko se neko je viknuo, "Sta mislite o AltoVista!". To je vodilo do imena AltaVista, koje znaci pogled odozgo. U druge znacajne izume AltaViste ulaze sledeci:

- prve vise jezicke mogucnisti pretrage na Internetu;
- prva tehnologija pretrazivanja koja je podrzavala Kineski, Japanski i Korejanski jezik preko prevodioca Bejbl Fisa (Babel Fish);
- prvi Internetov masinsko-prevodilacki servis koji je mogao da prevede reci, fraze i cele sajtove na Engleski sa Spanskog, Francuskog, Nemackog, Portugalskog, Italijanskog i Ruskog, i obrnuto;

Alta ima i mogucnost multimedijalne pretrage weba za fotografije, video fajlove ili muzicke fajlove, sa procenama indexa preko 90 miliona multimedijalnih objekata. Dodatkom "Ask Jeeves" tehnologije pretrage, 1999 godine, sad se umesto kljucnih reci mogu postaviti i pitanja. Kratak period svog postojanja je saradjivao sa Yahoo-om(1997.), prosledjujuci mu rezultate. Takodje je saradjivao i sa MSN-om koga je kratko "hranio", ali je sada 100% svoj pretrazivac. Alta vista tvrdi da je nagradjivan, za patente u okviru pretraga, vise od bilo koje kompanije na svetu.

"Vlasnici site-ova često prijavljuju veliki broj stranica AltaVisti, sa željom da povećaju broj pojavljivanja site-a, u rezultatima upita korisnika AltaViste. Najčešće prijavljuju stranice sa velikim brojem ključnih reči, ili sa ključnim rečima, koje se ne odnose na stvarni sadržaj stranica site-a. Takođe, naš spider pronalazi sadržaje strana, koji se razlikuju od sadržaja vidljivih u browser-u. Nastojimo da strogo demotivišemo takve tehnike ...

AltaVista je indeks, a ne mesto za stranice bezvrednih i netačnih informacija. Pokušaji, da se u naš indeks uvrste netačni sadržaji, ili da se naš indeks iskoristi za promociju strana Web site-a, umanjuju našim korisnicima vrednost indeksa.

Zato, nećemo prihvatiti URL adrese, za koje utvrdimo da narušavaju pouzdanost indeksa. U ekstremnim slučajevima, isključićemo sve strane takvog site-a ..."

Ukratko, AltaVista ne dozvoljava:

- Ciklično ponavljanje prijavljivanja
- Meta tagove sa ključnim rečima, koje ne odgovaraju sadržaju site-a

- Višestruko prijavljivanje istih stranica
- Neusaglašenost sadržaja strane, koji spider pronalazi, sa sadržajem koji browser prikazuje (skriveni sadržaji).

AltaVista spider je deep search spider, što znači da će indeksirati sve stranice prijavljenog site-a, ukoliko se prijavljuje Home page. Izuzetak su stranice koje se dinamički kreiraju, korišćenjem CGI.

AltaVista će indeksirati najveći mogući deo HTML koda, uključujući:

- tekstualne komentare slika;
- naslove;
- URL adrese;
- imena direktorijuma;
- vidljivi tekst;
- image mape i
- meta tagove.

AltaVista ignoriše HTML komentare!

Koristi meta keywords tag i meta description tag. Ključne reči koriste se za usaglašavanje sadržaja strane sa pojmom za pretraživanje, koji zadaje korisnik. AltaVista koristi opis umesto vidljivog teksta strane.

"Na koji način AltaVista određuje redosled strana u rezultatima pretraživanja?”

Svaki dokument dobija ocenu, u zavisnosti od broja ponavljanja tražene reči, pozicije reči u dokumentu, i zavisno od njihovih međusobnih veza. Nesvrsishodno ponavljanje reči, poznato kao "spamming", negativno utiče na rangiranje site-a. Spamming otkrivaju posebni programi. Kada se utvrdi da se spamming koristi na site-u, taj site biće uklonjen iz AltaVista indeksa ..."

Jasno je da to nisu jedini metodi, koje AltaVista koristi za određivanje ranga. Ali su, svakako, izuzetno značajni, i o njima se mora voditi računa pri kreiranju stranica.

Izuzetno je važno obratiti pažnju na njihov komentar o ponavljanju reči. Naime, ne precizira se na koje se delove HTML dokumenta konstatacija odnosi. Da li je to ceo dokument, ili se radi o meta tagovima? Svakako je jesno da ponavljanje pojmova ne mora da bude loše. Kao primer, može da posluži i činjenica da se, u rezultatima pretraživanja, može pronaći mnogo strana, koje sadrže ponavljanje. Ponavljanja se uočavaju i u okviru meta tagova, i u okviru

drugih delova HTML dokumenta.

Osim navedenih kriterijuma, AltaVista pri rangiranju koristi i popularnost Web site-a. Naime, što je više drugih site-ova, sa linkom na naš site, utoliko će rang biti viši. Po ovom kriterijumu se potiskuju noviji site-ovi!

Moze se slobodno reci da je AltaVista po značaju druga mašina za pretraživanje na Internetu. Nudi dobre servise, a politikom kontrole svog indeksa, nastoji da pruži objektivne rezultate pretražavanje i konzistentne sadržaje.

Size	oko 140 miliona
Spider Class	Deep
Meta Tag support	Yes
Frame support	Yes
Image Map support	Yes
Alternative Text support	Yes
HTML Comments	No
URL Searching	Yes
Embedded Directory	Yes
Submission URL	Submit

Tabela 2.Karakteristike pretrazivaca AltaVista

Takodje ima i strane sektore, a oni se odnose na zemlje: Svedsku, Veliku Britaniju, Francusku, Holandiju, Italiju i Indiju.

EXCITE

Osnivaci Mark Van Haren, Ryan McInture, Ben Lutch, Joe Kraus, Graham Spencer i Martin Reinfried, pet hakera i istaknuti politicki naucnik. Svoja istrazivanja su vrsili i Stenford biblioteci, da bi decembra 1994., Kleiner Perkins Caulfield and Byers i Institutional Ventur Partners su investirali u Excite sa kupovinom hard diskova vrednim 4000$.

Nakon godinu dana, decembra 1995 godine, su izasli na Internet. Sredinom 1996. uzeli Magelana i pri kraju 1996. kupuju WebCrawler.

Korisnici mogu da pretrazuju specificne vesti, ali one nisu arhivirane i nestaju sa sistema nakon nekoliko nedelja.

Size	oko 55 miliona
Spider Class	Shallow
Meta Tag support	Partial
Frame support	No
Image Map support	No
Alternative Text support	No
HTML Comments	No
URL Searching	No
Embedded Directory	Yes
Submission URL	Submit

Tabela 3.Karakteristike pretrazivaca Excite

Exite nije javno deklarisao svoju politiku u odnosu na spamming, niti su ograničili broj i intervale ponavljanja prijavljivanja. To, međutim, ne znači da se spamming i naknadna prijavljivanja ne penališu.

Excite indeksira vidljivi tekst na strani, ali ignoriše alternativne tekstove slika, komentare i keywords Meta tag. Ne podržava frame-ove i ne pretražuje image mape na strani klijenta.

Nedavno je radikalno promenio politiku u odnosu na Meta tagove. U prvobitnoj verziji ignorisao je keywords i description Meta tagove. Umesto pretraživanja teksta strane, koristio je algoritme veštačke inteligencije, sa ciljem da utvrdi sadržaj strane. Dobijene rezultate koristio je za formiranje sopstvenog sadržaja strane.

Od 1999. godine, pretražuje Meta description tag. Ukoliko naša strana ne sadrži Meta description tag, Exite će primeniti rutine veštačke inteligencije, a rezultat će za nas biti krajnje neizvestan.

"Naš savet je jednostavan. Tekstove, koji nisu direktno vezani za osnovni sadržaj strane, locirajte ne drugu, odgovarajuću stranu. Nastojte da održite jednostavnost i konciznost ..."

Logika je jednostavna, ali postavljeni zahtev nije uvek lako postići. Exite želi strane na kojima je sadržaj konzistentan i kratak.

Drugi važan kriterijum je broj linkova, usmeren ka našoj strani.

Iako je mali u odnosu na druge mašine za pretraživanje, Excite je strateški pozicioniran, pružajući podršku za AOL, što je već samo po sebi od značaja.

HOTBOT

Poceo sa radom u Maju 1996 godine. HotBot je dualni sistem. Sadrži Inktomi engine i directory system. Vlasnik i upravljac HotBot-a je bio Wired Magazine, ali je nedavno Wired Magazine postao vlasnistvo Lycos-a. Rezultati pretrazivanja su dobavljani iz Inktomi baze podataka. Pre je koristio Look Smart za kategorizaciju direktorijuma listi sajtova, ali se prebacio na The Open Directory Project sredinom 1999.

Size	oko 110 miliona
Spider Class	Deep
Meta Tag support	Yes
Frame support	No
Image Map support	No
Alternative Text support	No
HTML Comments	Yes
URL Searching	Yes
Embedded Directory	Yes
Submission URL	Submit

Tabela 4.Karakteristike pretrazivaca HotBot

HotBot ne daje previše informacija u pogledu politike prijavljivanja. Dostupne informacije se uglavnom odnose na spamming.

"Poznato nam je da dosta Web site-ova koristi spamming, što dovodi do toga da mašine za pretraživanje, u svojim rezultatima, ukazuju na stranice koje nisu u kontekstu zadatog pojma za pretraživanje. Spamming tehnike koriste bezosećajno ponavljanje reči, meta tagove neusaglašene sa tekstom, ili reči koje se ne mogu pročitati u browser-u, zbog male veličine fonta, ili boje fonta.

Ukoliko HotBot prepozna neku od navedenih tehnika, strogo će penalisati rang site-a ..."

HotBot ne podržava frame-ove, image mape i alternativne tekstove slika. Indeksira vidljive tekstove strane, Meta tagove i HTML komentare.

I pored deklaracije da se radi o Deep Spider klasi, HotBot ne pretražuje poddirektorijume, pa je potrebno prijaviti i URL adrese direktorijuma, koji se nalaze na nižem nivou.

I još jedna karakteristika: ukoliko se izostavi description Meta tag, HotBot će prihvatiti URL naziv domena, kao naslov za rangiranje.

"Osnovni elementi koji utiču na rang strane su: reči iz naslova, keywords Meta tags, frekvencija reči u dokumentu i dužina dokumenta ..."

Ponovo, kao AltaVista i InfoSeek, broj linkova sa drugih site-ova, prema našem site-u, podićiće rang našeg site-a.

HotBot je dualni sistem - mašina za pretraživanje i direktorijum. Direktorijum je deo otvorenog projekta, što znači da upis moramo izvršiti sami. Odnosno, site treba prijaviti i mašini za pretraživanje i upisati se u direktorijum.

INFOSEEK

Nakon igranja sa modelom direktorijuma sajta , Infoseek i dalje pretrazuje povremeno. Takodje pravi veliki direktorijum sajtova unakrsno linkovanih sa rezultatima pretraga.

Poceli sa radom u Avgustu 1995. kao direktorijum servis. Medjutim, na zimu 1996., nova masina za pretrazivanje zvana Ultra pocela je sa radom zajedno sa 25 miliona URL-ova. Godine 1999, 45% udela Infoseek-a je kupljeno od strane Disney-a i on je ulazi u procesu pravljenja novog sajta zvanog GO.com. Postojale su glasine da Infoseek kao samostalan pretrazivac nece opstati sa startom GO-a, sto se nije obistinilo.

Size	oko 30 miliona
Spider Class	Shallow
Meta Tag support	Yes
Frame support	No
Image Map support	Yes
Alternative Text support	Yes
HTML Comments	No
URL Searching	Yes
Embedded Directory	Yes
Submission URL	Submit

Tabela 5.Karakteristike pretrazivaca Infoseek

"Reči koje se koriste za opis stranica moraju tačno da interpretiraju sadržaj dokumenta. InfoSeek detektuje uobičajene spamming tehnike i penališe strane na kojima se koriste. Nažalost, u nekim slučajevima je penalizacija nedovoljno selektivna, pa se može odraziti i na pojedine, potpuno korektne strane. Da bi se postigli najbolji rezultati, treba izbegavati:

- ponavljanje ključnih reči;
- korišćenje ključnih reči van konteksta sadržaja dokumenta;
- korišćenje tekstova, koji su iste boje kao i pozadina strane;
- dupliciranje strana na različite URL adrese;

- korišćenje različitih strana, koje su preusmerene na istu URL adresu...

InfoSeek registruje i druge oblike spamming tehnika. Bilo koje narušavanje postavljenih pravila dovodi do isključivanja strane iz InfoSeek indeksa ..."

Jedan od problema, ili ograničenja, kaja važe za InfoSeek su tehnike automatskog prijavljivanja sadržaja. Ograničenja se prvenstveno odnose na popularne Web site-ove, na kojima se moguć besplatan hosting (besplatno korišćenje računarskih resursa za smeštaj naše prezentacije). Najpopularniji su: GeoCities.com, AngelFire.com, Tripod.com ...).

Imajući u vidu ovo ograničenje, može se definisati:

Ograničenja automatskog prijavljivanja uvode i ostale mašine za pretraživanje. Osim toga, procedura ručnog prijavljivanja dozvoljava da se u periodu od 24h prosledi samo jedna prijava iste URL adrese.

InfoSeek sistem za prijavljivanje održava se offline, pa se može dogoditi, da u momentu, kada želimo da prijavimo naš site, on nije na raspolaganju.

InfoSeek Spider indeksira celokupan vidljivi tekst, podržava alternativne tekstove slika, Meta tagove. Ne čita HTML komentare.

"Koristi opširne deskriptivne naslove, uključuje description Meta tag i kreira keywords Meta tagove, odvojene zapetama. Koristi i sinonime, kako bi što bolje opisao sadržaj site-a. Nekontrolisano ponavljanje reči i fraza može dovesti do lošijeg ranga (relevancy score), pa i do isključenja iz InfoSeek indeksa ..."

Iskustvo je pokazalo da, pored navedenih elemenata, InfoSeek koristi još nekoliko kriterijuma za preciznije rangiranje:

- popularnost site-a;
- Meta Tags;
- usaglašenost Meta tagova;
- podatke koje InfoSeek Spider dobije nakon naknadnih poseta site-u.

LYCOS

Osnovan u Januaru 1994, poceo sa radom u junu 1994. Ime Lycos dolazi iz latinskog jezika i znaci "vucji pauk". Postoje standardni rezultati pretrage preko Lycos Pro-a i kategorizovane kliste preko WiseWire-a. Lycos je rodjen kao istrazivacki projekat na Cernegie Mellon Univerzitetu, kreiran od strane Dr. Michael Mauldin. Infoseek je prva Internet kompanija koja je bazirala svoju relamu na CPM-u (cost per thousand page views) koji je sada postao industrijski standard.

Aprila 1996, Lycos Inc., je postala javna trgovacka kompanija na NASDAQ-ovom robnom market sistemu, pod simbolom LCOS. Postavsi javna trgovacka kompanija samo deset meseci nakon svog osnivanja, Lycos je najmladja kompanija koja je postala javna u NASDAQ-ovoj istoriji.

Aprila 1998., Lycos je zadobila WiseWire Corporation, koja sada hrani Lycos Web Guides, koji su automatski i saradnicki napravljeni preko korisnickog ulaza.

U skorijem periodu su kupili Wired Digital, dobivsi i HotBot-a u procesu. Nakon divljeg osvajanja 97-98., LycosNetwork sada sadrzi: Gamesville, Tripod, WhoWhere, Lycos Communications Angelfire, Hotbot, Hotwired, Wired News, Quote, Sonique, and Webmonkey. Kancelarije u : Waltham, Mass. (glavni stab); Njujorku, N.Y.; Mountajn Vju, Kalifornija; Dalas, Teksas; Los Andjeles, Kalifornija.; San Francisko, Kalifornija; i Cikago i Majami. Internacionalne kancelarije su locirane u Brazilu, Nemackoj, Italiji, Francuskoj, Japanu, Koreji, Meksiko, Ujedinjenom Kraljevstvu, Spaniji i Holandiji.

Size	oko 30 miliona
Spider Class	Shallow
Meta Tag support	No
Frame support	No
Image Map support	No
Alternative Text support	Yes
HTML Comments	No
URL Searching	Yes
Embedded Directory	Yes
Submission URL	Submit

Tabela 6.Karakteristike pretrazivaca Lycos

Lycos je jedna od mašina za pretraživanje, čija politka nije najjasnije definisana. Navešće se, zato, nekoliko zanimljivih napomena:

možete prijaviti više URL adresa sa svoga site-a, pod uslovom da se na njima ne nalaze isti sadržaji;

Lycos Spider probaće da proprati linkove site-a, ali najverovatnije samo na prvom hijerarhijskom nivou;

URL adrese sa specijalnim karakterima (&, %) neće biti dodati ...

Raspolaže relativno siromašnim funkcijama. Lycos Spider indeksira celokupan vidljivi tekst i alternativne tekstove slika. Ne podržava Meta tagove, HTML komentare, frame-ove i imagemaps.

Odaberite dve-tri reči ili fraza, koje treba da privuku pažnju. Smestite ih na početak HTML koda. Važne reči treba češće da budu zastupljene u većim naslovima (H1, H2), bliže vrhu ekrana, ukoliko se već ne nalaze u samom naslovu site-a i strane. Uvodni paragraf, sa tekstom u kome se pominju najvažnije ključne reči, pomoćiće Lycos Spider-u da uradi bolji apstrakt našeg site-a.

Grafičke strane se teško indeksiraju, bez obzira na kvalitet slika. Alternativni tekstovi se prihvataju. Znači, najbolje je imati što više teksta.

Ukoliko se na site-u koriste frame-ovi, obavezno koristiti i <NOFRAMES> sekciju, koju će Lycos Spider pretražiti i indeksirati.

Lycos nastavlja da koristi stari metod rangiranja, zasnovan na analizi elemenata. Ova tehnika danas nije suviše dobra, zbog velikog broja site-ova koje treba indeksirati.

Važno je navesti da Lycos sistem prvenstveno analizira prvih 200 vidljivih karaktera sadržaja dokumenta, na osnovu kojih generiše sumarni sadržaj site-a.

Lycos je jedna od najstarijih mašina za pretraživanje. Rangiranja vrši kombinovanjem elemenata i teksta iz prvog paragrafa HTML dokumenta.

GOOGLE

Poceo kao istrazivacki projekat na Univerzitetu Stanford, Google je zastupljen na Internetu od 1997. Sredinom 1999 godine dobija 20 miliona americkih dolara vrednu investiciju koja mu je pomogla da se plasira na prvo mesto Netscape Netcenter.

Google nudi neke od najjedinstvenijih rezultata od masina za pretrazivanja. Koristeci sistem zvan PageRank, Google filtrira veliku porciju irelevantnih rezultat. On takodje naklonjeniji prema EDU i GOV sajtovima sto je opustajuca promena za razliku od drugih .com pretrpanih spamom masina za pretrazivanje. Google trenutno sadrzi 25 miliona stranica indeksiranih u svojoj bazi podataka i napreduje sa teznjom da prebaci granicu od 100 miliona stranica. Treceg juna, 1999 godine, Google je primio priliv u iznosu 25 miliona USD-a. Takodje su presekli sve veze sa Stanford Univerzitetom i sada upravljaju potpuno samostalno. Sredinom 2000. je bio izabran za glavnog provajdera rezultata pretraga na Yahoo.

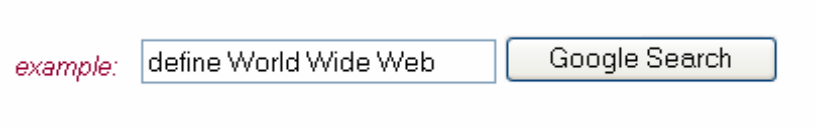
Da bi omogucio lak I brz pristup web stranicama (vise od 8 milijardi), Google ima mnogo specijalnih features, kako bi omogucio da nadjete tacno ono sto I trazite.

- Cashed Links; Google uzima snapshot svake stranice u procesu gmizenja po web-u I kesira ih kao back-up u slucaju da je originalna stranica nedostupna. Kesiran sadrzaj je sadrzaj koji Google koristi da odluci da li je stranica relevantna Ili nije za vasu pretragu. Kada se kesirana stranica prikaze, sadrzace header na vrhu koji sluzi kao podsetnik da to nije zasigurna najnovija verzija stranice. Kesirani linkovi ce nedostajati ukoliko za sajtove koji nisu indeksirani, takodje I za sajtove cioji vlasnici zabranili kesiranje njihovog sadrzaja.

- Calculator; Da bi koristili Google ugradjenu funkciju digitron, prosto treba uneti podatke u search box I pritisnuti Enter (ili kliknuti na dugme search). Digitron je sposoban za resavanja zadataka koji sadrze osnovne aritmeticke operacije, komlikovaniju matematiku, jedinice za merenje I konverzije I fizicke konstante.

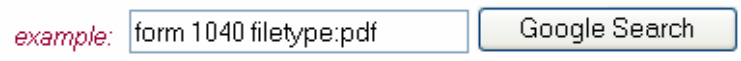
Primeri: 5+2*2; 2^20; sqrt(-4); half a cup in teaspoons; 160 pounds * 4000 feet in Calories

- Definitions; Da bi se videla definiciju za rec ili frazu, ukuca se rec "define," onda space, pa rec koju zelimo da definisemo. Ako je Google vide definiciju za rec ili frazu na na Web-u, vratice informaciju I prikazace je na vrhu vaseg rezultata pretrazivanja.



- File Types; Google je prosirio broj fajlova tipa non-HTML, koje prertrazuje na 12. U dodatku PDF dokumenata, Google sada pretrazuje Microsoft Office, PostScripty, Corel WordPerfect, Lotus 1-2-3 I ostale. Novi tipovi fajlova se jednostavno prikazuju u Google-ovim rezultatima pretrage, kada god su relevantni za korisnicku pretragu.
  Google takodje pruza korisniku mogucnost "View as HTML", dozvoljavajuci korisniku da pretrazi sadrzaj ovih fajlova, iako odgovarajuca aplikacija nije instalirana. Ova opcija takodje dozvoljava korisnicima da izbegnu viruse koji su ponekad nalaze unutar odredjenih fajl formata.

  Ukoliko zelja da se vidi partikularan set resenja za specificne fajl tip (na primer, PDF linkovi), ukuca se filetype:[extension] (na primer, filetype:pdf) u search box-u zajedno sa uslovima pretrage.

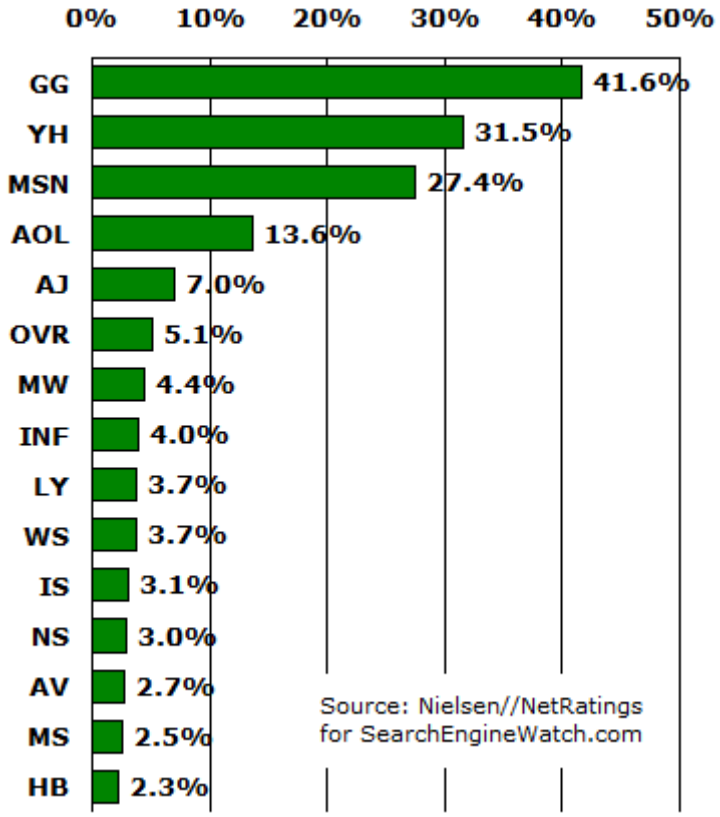


- Froogle; Ako trazite neki proizvod koristeci Google, mozete da vidite relevantne informacije o proizvodima. Ovi proizvodi su povezani sa sajtovima prodavaca koji saradjuju sa Google-om.

- I'm Feeling Lucky; Dugme "I'm feeling Lucky" vodi direktno do prve stranice koju je Google stavio u svom rezultatu pretrage. Ukratko, potrosice te manje vremena pretrazujuci web stranice, a vise ce te da gledate u njih.
- Local Search; Google Local omogucava da pretrazite web, trazeci prodavnice u odredjenom kraju.
- New Headlines; Kada pretrazujete Google-om mozete da vidite linkove na vrhu rezultata oznaceni kao "News". Ti linkovi vas povezuju sa vestima izvucenim iz vise servisa koje Google prati.
- PhoneBook; Google je dodao opciju pretrage adrese ulica SAD-a i telefonski brojeva.
- Site Search; Rec "site" pracena adresom omogucava da pretragu svedete na specifican sajt.
- Street Maps; Da bi koristili Google kako bi nasli mape ulica, potrebno je ukucati adresu ulice u SAD-u, ukljucujuci zip code ili drzavu.
- Travel Information; Da bi videli odlozene letove ili trenutno vreme na odredjenim aerodromima, potrebno je ukucati tri slova aerodroma i rec "airport".
- Web Page Translation; Koristeci masinsku tehnologiju prevodjenja, Google sada daje mogucnost da se na engleski jezik prevedu stranice na Italijanskom, Francuskom, Jaanskom, Nemackom i Portugalskom.

RANGIRANJE INTERNET PRETRAZIVACA

Evo nekih rezultata koji prikazuju poredjenja izmedju pretrazivaca.



Slika 1.Rezultati popularnosti Internet pretrazivaca

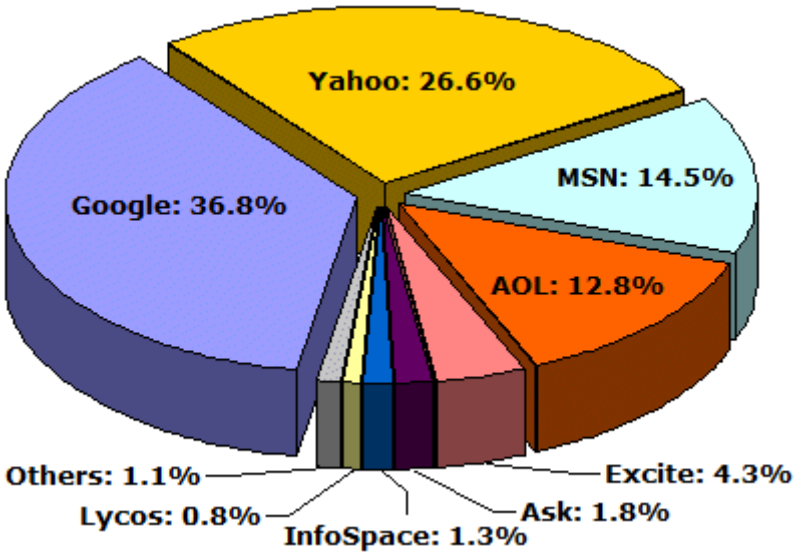
Skracenice: GG=Google, YH=Yahoo, MSN=MSN, AOL=AOL, AJ=Ask Jeeves. OVR=Overture, MY=MyWay.com, INF=Information.com, LY=Lycos Networks, WS=WebSearch.com, IS=InfoSpace Networks, NS=Netscape Search, AV=AltaVista, MS=Microsoft.com, HB=HighBeam.com.

Postoji drugaciji nacin racunanja popularnosti pretrazivaca, a to je vreme koje korisnik utrosi za pretrazivanje u toku jednog meseca. Rezultati za jun 2004. su:

Pretrazivac	Prosecno korisnicko vreme
Google	0:29:57
AOL Search	0:28:28
Netscape	0:13:09
InfoSpace	0:13:09
Yahoo	0:11:04
Web Search	0:08:06
MSN Search	0:07:39
Ask Jeeves	0:06:29
AltaVista	0:06:27
My Way Search	0:05:11
Overture	0:03:25
Lycos Network	0:02:53
Microsoft Search	0:02:22
HighBeam Research	0:01:36

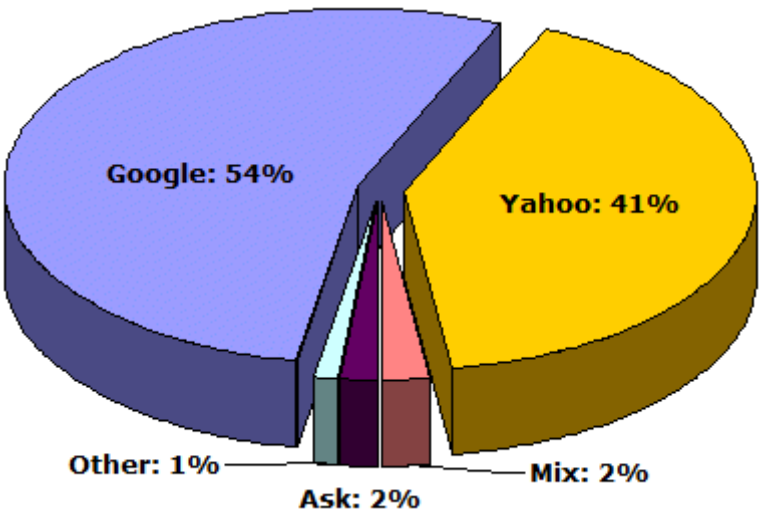
Tabela 7.Rezultati popularnosti Internet pretrazivaca po utrosenom vremenu

Pie grafikon pokazuje rezultate ucinjenih pretraga u procentima (maj 2004):



Slika 2.Rezultati popularnosti Internet pretrazivaca(maj. 2004.)

Rezultati na grafikonu iznad pokazuju rezultate pretrazivaca koji koriste svoju, ali i tudju tehnologiju; Grafikon ispod pokazuje udeo pretrazivaca kao distributera rezultata drugim pretrazivacima:



Slika 3.Rezultati distribucije podataka izmedju pretrazivaca

ZAKLJUCAK

Internet pretrazivaci i pored pospesivanja napredka i popularnosti WWW-a, doprinose i njegovoj kategorizaciji i organizaciji. Kad bih svi Internet pretrazivaci prestali u trenutku sa radom, nastupio bih opsti haos na Internetu. Jednostavno receno, bez Internet pretrazivaca ne bi bilo ni pretrazivanja, vec samo pukog trazenja igle u plastu sena.

SADRZAJ:

1. searchengine**watch.com/**

2. **computer.howstuff**works**.com/**search**-**engine**.htm**

3. **www.lib.berkeley.edu/TeachingLib/ Guides/Internet/SearchEngines.**

4. **www.webopedia.com/TERM/s/search_engine.html**

5. searchengine**watch.com/reports/**

6. **www.sofotex.com/ Search-Engine:-**statistics**-of-requests-download_L19619.html**